Technical University of Denmark - Informatics and Mathematical Modelling
Technical Report IMM-2007-02 (16 January 2007)

Evaluation of Nonparametric Probabilistic Forecasts of Wind Power

Pierre Pinson*, **Jan K. Møller, Henrik Aa. Nielsen, H. Madsen**
Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark

George N. Kariniotakis
Centre for Energy and Processes, Ecole des Mines de Paris, Sophia Antipolis, France

Abstract

Predictions of wind power production for horizons up to 48-72 hour ahead comprise a highly valuable input to the methods for the daily management or trading of wind generation. Today, users of wind power predictions are not only provided with point predictions, which are estimates of the most likely outcome for each look-ahead time, but also with uncertainty estimates given by probabilistic forecasts. In order to avoid assumptions on the shape of predictive distributions, these probabilistic predictions are produced from nonparametric methods, and then take the form of a single or a set of quantile forecasts. The required and desirable properties of such probabilistic forecasts are defined and a framework for their evaluation is proposed. This framework is applied for evaluating the quality of two statistical methods producing full predictive distributions from point predictions of wind power. These distributions are defined by 18 quantile forecasts with nominal proportions spanning the unit interval. The relevance and interest of the introduced evaluation framework are consequently discussed.

Key words: wind power, uncertainty, probabilistic forecasting, quantile forecasts, quality evaluation, reliability, sharpness, resolution, skill.

* *Corresponding author:*
P. Pinson, Informatics and Mathematical Modelling, Technical University of Denmark,
Richard Petersens Plads (bg. 321 - 020), DK-2900 Kgs. Lyngby, Denmark.
Tel: +45 4525 3349, fax: +45 4588 2673, email: pp@imm.dtu.dk, webpage: www.imm.dtu.dk/~pp

Contents

1 Introduction

Wind power is the fastest-growing renewable electricity-generating technology. The targets for the next decades aim at high share of wind power in electricity generation in Europe (Zervos, 2003). However, such a large scale integration of wind generation capacities induces difficulties in the management of a power system. Also, a present challenge is to conciliate this deployment with the process of the European electricity markets deregulation. Increasing the value of wind generation through the improvement of prediction systems' performance is one of the priorities in wind energy research needs for the coming years (Thor and Weis-Taylor, 2002). A state of the art on wind power forecasting has been published by Giebel et al. (2003).

Most of the existing wind power prediction methods provide end-users with point forecasts. The parameters of the models involved are commonly obtained with minimum least square estimation. Write p_{t+k} the measured power value at time $t+k$, which can be seen as a realization of the random variable P_{t+k}. Then, denote by $\hat{p}_{t+k|t}$ a point forecast issued at time t for lead time $t+k$, based on a model M, its parameters ϕ_t, and the information set Ω_t gathering the available information on the process up to time t. Estimating the model parameters with minimum least squares makes that $\hat{p}_{t+k|t}$ corresponds to the conditional expectation of P_{t+k}, given M, Ω_t and ϕ_t:

$$\hat{p}_{t+k|t} \;=\; \mathbb{E}[P_{t+k}|\mathbf{M}, \phi_t, \Omega_t] \tag{1}$$

A large part of the recent research works in wind power forecasting has focused on associating uncertainty estimates to these point forecasts. Pinson and Kariniotakis (2004) have described two complementary approaches that consist in providing forecast users with skill forecasts (commonly in the form of risk indices) or alternatively with probabilistic forecasts. The present paper focuses on the latter form of uncertainty estimates, which may be either derived from meteorological ensembles (Nielsen et al., 2004, 2006b), based on physical considerations (Lange and Focken, 2005), or finally produced from one of the numerous statistical methods that have appeared in the literature (Bremnes, 2006; Gneiting et al., 2006; Møller et al., 2006; Nielsen et al., 2006a; Pinson, 2006). They may take the form of quantile, interval or density forecasts. If appropriately incorporated in decision-making methods, they permit to significantly increase the value of wind generation. Recent developments in that direction include among others methods for dynamic reserve quantification (Doherty and O'Malley, 2005), for the optimal operation of combined wind-hydro power plants (Castronuovo and Pecas Lopes, 2004), or finally for the design of optimal trading strategies in liberalized electricity pools (Pinson et al., 2006a).

A set of standard error measures and evaluation criteria for the verification of point forecasts of wind has been described by Madsen et al. (2005). However, evaluating probabilistic forecasts is more complicated than evaluating point predictions. While it is easy to appraise a single point forecast as being false because the deviation between predicted and real values is non-negligible, an individual probabilistic forecast cannot be deemed as incorrect. Indeed, when an interval forecast states there is a 50% probability that expected power generation (for a given horizon) would be between 1 and 1.6MW and that the actual outcome equals 0.9MW, how to tell if this case should be part or not of the 50% of cases for which intervals miss?

The aim of the present report is to identify the required properties of probabilistic forecasts

of wind power, and to propose a framework for evaluating these forecasts in terms of their statistical performance (referred to as their 'quality'). The 'value' of the probabilistic forecasts, which relates to the increased benefits (i.e. monetary, CO_2 savings or others) for forecasts consumers from the use of such predictions, is not dealt with here. For a discussion on these two aspects of quality and value, we refer to Pinson et al. (2006b). Such an evaluation framework may allow forecast users to evaluate and compare rival approaches for wind power probabilistic, and forecasters to identify weak points of their methods, which will require further developments. In an operational environment the proposed criteria can be used for monitoring forecast performance.

The report is structured as follows. Section 2 concentrates on giving some definitions regarding the type of forecasts considered in the present paper. The proposed framework for probabilistic forecast evaluation is described in Section 3, with focus on practical definitions of the different aspects encompassed in the term 'quality' for probabilistic forecasts of wind power, as well as methods for their evaluation. This framework is consequently applied (Section 4) for comparing the quality of two competing methods for providing probabilistic predictions of wind power on the test case of a real-world wind farm over a period covering almost 2 years. These two methods are adaptive quantile regression (Møller et al., 2006) and adapted resampling (Pinson, 2006, Ch. 4). This case-study allows us to comment on the relevance of the described framework and evaluation criteria. Section 5 discusses some specific issues related to the sensitive aspect of reliability evaluation, while Section 6 ends the report by drawing general conclusions on the proposed evaluation framework.

2 Nonparametric probabilistic forecasts: some definitions and remarks

Write f_t the probability density function of the random variable P_t, and denote by F_t the related cumulative distribution function. Formally, provided that F_t is a strictly increasing function, the quantile $q_t^{(\alpha)}$ with proportion $\alpha \in [0,1]$ of the random variable P_t is uniquely defined as the value x such that

$$\mathrm{P}(P_t < x) \;=\; \alpha \tag{2}$$

or equivalently as

$$q_t^{(\alpha)} \;=\; F_t^{-1}(\alpha) \tag{3}$$

Then, a quantile forecast $\hat{q}_{t+k|t}^{(\alpha)}$ with nominal proportion α is an estimate of $q_{t+k}^{(\alpha)}$ produced at time t for lead time $t+k$, given the information set Ω_t up to time t. Note that only the aspects of evaluating the skill of marginal probabilistic forecasts are treated here. Marginal probabilistic forecasts are produced on a per-horizon basis, in contrast with simultaneous probabilistic forecasts, i.e. for which probabilities are defined over the whole forecast length.

Interval forecasts (equivalently referred to as prediction intervals) give a range of possible values within which the true effect p_t is expected to lie with a certain probability, its nominal coverage rate $(1-\beta)$, $\beta \in [0,1]$. A prediction interval $\hat{I}_{t+k|t}^{(\beta)}$ produced at time t for time

$t+k$ is defined by its lower and upper bounds, which are indeed quantile forecasts,

$$\hat{I}^{(\beta)}_{t+k|t} \;=\; [\hat{q}^{(\alpha_l)}_{t+k|t},\; \hat{q}^{(\alpha_u)}_{t+k|t}] \tag{4}$$

whose nominal proportions α_l and α_u are such that

$$\alpha_u - \alpha_l \;=\; 1-\beta \tag{5}$$

This general definition of prediction intervals makes that a prediction interval is not uniquely defined by its nominal coverage rate. It is thus also necessary to decide on the way they should be centred on the probability density function. Commonly, it is chosen to centre (in probability) the intervals on the median, so that there is the same probability that an uncovered true effect p_{t+k} lies below or above the estimated interval. This translates to:

$$\alpha_l \;=\; 1-\alpha_u = \frac{1-\beta}{2} \tag{6}$$

Such prediction intervals are then referred to as central prediction intervals.

If considering (assumed) Normally distributed processes, or more generally symmetric target distributions, estimated prediction intervals are centred on the point prediction $\hat{p}_{t+k|t}$ itself and give the equally probable (given $(1-\beta)$) upward and downward margins in which the true effect p_{t+k} may lie. Owing to symmetry, the mean and median of these target distributions are equal. Moreover, the upper and lower sides of the intervals have the same size. Therefore, whatever the nominal coverage rate, the point forecast $\hat{p}_{t+k|t}$ is covered by the interval forecast it is associated to. For a nonlinear and bounded process such as wind generation, probability distributions of future power output may be skewed and heavy-tailed (Pinson, 2006; Lange, 2005). For these asymmetric distributions, the median may significantly differ from the mean, and thus central prediction intervals (for rather low nominal coverage rate) may not even cover the point forecast value.

For most forecasting applications, an important question concerning the intervals arises: how to choose an optimal nominal coverage rate? This question is also valid for the case of forecast users that would be provided with a unique quantile forecast of given nominal proportion. Bremnes (2004) states that revenue-maximization strategies for trading wind generation on the Nord Pool electricity market only require a single quantile forecast only, whose nominal proportion can be directly determined from the characteristics of the market (and also provided that independence is assumed between volumes of wind generation on the market and market prices). Though, for more general trading strategies i.e. including the risk aversion of the market participant, and for which the loss function of the forecast user is more complex, the proportion of this 'optimal' quantile may be more difficult to determine, and may vary over time (Pinson et al., 2006a). Back to the case of prediction intervals, they can be seen as embarrassingly wide when the nominal coverage is set at a value of 90% or larger, since they would cover extreme prediction errors (or even outliers). In addition, working with high-coverage intervals means that one aims at modelling the very tails of distributions. Obviously, the robustness of the prediction methods becomes a critical aspect. In contrast, if one sets a low nominal coverage rate, say 50%, intervals will be more narrow and more robust with respect to extreme prediction errors. But, such low nominal coverage rate will translate to future power values being equally likely to lie inside or outside these bounds. In both cases, prediction intervals appear hard to handle and that is why an intermediate degree of confidence (75-85%) seems a good compromise (Chatfield, 2000). Consequently, instead of focusing on a particular nominal coverage rate,

producing a forecast of the whole probability distribution of expected generation may be a relevant alternative. In practice, if no assumption is made about the shape of the target distributions, a nonparametric forecast $\hat{f}_{t+k|t}$ of the density function of the variable of interest at lead time $t+k$ can be produced by gathering a set of m quantiles forecasts such that

$$\hat{f}_{t+k|t} \;=\; \{\hat{q}^{(\alpha_i)}_{t+k|t}, i = 1, \ldots, m \mid 0 \leq \alpha_1 < \alpha_2 < \ldots < \alpha_m \leq 1\} \tag{7}$$

that is, with chosen nominal proportions spread on the unit interval. These types of probabilistic forecasts are hereafter referred to as predictive distributions.

3 A framework for evaluating nonparametric probabilistic forecasts

Since it has been observed it was not reasonable to formulate assumptions regarding the shape of predictive distributions of wind power, the majority of probabilistic forecasting methods described in the literature avoid making such an assumption (Bremnes, 2006; Nielsen et al., 2006a; Pinson, 2006). This motivates the introduction of a specific framework dedicated to the evaluation of wind power probabilistic forecasts, whatever the model involved.

An evaluation set consists of series of quantile forecasts, for a unique or various nominal proportions, and observations. Let us say that this evaluation set is composed by N forecast series with forecast length $k_{\max}$. One can then apply the measures and scores introduced hereafter to this dataset, regardless of any classification. This will translate to an unconditional evaluation of the prediction quality. Though, there may be several variables that one would suspect to influence the quality of the intervals. The evaluation can then be made conditional to these variables in order to reveal their influence. For instance, it is straightforward to consider that the evaluation should be made conditional to the forecast horizon — it is indeed the case hereafter. Also, one may consider other variables e.g. level of predicted power, which are expected to impact the forecast quality. The proposed evaluation framework allows for conditional quality evaluation as illustrated in a following section.

3.1 Approach proposal: required and desirable properties

A requirement for probabilistic forecasts is that the nominal probabilities, i.e. the nominal proportions of quantile forecasts, are respected in practice. Over an evaluation set of significant size, the empirical (observed) and nominal probabilities should be as close as possible. Asymptotically, this empirical coverage should exactly equal the pre-assigned probability. That first property is commonly referred to as reliability by meteorologists (Atger, 1999). In contrast, statisticians refer to the difference between empirical and nominal probabilities as the bias of a probabilistic forecasting method (Granger et al., 1989; Taylor, 1999). Consequently, this requirement of reliability of a given method translates to the probabilistic predictions being unbiased.

Besides this requirement, it is highly desirable that probabilistic predictions provide fore-

cast users with a situation-dependent assessment of the prediction uncertainty. Their size should then vary depending on various external conditions. For the example of wind power forecasting, it is intuitively expected that prediction intervals (for a given nominal coverage rate) should not have the same size when predicted wind speed equals zero and when it is near cut-off speed. In the meteorological literature, the sharpness of probabilistic forecasts is defined as the ability of these forecasts to deviate from the climatological mean probabilities, whereas resolution stands for the ability of providing different conditional probability distributions depending on the level of the predictand (Stephenson, 2003). For probabilistic forecasts with perfect reliability, these two notions are equivalent (Toth et al., 2003).

Note that our proposal for the evaluation of sharpness and resolution will derive from a more statistical point of view with focus to the shape of predictive distributions. Resolution is more generally considered as the ability of providing probabilistic forecasts conditional to the forecast conditions. This is because for a weather-related process such as wind generation, not only the level of the predictand but also some other explanatory variables e.g. wind direction may have an influence on the prediction uncertainty. In parallel, sharpness is seen as the property of concentrating the probabilistic information about future outcome. This definition derives from the idea that reliable predictive distributions of null width would correspond to perfect point predictions. A similar definition has been given by Gneiting and Raftery (2004) when discussing the skill of probabilistic forecasts, and this definition is implicit in the proposal by Roulston and Smith (2002) of using the ignorance score which is based on the entropy of predictive distributions.

The framework proposed by Christoffersen (1998) for interval forecast evaluation, and which is widely used among the econometric forecasting community (Wallis, 2003; Clements, 2005), consists in testing the hypothesis of correct conditional coverage of the prediction intervals. Such framework has been introduced for the specific case of one-step ahead prediction intervals. It can be easily shown that this is equivalent to testing the correct unconditional coverage of the intervals, as well as their independence. However, for the case of wind power forecasting, one has to consider multi-step ahead predictions for which there exists a correlation among forecasting errors.[1] Prediction intervals hence cannot be independent. Instead of applying Christoffersen's framework, it appears preferable to develop an evaluation framework based on an alternative paradigm: reliability is seen as a primary requirement while sharpness and resolution represent the inherent value of the method. While reliability can be increased by using some re-calibration methods (e.g. conditional parametric models (Nielsen et al., 2006b) or smoothed bootstrap (Hall and Rieck, 2001)), sharpness and resolution are invariant properties that cannot be enhanced by applying post-processing method (Toth et al., 2003).

3.2 Reliability

Nonparametric probabilistic predictions as defined above either comprise a single quantile forecast, or consist in a collection of quantile forecasts for which the nominal proportions are known. Hence, evaluating the reliability of probabilistic predictions is achieved by

[1] The correlation among forecasting errors mainly originates from the inertia in the meteorological prediction uncertainty. In addition, if the wind power prediction model includes an autoregressive part, it will also contribute to the correlation of errors in forecasts for successive look-ahead times. For the class of statistical stuctural models, the dependency among forecasting errors can be explicitly formulated, see (Madsen, 2006) for instance.

verifying the reliability of each individual quantile forecast.

Let us in a first stage introduce the indicator variable $\xi_{t,k}^{(\alpha)}$. Given a quantile forecast $\hat{q}_{t+k|t}^{(\alpha)}$ issued at time t for lead time $t+k$, and the actual outcome p_{t+k} at that time, $\xi_{t,k}^{(\alpha)}$ is given by

$$\xi_{t,k}^{(\alpha)} \;=\; \mathbf{1}_{p_{t+k}<\hat{q}_{t+k|t}^{(\alpha)}} \;=\; \begin{cases} 1, & \text{if} \quad p_{t+k} < \hat{q}_{t+k|t}^{(\alpha)} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The time-series $\{\xi_{t,k}^{(\alpha)}\}$ ($t = 1, \ldots, N$) of indicator variable is then a binary sequence that corresponds to the series of 'hits' (if the actual outcome p_{t+k} lies below the quantile forecast) and 'misses' (if otherwise) over the evaluation set. It is by studying $\{\xi_{t,k}^{(\alpha)}\}$ that one can assess the reliability of a time series of quantile forecasts. Indeed, an estimate $\hat{a}_k^{(\alpha)}$ of the actual coverage $a_k^{(\alpha)} = \mathbb{E}[\xi_{t,k}^{(\alpha)}]$, for a given horizon k, is obtained by calculating the mean of the $\{\xi_{t,k}^{(\alpha)}\}$ time-series over the test set:

$$\hat{a}_k^{(\alpha)} \;=\; \frac{1}{N}\sum_{t=1}^{N_T} \xi_{t,k}^{(\alpha)} \;=\; \frac{n_{k,1}^{(\alpha)}}{n_{k,0}^{(\alpha)} + n_{k,1}^{(\alpha)}} \tag{9}$$

where $n_{k,1}^{(\alpha)}$ and $n_{k,0}^{(\alpha)}$ correspond to the sum of hits and misses, respectively. They are calculated with:

$$n_{k,1}^{(\alpha)} \;=\; \#\{\xi_{t,k}^{(\alpha)} = 1\} \;=\; \sum_{t=1}^{N} \xi_{t,k}^{(\alpha)} \tag{10}$$

$$n_{k,0}^{(\alpha)} \;=\; \#\{\xi_{t,k}^{(\alpha)} = 0\} \;=\; N - n_{k,1}^{(\alpha)} \tag{11}$$

This measure of empirical coverage serves as a basis for drawing reliability diagrams, which give the empirical probabilities versus the nominal ones for various nominal proportions. The closer to the diagonal the better. In the present paper, reliability diagrams instead give the deviation from the 'perfect reliability' case for which empirical proportions would equal the nominal ones. They then give the bias of the probabilistic forecasting method for the nominal proportion α, calculated as the difference between these two quantities:

$$b_k^{(\alpha)} \;=\; \alpha - \hat{a}_k^{(\alpha)} \tag{12}$$

This idea is similar to the use of Probability Integral Transform (PIT) histograms as proposed by Gneiting et al. (2005) except that reliability diagrams directly provide that additional information about the bias of the method considered.

In addition, these diagrams allow one to summarize the reliability assessment of various quantile forecast series with different nominal proportions, and thus to see at one glance if a given method tends to systematically underestimate (or overestimate) the uncertainty. Figure 1 depicts an example of a reliability diagram that may serve for assessing the reliability of predictive distributions produced by a state-of-the-art method. Bias values are calculated for each quantile nominal proportion, as an average over the forecast length,

$\bar{b}^{(\alpha)} = 1/k_{max} \sum_k b_k^{(\alpha)}$. For instance, the bias is of 0.9% for the quantile with nominal proportion 0.6. In other words, the observed coverage for that quantile is of 59.1% instead of the required 60%. For the example of Figure 1, the reliability of the quantile forecasts can be appraised as rather good since all deviations are lower than 2%. However, the fact that quantiles are slightly overestimated for proportions lower than 0.5 and slightly underestimated for proportions above that value indicates that corresponding predictive distributions are slightly too narrow. Note that if calculating the overall bias $\bar{\bar{b}}$ of predictive distributions for this test-case, it would clearly be close to 0. Such calculation would dilute the information relative to each single quantile, which does not appear desirable. This remark is also valid for the case of evaluating the reliability of nonparametric prediction intervals: only checking if the nominal coverage rate of the intervals is respected is not sufficient. It is indeed necessary to verify that both quantiles defining the interval are unbiased.

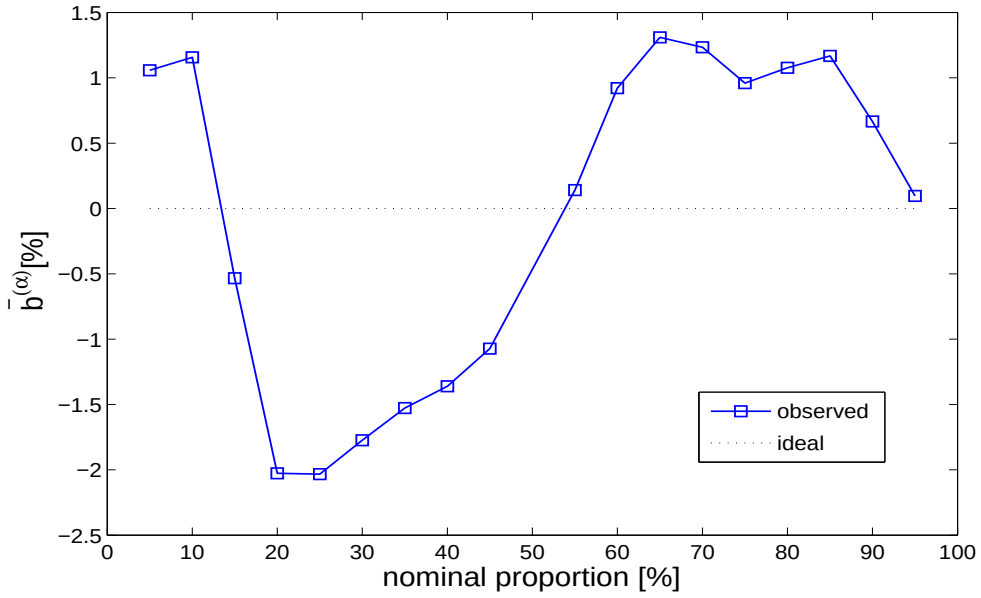

Figure 1: *Example of a reliability diagram depicting deviations as a function of the nominal proportions, for the reliability evaluation of a method providing probabilistic forecasts of wind generation.*

When focusing on point forecasting for non-linear processes, Tong (1995) explains that the quality of point prediction methods may significantly be driven by some external factors, and thus that the quality of such methods should be evaluated as a function of the level of explanatory variables, for different subperiods of the evaluation set, etc. A similar approach should be applied here with the aim of evaluating the correct conditional coverage of a given method. Correct conditional coverage can therefore be defined by: "whatever the chosen grouping of the forecast/observation pairs from the evaluation, probabilistic predictions should be reliable". The interest of using such definition of correct conditional coverage will be illustrated in a following section.

3.3 Sharpness and resolution

Remember that the proposed definition for sharpness corresponds to the ability of probabilistic forecasts to concentrate the probabilistic information about future outcome. Hence, an intuitive approach to the evaluation of sharpness for the case of interval forecast relates to studying the distribution of their size over the evaluation set. For instance, Bremnes (2006) summarizes these distributions with boxplots. Our proposal, following previous analyses by Nielsen et al. (2006b) and Pinson et al. (2006c), is to focus on the mean size of the intervals only. If writing

$$\delta_{t,k}^{(\beta)} = \hat{q}_{t+k|t}^{(1-\beta/2)} - \hat{q}_{t+k|t}^{(\beta/2)} \tag{13}$$

the size of the central interval forecast (with nominal coverage rate $(1-\beta)$) estimated at time t for lead time $t+k$, a measure of sharpness for these intervals and for horizon k is given by $\bar{\delta}_k^{(\beta)}$, the mean size of the intervals:

$$\bar{\delta}_k^{(\beta)} = \frac{1}{N}\sum_{t=1}^{N}\delta_{t,k}^{(\beta)} = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{q}_{t+k|t}^{(1-\beta/2)} - \hat{q}_{t+k|t}^{(\beta/2)}\right) \tag{14}$$

Obviously, this measure cannot be used if aiming at evaluating one quantile forecast only. For the case of predictive distributions, for which forecasts are defined by a set of quantile forecasts, one can gather quantile forecasts by pairs, in order to obtain a set of central prediction intervals with different nominal coverage rates. One can then use summarize the evaluation of the sharpness of predictive distributions with δ-diagrams, which give $\bar{\delta}_k^{(\beta)}$ as a function of the nominal coverage rate of the intervals. Such diagrams permit to better appraise the shape of predictive distributions.

δ-diagrams can be drawn over the whole forecast length, i.e. by depicting $\bar{\bar{\delta}}^{(\beta)} = 1/k_{max}\sum_k \bar{\delta}_k^{(\beta)}$ as a function of the nominal coverage rate of the intervals. However, as it is known that the uncertainty of power predictions is significantly influenced by the forecast horizon, it is commonly accepted that a specific uncertainty estimation model should be setup for each look-ahead time, and that their evaluation should be carried out similarly. Wind power generation is a process for which the prediction uncertainty is situation-specific and highly variable. More than the forecast horizon, this uncertainty may be influenced by several explanatory variables such as the level of predicted power or wind speed for instance. The resolution property has been defined as the ability to generate different probabilistic information depending on the forecast conditions. Note that predictive distributions must still be reliable. Thus, resolution can then be further defined as the ability of providing different predictive distributions under the requirement of conditional reliability. For its evaluation, one can draw δ-diagrams for different groupings of the forecasts conditions, and compare the average shape of predictive distribution.

3.4 A unique skill score

As for point-forecast verification, it is often demanded that a unique skill score would give the whole information on a given method performance. Such a measure would be given by scoring rules that associate a single numerical value $\mathrm{Sc}(\hat{f}, p)$ to a predictive distribution $\hat{f}$

if the event p materializes. Then, we can define as

$$\text{Sc}(\hat{f}', \hat{f}) = \int \text{Sc}(\hat{f}'(p), p)\hat{f}(p)dp \tag{15}$$

the score under $\hat{f}$ when the predictive distribution is $\hat{f}'$.

Even if sharpness and resolution as introduced above are intuitive properties that can be visually assessed with diagrams, they can only contribute to a diagnostic evaluation of the method. They cannot allow one to objectively conclude on a higher quality of a given method. In contrast, a scoring rule such as that defined above, if proper, would permit to do so. The propriety of a scoring rule reward a forecaster that expresses her true beliefs. Murphy (1993) refers to that aspect as the forecast 'consistency' and states that a forecast (probabilistic or not) should correspond to the forecaster's judgment. If we assume that a forecaster wishes to maximize her skill score over an evaluation set, then a scoring rule is said to be proper if for any two predictive distributions $\hat{f}$ and $\hat{f}'$ we have

$$\text{Sc}(\hat{f}', \hat{f}) \leq \text{Sc}(\hat{f}, \hat{f}), \quad \forall \hat{f}, \hat{f}' \tag{16}$$

The scoring rule Sc is said to be strictly proper if equation (16) holds with equality if and only if $\hat{f}' = \hat{f}$. Hence, if $\hat{f}$ corresponds to the forecaster's judgment, it is by quoting this particular predictive distribution that she will maximize her skill score. The propriety of various skill scores defined for continuous density forecasts is discussed by Bröcker and Smith (2006b).

If producing nonparametric probabilistic forecasts by quoting a set of m quantiles with various nominal proportions (cf. equation (7)), it can be shown that any scoring rule of the form

$$\text{Sc}(\hat{f}, p) = \sum_{i=1}^{m} \left(\alpha_i s_i(\hat{q}^{(\alpha_i)}) + (s_i(p) - s_i(\hat{q}^{(\alpha_i)}))\xi^{(\alpha_i)} + h(p) \right) \tag{17}$$

with $\xi^{(\alpha_i)}$ the indicator variable for the quantile with proportion α_i, s_i non-decreasing functions and h arbitrary, is proper for evaluating this set of quantiles (Gneiting and Raftery, 2004). If $m = 1$, this resumes to evaluating a single quantile with nominal proportion α, while the case $m = 2$ with $\alpha_1 = \beta/2$ and $\alpha_2 = 1 - \beta/2$ relates to the evaluation of a prediction interval with nominal coverage rate $(1 - \beta)$. $\text{Sc}(\hat{f}, p)$ is a positively rewarding score: a higher score value stands for an higher skill. In addition, the skill score introduced above generalizes scores that are already available in the literature. For instance, for the specific case of central prediction intervals with nominal coverage rate $(1 - \beta)$, one retrieves an interval score that has already been proposed by Winkler (Winkler, 1972) by putting $\alpha_1 = \beta/2$ and $\alpha_2 = 1 - \beta/2$, $s_i(p) = 4p$, $(i = 1, 2)$, and $h(p) = -2p$. In parallel, if focusing on a single quantile only, the scoring rule given by equation (17) generalizes the loss functions considered for model estimation in quantile regression (Koenker and Basset, 1978; Nielsen et al., 2006a; Møller et al., 2006) and local quantile regression (Bremnes, 2006). This loss function is used here for defining the scoring rule for each quantile, i.e. with $s_i(p) = p$, and $h(p) = -\alpha p$. Consequently, the definition of the skill score introduced in equation (17) becomes

$$\text{Sc}(\hat{f}, p) = \sum_{i=1}^{m} (\xi^{(\alpha_i)} - \alpha_i)(p - \hat{q}^{(\alpha_i)}) \tag{18}$$

This score is positively oriented and admits a maximum value of 0 for perfect probabilistic predictions.

Using a unique proper skill score allows one to compare the overall skill of rival approaches, since scoring rules such as that given above encompass all the aspects of probabilistic forecast evaluation. However, a unique score cannot tell what are the contributions of reliability or sharpness and resolution to the skill (or to the lack of skill).[2] The skill score given by equation (17) cannot be decomposed as this can be done for the case of the continuous ranked probability score (Hersbach, 2000). Though, if reliability is verified in a prior analysis, relying on a skill score permits to carry out an assessment of all the remaining aspects, namely sharpness and resolution.

4 Application results

In the above sections, the framework for the evaluation of nonparametric probabilistic forecasts in the form of a single quantile forecasts, or of a set of quantile forecasts, has been described. The case study of a wind farm for which probabilistic forecasts are produced with two competing methods is considered. The various properties making the quality of the methods considered are studied here.

4.1 Description of the case-study

Predictions are produced for the Klim wind farm, which is a 21MW wind farm located in the North of Jutland, in Denmark. The nominal power of that wind farm is hereafter denoted by P_n. The period for which point predictions are generated goes from March 2001 until end of April 2003. Hourly power measurements for that wind farm are also available over the same period. The point predictions result from the application of the WPPT method (Nielsen et al., 2002), which uses meteorological predictions of wind speed and direction (with an hourly temporal resolution) as input, as well as historical measurements of power production. Meteorological predictions have a forecast length of 48 hours and are issued every 6 hours from midnight onwards. But then, points predictions of wind power are issued every hour: they are based on the most recent meteorological forecasts and are updated every time a new power measure becomes available. They thus have a varying forecast length: from 48-hour ahead for power predictions generated at the moment when meteorological predictions are issued, down to 43-hour ahead for those generated 5 hours later. In order to have the same number of forecast/observation pairs for each look-ahead time, the study is restricted to horizons ranging from 1- to 43-hour ahead. All predictions and measures are normalized by the nominal power P_n of the wind farm, so that that they are all expressed in percentage of P_n.

Two competing methods are used for producing probabilistic forecasts of wind generation. These methods are the adapted resampling method described by Pinson (2006) and the adaptive quantile regression method introduced by Møller et al. (2006). They both use the level of power predicted by WPPT as unique explanatory variable. A specific model is set up for each look-ahead time. The memory length allowing time-adaptivity of the methods is

[2] This has already been stated by Roulston and Smith (2002) when introducing the 'ignorance score', which despites its many justifications and properties has no ability to tell why a given method is better than another.

chosen to be of 300 observations. In order to obtain predictive distributions of wind power, each method is used to produce 9 central prediction intervals with nominal coverage rates of 10, 20, ..., and 90%. This translates to providing 18 quantile forecasts with nominal proportions going from 5 to 95% by 5% increments, except for the median. Figure 2 gives an example of such probabilistic forecasts of wind generation, in the form of a fan chart.

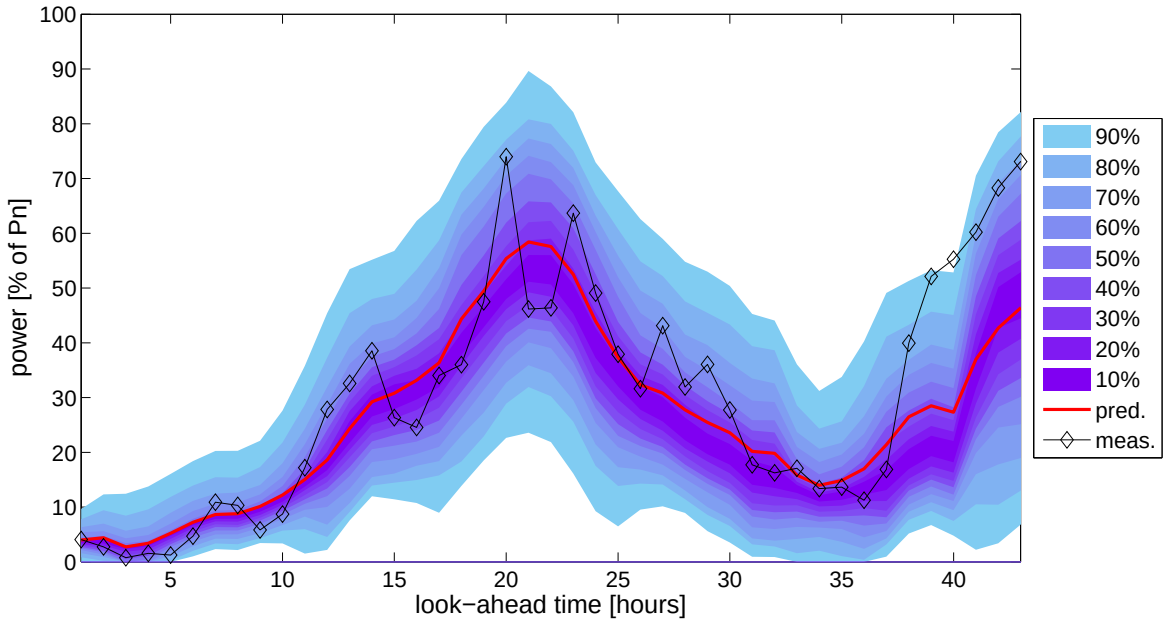

Figure 2: *Example of probabilistic predictions of wind generation in the form of nonparametric predictive distributions. Point predictions are obtained from wind forecasts and historical measurements of power production, with the WPPT method. They are then accompanied with interval forecasts produced by applying the adapted resampling method. The nominal coverage rates of the prediction intervals are set to 10, 20, ..., and 90%.*

The first 3 months of data are utilized for initializing the methods and estimating the necessary parameters. The remaining of the data is considered as an evaluation set. After discarding missing and suspicious forecast/observation pairs, this evaluation set consists of 14685 series of hourly predictions.

4.2 Reliability assessment

Reliability is assessed first, since it has been defined as a primary requirement. Time-series of indicator variables are generated by separately considering time-series of quantile forecasts for each method, for each look-ahead time, and for each nominal proportion. By calculating the overall bias $\bar{\bar{b}}$ for both methods, i.e. over the whole range of nominal proportions and look-ahead time, one obtains the values given in Table 1. These bias values are very low, indicating the ability of the methods to globally respect the nominal probabilities. Though, this single value may dilute the information about a method's reliability, and this property should then be evaluated conditionally to some variables. Here, the reliability of the methods is studied for each nominal proportion (Figure 3), and also as a function of the look-ahead time (Figure 4).

Table 1: *Overall bias for both the adapted resampling and adaptive quantile regression methods. The bias is calculated as the mean deviation from perfect reliability over the whole range of forecast horizons, and over the whole range of nominal proportions.*

Method	Adapted resampling	Adaptive quantile regression
$\bar{b}[\%]$	0.218	0.082

The deviations from perfect reliability are small for both methods over the whole range of nominal proportions, except for the very low ones (5 and 10%). Since distributions of power output are highly right-skewed for low levels of predicted power, it is more difficult to predict in a very reliable way quantiles whose values are close to 0. It is interesting to see that the adapted resampling method tends to underestimate the quantiles with very low proportions while the adaptive quantile regression method tend to overestimate them. On a more general basis, predictive distributions are slightly too narrow. Note that these very low bias values are to be related to the size of the evaluation set. Since this set is large it is expected to witness low bias values.

For the two methods considered in the present paper, a specific model is used for each look-ahead time. Evaluating reliability as a function of the look-ahead time may allow one to detect some undesirable behaviour of the chosen method for probabilistic forecasting. From Figure 4, one sees that the bias of both methods is small over the whole forecast length, and that there is no trend that would consist in the bias increasing as the forecast lead time gets further. Though, the bias for the adapted resampling method is significantly positive for all look-ahead times, which is due to the relatively large positive bias values for nominal proportions 0.05 and 0.1 (cf. Figure 3). Due to the varying maximum forecast length of the prediction series, the amount of data for evaluation of reliability is 1/6th of the length of the evaluation set for look-ahead time 48, 1/3rd for look-ahead time 47, etc. This has to be taken into account when appraising the values of the evaluation criteria in the present study.

4.3 Evaluation of the quality of the methods

A necessary statement before to carry on with the evaluation of sharpness or of the overall quality of the methods is that they are reliable. This statement appears to be reasonable in view of the reliability assessment carried out in the above paragraph.

Focus is now given to the sharpness of the predictive distributions produced from both methods. Figure 5 gathers δ-diagrams drawn for specific forecast horizons, i.e. those related to 1-hour ahead, 12-hour ahead and 30-hour ahead predictions, as well as an average over the forecast length. An example information that can be extracted from these δ-diagrams is that for 1-hour ahead predictions, both methods generate prediction intervals of nominal coverage 90% — which has been considered as unconditionally reliable — that have a size of 19% of P_n. This information on the size of the intervals is of particular importance for practitioners who will use these intervals for making decisions. By comparing the δ-diagrams for the three different look-ahead times, one sees that predictive distributions are less sharp for further look-ahead time, reflecting that point predictions are less accurate. The sharpness of both methods is very similar, with the adapted resampling method

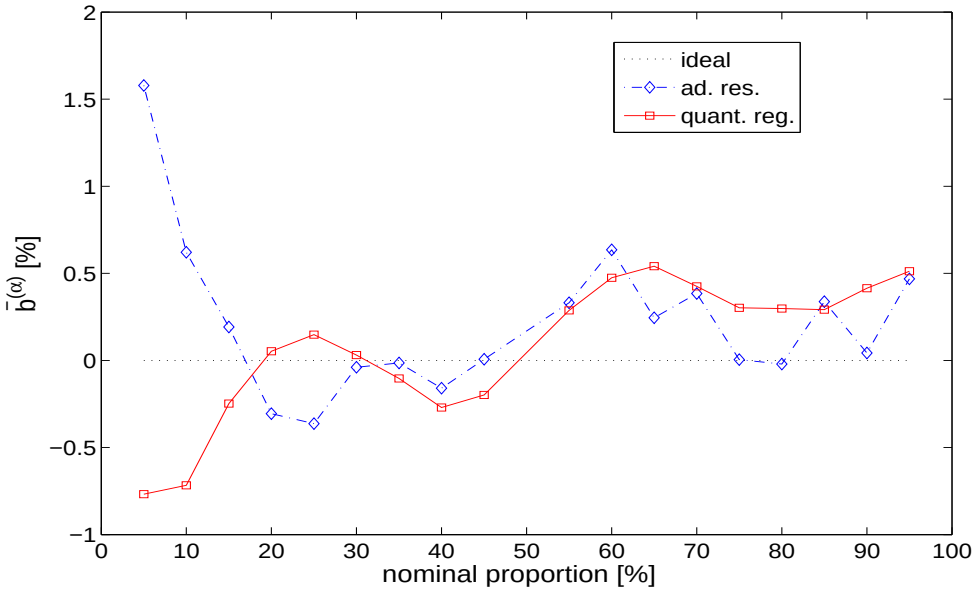

Figure 3: *Reliability evaluation: bias values for each of the quantile nominal proportion, for both the adapted resampling and adaptive quantile regression method. Bias values are given as averages over the forecast length.*

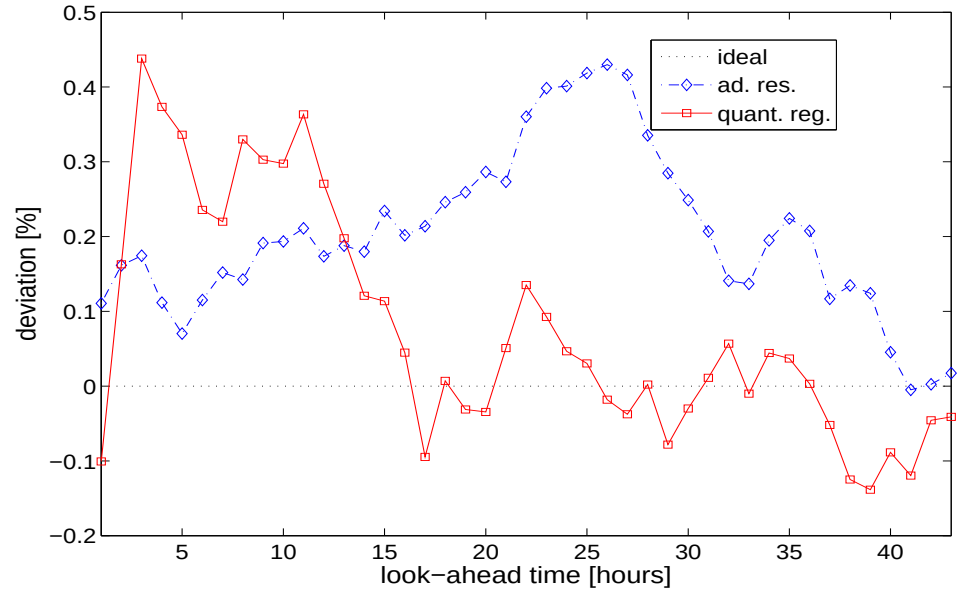

Figure 4: *Reliability evaluation: bias as a function of the look-ahead time, for both the adapted resampling and adaptive quantile regression method. Bias values are given as averages over the 18 different quantile nominal proportions.*

being sharper in the central part of the predictive distributions and adaptive quantile regression sharper in the tail part. This may indicate that the adaptive quantile regression method is more robust with respect to extreme prediction errors or outliers.

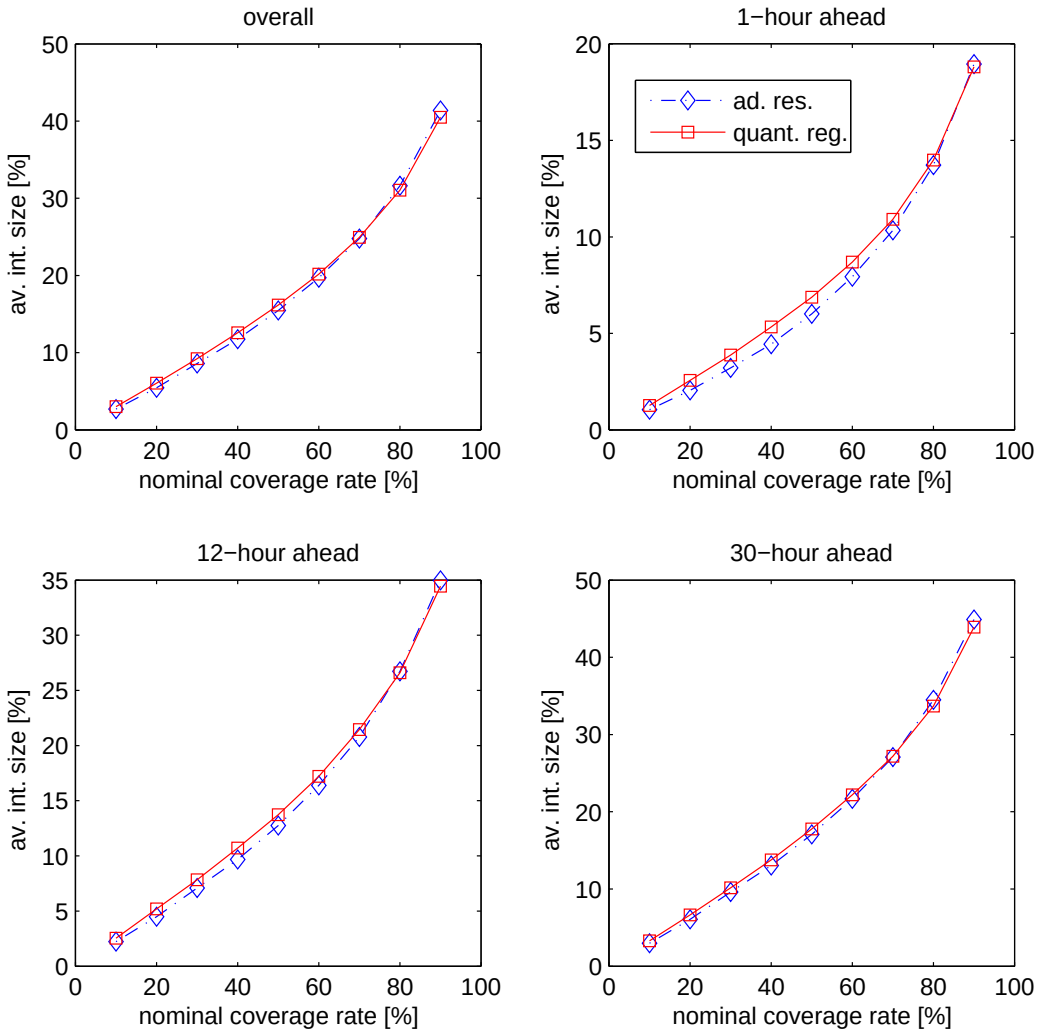

Figure 5: *Sharpness evaluation: δ-diagrams giving the sharpness of predictive distributions produced from the adapted resampling and adaptive quantile regression method. These diagrams are for 1-hour ahead, 12-hour ahead and 30-hour ahead forecasts, as well as an average over the forecast length.*

The overall quality of predictive distributions obtained from the adapted resampling and adaptive quantile regression methods is then evaluated by using the skill score given by equation (18). Skill score values are calculated at each forecast time and for each forecast horizon. When averaged over the evaluation set, the skill score as a function of the look-ahead time is obtained, as depicted in Figure 6. The overall skill score value, summarizing the overall quality of the methods by a unique numerical value, equals -0.65 for adapted resampling and -0.64 for adaptive quantile regression. This tells that the latter method globally has a higher skill than the former one. In addition, Figure 6 shows the skill of adaptive quantile regression (for this test case) is slightly higher for each individual look-ahead time. This appear reasonable in regard to our comments such that adaptive quantile regression was globally more reliable and such that both methods had similar sharpness.

However, when focusing on prediction intervals with a 50% nominal coverage rate, adapted resampling has been found more reliable and sharper than adaptive quantile regression, but the latter method still has a higher skill score than the former one. This may appear surprising, but actually the decisions on acceptable reliability and higher sharpness from reliability and δ-diagrams are subjective. They do not have the strength of the propriety of the skill score. This finding indicates that some behaviours of the methods (desirable or unwanted) are not visible from such global evaluation. A conditional evaluation of the quality of the methods will permit to reveal these aspects.

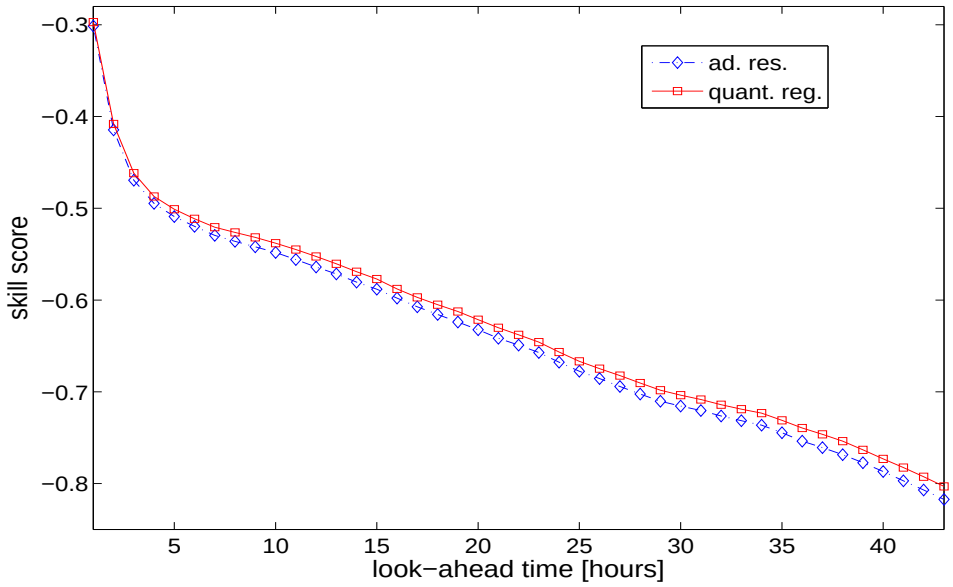

Figure 6: *Evaluation of the quality of the two methods with the skill score. This score is calculated for the whole predictive distributions and depicted as a function of the look-ahead time.*

4.4 Resolution analysis from a conditional evaluation

Both probabilistic forecasting methods considered here use point predictions of wind power as explanatory variable. The resulting probabilistic predictions should be conditional to the level of this variable and still reliable. This relates to the wanted resolution property of the probabilistic forecasting methods. Reliability of predictive distributions is hereafter further assessed as a function of the level of the predictand. The conditional reliability of probabilistic predictions is highly desirable. If the process considered was homoskedastic, this conditional evaluation of reliability would not appear as necessary. It could also be of interest here to study the conditional reliability of predictive distributions given some other explanatory variable e.g. predicted wind speed or direction. This may give some insight on additional variables to consider as input to the probabilistic forecasting methods. However, the aim of the present paper is to illustrate the interest of such evaluation and not to carry out the full evaluation exercise.

Because values of predicted quantiles (depending on the nominal proportion) may not span the whole range of possible power production values, it is decided to split the evaluation set in a number n_{bin} of equally populated classes of point prediction values. This contrasts

with the possibility of defining classes from threshold power values, which could result in evaluating reliability over power classes with very few pairs of forecast/obervation. This exercise is carried out with $n_{\text{bin}} = 10$. Table 2 gives the minimum, maximum and mean predicted power values for every classes. One clearly sees from this Table that the distribution of predictions is concentrated on low power values. The 10% smallest power prediction values are comprised between 0 and 1.48% of P_n, while the 10% largest values are between 52.92% an 94.67% of P_n. Bias values are calculated for each nominal proportions, but over the whole forecast length since no specific behaviour that would be related to the forecast horizon has been observed. Figure 7 depicts the results of this exercise for 4 out of the 10 the power classes, i.e. the classes 2, 6, 8 and 9. The reliability diagrams for all power classes are gathered in Figures 10 and 11 in the Appendix.

Table 2: *Characteristics of the equally populated classes of predicted power values used for the conditional evaluation of the probabilistic forecasting methods. Each class contain 10% of the predicted power values.*

Class	Min. power value [% P_n]	Mean. power value [% P_n]	Max. power value [% P_n]
1	0	0.38	1.48
2	1.48	2.97	4.49
3	4.49	5.97	7.43
4	7.43	9.12	10.98
5	10.98	13.22	15.58
6	15.58	18.28	21.19
7	21.19	24.56	28.36
8	28.36	32.87	37.91
9	37.91	44.70	52.92
10	52.92	66.21	94.67

The size of the dataset used for drawing each of these reliability diagrams is only 10% of that used for drawing the reliability diagram of Figure 3. Therefore, larger deviations from perfect reliability may be considered as more acceptable. Still, the dataset contains 1485 forecast/observation pairs each, and bias values such as those witnessed for the power class 2 are significantly large. For this class of predicted power values, bias values are up to 16% for the adapted resampling method. They do not reach such level for adaptive quantile regression, but they are nonetheless significant (up to 10%). An interesting point is that the adapted resampling method largely underestimate the quantiles with low nominal proportions, i.e. they are too close to the zero-power value, while the other method does the inverse. Note that power predictions for this power class are contained between 1.48% and 4.49% of P_n. For such power prediction values, distributions of wind power output are highly right-skewed and with a high kurtosis. In other words, they are very peaked and sharp close to the zero-power value with a long thin tail going towards positive power values. In such case, it is very difficult to accurately predict the quantiles with low nominal proportions. In addition, such deviations from perfect reliability express deviations in terms of probabilities. In terms of numerical values, predicted quantiles must be very close to the real ones in this range of predicted power values.

Concerning the other reliability diagrams of Figure 7, the power classes considered are more related to the linear part of the power curve, for which predictive distributions are

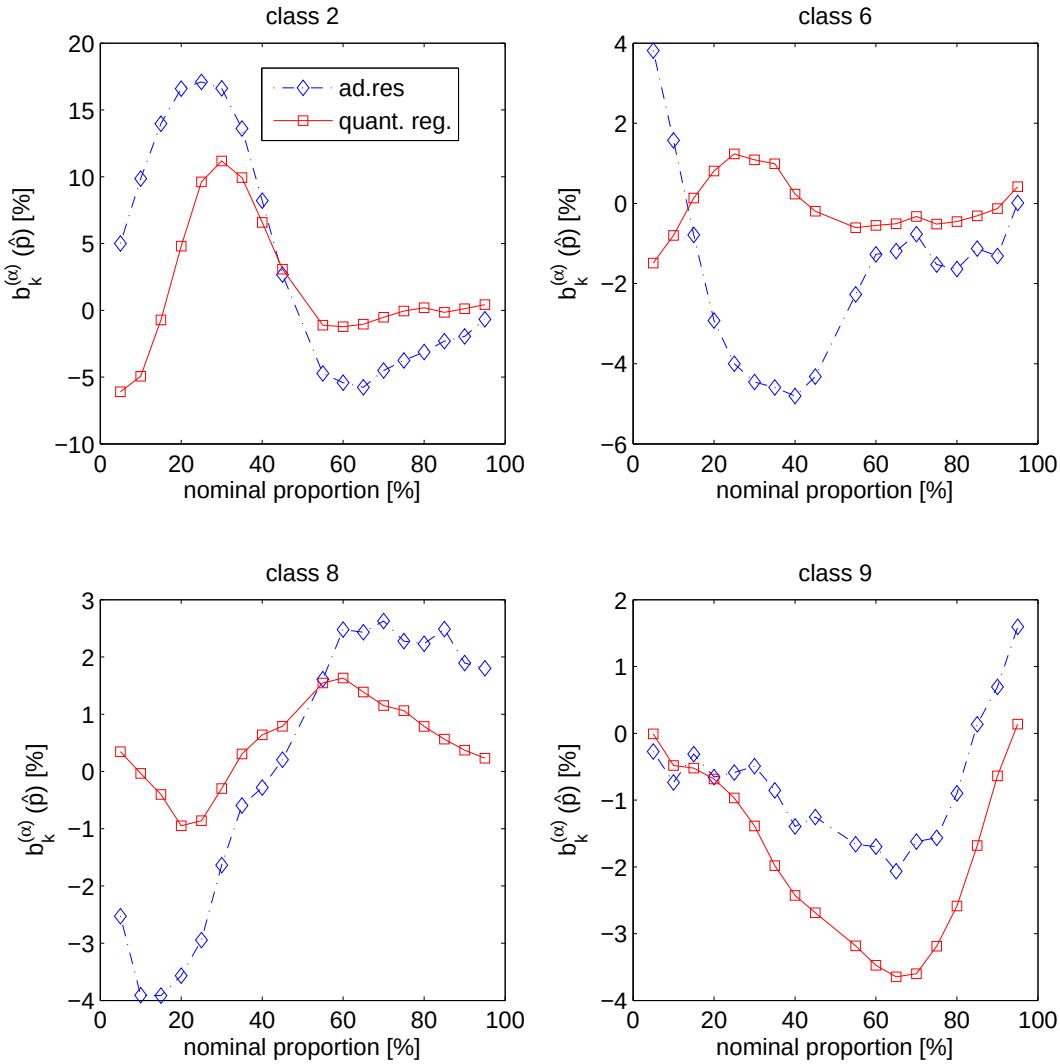

Figure 7: *Conditional reliability evaluation: reliability is assessed as a function of the level of predicted power. Forecast/observation pairs are sorted in 10 equally populated classes of predicted power values. Reliability diagrams are given for power classes 2, 6, 8 and 9.*

more symmetric and less peaked. The reliability diagram related to the power class 9 gives an example of adapted resampling being more reliable that adaptive quantile regression for some range of power values. But actually, for 7 out of the 10 power classes (cf. Figures 10 and 11 in the Appendix), the latter method has been found to be more reliable than the former one, i.e. with lower bias values over the whole range of quantile nominal proportions. This tells that for this test case adaptive quantile regression is actually more conditionally reliable than adapted resampling. This is particularly valid for the power classes related to low predicted power values (power classes from 1 to 5 in Figure 10). In this range of predicted power values, the deviations from perfect reliability for adapted resampling reach very high levels, while those for quantile regression are contained in a reasonable range (except for power class 2, surprinsingly).

The conditional evaluation of sharpness and skill (conditional to the level of predicted power) is given in Figures 8 and 9, respectively. Figure 8 depicts the δ-diagrams for the 4 power classes considered above. Sharpness is calculated as an average over the whole forecast length, and is representative of the evaluation that could be carried for each look-ahead time. Figure 9 shows skill diagrams that give the value of the skill score for each quantile separately, averaged over the whole forecast length. All the results related to the conditional evaluatio of sharpness are gathered in Figures 12 and 13, while those for the conditional evaluation of skill are gathered in Figures 14 and 15 .

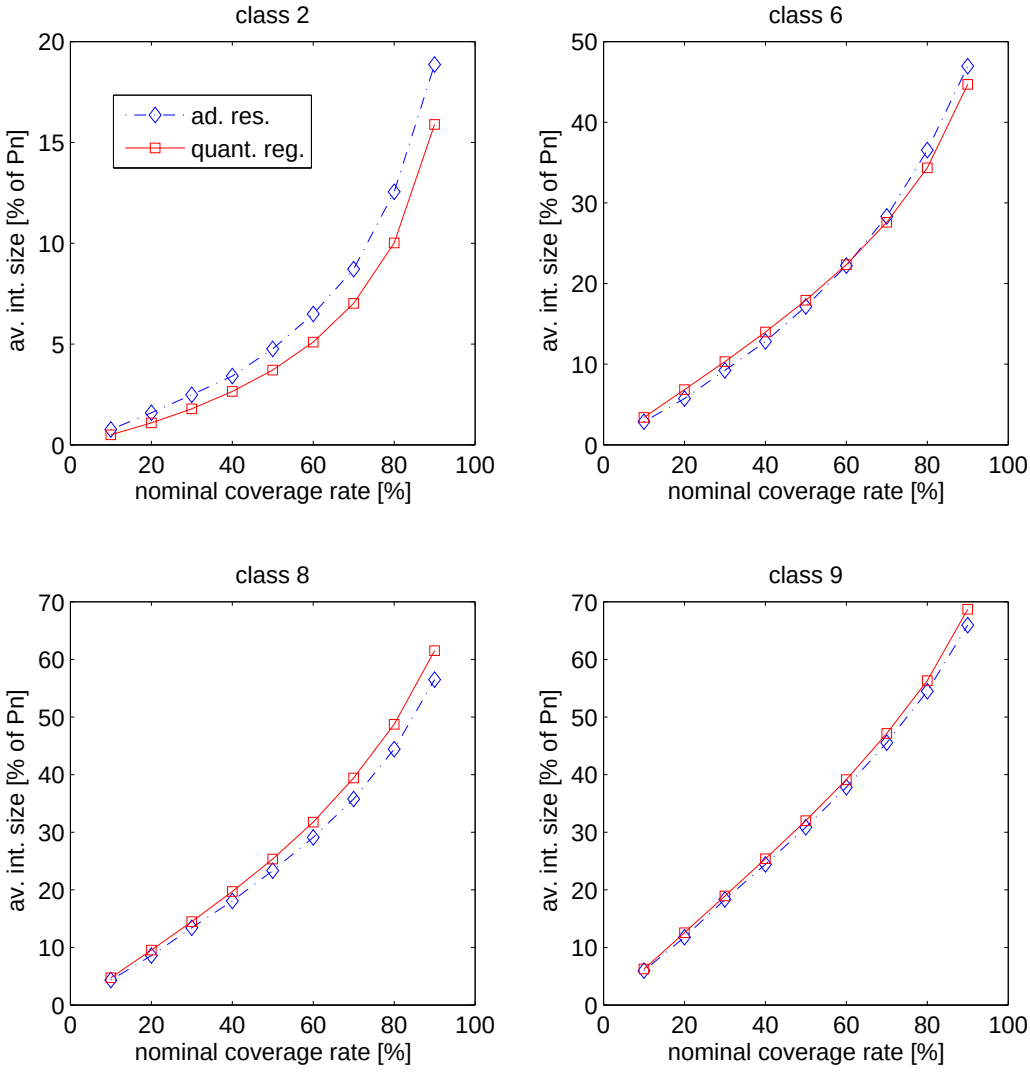

Figure 8: *Conditional sharpness evaluation: sharpness is evaluated as a function of the level of predicted power. Forecast/observation pairs are sorted in 10 equally populated classes of predicted power values. δ-diagrams are given for power classes 2, 6, 8 and 9.*

Let us focus on the power class 2 in a first stage. It has been explained above that adaptive quantile regression was more reliable for this power class, especially for low quantile nominal proportions. In addition, one sees that the predictive distributions produced with this

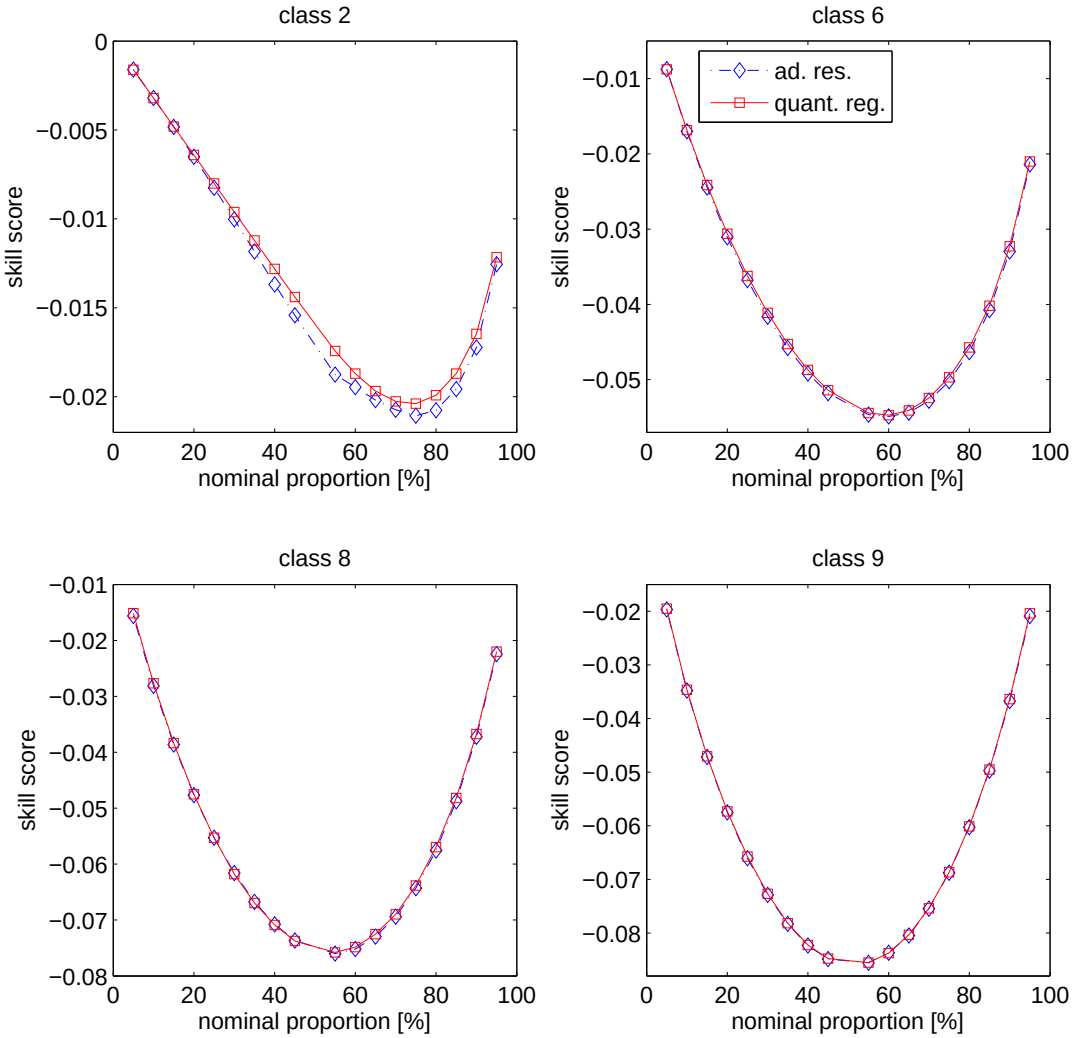

Figure 9: *Conditional skill evaluation: the skill of predictive distributions is evaluated as a function of the level of predicted power. Forecast / observation pairs are sorted in 10 equally populated classes of predicted power values. Skill diagrams, giving the skill score values for each quantile nominal proportions, are depicted for power classes 2, 6, 8 and 9.*

method appear to be sharper. Though, skill score values are very similar for low quantile nominal proportions, supporting our comment such that the large deviations from perfect reliability are to be counterbalanced by the fact that the numerical difference between predicted and 'true' quantiles must be very small. In this class, it is pretty clear that adaptive quantile regression is more skilled. For the others, the difference in skill is very small, but adaptive quantile regression is found more skilled for all of them. This is even valid for power classes such as power class 9, for which adapted resampling is found to be more reliable, and generates sharper predictive distributions. From a general point of view, the significantly higher conditional reliability of quantile regression explains its higher skill.

δ-diagrams are informative on the shape of predictive distributions: here, they show that

the two methods behave differently depending on level of predicted power, either on the whole range of nominal proportions, or on specific parts of predictive distributions. E.g. in power class 6, adaptive resampling is sharper in the central part of predictive distributions but not in the tail part. Though, one must understand that this sharpness criterion does not allow to conclude on a higher skill of such or such method. Finally, the δ-diagrams of Figure 8 shows that the shape of predictive distributions varies depending on the level of predicted power by the WPPT method. Especially, they are very sharp with thin tails for low power values (class 2), and more wide with thicker tails for power values in the linear part of the power curve (classes 6, 8 and 9). This demonstrates the ability of the two statistical methods to provide different — and still reliable for quantile regression — probabilistic information depending the forecast conditions, which are here characterized by the level of predicted power only.

5 Discussion on reliability assessment

The interest of reliability diagrams lies in their direct visual interpretation. However, this visual comparison between nominal and empirical probabilities introduces subjectivity, since the decision of whether probabilistic predictions can be considered as reliable or not is left to the analyst. This has been illustrated by the conditional evaluation exercise. This visual assessment of reliability contrasts with the more objective framework based on hypothesis testing used by the econometric forecasting community. Initially, Christoffersen (1998) proposes a likelihood ratio χ^2-test for evaluating the unconditional coverage of interval forecasts of economic variables, accompanied by another test of independence. Actually, the use of hypothesis testing is also not appropriate in this case. This is because one formulates a null hypothesis such that "the considered method is reliable", and consequently uses the inability to reject this null hypothesis for concluding on acceptable reliability. However, this ability to reject a null hypothesis in that manner is an inconclusive result (Ross, 2004, pp. 291-350). Instead, rejecting a null hypothesis formulated as "the considered method is *not* reliable" would permit to conclude on an acceptable reliability.

A similar application of hypothesis tests in the area of wind power forecasting relates to Bremnes (2006, 2004). He describes a Pearson χ^2-test for evaluating the reliability of the quantiles produced from a local quantile regression approach. However, χ^2-tests rely on an independence assumption regarding the sample data. Owing to the correlation of wind power forecasting errors, it is expected that series of interval hits and misses can come clustered together in a time-dependent fashion. This actually means that independence of the indicator variable sequence cannot be assumed in our case. Consequently, serial correlation invalidates the significance level of hypothesis tests. In general, it is known that statistical hypothesis tests cannot be directly applied for assessing the reliability of probabilistic forecasts due to the either serial or spatial correlation structures (Hamill, 2000). Pinson et al. (2006c) illustrate this result by the use of a simple simulation experiment where a quantile forecast known to be reliable is considered. It is shown that, except for 1-step ahead forecasts, the correlation invalidates the level of significance of the tests. It is demonstrated that this is because the correlation inflates the uncertainty of the estimate of actual coverage. Therefore, statistical hypothesis tests cannot be directly applied unless the correlation structure in the time series of indicator variable is previously removed.

An alternative to the use of hypothesis testing (and which is more appropriate, owing to our comment on a wrong use of hypothesis testing) consists in adding confidence bars in re-

liability diagrams (Bröcker and Smith, 2006a). This permits to inform on how to interpret the reliability estimates in regard to the characteristics of the evaluation set. In addition, this nicely goes along with the idea of the visual assessment of reliability via reliability diagrams. However again, for the specific case of multi-step ahead probabilistic forecasts of wind generation, the correlation structure needs to be considered for associating these bars to the reliability estimates. This may be done by using nonparametric methods for dependent data, as described by Lahiri (2003) for instance, and will be the focus of further developments.

6 Concluding remarks

Probabilistic predictions are becoming a common output of wind power prediction systems. They aim at giving an information on the forecast uncertainty in addition to the more classical point predictions. The question of how to evaluate probabilistic forecasts of wind power needs to be discussed, with consideration given to specific aspects of wind power forecasting. It has been explained why the existing frameworks introduced for some other forecasting applications are not appropriate for the wind power case. This paper comprises a proposal directed towards diagnostic evaluation of probabilistic predictions of wind power. The described evaluation framework is composed of measures and diagrams, with the aim of providing useful information on each of these properties, namely reliability, sharpness, resolution and skill. The use of the proposed evaluation framework for appraising the quality of two state-of-the-art methods for wind power probabilistic forecasting on a real-world case-study has allowed us to illustrate the relevance of these criteria, and to comment on the proper way to assess a method's quality. The importance of carrying out this evaluation conditional to the level of some explanatory variables has also been underlined. This is because wind power generation is a complex stochastic process for which the forecast uncertainty is influenced by a large number of external factors.

The decision of whether a given probabilistic forecasting method is reliable or not is subtle and further developments of the framework are needed for better concluding on that aspect. In parallel, the intuitive measure of sharpness based on the size of interval forecasts is very informative. Though, it has been explained that it cannot permit — even if it is often done in practice — to conclude on a higher skill of a given method. For that purpose, it is indeed more appropriate to rely on proper skill scores, which have nice theoretical properties insuring that a higher skill score value corresponds to a higher quality. Finally, appraising the resolution of a probabilistic forecasting method necessitates a conditional evaluation of the other properties. For the specific case of the wind power application, a higher resolution of probabilistic forecasts will be achieved by better understanding and including the influence of external factors e.g. related to meteorological conditions, on the forecast uncertainty. Statistical methods such as those considered in the present paper may be straightforwardly enhanced for including more explanatory variables known to impact on forecast uncertainty. Alternatively, it is expected that probabilistic predictions derived from meteorological ensemble forecasts would have a higher resolution, though their reliability is still a sensitive aspect. The proposed framework will be used as a basis for comparing these competing approaches to probabilistic forecasting of wind generation.

Focus has been given here to the quality of probabilistic predictions, i.e. to their statistical performance. While increasing this quality is the main focus of forecasters, forecast users are mainly interested in their value, i.e. the benefits resulting from the use of predictions in

decision-making. It will be of particular importance to show how a higher quality of probabilistic predictions translates to a higher value. More particularly, the role of increased reliability, sharpness or resolution in providing (or not) additional value should be highlighted. This issue is obviously problem-dependent, as a trader or a transmission system operator will not make the same use of the probabilistic forecasts of wind generation.

Acknowledgements

The results presented have been generated as part of the 'Forbedret Vindkraftforudsigelser' project supported by the Danish PSO fund under the contract number PSO-5766. The Danish PSO fund is hereby greatly acknowledged. The authors would also like to acknowledge Elsam (now part of Dong Energy) for providing the data for the Klim wind farm.

References

Atger, F., 1999. The skill of ensemble prediction systems. Monthly Weather Review 127, 1941–1957.

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

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

Bröcker, J., Smith, L. A., 2006a. Increasing the reliablity of reliablity diagrams. Weather and Forecasting, (submitted).

Bröcker, J., Smith, L. A., 2006b. Scoring probabilistic forecasts: the importance of being proper. Weather and Forecasting, in press.

Castronuovo, E. D., Pecas Lopes, J. A., 2004. On the optimization of the daily operation of a wind-hydro power plant. IEEE Transactions on Power Systems 19 (3), 1599–1606.

Chatfield, C., 2000. Time-Series Forecasting. Chapman & Hall/CRC.

Christoffersen, P. F., 1998. Evaluating interval forecasts. International Economic Review 39 (4), 841–862.

Clements, M. P., 2005. Evaluating Econometric Forecasts of Economic and Financial Values. Palgrave Macmillan.

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

Giebel, G., Kariniotakis, G., Brownsword, R., 2003. State of the art on short-term wind power prediction, ANEMOS Deliverable Report D1.1, available online: http://anemos.cma.fr.

Gneiting, T., Balabdaoui, F., Raftery, A. E., 2005. Probabilistic forecasts, calibration and sharpness. Tech. rep., University of Washington, Department of Statistics, technical report no. 483.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., Aldrich, E., 2006. Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space-time method. Journal of the American Statistical Association 101 (475), 968–979, Applications and Case-studies.

Gneiting, T., Raftery, A. E., 2004. Strictly proper scoring rules, prediction, and estimation. Tech. rep., University of Washington, Department of Statistics, technical report no. 463.

Granger, C. W. J., White, H., Kamstra, M., 1989. Interval forecasting: an analysis based upon ARCH- quantile estimators. Journal of Econometrics 40, 87–96.

Hall, P., Rieck, A., 2001. Improving coverage accuracy of nonparametric prediction intervals. Journal of the Royal Statistical Society B 63 (4), 717–725.

Hamill, T. M., 2000. Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review 129, Notes and Correspondence.

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

Koenker, R., Basset, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Lahiri, S. N., 2003. Resampling methods for dependent data. Springer Verlag.

Lange, M., 2005. On the uncertainty of wind power predictions - Analysis of the forecast accuracy and statistical distribution of errors. Trans. ASME, Journal of Solar Energy Engineering 127 (2), 177–184.

Lange, M., Focken, U., 2005. Physical Approach to Short-Term Wind Power Prediction. Springer.

Madsen, H., 2006. Time Series Analysis (second edition). Technical University of Denmark, Kgs. Lyngby, (ISBN 87-643-0098-6).

Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H. A., Nielsen, T. S., 2005. Standardizing the performance evaluation of short term wind power prediction models. Wind Engineering 29 (6), 475–489.

Møller, J. K., Nielsen, H. A., Madsen, H., 2006. Time adaptive quantile regression. Computational Statistics and Data Analysis, (submitted).

Murphy, A. H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and Forecasting 8, 281–293.

Nielsen, H. A., Madsen, H., Nielsen, T. S., 2006a. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. Wind Energy 9 (1-2), 95–108.

Nielsen, H. A., Nielsen, T. S., Madsen, H., Badger, J., Giebel, G., Landberg, L., Sattler, K., Voulund, L., Tøfting, J., 2006b. From wind ensembles to probabilistic information about future wind power production - results from an actual application. In: Proc. PMAPS 2006, 'Probabilistic Methods Applied to Power Systems', IEEE Conference, Stockholm.

Nielsen, H. A., Nielsen, T. S., Madsen, H., Sattler, K., 2004. Wind power ensemble forecasting. In: Proc. Global WindPower 2004, Chicago, Illinois (USA).

Nielsen, T. S., Nielsen, H. A., Madsen, H., 2002. Prediction of wind power using time-varying coefficient functions. In: Proc. IFAC 2002, 15th World Congress on Automatic Control, Barcelona, Spain.

Pinson, P., 2006. Estimation of the uncertainty in wind power forecasting. Ph.D. thesis, Ecole des Mines de Paris, Paris, France, www.pastel.paristech.org/bib.

Pinson, P., Chevallier, C., Kariniotakis, G., 2006a. Trading wind generation with short-term probabilistic forecasts of wind power. IEEE Transactions on Power Systems, (submitted).

Pinson, P., Juban, J., Kariniotakis, G., 2006b. On the quality and value of probabilistic forecasts of wind generation. In: Proc. PMAPS 2006, 'Probabilistic Methods Applied to Power Systems', IEEE Conference, Stockholm.

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

Pinson, P., Kariniotakis, G., Nielsen, H. A., Nielsen, T. S., Madsen, H., 2006c. Properties of quantile and interval forecasts of wind generation and their evaluation. In: Proc. EWEC 2006, 'European Wind Energy Conference', Scientific Track, Athens.

Ross, S. M., 2004. Introduction to Probability & Statistics for Engineers and Scientists. Elsevier Academic Press, Amsterdam.

Roulston, M. S., Smith, L. A., June 2002. Evaluating probabilistic forecasts using information theory. Monthly Weather Review 130, 1653–1660, Notes and Correspondence.

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.

Taylor, J. W., 1999. Evaluating volatility and interval forecasts. Journal of Forecasting 18, 111–128.

Thor, S.-E., Weis-Taylor, P., 2002. Long-term research and development needs for wind energy for the time frame 2000-2020. Wind Energy 5, 73–75.

Tong, H., 1995. A personal overview of nonlinear time-series analysis from a chaos point of view. Scandinavian Journal of Statistics 22, 399–445.

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.

Wallis, K. F., 2003. Chi-squared tests of interval and density forecasts, and the bank of england's fan charts. International Journal of Forecasting 19, 165–175.

Winkler, R. L., 1972. A decision-theoretic approach to interval estimation. Journal of the American Statistical Association 67, 187–191.

Zervos, A., 2003. Developing wind energy to meet the Kyoto targets in the European Union. Wind Energy 6 (3), 309–319.

Appendix

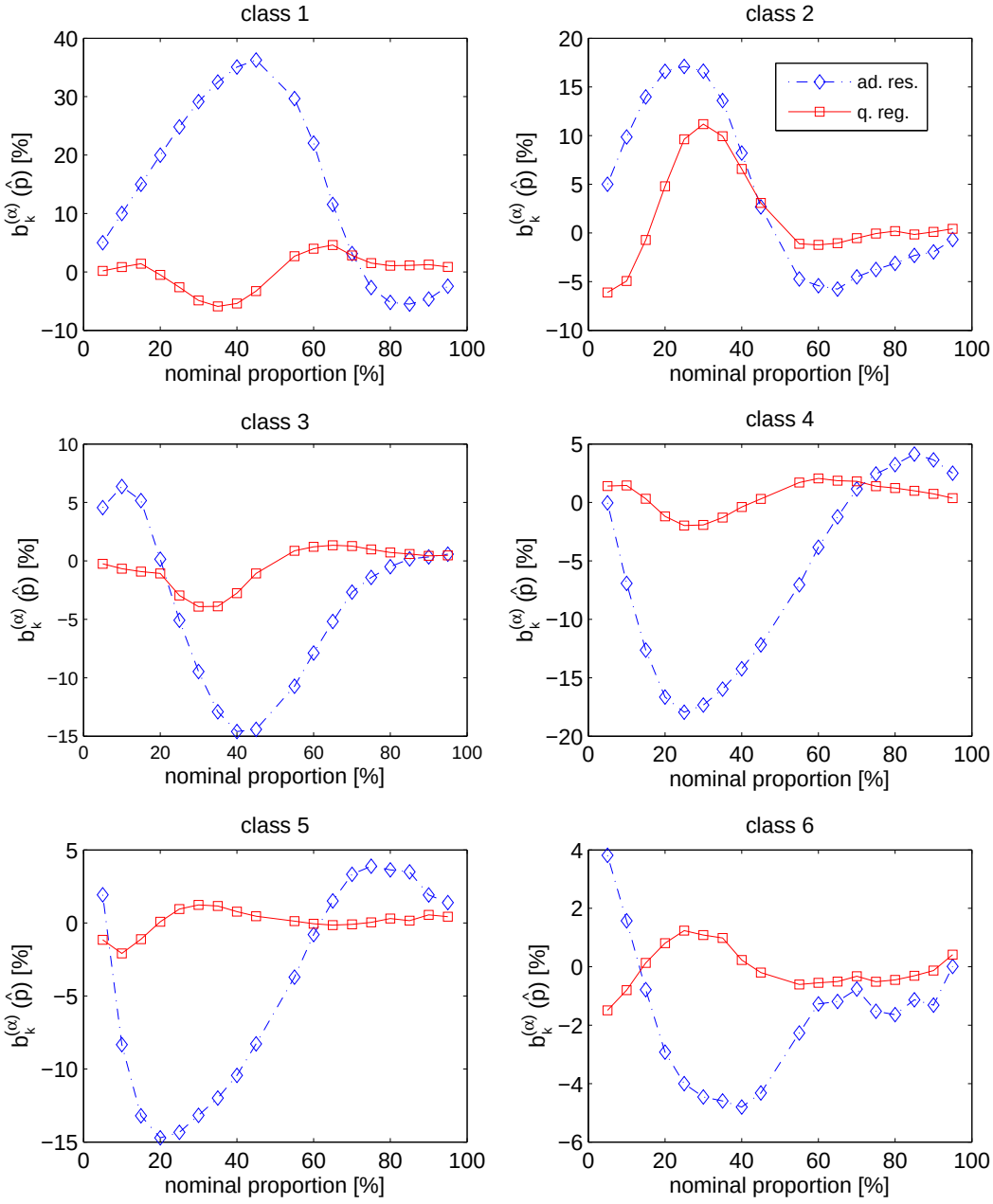

Figure 10: *Conditional reliability evaluation: reliability is assessed as a function of the level of predicted power. Forecast/observation pairs are sorted in 10 equally populated classes of predicted power values. Reliability diagrams are given for power classes 1 to 6.*

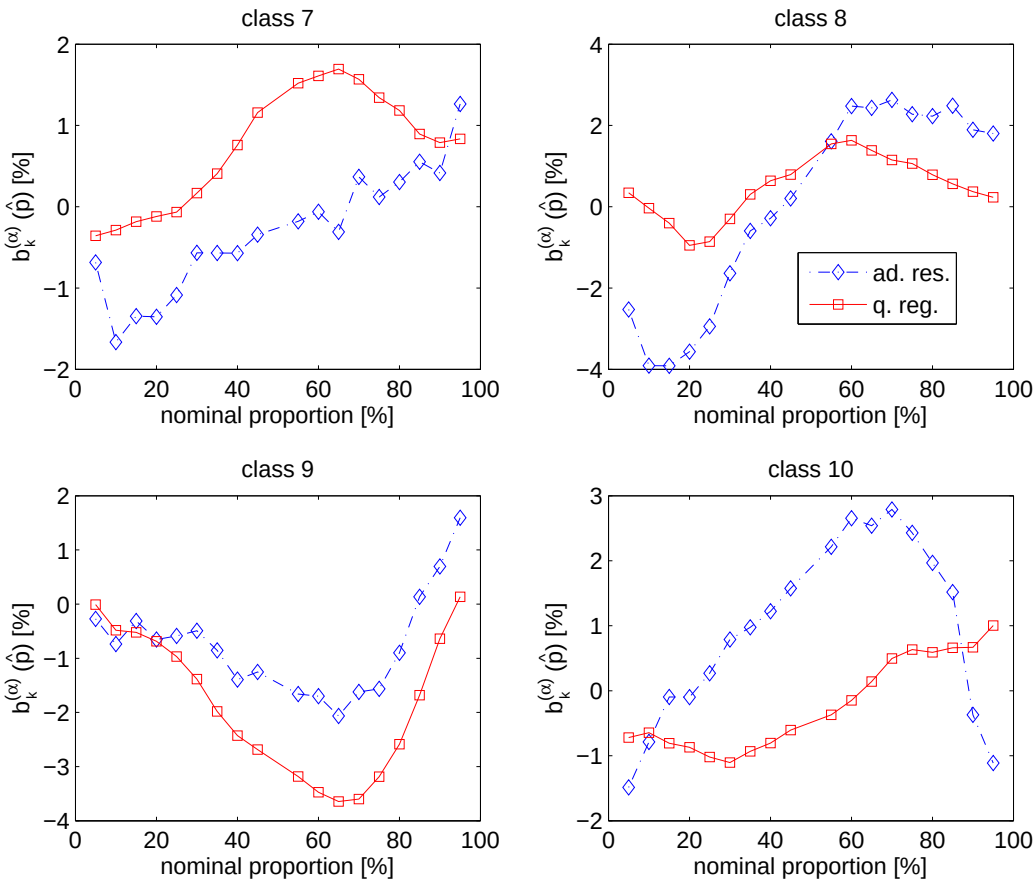

Figure 11: *Conditional reliability evaluation: reliability is assessed as a function of the level of predicted power. Forecast/observation pairs are sorted in 10 equally populated classes of predicted power values. Reliability diagrams are given for power classes 7 to 10.*

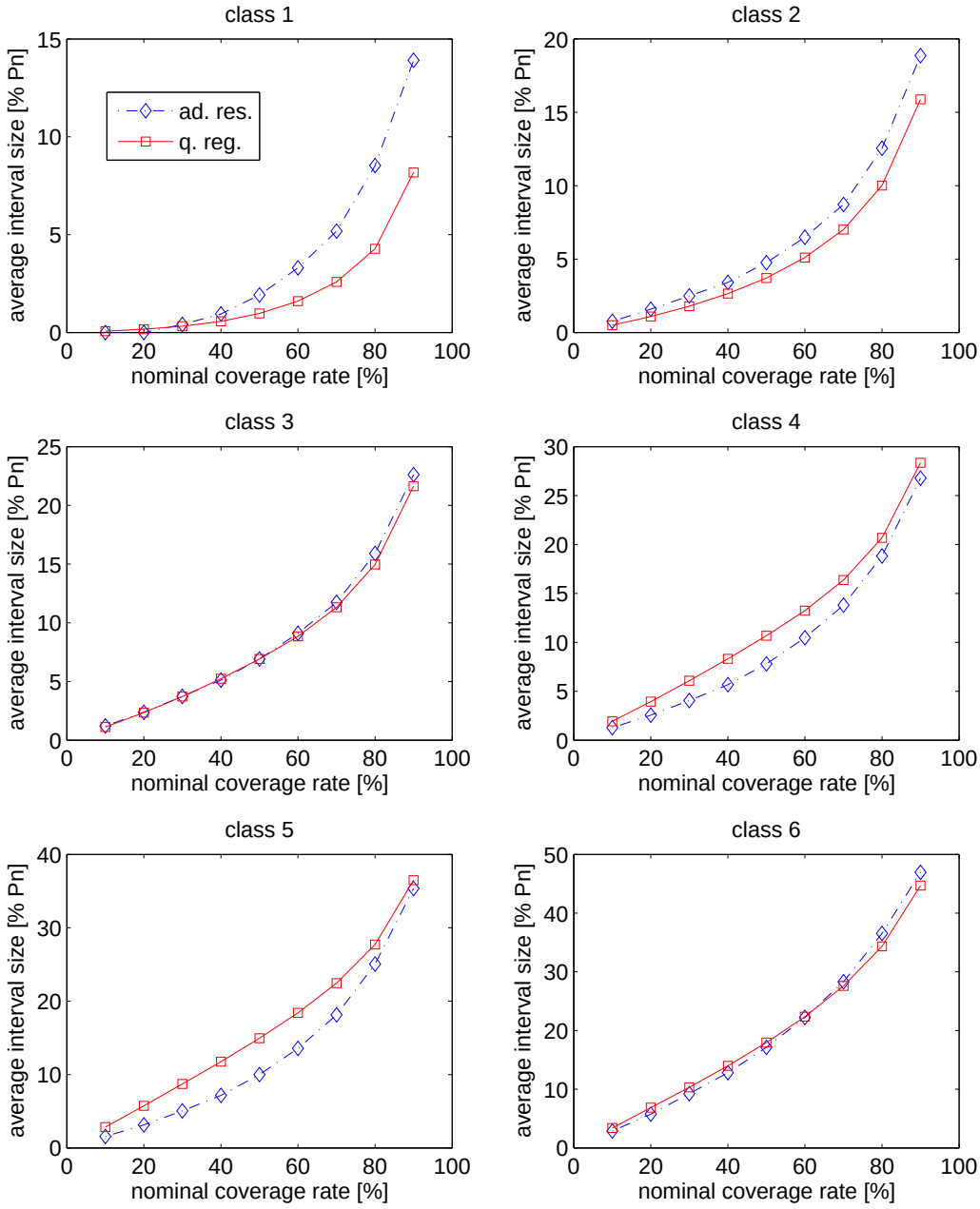

Figure 12: *Conditional sharpness evaluation: sharpness is evaluated as a function of the level of predicted power. Forecast/observation pairs are sorted in 10 equally populated classes of predicted power values. δ-diagrams are given for power classes 1 to 6.*

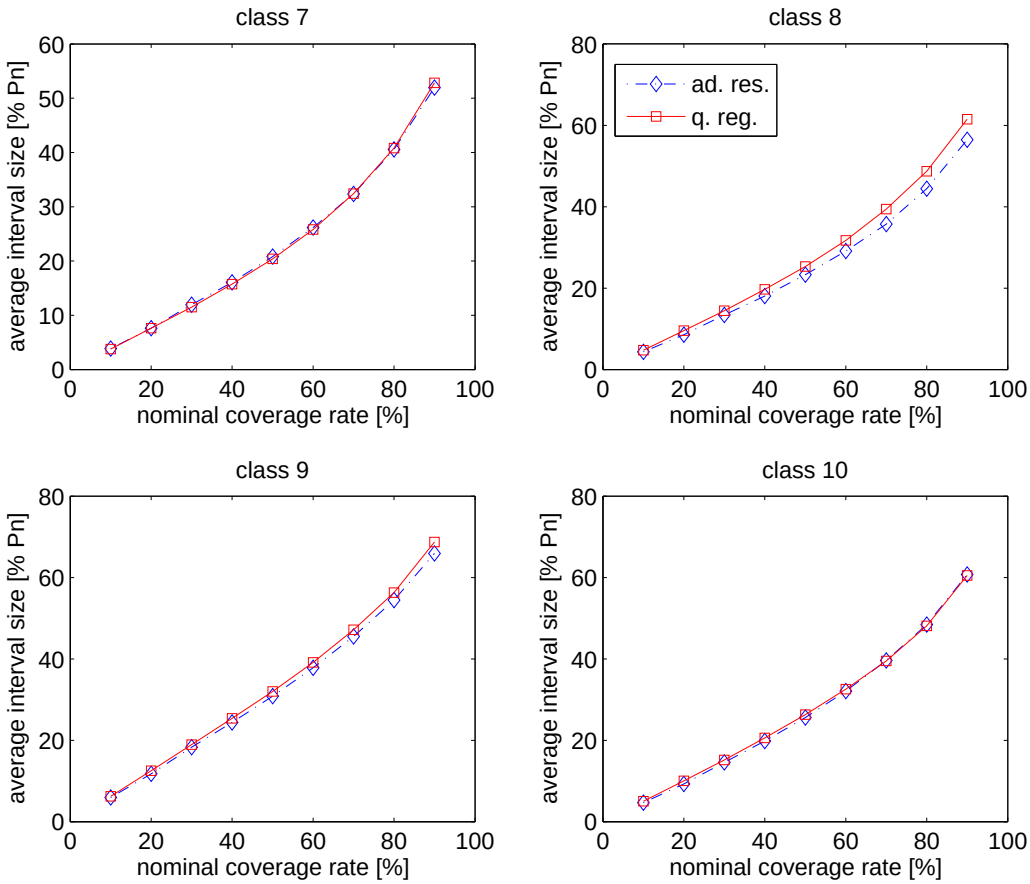

Figure 13: *Conditional sharpness evaluation: sharpness is evaluated as a function of the level of predicted power. Forecast/observation pairs are sorted in 10 equally populated classes of predicted power values. δ-diagrams are given for power classes 7 to 10.*

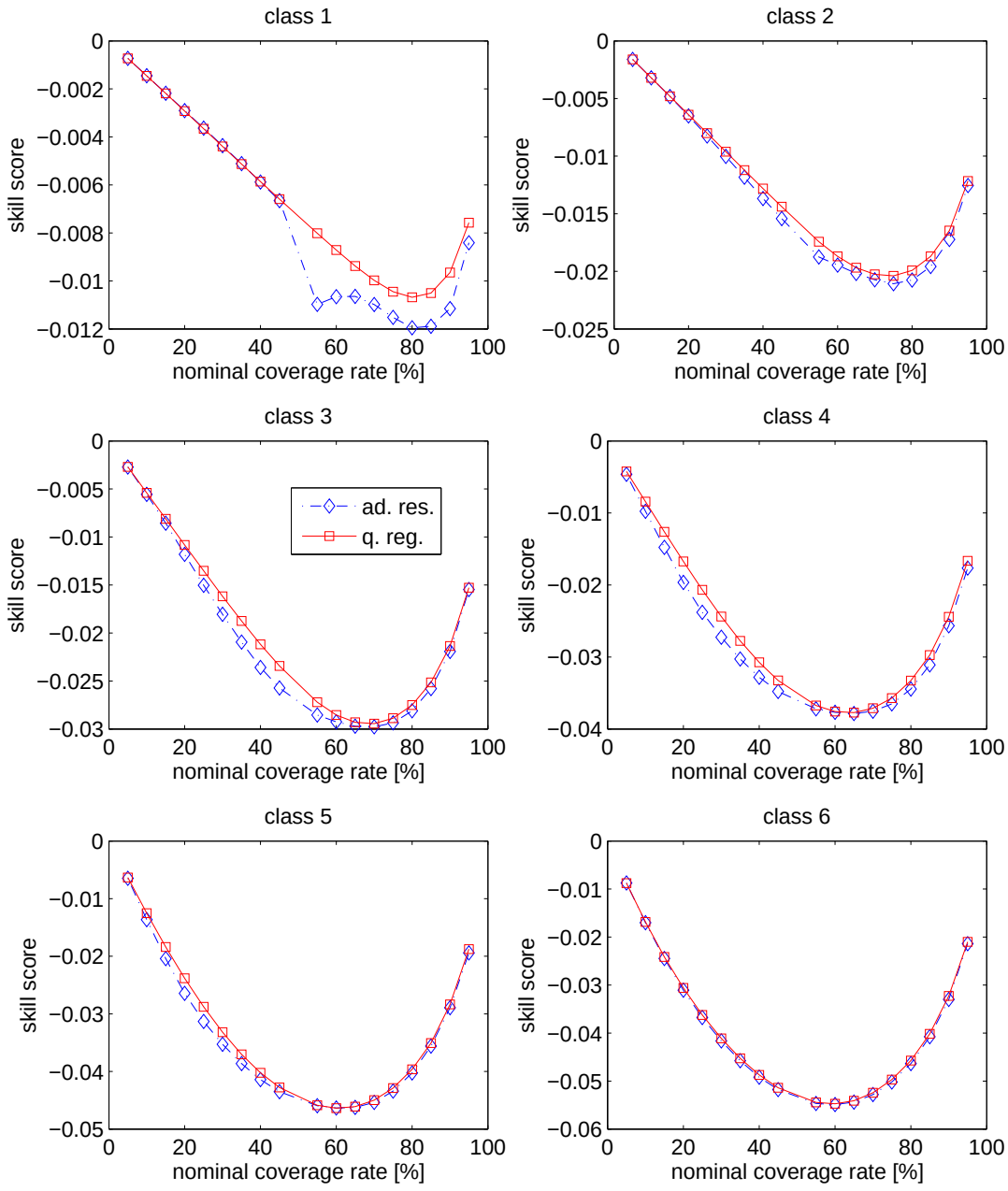

Figure 14: *Conditional skill evaluation: the skill of predictive distributions is evaluated as a function of the level of predicted power. Forecast/observation pairs are sorted in 10 equally populated classes of predicted power values. Skill diagrams, giving the skill score values for each quantile nominal proportions, are depicted for power classes 1 to 6.*

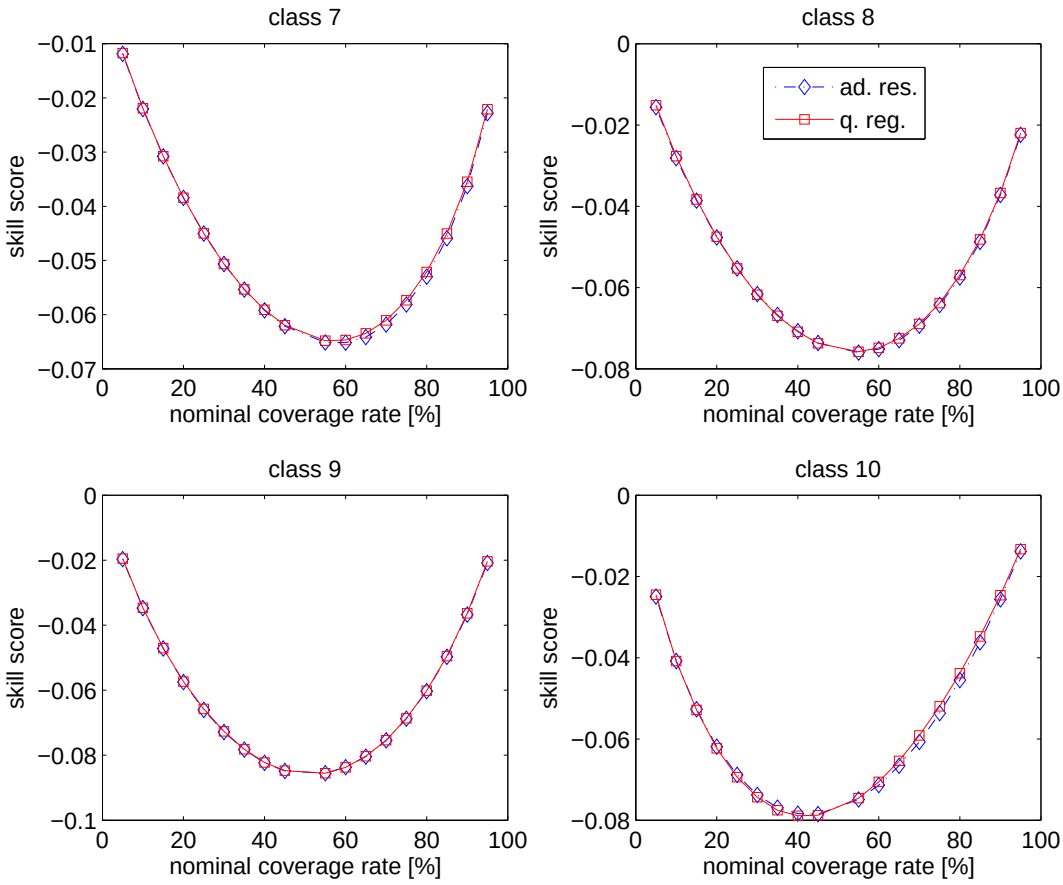

Figure 15: *Conditional skill evaluation: the skill of predictive distributions is evaluated as a function of the level of predicted power. Forecast/observation pairs are sorted in 10 equally populated classes of predicted power values. Skill diagrams, giving the skill score values for each quantile nominal proportions, are depicted for power classes 7 to 10.*