SafeWind

Collaborative project funded by the European Commission under the 7th Framework Program, Theme 2007-2.3.2: Energy

"Multi-scale data assimilation, advanced wind modelling & forecasting with emphasis to extreme weather situations for a safe large-scale wind power integration"

Grant Agreement N°: 213740

Deliverable Dp-6.2

"Methodology for the evaluation of probabilistic forecasts"

DOCUMENT TYPE	Report
DOCUMENT NAME:	swind_deliverable_Dp-6.2_Forecast_Verification_v2.1.pdf
VERSION:	V2.1(*)
DATE:	2009.10.15
CLASSIFICATION:	R0: General public
STATUS:	Released

Abstract: This Deliverable of the SafeWind project provides a consistent methodology for evaluating probabilistic forecasts of wind power production. We provide guidelines that will help to standardise the evaluation process in order to guarantee best practice in the commercial application of wind power forecasting. These guidelines will also allow the definition of an evaluation framework for new probabilistic forecasting methods to be developed. Obtaining a probabilistic forecast provides the end user with greater flexibility in terms of optimising their specific utility function. Probabilistic forecasts include ensemble prediction systems and density forecasts that describe the evolution of the full distribution of outcomes and quantile forecasts that can provide information about the dynamics of the tails of the distribution. As many researchers work with wind speed time series, we provide a standardised wind power curve in order to facilitate comparisons when wind power is not available. We stress the need for appropriate benchmarks for probabilistic forecasts such as persistence and climatology. These benchmarks are required for visual comparison and also for quantitative assessment in situations where the wind power time series vary greatly depending on the time and spatial scale. Finally we describe techniques for conditional evaluation which allows the end user to emphasise the forecast performance at specific levels of wind power, wind speed and wind direction. The classification analysis of extreme events such as ramps, turbine shutdown and extreme weather patterns are also discussed.

AUTHORS[1], REVIEWERS				
MAIN AUTHOR/EDITOR:	Patrick E McSharry (UOXF)			
AFFILIATION:	Oxford University			
ADDRESS:	Smith School of Enterprise and the Environment, Hayes House, 75 George Street, Oxford OX1 2BQ			
TEL.:	+44 1865 270744			
EMAIL:	patrick@mcsharry.net			
FURTHER AUTHORS:	Pierre Pinson (DTU), Robin Girard (ARMINES)			
PEER REVIEWERS:	Michael Denhard (ECMWF)			
REVIEW APPROVAL:	Approved :	X	Rejected (improve as indicated below) :	
SUGGESTED IMPROVEMENTS:	*This is a really nice and well written overview of the verification methodology ! Many thanks to the authors !*			
APPROVER:	G. Kariniotakis			

VERSION HISTORY			
VERSION[2]:	DATE:	COMMENTS, CHANGES, STATUS:	PERSON(S):
V0.1	2009-09-24	Agreement of contents at 3rd Contractors meeting in Belfast	Patrick E McSharry
V0.2	2009-10-08	Receipt of case study material from Pierre Pinson	Patrick E McSharry
V0.3	2009-10-12	First complete draft version before review	Patrick E McSharry
v0.4	2009-10-12	Few corrections and additions	Robin Girard
v0.5	2009-10-13	Few corrections and additions	Pierre Pinson
v0.6	2009-10-13	Few corrections and additions (refs to other safewind reports, improvement of figures, and text on reliability diagrams)	Pierre Pinson
v0.7	2009-10-15	Final review of text	Patrick E McSharry
v0.8	2009-10-23	Corrections resulting from Michael Denhart's review	Patrick E McSharry

STATUS, CONFIDENTIALITY, ACCESSIBILITY									
STATUS:				CONFIDENTIALITY:				ACCESSIBILITY:	
S0	Approved/Released	X		**R0**	General public	**X**		Private web site	
S1	Reviewed			**R1**	Restricted to project members			Public web site	**X**
S2	Pending for review			**R2**	Restricted to European Commission			Paper copy	
S3	Draft for comments			**R3**	Restricted to WP members + PL				
S4	Under preparation			**R4**	Restricted to Task members +WPL+PL				

PL: *Project leader* **WPL:** *Work package leader* **TL:** *Task leader*

[1] **The authors of this document are solely responsible for its content, which does not represent the opinion of the European Community and the European Community is not responsible for any use that might be made of data appearing therein.**

[2] **VERSION NAMING : V0.x** draft before peer-review approval, **V1.0** at the approval, **V1.x** minor revisions, **V2.0** major revision

Contents

1. Introduction

Accurate short-term forecasts of wind farm power production are required for reliable and efficient integration of wind energy onto electrical power systems. This is especially the case for liberalised electricity markets where wind power forecasts can improve the feasibility of wind power generation. The relevant forecast horizon depends on the specific application. Forecast horizons of several hours are required for scheduling power generation, day-ahead forecasts are needed for electricity markets and forecasts over periods from days to weeks are necessary for maintenance planning. In practice, most commercial technologies for wind power forecasting rely on combinations of numerical weather predictions and statistical downscaling models (Kariniotakis *et al.*, 2006; Costa *et al.*, 2008).

Anemos, a European Union R&D project, investigated the performance of over ten different prediction systems for forecasting wind power using both statistical and physical models (www.anemos.cma.fr). As part of the project, integrated software was developed to host a number of models and these have been employed by various utilities. During the Anemos project, Madsen *et al.* (2005) developed a protocol, consisting of a set of criteria for evaluating short-term point forecasts of wind power production. The provision of a standardized approach for evaluating wind power forecasts was motivated by the economic impact of wind power and the commercial importance of being able to correctly assessing competing technologies.

SafeWind is a European Union R&D project consisting of 21 partners from 10 countries. It focuses on probabilistic forecasts of wind power production. Many weather forecasting centres now provide ensemble predictions which attempt to account for uncertainty in initial conditions and the imperfections of the equations that represent the dynamics of the atmosphere. Leutbecher and Palmer (2008) provide an overview of ensemble forecasting and recent developments in probabilistic forecasting within the meteorological community. Along with the ensemble predictions systems provided by the European Centre for Medium-ranged Weather Forecasting (ECMWF) and the National Center for Environmental Prediction (NCEP), many practitioners have become interested in employing ensembles consisting of point forecasts issued by different weather services (Gneiting *et al*., 2006).

Numerous disciplines have begun to appreciate the benefits of probabilistic forecasting. In addition to meteorological innovations, probabilistic forecasts have been studied in finance and economics (Abramson and Clemen, 1995; Tay and Wallis, 2000; Timmermann, 2000). Taylor and Buizza (2006) demonstrate the advantages of pricing weather derivatives based on density forecasts. Pinson *et al.* (2007) found that optimal management and trading of generated energy should be based on probabilistic forecasts. In practice, many practitioners and decision-makers have been reluctant to move beyond point forecasts. This may be due to a number of factors including the additional financial cost and complexity of storing and processing ensemble forecasts and the significant time needed for developing the statistical expertise to understand, implement and evaluate probabilistic forecasts. Demonstrating the economic benefits of probabilistic forecasts is of great importance and is one of the challenges of the SafeWind project.

This surge of interest in probabilistic forecasting has led to numerous publications on the most appropriate methodologies for quantifying the performance of these forecasts which may be delivered as density forecasts, quantile forecasts or prediction intervals. Jolliffe and Stephenson (2003) provide an overview of the techniques available for verifying probabilistic forecasts of categorical and continuous variables in the atmospheric sciences.

This report provides a protocol for evaluating probabilistic forecasts. Particular emphasis is placed on suitable techniques for wind power probabilistic forecasts. Note that the present report is closely related to another report, with SafeWind reference DP5.1, which focuses on the evaluation suite developed at ECMWF (Denhard *et al*. 2009). In this report, Section 2 provides an overview of the statistical measures used for quantifying the performance of point forecasts. Section 3 introduces probabilistic forecasting and describes the specific mechanisms for conveying the resulting information such as density forecasts, quantile forecasts and prediction intervals. Section 4 focuses on establishing the performance of probabilistic forecasts through particular attributes such as reliability and sharpness. We stress the need for appropriate benchmarks relating to persistence and the unconditional density or climatology when comparing probabilistic forecasts. Section 5 provides a set of guidelines in order to establish criteria for guaranteeing best practice in the commercial application

of wind power forecasting. Section 6 concludes the report with a summary of the key points to remember when evaluating probabilistic forecasts.

2. Point forecasts

The most common format for a forecast is to provide a best guess of the future value of the wind power production. Let the time series formed by observations of the wind power production at discrete times *t* be given by y_t. Any point forecast for a horizon of k steps ahead may be written as

$$\hat{y}_{t+k|t} = f(\Omega_t, k),$$

where Ω_t is the information set at time *t* consisting of all observations available up to and including this time.

The concept of this point forecast representing the forecaster's best guess of the future wind power production implies that there exists some notion of the utility function of the practitioner that will act on the information conveyed by the forecast. Another way of viewing this is to consider the implications of forecast errors of different magnitudes and sign. When selecting an appropriate model for generating point forecasts, the modeller should take account of the specific measure of forecast performance. The objective of the forecaster should be to deliver forecasts that maximise the relevant measure of forecast skill specified by the practitioner.

The presentation of forecast results is often a visual exercise, whereby competing models are compared using a selected measure of forecast accuracy (see Section 4). In order to facilitate the comparison of models, it is useful to have access to some simple benchmarks that help to establish what level of performance should be expected. We propose using the following appropriate benchmarks for this purpose.

2.1 Persistence

The persistence benchmark, also know as the random walk forecast, refers to the forecast obtained by issuing the last observation as the forecast for all future horizons:

$$\hat{y}^{per}_{t+k|t} = y_t.$$

For short forecast horizons, persistence provides a strong benchmark.

2.2 Simple moving average

For wind power time series displaying a high level of variability, it may be possible to improve the persistence benchmark by taking a simple moving average of the observations recorded during the last *m* time steps:

$$\hat{y}^{sma}_{t+k|t} = \frac{1}{m}\sum_{i=1}^{m} y_{t-i+1}.$$

2.3 Unconditional mean

A special case of the simple moving average, known as the unconditional mean and denoted by $\bar{y}$ is obtained when *m* is equal to the length of the available time series. This benchmark forecast corresponds to the long-term global average. In meteorology this benchmark is often referred to as the climatology forecast. For long-range forecast horizons, the unconditional mean provides a strong benchmark.

2.4 Weighted benchmark

As the persistence is an appropriate benchmark for short horizons and the unconditional mean is competitive at the long horizons, a stronger benchmark for intermediate horizons can be constructed by taking a weighted average where the weights are a function of the forecast horizon. Following Nielsen et al. (1998), we may write this weighted benchmark as

$$\hat{y}_{t+k|t}^{wb} = \alpha_k y_t + (1-\alpha_k)\bar{y}.$$

The parameters, α_k, should be estimated using the available training data set.

3. Probabilistic forecasts

Accurate forecasting of the wind resource up to two days ahead is recognised as a major contribution for reliable large-scale wind power integration. Especially, in a liberalised electricity market, prediction tools enhance the position of wind energy compared to other forms of dispatchable generation.

The **SafeWind** project aims to develop advanced forecasting models that will substantially outperform current methods. Emphasis is given to situations like complex terrain, extreme weather conditions, as well as to offshore prediction for which no specific tools currently exist. The prediction models are implemented in a software platform and installed for online operation at onshore and offshore wind farms by the end-users participating in the project. The project demonstrates the economic and technical benefits from accurate wind prediction at different levels: national, regional or at single wind farm level and for time horizons ranging from minutes up to several days ahead.

3.1 Ensemble forecasts

Many operational weather forecast providers now produce multiple simulations of their numerical weather prediction (NWP) model, resulting in an ensemble forecast. The ensemble provides distinct scenarios that are equally likely given the dynamical model of the atmosphere and the available initial conditions. Different members of the ensemble attempt to reflect the propagation of uncertainty in the modelling process. Such uncertainty arises from missing/erroneous observations, parametrical uncertainty and model error. There are a number of different techniques available for constructing ensemble forecasts such as singular vectors and bred vectors and an excellent review is provided by Leutbecher and Palmer (2008). The European Centre for Medium-Range Weather Forecasting (ECMWF) currently produces an ensemble forecast consisting of 50 members, a control and one deterministic high-resolution prediction. The Natural Centers for Environmental Prediction (NCEP) produces a global ensemble forecast system with 16 members.

Ensemble forecasts are often treated as the raw output from the NWP model and may require calibration to improve the statistical properties. Calibration is particularly important when employing ensemble forecasts for wind power production (Roulston *et al*., 2003; Taylor *et al*., 2009; Pinson and Madsen, 2009).

3.2 Density forecasts

A density forecast refers to a continuous probability density function for the wind power production. It provides a comprehensive description of the future for a given lead time. The variance of the density forecast can be used to convey the uncertainty associated with the forecast. We use $\hat{f}_{t+k|t}(y)$ to represent a density forecast for the wind power production issued at time t for lead time t+k. Similarly, $\hat{F}_{t+k|t}(y)$ is employed to denote the corresponding cumulative distribution function (CDF) of a probabilistic forecast and this will be referred to as a CDF forecast. A point forecast can be computed from the density forecast by calculating the mean.

3.3 Quantile forecasts

If the cumulative distribution function, $F_{t+k|t}(y)$, is a strictly increasing function, the quantile $q_{t+k|t}^{(\alpha)}(y)$ with proportion $\alpha \in [0, 1]$ of the random variable, y_{t+k}, is uniquely defined by $P(y_{t+k} < q_{t+k}^{(\alpha)}) = \alpha$ or $q_{t+k}^{(\alpha)} = F^{-1}(\alpha)$. Quantile regression was introduced by Koenker and Bassett (1978) and provides a means of estimating conditional quantile functions. The estimation of conditional quantile functions

may be achieved by regressing on observed covariates. We use $\hat{q}^{(\alpha)}_{t+k|t}(y)$ to denote the forecast for the quantile with nominal proportion α issued at time t for lead time t+k. Pinson *et al.* (2007) have employed non-parametric approaches to generate quantile forecasts of wind power. Density forecasts may be constructed by defining a family of quantile forecasts with nominal proportions that span the unit interval. A continuous density forecast may be constructed by interpolating through the family of quantiles. Gneiting *et al.* (2007) has shown that for many practitioners, their costs functions are such that optimal decision-making can be directly related to a particular quantile of the density forecast.

3.4 Prediction intervals

Prediction intervals are often employed to convey a range of values within which the verification is expected to occur with a given probability. This probability may be specified as a coverage rate $(1-\beta)$ such that $\beta \in [0, 1]$. Two selected quantiles of the density forecast may be used for this purpose. A prediction interval, $\hat{I}^{(\beta)}_{t+k|t}$, issued at time t for lead time t+k is defined by its lower and upper bounds corresponding to quantile forecasts,

$$\hat{I}^{(\beta)}_{t+k|t} = \left[\hat{q}^{(\alpha_l)}_{t+k|t}, \hat{q}^{(\alpha_u)}_{t+k|t}\right]$$

whose proportions α_l and α_u are related through $\alpha_l - \alpha_u = 1-\beta$. This definition does not uniquely define a prediction interval given a coverage rate. To provide a unique definition we must decide on how to centre the prediction interval on the forecast density. A typical approach is to define central prediction intervals by centring the intervals on the median, which ensures that there is an equal probability that the verification lies above or below the prediction interval. This places an additional constraint on the parameters in the form of $\alpha_l = 1 - \alpha_u = (1-\beta)/2$.

3.5 Risk indices

Pinson and Kariniotakis (2004) define a 'meteo-risk index' (MRI) to measure the spread of the weather forecasts at a given time. This is achieved by measuring the variance of recent forecasts. A related approach is to relate the spread of the wind power ensembles to the error of the wind power control forecast. From initial investigations of how practitioners might employ probabilistic forecasts for decision-making, it appears that such risk indices may be beneficial for conveying the associated level of uncertainty in the forecast. The approach has been further developed by Pinson *et al.* (2009a), where different types of ensemble forecasts of meteorological variables are considered as input. They include the ECMWF and NCEP ensemble forecasts, as well as a lagged-average alternative consisting of the lagged ECMWF control forecasts (5 members). They all are converted to ensemble forecasts of wind power before the risk indices are calculated. It is found by Pinson *et al.* (2009a) that the lagged-average forecasts already permit to resolve among situations with various levels of forecast uncertainty. The NCEP-based and ECMWF-based ensemble risk indices have a higher resolution, with an advantage to the latter ones.

3.6 Probabilistic benchmarks

As for the case of point forecasts, it is convenient to define appropriate benchmarks to facilitate the comparison of performance across competing methods. The following are recommended benchmarks for probabilistic forecasts.

Persistence distribution

The persistence distribution is constructed to reflect the density forecast that corresponds to issuing the persistence point forecast. While the single most recent observation is sufficient to provide the persistence point forecast, a longer record of observations is required to describe the density. For the case of wind power forecasting, we recommend using the distribution of observations occurring in the previous 12 hours to construct a benchmark for persistence.

Unconditional distribution

The unconditional distribution is formed by using all past observations to construct a probabilistic forecast. This benchmark represents the idea that the local temporal information does not provide any predictability and that the time ordering of the observations is irrelevant when constructing a forecast. Within the meteorological community, the term climatology is often used to denote the long-term average of a particular meteorological variable. The motivation behind this benchmark is similar to that of the unconditional distribution but climatology may also depend on the specific spatial location and on the time of year. In practice, long records of 25 to 40 years are employed to define these benchmarks.

Uniform distribution

The uniform distribution is chosen to lie between the minimum wind power production of zero and the maximum normalised wind power production of one. This benchmark density forecasts implies that there is an equal probability of taking on any value of wind power generation between zero and one.

4. Forecast performance

The ability to provide accurate forecasts of wind power production up to two days ahead is crucial for efficient and sustainable operation of the power system with large amounts of wind capacity. Many important planning and management decisions require knowledge of the likely performance of wind power forecasting techniques. The following sections review the general methodology and various statistical measures that may be employed for quantifying forecast performance.

4.1 Training and testing data

When constructing mathematical/statistical models for forecasting it is important to avoid over-fitting when estimating model parameters. An over-fit model is one where the estimated parameters have been tuned to provide a close fit of one particular data set with the problem that this level of accuracy cannot be sustained when facing new data sets. The objective of a forecasting model is to fit the dynamics underlying the available data but not the noise resulting from observational uncertainty such as measurements errors. This aim is particularly difficult since we are required to learn the dynamics from one particular realisation of the data. For this reason it is necessary to construct and evaluate models on different data sets in order to establish how well the model might generalise to new data. This may be achieved by using a training data set for estimation of model parameters and a testing data set for evaluation.

4.2 Point forecast evaluation

Suppose that $\hat{y}_{t+k|t}$ is a point forecast issued at time t with lead time t+k. If the actual observed value, also known as the verification, is y_{t+k}, then the corresponding prediction error is given by $\varepsilon_{t+k|t} = \hat{y}_{t+k|t} - y_{t+k|t}$.

Forecast bias

The bias of the forecast refers to the systematic error in the forecast. This quantity is estimated as the average error over the evaluation period and is computed separately for each forecast horizon:

$$BIAS(k) = \bar{\varepsilon}_k = \frac{1}{N}\sum_{t=1}^{N} \varepsilon_{t+k|t}\,.$$

Root mean square error

The root-mean-square-error (RMSE) is defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\varepsilon_{t+k|t}^{2}}\ .$$

The RMSE is associated with the assumption of independent and normally distributed forecast errors. If the model produces normally distributed forecast errors, then the maximum likelihood estimate of the parameters will coincide with the parameter values obtained by minismising the RMSE using the training data set. If one wants to minmise the RMSE error, this can be achieved by issuing the mean of the forecast density as the point forecast.

Mean absolute error

The mean absolute error (MAE) is calculated using,

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|\varepsilon_{t+k|t}\right|.$$

If one wants to minimise the MAE, then the point forecast should correspond to the median of the forecast density.

Normalised scores

Many different wind farm operators may wish to compare the forecast accuracy of competing techniques across wind farms of varying size. For this reason it is useful to consider the normalised forecast errors expressed as a fraction of the installed capacity. The corresponding values for the BIAS, RMSE and MAE can then be compared in a meaningful way as they will convey the average error as a fraction of the installed capacity.

4.3 Probabilistic characteristics

There are a number of desired characteristics that are expected from a good forecast. In the following we describe these characteristics for the specific case of a probabilistic forecast.

Reliability

Reliability refers to the degree of similarity between the forecasts and the observations. For probabilistic forecasts, one may think of the reliability as a measure of the bias of a probabilistic forecasting system. We expect that the empirical coverage achieved by each quantile forecast should equal the specified proportion. This implies that y_{t+k} should fall under the quantile forecast $\hat{q}_{t+k|t}^{(\alpha)}(y)$ in α% of the forecasts. Forecast reliability should be tested for each specific forecast horizon k. Reliability is not sufficient to characterise the quality of a probabilistic forecast since a forecast based on climatology is perfectly reliable and yet has no skill. Reliability is a fundamental property for decision-making and can be improved by recalibration methods.

Sharpness

Sharpness refers to the degree of concentration of the distribution of the probabilistic forecast. If the density forecast takes the form of a delta function, this would have maximum sharpness in that it suggests that the forecaster believes that one particular value will occur with complete certainty. This would equate with the idealised notion of a perfect point forecast (Gneiting and Raftery, 2007). In contrast, the unconditional distribution is not sharp since there is a probability that the future observation can take on any value that has been observed in the past. A uniform distribution covering

the range of previously observed values will have the lowest sharpness. Unlike reliability, sharpness is an inherent property of the forecast system and cannot be improved by recalibration techniques (Pinson *et al*., 2009b). Reliability is related to sharpness the same way bias is related to variance in deterministic evaluation. Indeed, one can artificially increase sharpness by applying a scale factor to reduce the dispersion of the forecasted distribution; this will increase sharpness but will lead to a loss of reliability.

Resolution

Resolution provides a measurement of the forecast accuracy conditional on one or more explanatory variables. In the case of wind power generation, such variables might include wind speed, wind direction and the current or predicted level of wind power production. As is the case for sharpness, resolution is an inherent characteristic of the forecast system. In meteorology, sharpness measures the ability of forecasts to deviate from the climatology, whereas resolution reflects the ability to provide different density forecasts conditional on the level of the predictand (Stephenson, 2003). In the idealised case of perfectly reliable probabilistic forecasts, these two notions are equivalent (Toth *et al.,* 2003).

Skill

Skill quantifies the overall quality of the forecast and should include measures of reliability, sharpness and resolution. While simplifying the comparison of forecasting systems, each measure of skill tends to focus differently on each of the above characteristics. We will recommend some skill measures for evaluating probabilistic wind power forecasts in Section 4.4.

Economic value

For many practitioners the appropriate measure of forecast accuracy may be directly related to the end use of the forecasting system. The forecasts may be part of a decision-making process that can be evaluated in terms of the added economic value of using one forecast over another. From this perspective one might view the forecasts as part of a cost-benefit analysis whereby the ultimate metric is the financial cost or reward measured in monetary terms. It is worth noting that each decision-making process will imply a cost function with some mathematical structure that is likely to be different from the commonly used quadratic cost function which underpins the RMSE metric. In many cases it may be impossible to describe the cost function using an analytic mathematical relationship and numerical optimisation may be the only means of selecting the best forecasting system in order to ensure an optimal decision-making algorithm. One approach to evaluating the economic value of a forecast system is to play weather roulette whereby the players can bet on their forecasts (Hagedorn and Smith, 2008).

4.4 Probabilistic evaluation

As wind power forecasting systems become increasingly available, practitioners are now faced with the decision of selecting one or more service providers. Naturally obtaining forecasts with good historical performance is one key requirement. Whilst selecting from among a set of point forecasts may be a relatively straightforward process, the decision is more complicated when considering probabilistic forecasts. The following sections outline important diagnostic tools for evaluation the characteristics of a probabilistic forecasting system and concludes with some appropriate measures of skill.

Reliability diagram

The reliability diagram provides a means of visualising the (probabilistic) bias of the probabilistic forecasting system. It may be built in different ways if considering multi-categorical events or

continuous variables. In the first case, the diagram is constructed by plotting the observed frequency of the event against the forecasted probability, where the range of forecasted probabilities is divided into bins (for example, 0-5%, 5-10%, 10-15%, etc.). The diagonal line indicates perfect reliability (the average observed frequency is equal to the predicted probability for each category), and the horizontal line represents the climatological frequency. Sometimes sample sizes are plotted either as a histogram, or as numbers next to the data points. In the scond case, when dealing with continuous variables, reliability diagrams are similar to quantile-quantile plots, in that they give the observed proportion of the various quantiles composing the predictive densities, against the nominal ones. The following evaluation approach (rank histogram) gives exactly the same information, with the difference being that it is designed for ensemble forecast evaluation.

Rank histogram

Another way of analysing the calibration of a probability forecast of an ensemble system is to construct a rank histogram. Rank histograms are usually generated for ensemble systems with a limited number of members. If the probability forecast of such an ensemble is well calibrated, the observation is equally likely to lie between any two ordered adjacent members, including the cases when the observation will be outside the ensemble range on either side of the distribution. Then the rank histogram should be flat with the same number of verifications in each interval. Especially due to the limited size of the ensemble, the observation may lay outside the ensemble range. For an ensemble system with 51 members as is the case for ECMWF this will happen 2/51 (~4%) of the time.

Probability integral transform

For continuous density forecasts, an alternative method to the rank histogram is required for assessing its calibration. The probability integral transform (PIT) corresponding to the CDF forecast, $\hat{F}_{t+k|t}(y)$, and actual observed wind power y_{t+k} is given by $z_{t,k} = \hat{F}_{t+k|t}(y_{t+k})$. This approach to evaluating probabilistic forecasts is attributed to Rosenblatt (1952). If the probabilistic forecasts correspond to the true predictive density, then the series of $z_{t,k}$ values will be iid U[0,1] (Diebold *et al*., 1998). We can therefore test whether there exists statistically significant evidence that the $z_{t,k}$ values are not iid U[0,1]. Rather than assess the $z_{t,k}$ values directly, it is advisable to define $z'_{t,k} = \Phi^{-1}(z_{t,k})$ where Φ^{-1} is the inverse normal transformation, such that the null becomes $z'_{t,k}$ iid N(0,1) which is more amenable to statistical tests (Berkowitz, 1999). This analysis of the $z_{t,k}$ values disregards their time ordering and is therefore an unconditional measure of forecast reliability in contrast with the conditional calibration introduced by Christoffersen (1998).

Logarithmic score

The logarithmic score was first proposed by Good (1952) and has been widely used under a number of names such as the predictive deviance (Knorr-Held and Rainer 2001) and ignorance score (Roulston and Smith, 2002). The log likelihood for a probabilistic forecast issued at time t with a lead time of t + k is given by $L_{t+k|t} = \ln \hat{f}_{t+k|t}(y_{t+k})$ where $\hat{f}_{t+k|t}(y_{t+k})$ is the probability estimate, provided by the density forecast, $\hat{f}_{t+k|t}(y)$, evaluated at the particular value of the observed wind power y_{t+k}. This can be calculated empirically by estimating the derivative of the CDF forecast, $\hat{F}_{t+k|t}(y)$, of wind power (Taylor *et al*., 2009). Suppose $\hat{F}_{t+k|t}(y)$ is constructed using M sampled values from the density forecast. This implies that there is an equal probability of 1/(M+1) of the actual observation falling between any two neighbouring values, below the minimum value or above the maximum value. Let Δz be the difference between the nearest values on either side of the actual observed wind power. The empirically calculated log likelihood is given by ln 1/((M + 1)Δz). The average of these log likelihoods over each forecast/verification pair provides a score for each forecast horizon,

$$LS(k) = \frac{1}{N}\sum_{t=1}^{N} L_{t+k|t} \,.$$

Continuous ranked probability score

The continuous ranked probability score (CRPS) has gained popularity as a means of evaluating probabilistic forecasts (Matheson and Winkler 1976; Hersbach, 2000; Gneiting *et al*., 2005). CRPS for a CDF forecast, $\hat{F}_{t+k|t}(y)$, and corresponding verification, y_{t+k}, is defined by taking the integral of the Brier scores for the associated binary probability forecasts at all real-valued thresholds,

$$\text{crps}(\hat{F}_{t+k|t}(y), y_{t+k}) = \int_{-\infty}^{\infty} \left(\hat{F}_{t+k|t}(y) - I(y \geq y_{t+k})\right)^2 dy \,,$$

where I(·) is an indicator function that equals 1 if the event inside the brackets is true and 0 otherwise. The average of these CRPS values over each forecast/verification pair provides a score for each forecast horizon,

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

CRPS provides what is known as a proper score, in that the forecaster minimises the expected score for an observation drawn from the probabilistic forecast, $\hat{F}$, by issuing $\hat{F}$ rather than any other competing probabilistic forecast. Another useful property of the CRPS score arises from the fact that for point forecasts CRPS reduces to the mean absolute error (MAE).

Quantile loss function

The quantile loss function, also known as the pinball loss or check function, is typically used to define a specific quantile of a distribution (Koenker and Bassett, 1978). For a particular proportion $\alpha \in [0, 1]$, the quantile loss function is a piecewise linear function given by

$$\rho_\alpha(u) = u\left(\alpha - I(u < 0)\right),$$

where u is the difference between the observed value and the estimated value.
The problem of estimating the quantile with proportion α may be written as

$$\hat{q}^{(\alpha)} = \min_q \sum_{t=1}^{N} \rho_\alpha(y_t - q) \,.$$

In addition to the quantile loss function being used for estimating the quantile in this way, it can also be employed for the evaluation of quantile forecasts. A series of quantile forecasts, $q^{(\alpha)}_{t+k|t}(y)$ issued at times t with horizon k and proportion α may be evaluated using

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

In the simple case of α=1/2, this score reduces to one half of the MAE.

Conditional evaluation

The quality of forecasts may vary significantly depending on a range of external factors. For this reason, it is important to consider the evaluation of probabilistic forecasts conditional on the level of relevant explanatory variables such as wind speed, wind direction and wind power. In addition, it may be informative to provide results conditioned on different times of the year since the forecast performance may vary throughout the year. Such conditional evaluations are important for identifying additional explanatory factors that might be employed to improve the forecast system.

4.5 Quantile forecast evaluation

We recommend evaluating the post-sample quantile forecasts using the hit percentage which assesses the unconditional coverage of the quantile forecasts. For a particular quantile forecast,

$\hat{q}_{t+k|t}^{(\alpha)}$, issued at time t with lead time t+k with verification, y_{t+k}, we define an indicator variable, $\xi_{t,k}^{(\alpha)} = I(y_{t+k} < \hat{q}_{t+k|t}^{(\alpha)})$. The time series of $\xi_{t,k}^{(\alpha)}$ yields a binary sequence that corresponds to "hits" if the verification lies below the quantile forecast and otherwise is recorded as a "miss". An analysis of $\xi_{t,k}^{(\alpha)}$ is required to assess the reliability of the quantile forecast. For each horizon, k, we can calculate the actual coverage of the quantile forecast by taking an average over the evaluation set,

$$\hat{a}_k^{(\alpha)} = \frac{1}{N}\sum_{t=1}^{N} \xi_{t,k}^{(\alpha)}.$$

In order to quantify reliability, we can measure the bias of the forecasting system by

$$b_k^{(\alpha)} = \alpha - \hat{a}_k^{(\alpha)}.$$

When summarizing forecast performance, it may be useful to provide bias values for each quantile nominal proportion, as an average over the entire length of relevant forecast horizons,

$$\overline{b}^{(\alpha)} = \frac{1}{k_{\max}}\sum_{k=1}^{k_{\max}} b_k^{(\alpha)}.$$

Engle and Manganelli (2004) introduced a dynamic quantile (DQ) test to evaluate the dynamic properties of a conditional quantile such as a quantile forecast. The DQ test involves the joint test of whether the hit variable,

$$h_t = I(y_{y+k} < q_{t+k+t}^{(\alpha)}) - \alpha,$$

is distributed iid Bernoulli with probability α and is independent of the conditional quantile estimator, $\hat{q}_{t+k|t}^{(\alpha)}$. If the probabilistic forecasting system is perfect, the time series of h_t values will have zero unconditional and conditational expectations. Engle and Manganelli (2004) and Taylor (2008) suggest using four lags of h_t in the test's regression to construct a DQ test statistic, which, under the null hypothesis of perfect unconditional and conditional coverage, is distributed $\chi^2(6)$.

4.6 Prediction interval evaluation

As a prediction interval comprises two quantiles, all of the discussion relating to a single quantile carries over to assessing the performance of a prediction interval. Indeed simply checking the coverage provided by the prediction interval is not sufficient. It is important to assess whether both quantiles required for defining the prediction interval are unbiased. One approach to testing the sharpness of the prediction intervals is to focus on the width of the intervals. For prediction intervals centred on the median with coverage rate (1-β), their width is given by

$$\delta_{t,k}^{(\beta)} = \hat{q}_{t+k|t}^{(1-\beta/2)} - \hat{q}_{t+k|t}^{(\beta/2)},$$

and a measure of the sharpness of these intervals at horizon k, is given by the average,

$$\overline{\delta}_k^{(\beta)} = \frac{1}{N}\sum_{t=1}^{N} \delta_{t,k}^{(\beta)}.$$

It may also be useful to summarise this information about the sharpness of the prediction intervals over a range of forecast horizons by taking the average,

$$\overline{\overline{\delta}} = \frac{1}{k_{\max}}\sum_{k=1}^{k_{\max}} \overline{\delta}_k^{(\beta)}.$$

4.7 Extreme event evaluation

SafeWind is particularly concerned with the use of probabilistic forecasting methods for assessing the likelihood of extreme events. Examples of these extreme events are high wind speeds leading to the wind turbines being shut down or to low wind speeds where no wind power is produced. In the following we review different measures for quantifying the skill of a forecasting system that aims to forecast specific events.

Brier score

The Brier score (*BS*) introduced by Brier (1950) measures the difference between the forecasted probability of an event and its occurrence, which is expressed as one if the event has occurred or zero otherwise. Suppose that an event is defined by the observable, y_{t+k}, falling below a threshold value, y, then the Brier score for the CDF forecast, $\hat{F}_{t+k|t}(y)$, issued at time t with lead time t+k, is given by:

$$BS(k,y) = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{F}_{t+k|t}(y) - I(y_{t+k} \le y)\right)^2 .$$

The lower the Brier score the better is the probabilistic forecasting system.

Classification analysis

Forecasts are often employed for providing the basis of a classification system regarding certain specific events. For example, we may want to know the probability that the wind speed will be greater than a particular threshold that might lead to shutdown of the wind turbine. In this case it is important to know whether or not the wind speed exceeds the cut-off speed of the turbine. In contrast to a deterministic forecast system, which predicts an event (e.g. wind speed > 25m/s) by a yes/no decision, a probabilistic forecast system assigns a probability *p* between 0 and 1 to the event. Nevertheless, users can generate dichotomous (yes/no) forecasts by specifying a threshold for the forecasted probability such that the probabilistic forecast system must then predict the event with at least this probability in order to generate a warning. To assess the forecast skill under these conditions the forecast is simplified to a yes/no statement (categorical forecast). Similarly, the observation is assigned to one of the two categories: event observed/not observed.

Contingency tables

The classification analysis forms a two-by-two matrix representing four possible outcomes for each forecast/verification pair. We then count the number of forecast/verification pairs yielding each outcome. Let *h* denote the number of "hits" corresponding to correct yes-forecasts (the event is predicted to occur and it does occur); *f* "false alarms" corresponding to incorrect yes-forecasts, *m* "misses" (incorrect no-forecasts); and *z* "zeros" for correct no-forecasts. If there are *N* forecasts in total, then $h + f + m + z = N$. A perfect forecast implies that both *f* and *m* are zero. Table 1 summarises these numbers in what is known as a contingency table. Many verification scores can be computed from these numbers and we will describe those that are relevant for wind and wind power forecast evaluation.

Table 1: *A contingency table for comparing forecasts and verifications.*

Data	observed	not observed	Total
forecasted	*h*	*F*	forecasted yes
not forecasted	*m*	*Z*	forecasted no
Total	observed yes	observed no	N

Frequency Bias (Bias score): FBIAS = (*h* + *f*) / (*h* + *m*)
compares the frequency of forecasted events to the frequency of observed events. **Range:** 0 to infinity. **Perfect score:** 1. Indicates whether the forecast system has a tendency to under-forecast (FBIAS < 1) or over-forecast (FBIAS > 1) events.

Probability of Detection (Hit Rate): POD = *h* / (*h* + *m*)
is the fraction of observed events that were correctly forecast. **Range:** 0 to 1. **Perfect score:** 1. It ignores the false alarms and can therefore be artificially improved by issuing more "yes" forecasts to increase the number of hits (over-forecast).

False Alarm Ratio: FAR = f / (h + f)
gives the fraction of the forecasted "yes" events that were false alarms. **Range:** 0 to 1. **Perfect score:** 0. Sensitive to false alarms, but ignores misses.

Probability of False Detection (False Alarm Rate): POFD = f / (z + f),
is the fraction of false alarms given the event did not occur (relative to observed "no" events). **Range:** 0 to 1. **Perfect score:** 0. It can be artificially improved by issuing fewer "yes" forecasts to reduce the number of false alarms.

Threat score (Critical Success Index): TS = h / (h + m + f)
measures the fraction of the observed and forecasted "yes" events that were correctly predicted - ignoring correct negatives. **Range:** 0 to 1, 0 indicates no skill. **Perfect score:** 1. TS is only concerned with forecasts that count. Depends on climatological frequency of events (poorer scores for rarer events) since some hits can occur purely due to random chance (-> ETS).

Equitable Threat score: ETS = (h - h_{ran}) / (h + m + f - h_{ran}),
where h_{ran} = (h + m)(h + f)/N are the hits due to random chance. **Range:** -1/3 to 1, 0 indicates no skill. **Perfect score:** 1. It corrects the TS for hits associated with random chance.

Heidke skill score: HSS = (h + z - hz_{ran}) / (N - hz_{ran}),
where hz_{ran} = [(h+m)(h+f) + (z+m)(z+f)] / N are the expected correct forecasts due to random chance. **Range:** -∞ to 1, 0 indicates no skill. **Perfect score:** 1. Measures the fraction of correct forecasts after eliminating those forecasts which would be correct due to random chance.

True skill statistic: TSS = POD - POFD
measures how well did the forecast separate the "yes" events from the "no" events. **Range:** -1 to 1, 0 indicates no skill. **Perfect score:** 1. The TSS does not depend on the climatological frequency of the event, but for rare events the TSS is weighted towards the POD term, because then most forecasts will be correct negatives and the second term (POFD) is close to zero.

Extreme Dependency Score (EDS): EDS = 2 log((h + m)/N)/(log(h/N)) - 1,
compares the fraction of the observed events with the fraction of the correctly forecasted events (see Stephenson *et al.*, 2008). **Range:** -1 to 1. **Perfect score:** 1. It has been suggested as an alternative to more common contingency table scores, since the EDS does not tend to zero for rare events.

Relative operating characteristics

In order to use a probabilistic forecasting system for providing warnings of predefined events it is important to test the skill for a particular level of probability thresholds. Using a set of increasing probability thresholds to make the yes/no decision and plotting the Hit Rates (POD) vs. the False Alarm Rates (POFD) generates the two-dimensional Relative Operating Characteristics or ROC-diagram. A point in the ROC diagram for a given probability threshold is defined by the POFD value on the horizontal axis and the POD value on the vertical axis.

The upper left corner of the ROC-diagram represents a perfect forecast system where there are no false alarms and only hits. The closer the point is to this upper left corner the higher the skill. The lower left corner, where both hit and false alarm rate are zero, represents a system which never provides any warnings of an event. The upper right corner represents a system where the event never occurs. In reality, a non-perfect system will have its values on a long convex curve pointing to the upper-left corner (the "ROC curve"). The area under the ROC curve is often used as a measure of forecast skill.

The ROC curve enables a comparison between a probabilistic and a deterministic forecast system. If the deterministic value lies above the ROC curve, the deterministic system is more skilful than the probabilistic system. In terms of utility, however, greater advantages may be gained from the probabilistic information. Only probabilistic forecast systems enable the minimisation of a user specific

cost-function. Extremely good deterministic forecasts would be required to be more useful than a probabilistic forecast. It may, however, be possible to add a good deterministic forecast to a probabilistic forecast system in order to make the combined system more beneficial than any of its individual components.

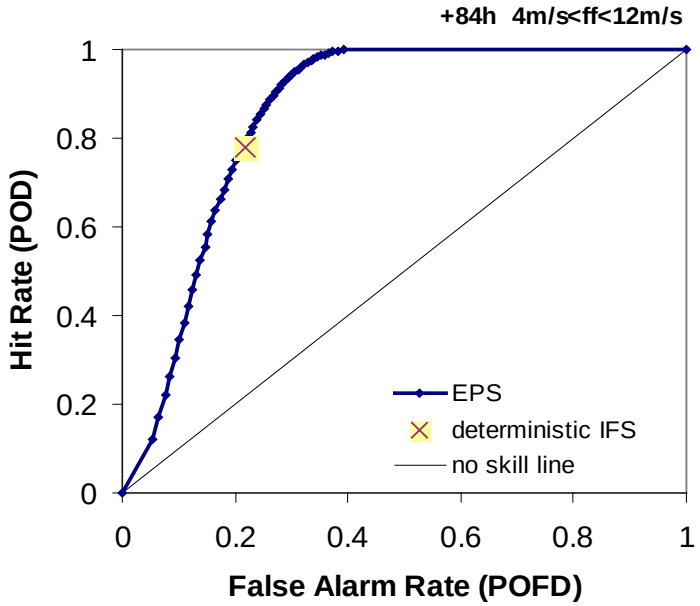


Figure 1: *Example of a ROC-diagram for hourly wind speeds in approximately 105m height between 4 and 12 m/s over Europe for the December/January/February period 2008 plotted against the probability threshold. For every probability threshold a contingency table is generated. A correct forecast (hit) was issued, if the probability of wind speeds between 4 and 12m/s exceeds the probability threshold. The squares with the crosses mark the score values for the deterministic integrated forecast system (IFS) forecast. The black vertical lines are two possible user defined thresholds for decision-making. The right line marks a threshold of 55% with maximum HSS and the left line a probability threshold of 27% with maximum TSS.*

4.8 Model comparison

It will often be necessary to quantify the gain of some advanced forecasting system to a chosen reference system. This gain, denoted as an improvement with respect to the considered reference forecast system is called a “Skill Score” and is defined as:

$$\mathrm{SkillScore(k)} = \frac{ScoreREF(k) - Score(k)}{ScoreREF(k)} = 1 - \frac{Score(k)}{ScoreREF(k)},$$

where k is the lead time of the forecast and Score is the considered evaluation criterion, which can be either a deterministic or a probabilistic measure. An appropriate reference system may be an earlier version of the model or one of a number of suitable benchmarks (see Section 3.6). A very important reference is the forecast generated from the climatological distribution of the variable under consideration (e.g. wind speed or wind power).

5. Guidelines and recommendations

As an increasing amount of research is being devoted to wind power integration studies, it is important to establish some guidelines for collected time series, constructing models and evaluating the forecast performance of competing methods. Many researchers and end-users will wish to easily translate the

results of one study to their own particular area of expertise. Given the multidisciplinary nature of wind power research it is useful to provide a protocol for evaluating forecast accuracy.

5.1 Data sampling

Metered wind power production is typically measured at intervals ranging from 5 minutes to 15 minutes. Given the nature of the fluctuations of wind speed in the location of the wind farm it is important that the forecast results are provided at a similar temporal resolution. At present this high temporal resolution is not possible for numerical weather forecasts, even though the very high-resolution models may provide output every 15 minutes (e.g. COSMO-DE). Statistical models may be employed to increase the temporal resolution of the forecasts.

5.2 Minimum testing length

In many countries it is difficult to obtain long records of wind power production. Due to the intra-annual variation in wind speeds, we suggest employing an entire year of wind power observations in order to understand the degree of predictability for each month of the year. Furthermore electricity load varies dramatically throughout the year and results over a period of at least one year are required for many important statistical analyses in wind integration studies (TradeWind, 2009; IEA, 2009).

5.3 Forecast horizons

The range of forecast horizons that are important vary depending on the specific applications of the end-user. Forecast horizons of several hours are required for scheduling power generation. Forecasts for one day-ahead are typically required for electricity markets but this depends on the exact mechanism employed by the particular market. Long-term forecasts ranging from days to weeks are usually necessary for maintenance planning. We recommend providing forecasts ranging from zero to 72 hours ahead as a means of addressing the needs of the majority of end-users in the wind energy community.

5.4 Protocol

The purpose of providing a protocol is to help standardise the process of collecting and processing wind power data, estimating models and evaluating probabilistic forecasts. This will greatly facilitate the assessment of different forecasting approaches whether they have been evaluated using the same testing data set or not. An adequate description of the operational framework is required in order to convey the relevance of any study to other practitioners and the first part of the protocol is devoted to this aspect.

- Describe the size and type of the wind power resource (capacity, number of turbines, manufacturer).
- Provide explicit details about the location and geographical setting of the wind farm (latitude, longitude, altitude, on-shore/offshore, terrain).
- Provide details of known maintenance periods where one or more turbines were turned off
- Explain and describe what sources are used for the data such as SCADA or metered data.
- Describe the sampling strategy, explaining whether the data are instantaneous readings or represent an average over a specific period of time such as the last 10 minutes before the time stamp. This information should be provided for all observed variables.
- Specify the characteristics of NWP forecasts (frequency of delivery, delay in delivery, horizon, time step, resolution, grid values or interpolated at the position of the farm).
- Specify the frequency of forecast updates. Some models produce forecasts when the NWPs forecasts are updated (i.e. every 6, 12 or 24 hours) whereas other models operate with a sliding window (typically one hour) since they consider on-line production data as an input.
- Specify the exact time periods used for estimating models and evaluating forecasts
- Employ distinct data sets for estimation (training) and evaluation (testing)
- The length and period (beginning/end) of the test set should be clearly defined
- Assess the quality of the data, giving details of missing or erroneous data. This should be

undertaken before evaluating the forecast performance
- Calibrate models for producing forecasts using at least one year of wind power production
- Evaluate forecasts on a test set consisting of at least one year of wind power observations
- Calculate a number of evaluation scores for both point forecasts and probabilistic forecasts.
- Provide MAE and RMSE for point forecasts and CRPS for probabilistic forecasts for all horizons.
- Attempt to provide the most appropriate performance measures since these will often depend on the specific application of the forecast system.
- Computer evaluation scores at forecast horizons ranging from zero to 72 hours ahead.
- Provide all evaluation scores as a percentage of the installed wind capacity in order to facilitate comparisons across wind farms and regions.
- Provide the improvement scores for comparison between models and against benchmarks.
- Attempt to present results that are appropriate for a range of end-users. For example, the interaction of forecast errors with electricity load is important for operating power systems. One approach is to quantify the forecast performance for monthly sub-periods given that the forecast performance may vary throughout the year.

6. Test cases

We introduce a standard deterministic wind power curve for facilitating comparisons of wind power forecasting studies in situations where no wind power time series are available. To increase realism it may be necessary to define a stochastic wind power curve. In order to illustrate the application of the ideas presented in the report, we provide a case study based on the evaluation of a nonparametric method for probabilistic forecasting of wind power generation.

6.1 Deterministic power curve

In many situations, researchers will not have access to wind power observations either due to the fact that no wind farms are currently in operation or because the existing data is commercially sensitive. In this case, it is often necessary to employ wind speed observations that are usually more readily available. Following Taylor *et al*. (2009), we recommend using a standard deterministic wind power curve in order to facilitate comparisons across different wind farms and regions. The nonlinear relationship is similar to that used in other studies (e.g. Roulston *et al*., 2003). The curve is cubic up to what is known as the “nominal speed” beyond which the power generated is limited by the capacity of the turbine. At the “disconnection speed,” the turbine is shut down in order to prevent damage from excessively strong wind. The standard deterministic curve that we provide here has a maximum capacity of one and can be employed with a multiplicative factor in order to scale up to the appropriate capacity if necessary. The standard wind power curve for a wind speed, v, measured in metres per second, is given by

$$f(v)=\begin{cases} av^3, & v<11 \\ bv^2+cv+d, & 11\le v\le 12.5 \\ 1, & 12.5<v\le 21 \\ 0, & v>21 \end{cases},$$

where a = 6.24e-4, b = -7.55e-2, c = 1.887 and d = 10.793. An illustration of this standard deterministic wind power curve is given in Figure 2.

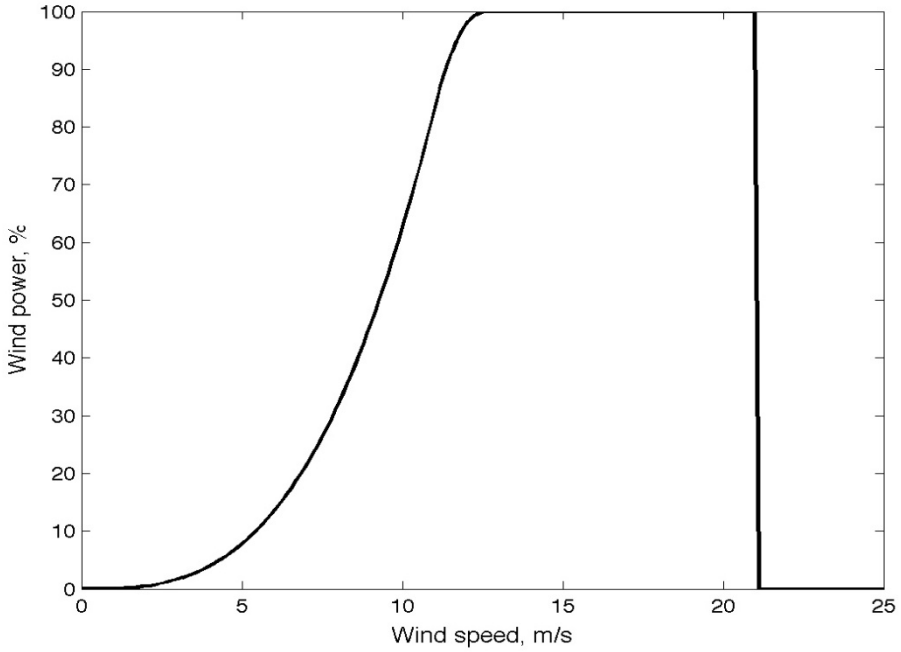


Figure 2: *Illustration of the standard deterministic wind power curve.*

6.2 Stochastic power curve

Many wind turbine manufacturers provide a power curve based on performance under ideal conditions. In practice, the power curve can vary depending on several factors, including other meteorological variables, the power control system of the specific turbine being employed and the location and topographical conditions of the wind farm. There exists substantial evidence that the wind power curve changes with these external factors and this may require a stochastic power curve as has been investigated by (Sanchez, 2006). For this reason, when comparing wind power forecasts, rather than use a single deterministic curve, it may be more realistic to employ a standard stochastic wind power curve. A stochastic power curve may be specified as $p = f(v) + \eta$ where $f(v)$ is the standard deterministic power curve from Section 6.1 and η follows an appropriate distribution to describe the stochastic variation in the power output. For example, we could sample η from a uniform band of size δ on either side of $f(v)$ such that $\eta \sim U[-\min(\delta,f(v)), \min(\delta,1-f(v))]$.

6.3 Case study

For illustration purposes, we consider here nonparametric density forecasts of wind power generation for a portfolio of wind farms located in the very North of Western Denmark (Jutland), with an installed capacity of 242.2MW. This portfolio is composed by a number of turbines of different types spread over a significant area in North Jutland. They all are located onshore, on flat terrain. No information is explicitely available regarding the maintenance planning for the turbines, though the changes in nominal capacity for the whole portfolio are reported in the dataset. Power measurements originate from SCADA systems, and correspond to hourly averages. All forecasts and measurements are normalized by this nominal capacity P_n, and therefore expressed in percentages of P_n. Forecasts are issued hourly, and have an hourly temporal resolution up to a forecast length of 43 hours. The point forecasts of wind power generation were provided by the Wind Power Prediction Tool (WPPT) as described in e.g. Nielsen *et al*. (2002), while the nonparametric density forecasts were generated based on the adapted resampling method described in Pinson and Kariniotakis (2009), while further details can be obtained from Pinson (2006). The period for which both measurements and forecasts are available goes from the 1st of January 2006 until mid-November 2007. Over this period, the first six months are used for initial training of the point and probabilistic forecasting methods, while the

remainder of the dataset is used for genuine forecasting and evaluation of forecast quality. The evaluation set considered below covers the period from July 2006 to September 2007 (both included), and includes 11,544 forecast series, over which 96.8% of the forecast-verification pairs are considered as valid values for the evaluation.

Figure 3 depicts an example with wind power point forecasts issued on the 3rd September 2007 at 1pm, related nonparametric density forecasts, as well as the corresponding measurements. Density forecasts take the form of a set of central prediction intervals (i.e. centred in probability around the median) with increasing nominal proportions from 10% to 90%. They thus are defined by 18 quantile forecasts with nominal proportions from 5% to 95% with 5% increment, except for the median.

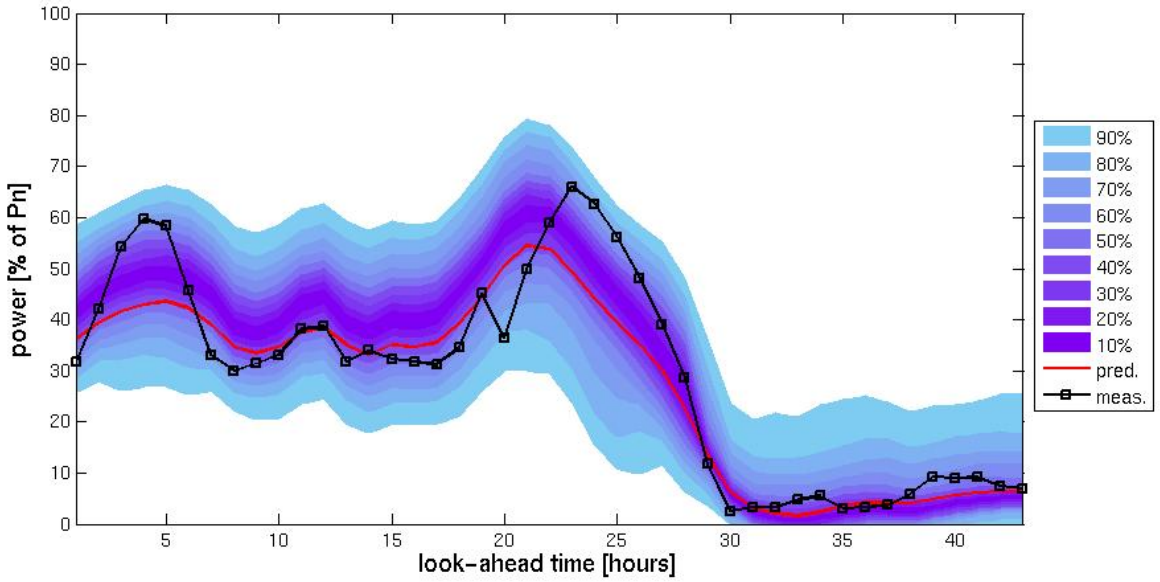


Figure 3: *Example of nonparametric density forecasts of wind power generation for a portfolio of wind farms in the North of Western Denmark (issued on the 3rd September 2007 at 1pm) in the form of a river of blood fan chart. Density forecasts are represented as a set of central prediction intervals with increasing nominal proportions. Power values are normalized by the total wind capacity Pn for the region. Measurements and point forecasts are also depicted.*

As a preliminary exploratory analysis, one may visualise a number of example plots similar to that depicted in Figure 3, in order to assess the quality of the probabilistic forecasts. This exploratory analysis is helpful for making some qualitative comments about the forecasts. A lack of sharpness/resolution may easily be identified, while reliability is more difficult to appraise visually from examples, except perahps for the quantiles with very high and very low nominal proportions, since measurements above or below such quantiles should be very rare events.

Focus on reliability

It is crucial to start the probabilistic forecast verification by assessing the probabilistic calibration of the nonparametric density forecasts of wind power generation. For that purpose, Figure 4 depicts reliability diagrams allowing visualising how the observed proportions of the quantiles defining nonparametric density forecasts deviate from the nominal ones. Two alternative visualisation tools are proposed, either as observed vs. nominal proportions, or as deviations from nominal proportions for each of the quantiles. These deviations are defined as the probabilistic bias of the quantile forecasts composing the full densities, the bias being defined here as the observed proportions minus the nominal ones. As argued by Pinson *et al*. (2007), the second visualisation may be preferred in the case for which nominal and observed quantile proportions appear very close (as is the case here). Such small deviations intuitively indicate an acceptable probabilistic calibration of the density forecasts, even though a closer look at the right hand panel in Figure 4 demonstrates deviations up to

-2.2% for low nominal proportions. This means that the quantile forecasts with nominal proportion 10% are in practice observed to be quantiles with a 8.8% proportion. An example of a qualitiative result that can be concluded from Figure 4 is that central prediction intervals appear to be too wide, since the probabilistic bias is negative for nominal proportions below 0.5, and respectively positive for those higher than 0.5.

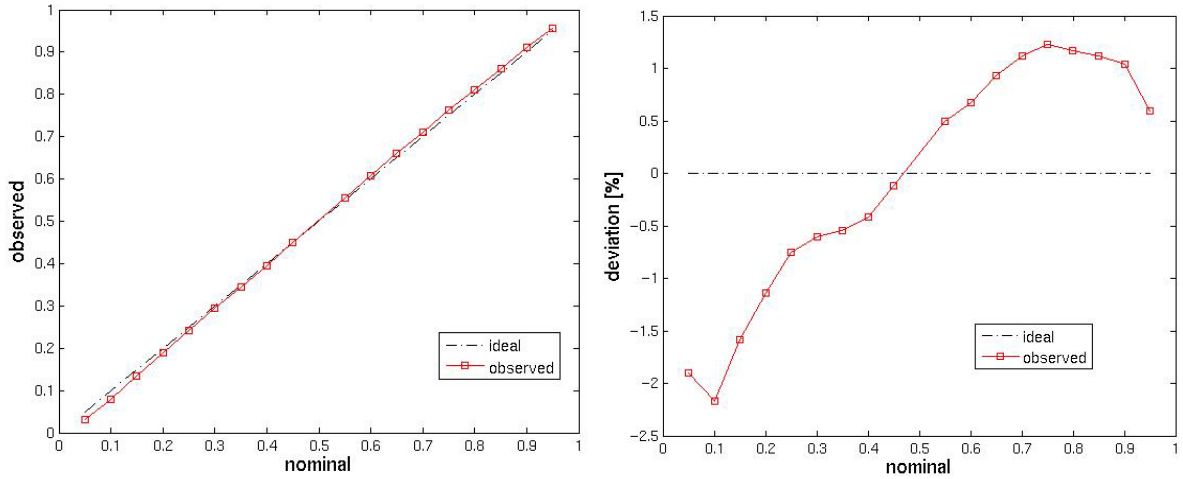


Figure 4: *Reliability diagrams for the reliability assessment of the nonparametric density forecasts of wind power generation for the period ranging from July 2006 until September 2007. Two alternatives visualisation are proposed, either as observed vs. nominal proportions, or as deviations from nominal proportions for each of the quantiles.*

The reliability evaluation performed above considers all look-ahead times indifferently. In the same way that one differentiates between nominal proportions, one may also differentiate depending on the look-ahead time, the month of the year, or other variables that might be considered as affecting the reliability of probabilistic forecasts. For nonlinear processes such as wind power generation, it is indeed intuitively expected that the shape of predictive densities will be different for low and high levels of predicted/measured power, and that this may hence affect the reliability of probabilistic forecasts. Our recommendation for wind power is that the reliability evaluation of probabilistic forecasts should be made conditional to a set of relevant explanatory/influential variables.

We first concentrate on the potential effect of the look-ahead time. The deviations from perfect reliability for all quantile forecasts considered above, and as a function of the look-ahead time, are shown in Figure 5. One clearly sees that the probabilistic bias of all quantile forecasts changes with the forecast horizon, despite being contained within an envelope of acceptable magnitude. Interestingly, there seems to be a trend such that the probabilistic forecasts are more reliable as the look-ahead time is increased.

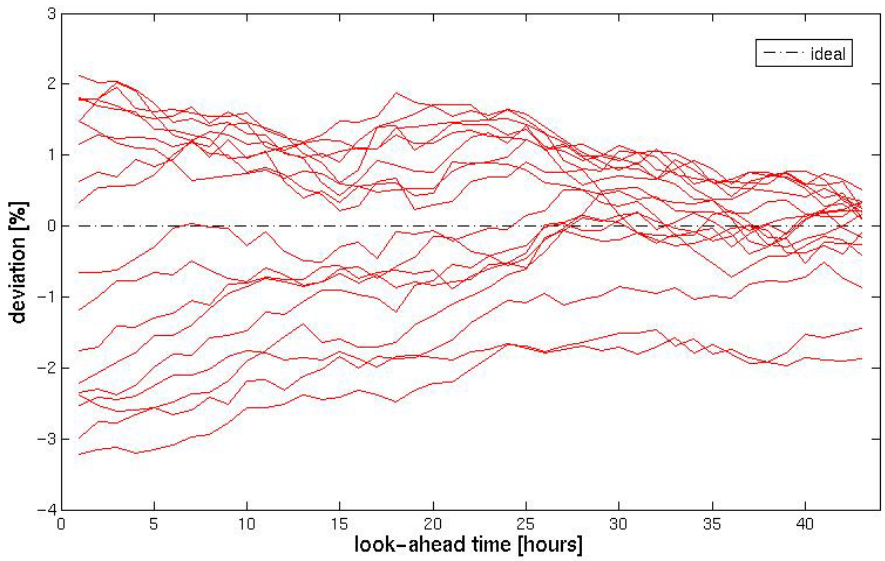


Figure 5: *Reliability diagrams (as deviations from nominal proportions) given as a function of the look-ahead time. This reliability assessment relates to the same case-study of nonparametric density forecasts of wind power generation for the period ranging from July 2006 until September 2007.*

Attention is then given to the potential effect of the level of predicted power to the reliability of the density forecasts obtained from adapted resampling. What we refer to as predicted power is the point forecasts that are used as input to the adapted resampling technique. This level of predicted power is split into five equally populated bins. A reliability diagram as in Figure 4 is produced for the set of density forecasts related to each of these bins, and then gathered in Figure 6. The figures in the legend imply that the first bin relates to the first 20% of the data, the second one to the following 20%, and so forth. As one can clearly observe from Figure 6, even if the density forecasts appear to be acceptable in terms of overall probabilistic calibration (cf. Figure 4), such densities are not reliable for low levels of predicted power. Deviations from prefect reliability of up to 20% are found in the present case. This is a deficiency of the probabilistic forecasting method that can only be identified when undertaking a conditional evaluation of the probabilistic forecast quality.

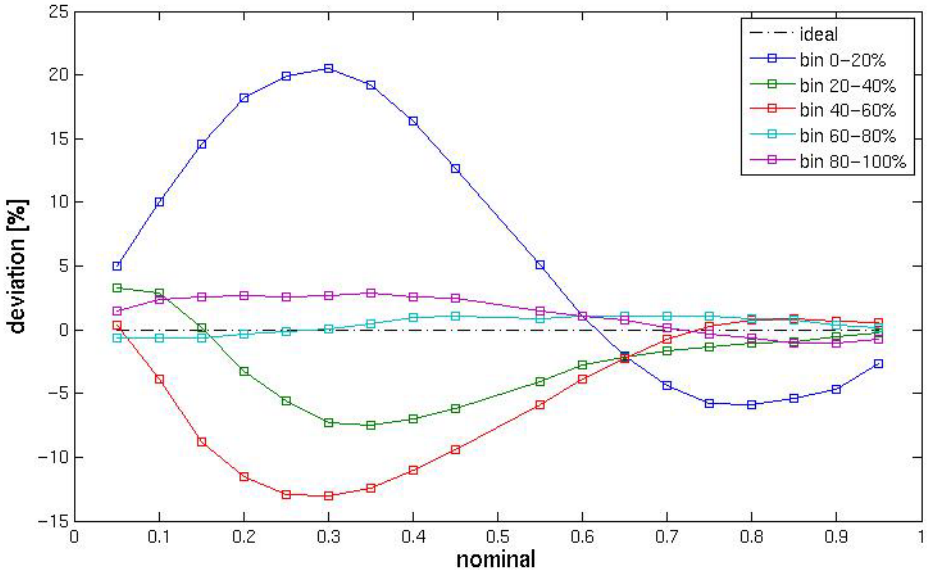
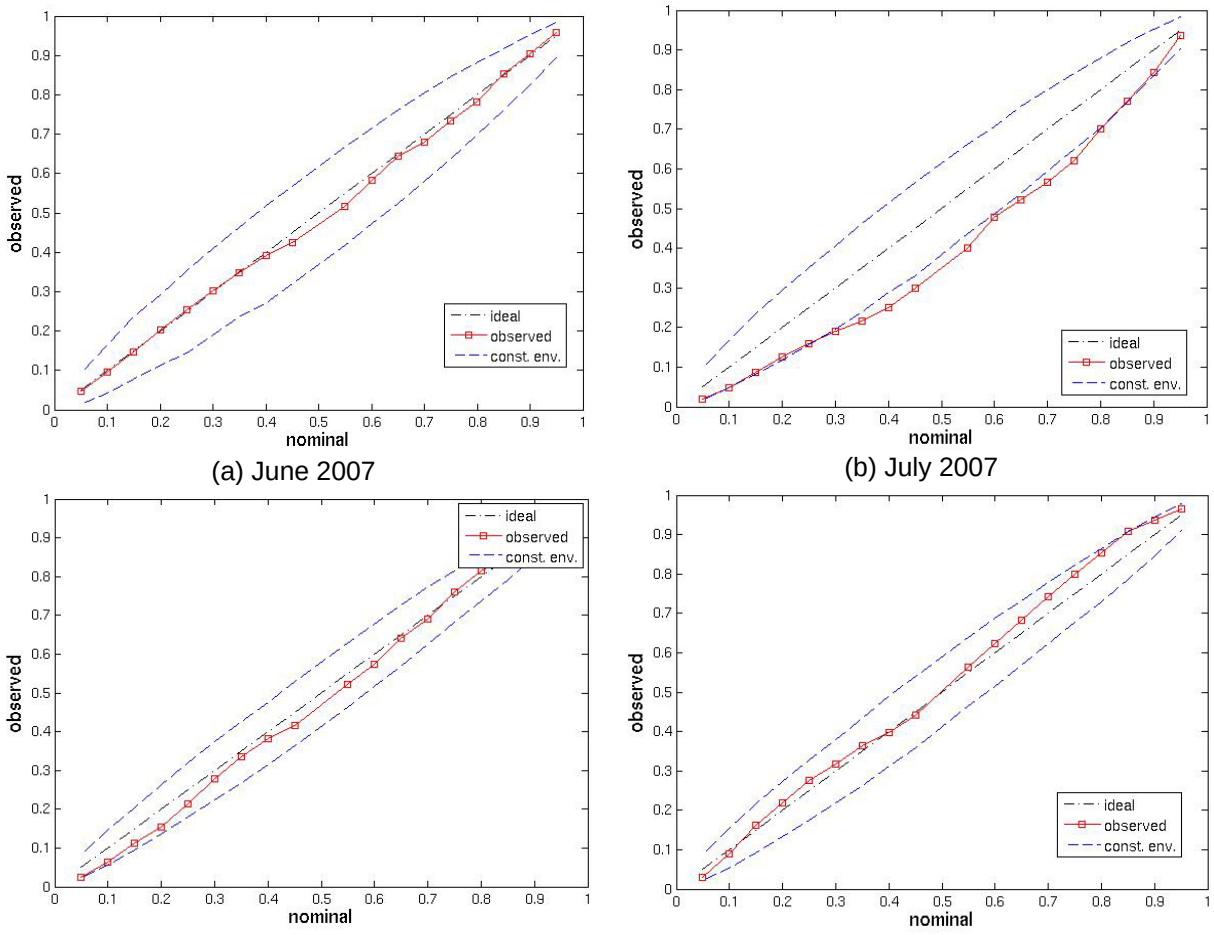


Figure 6: *Reliability diagrams (as deviations from nominal proportions, for each of the quantiles) given as a function of the level of predicted power (ie. given by the point forecasts used as input). The level of predicted power is split into five equally populated bins.*

In terms of conditional evaluation, it is also interesting to have some periodic reporting of the probabilistic forecast reliability. Having hourly updates of the wind power probabilistic forecasts in this case study, it appears reasonable to request monthly reporting of probabilistic forecast reliability. As an example, we will focus below 24-hour ahead forecasts on the period ranging from June 2007 to September 2007, for which the corresponding reliability diagrams are depicted in Figure 7. When issuing periodic reports, one may notice different levels of deviation between nominal and observed proportions of the quantile forecasts, as is the case in Figure 7. Even if probabilistic forecasts were perfectly reliable, the limited size of the samples used for forecast verification (here, between 720 and 744 hours), combined with the potential serial correlation in the sequence of forecast-verification pairs, make that deviations from the ideal diagonal case would be expected. The extent to which such deviations may be expected can be estimated e.g. by using the method described in Pinson *et al*. (2009b), allowing deriving consistency envelopes around the diagonal in reliability diagrams, for a given level of confidence. For the test case considered in Figure 7, the consistency envelopes have been determined for a 90% level of confidence. They mean that even if the sequence of forecast-verification pairs (composing the dataset for a given month) corresponded to a perfectly reliable forecast system, the observed reliability could lie anywhere within the consistency envelope, with a confidence level of 90%. In the present case when applying the surrogate consistency resampling method of Pinson *et al*. (2009b), the truncation point for the spectrum estimation has been set to 72 hours, while the number of resampling steps has been set to 1000.



(a) June 2007 (b) July 2007

(c) August 2007 (d) September 2007

Figure 7: *Reliability diagrams with 90%-consistency envelopes for the monthly reliability assessment of the nonparametric density forecasts of wind power generation. Results are produced for the months of June, July, August and September 2007.*

While the months of June and September contains the same number of forecast series and corresponding observations, the consistency envelopes are of very different magnitude. This is due to a somehow stronger serial correlation pattern present in the sequences of forecast-verification pairs for the month of June. A similar comment can be made when comparing the months of July and August: the consistency envelope for the former month is significantly wider than for the latter one.

For the months of June, August and September, the observed proportions of the quantile forecasts lie very close to the ideal diagonal. This is not the case over the month of July. For that particular month, the observed proportions even lie at the border or outside of the 90% consistency envelope, meaning that there would be very low probability of observing such proportions if the probabilistic forecasts were indeed reliable. As a conclusion, even though Figure 4 intuitively indicates that the probabilistic forecasts may be seen as reliable over the evaluation period of more than a year, the monthly reporting and evaluating carried out here shows that for the particular month of July 2007, there is strong indication that the probabilistic forecasts are not reliable.

Focus on overall skill

After having assessed the calibration of the probabilistic forecasts, we turn our attention to their overall skill, and comparison with benchmark probabilistic forecasting methods. The benchmarks we consider here are the naïve uniform and climatology ones. Remember that the naïve uniform corresponds to having no knowledge of the process of interest, and thus issuing a flat uniform density forecast at any time. And in parallel, the climatology density forecasts are issued based on a long record of measurements available onsite – here we use the whole dataset of measurements. One then issues, at any time, an unconditional density forecast that corresponds to the density of measurements. It is expected that, in terms of overall skill, the naïve uniform benchmark is the worst, while climatology contains some predictive information (and should be perfectly marginally calibrated). In meteorology, a probabilistic forecast is only considered as having some skill if its overall skill is better than that of climatology.

The overall skill of the adapted resampling method on the test case considered is evaluated using the Continuous Ranked Probability Score (CRPS). With this criterion, lower values indicate higher skill. It is depicted in Figure 8 as a function of the look-ahead time, for the adapted resampling method, as well as for the two benchmark methods employed. As mentioned above, the naïve uniform benchmark is the worst, while climatology is a more skilled benchmark. In comparison, adapted resampling appears to have high skill for look-ahead times ranging from 1 to 43-hours ahead, though decreasing as the forecast horizon increases.

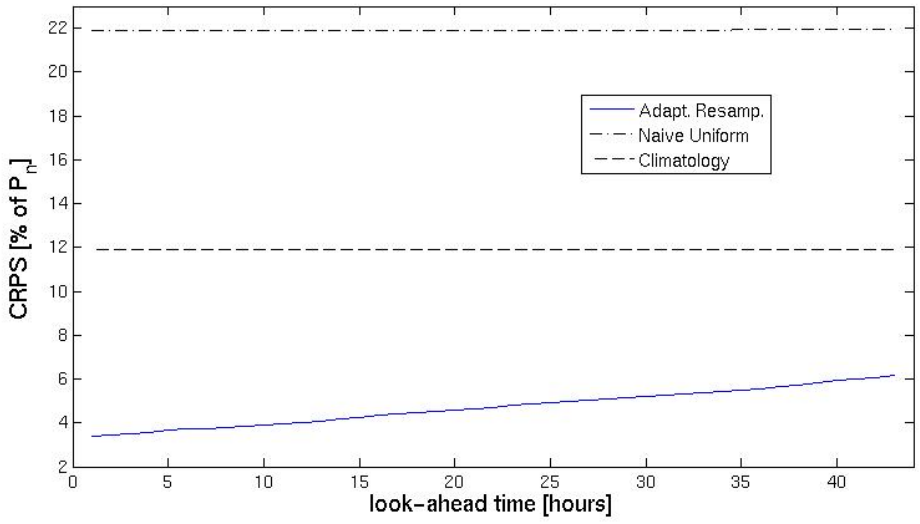


Figure 8: *Continuous Ranked Probability Score (CRPS) as a function of the look-ahead time, for the overall skill assessment of the nonparametric density forecasts of wind power generation over the period ranging from July 2006 until September 2007.*

A useful method for visualising the performance of the advanced probabilistic forecasting methods, with respect to the two benchmarks (naïve uniform and unconditional climatology densities), is to calculate the improvement in terms of CRPS (or some other criterion). This is known as a skill score and has been discussed earlier in Section 4.8. The magnitudes of the improvement measured against both benchmarks are depicted as a function of look-ahead time in Figure 9. They are highly significant, as already mentioned above, although they have a tendency to decrease with the look-ahead time, since the skill of the adapted resampling predictive densities decreases while that of the two benchmarks stays at a similar level.

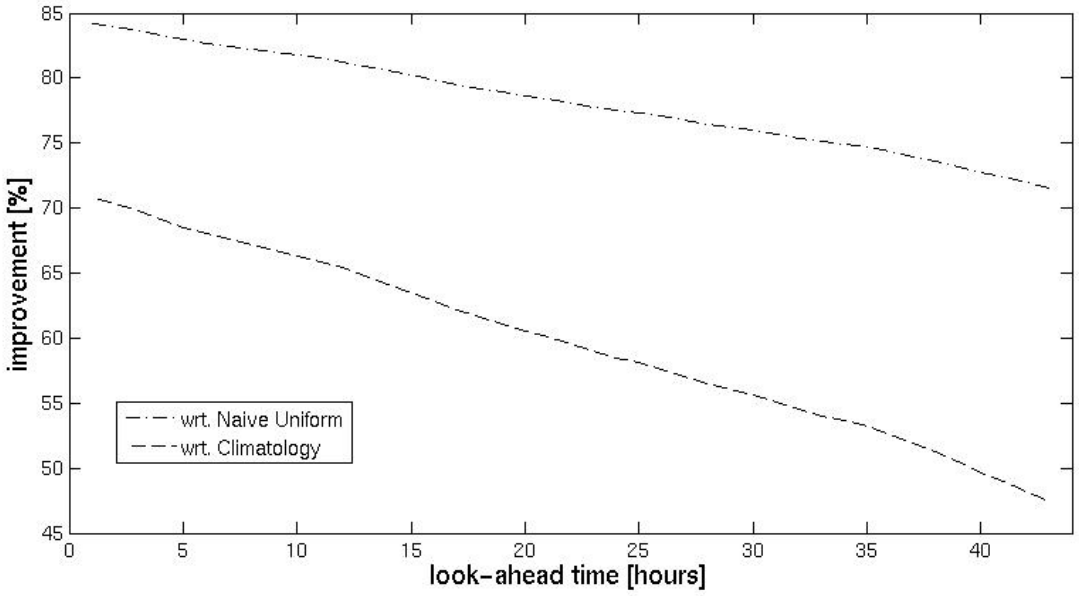
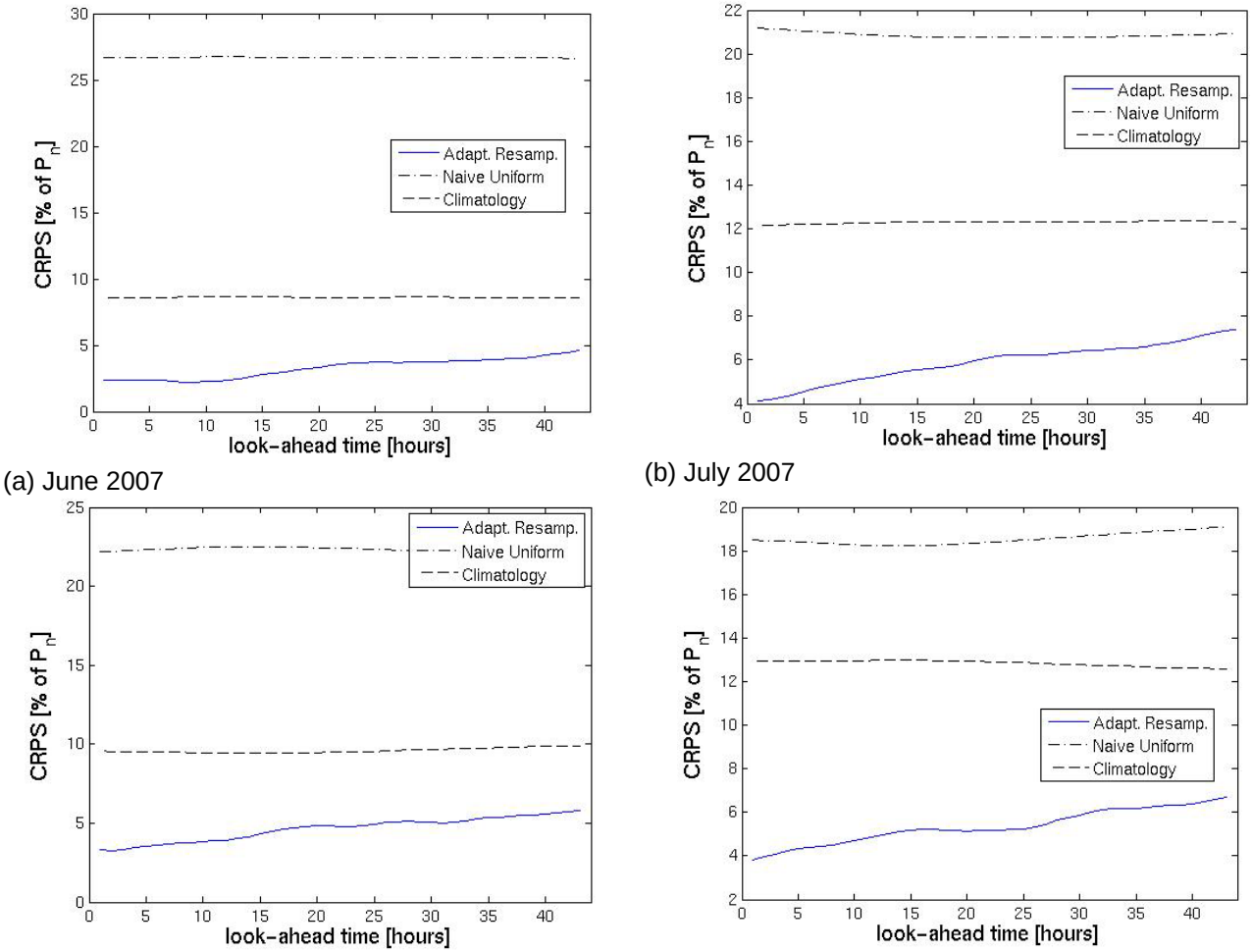


Figure 9: Improvement (in terms of *(CRPS)) shown by the adapted resampling probabilistic forecasts with respect to the two benchmarks, as a function of the look-ahead time. This evaluation relates to the whole period ranging from July 2006 until September 2007.*

As for the case of probabilistic forecast calibration, it is certainly interesting to perform a periodic evaluation/reporting of the overall skill of the probabilistic forecasting system considered here. This is achieved in Figure 10, by depicting the CRPS (as a function of the forecast horizon) for the adapted resampling method, as well as for the two benchmarks being considered, for the same months of June, July, August and September 2007.



(a) June 2007 (b) July 2007

(c) August 2007 (d) September 2007

Figure 10: *Continuous Ranked Probability Score (CRPS) as a function of the look-ahead time, for the overall skill assessment of the nonparametric density forecasts of wind power generation. This evaluation corresponds to a monthly reporting over the period ranging from June 2007 until September 2007.*

We show that for the benchmarks and the adapted resampling method, there are significant variations in the observed skill of the probabilistic forecasts from one month to the next, even though the CRPS values remain between 4 and 8% of the wind farm nominal capacity depending on the month and on the forecast horizon. As expected, the skill of probabilistic forecasts decreases as the look-ahead time increases. The observed increase in CRPS values is actually quite smooth and linear, similar to the overall skill evaluation shown in Figure 8. In terms of the comparison with the climatology benchmark, even though the skill of adapted resampling probabilistic forecasts is significantly higher, the improvements with respect to climatology vary from month to month, and depend upon the forecast horizon.

While the CRPS only gives an overall skill value for the densities, the use of a quantile skill score may allow one to focus on the skill of each quantile forecast separately (for a given nominal proportion, 50% for the median for instance). Gneiting and Ranjan (2008) have recently shown how such a decomposition with appropriate weights could be employed to obtain skill scores that focus on particular parts of the density forecast. For example, specific applications may benefit from the ability to concentrate on the central part of the forecast distribution, both tails or the individual tails. Note that a similar decomposition has been employed by Pinson *et al*. (2007) for the comparison of adapted resampling and time-adaptive quantile regression, which are two state-of-the-art methods for nonparametric probabilistic forecasting of wind power generation.

7. Conclusions

Wind power forecasting is becoming increasingly important as large amounts of wind energy are being integrated onto power systems. The ability to provide accurate probabilistic forecasts of wind power production is crucial for decision-making. There is currently a need for standardising the statistical measures being employed for quantifying the accuracy of probabilistic forecasts of wind power production. Furthermore it is crucial that a standard collection of benchmark forecasts are provided when presenting the results of studies that aim to compare wind power forecasts.

We have reviewed the considerable body of literature that exists for evaluating probabilistic forecasts. The key characteristics of a good probabilistic forecast system are discussed. Probabilistic forecasts may be employed in a number of formats including density forecasts, quantile forecasts and prediction intervals. We have shown how to construct, interpret and evaluate each of these probabilistic forecasts. The most appropriate format will depend on the objectives of the end-user of the forecasts and the specific decision-making process that relies on these forecasts.

Guidelines and recommendations for evaluating wind power predictions are presented and a minimum set of performance measures is described. When comparing the forecast performance of competing techniques, it is important not only to use appropriate performance measures, but also to use the same data since there will be large discrepancies in accuracy across data sets. One of the most important guidelines of any forecast evaluation is to select test data that has not been employed for estimating the model. This is crucial if the measured forecast performance is to be representative of future performance on new data.

In addition to the recommendations for the minimum set of performance measures, we also suggest that the resulting errors of each forecast system should be subjected to further exploratory analysis. For example, visualisation of the errors via histograms and scatter plots may help to explain the cause of errors with large magnitudes. It may be the case that these large errors result from some external influence, which can be identified using additional explanatory variables such as wind direction, wind speed, wind power or time of year. The benefits of such an exploratory analysis will lead to a deeper understanding of the forecast model and may suggest avenues for improving the model.

The case study demonstrates how one might undertake an evaluation of probabilistic forecasts of wind power production in practice. We start by describing a qualitative analysis of the probabilistic forecasts. The next step involves quantifying the reliability of the forecasts and we show how the reliability varies with both the forecast horizon and the level of the predicted wind power. Finally we employ the CRPS as a means of quantifying the skill of the probabilistic forecasts in comparison with two appropriate benchmarks. We demonstrate how the CRPS varies with the forecast horizon and the month of the year.

8. References

Abramson. B., Clemen. R. (1995). “Probability forecasting”. Int. J. Forecasting 11: 1–4.

Berkowitz, J. (2001). “Testing density forecasts, with applications to risk management”. J. Bus. Econ. Statist. 19: 465–474.

Brier, G. W. (1950). “Verification of forecasts expressed in terms of probability”. Monthly Weather Review, 78, 1–3.

Christoffersen, P.F. (1998). “Evaluating interval forecasts”. Int. Econ. Rev. 39: 841–862.

Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., Feitosa, E. (2008). “A review on the young history of the wind power short-term prediction”. Renew. Sust. Energ. Rev., 12, 1725–1744.

Denhard M. (2009). “Documentation of evaluation suite and probabilistic skill scores”, Technical Report, EU project SafeWind, Deliverable Report Dp-5.1.

Diebold, F.X., Gunther, T.A., Tay, A.S. (1998). “Evaluating density forecasts with applications to financial risk management”. Int. Econ. Rev. 39: 863–883.

Engle, R. F., Manganelli, S. (2004). “CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles.” J. Bus. Econ. Statist. 22, 367–381.

Good, I. J. (1952). “Rational Decisions,” Journal of the Royal Statistical Society, Ser. B, 14, 107–114.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005). “Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation”. Monthly Weather Review, 133, 1098-1118.

Gneiting, T., Larson, K., Westrick, K, Genton, M. G. and Aldrich, E. (2006). “Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time method”, Journal of the American Statistical Association, Vol. 101: 968-979.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). “Probabilistic forecasts, calibration and sharpness”. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 69: 243-268.

Gneiting T., Ranjan R. (2008). “Comparing density forecasts using threshold and quantile weighted scoring rules”. University of Washington, Department of Statistics, Technical Report no. 533.

Hagedorn, R. and Smith, L. A. (2008). “Communicating the value of probabilistic forecasts with weather roulette”. Meteorol. Appl., 16(2): 143-155.

Hersbach, H. (2000). “Decomposition of the continuous ranked probability score for ensemble prediction systems”. Weather and Forecasting 15, 559–570.

IEA (2009). IEA Wind Task 25 final report, phase 2006-2008, “Design and operation of power systems with large amounts of wind power”. Paris: International Energy Agency.

Jolliffe, I.T, Stephenson. D.B. (2003). “Forecast verification: a practitioner’s guide in atmospheric science”. Wiley: New York.

Kariniotakis G. (2006), "State of the art in wind power forecasting", 2nd International Conference on Integration of Renewable Energies and Distributed Energy Resources, Napa, California/USA, 4-8 December 2006, http://anemos.cma.fr

Knorr-Held, L., and Rainer, E. (2001). “Projections of Lung Cancer in West Germany: A Case Study in Bayesian Prediction,” Biostatistics, 2, 109–129.

Koenker, R. and Bassett, G. (1978). “Regression Quantiles.” Econometrica. January, 46 (1): 33–50.

Leutbecher, M., Palmer, T.N. (2008). “Ensemble forecasting”. J. Comput. Phys. 227: 3515–3539.

Matheson, J. E., and Winkler, R. L. (1976), “Scoring Rules for Continuous Probability Distributions,” Management Science, 22, 1087–1096.

Nielsen T.S., Madsen H., Nielsen H.Aa. (2002). “Prediction of wind power using time-varying coefficient functions”. Proceedings IFAC 2002, 15th World Congress on Automatic Control, Barcelona, Spain.

Petroliagis T. (2009). “Verification of probabilistic forecasts for extreme events”, Technical Report, EU project SafeWind, Deliverable Report Dp-5.1b.

Pinson, P. and Kariniotakis, G. (2004). “On-line assessment of prediction risk for wind power production forecasts”, Wind Energy 7(2): 119-132.

Pinson P. (2006). “Estimation of the uncertainty in wind power forecasting”. Ph.D. Thesis, Ecole des Mines de Paris, Paris, France.

Pinson, P., Nielsen, H.A., Møller, J.K., Madsen, H., Kariniotakis, G. (2007). “Nonparametric probabilistic forecasts of wind power: required properties and evaluation”. Wind Energy 10 (6), pp. 497-516.

P. Pinson, H.Aa. Nielsen, H. Madsen, G. Kariniotakis (2009a). “Skill forecasting from ensemble predictions of wind power“, Applied Energy 86(7-8), pp. 1326-1334.

Pinson, P. and Madsen, H. (2009). “Ensemble-based probabilistic forecasting at horns rev”. Wind Energy 12, 137-155.

Pinson P., Kariniotakis G. (2009). “Conditional prediction intervals of wind power generation”. IEEE Transactions on Power Systems, submitted.

Pinson P., McSharry P., Madsen H. (2009b). “Reliability diagrams for density forecasts of continuous variables: accounting for serial correlation”. Quarterly Journal of the Royal Meteorological Society, submitted.

Rosenblatt, M. (1952). “Remarks on a multivariate transformation”. Annals of Mathematical Statistics 23: 470-472.

Roulston, M. S., and Smith, L. A. (2002). “Evaluating probabilistic forecasts using information theory,” Monthly Weather Review, 130, 1653–1660.

Roulston, M.S., Kaplan, D.T., Hardenberg, J., and Smith, L.A. (2003), “Using medium-range weather forecasts to improve the value of wind energy production,” *Renewable Energy*, vol. 28, pp. 585–602, 2003.

Sanchez, I. (2006), “Short-term prediction of wind energy production”, International Journal of Forecasting, Vol. 22(1): 43-56.

Stephenson, D. B. (2003). Glossary. In: Jolliffe, I., Stephenson, D. (Eds.), “Forecast verification: a practitioner’s guide in atmospheric science”. John Wiley & Sons, Ltd, pp. 203–213.

Stephenson, D.B., Casati, B., Ferro, C.A.T., Wilson, C.A., (2008). “The extreme dependency score: a non-vanishing measure for forecasts of rare events”. Meteorological Applications 15(1): 41-50.

Tay, A.S., Wallis, K.F. (2000). “Density forecasting: a survey”. J. Forecasting 19: 235–254.

Taylor, J.W., Buizza, R. (2006). “Density forecasting for weather derivative pricing”. *Int. J. Forecasting 22:* 29–42.

Taylor, J.W. (2008). “Using exponentially weighted quantile regression to estimate value at risk and expected shortfall”. J. Fin. Econ. 6(3): 382-406.

Taylor, J.W., McSharry, P.E. and Buizza, R. (2009), “Wind power density forecasting using ensemble predictions and time series models”, IEEE Transactions on Energy Conversion 24(3): 775 – 782.

Timmermann, A. (2000). “Density forecasting in economics and finance”. J. Forecasting 19: 231–234.

Toth, Z., Tallagrand, O., Candille, G., Zhu, Y. (2003). Probability and ensemble forecasts. In: Jolliffe, I., Stephenson, D. (Eds.), Forecast Verification: A Practitioner’s Guide in Atmospheric Science. John Wiley & Sons, Ltd, pp. 137–164.

TradeWind (2009). “Integrating wind: developing Europe’s power market for the large-scale integration of wind power”, Final Report, February 2009. www.trade-wind.eu.