Properties of Quantile and Interval Forecasts of Wind Generation and their Evaluation

P. Pinson*, G. Kariniotakis
École des Mines de Paris, Centre for Energy and Processes
B.P. 207, 06904 Sophia-Antipolis, France
*pierre.pinson@ensmp.fr

H. Aa. Nielsen†, T. S. Nielsen, H. Madsen
Informatics and Mathematical Modeling, Technical University of Denmark
R. Petersens Plads, 2800 Lyngby, Denmark
†han@imm.dtu.dk

Abstract

Either for managing or trading wind power generation, it is recognized today that forecasting is a cornerstone. Traditionally, methods that are developed and implemented are point forecasting methods, i.e. they provide a single estimated value for a given horizon. As errors are unavoidable, several research teams have recently proposed uncertainty estimation methods in order to optimize the decision-making process (reserve quantification, bidding strategy definition, etc.). Here, focus is given to methods that quote quantiles or intervals from predictive distributions of wind generation. The paper describes what the required properties of appropriate uncertainty estimation methods are and how they can be evaluated. Finally, it is shown how the introduced evaluation criteria may be used for highlighting or optimizing the performance of current probabilistic forecasting methodologies.
Keywords: *Wind power, short-term forecasting, uncertainty estimation, probabilistic forecasting, evaluation methods*

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 [32]. 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 [30].

Traditional wind power prediction methods are point forecasting methods: they provide the 'most likely outcome' for a given look-ahead time. A part of the recent research works have focused on associating uncertainty estimates to these point forecasts or to produce fully probabilistic predictions, either based on meteorological ensembles [22], or produced from statistical methods [4, 21, 26]. This trend leads to a new generation of modeling approaches related to wind power forecasting, which, despites their apparent complexity, are expected to become an added value to operational prediction platforms.

A set of standard error measures and evaluation criteria has recently been described for the verification of point forecasts of wind generation [17]. However, evaluating probabilistic forecasts is more complicated than evaluating deterministic ones. While it is easy to say that a point forecast is false because the deviation between predicted and real values is of practical magnitude, 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 ?

Here, our aim is to describe what are the required properties of probabilistic forecasts and how these forecasts can be evaluated in terms of their statistical performance (referred to as their *quality*) in a non-parametrical framework. For that purpose, we will present relevant skill scores, measures and diagrams that have been introduced in the statistical and meteorological literature. The interest of this non-parametrical framework is that it can be used for evaluating predictive quantiles, interval or density forecasts. In a second part, we will use this set of criteria to show how it may serve for comparing rival approaches or how it may be used for optimizing a given method. This will allow us to draw conclusions on the quality of some of the current approaches for the probabilistic forecasting of wind generation.

2 Required properties for probabilistic forecasts

Prediction intervals give the range of possible values within which the true effect is expected to lie with a certain probability, its *nominal coverage rate*. Hereafter, intervals and quantiles are considered interchangeably since we concentrate on central prediction intervals $\hat{\mathcal{I}}^{(\alpha)}_{t+k/t}$, estimated at time t for look-ahead time $t+k$ with a nominal coverage rate $(1-\alpha)$, the bounds of which correspond to the $(\alpha/2)$ and $(1-\alpha/2)$ quantiles of the predictive distribution of expected events at that lead time:

$$\hat{\mathcal{I}}^{(\alpha)}_{t+k/t} \equiv [\hat{r}^{(\alpha/2)}_{t+k/t}, \hat{r}^{(1-\alpha/2)}_{t+k/t}]. \quad (1)$$

The first requirement for interval forecasts is that their empirical coverage should be close to the nominal one. Actually,

if considering infinite series of interval forecasts, that empirical coverage should exactly equal the pre-assigned probability. That first property is referred to as *reliability* or *calibration* in the literature [1, 8, 18].

Besides this first requirement, it is necessary that prediction intervals provide a situation-dependant assessment of the forecast uncertainty. Their size should then vary depending on various external conditions. For the example of wind prediction, 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. The most simple type of intervals is constant-size intervals (e.g. produced from climatology). Advanced methods for their estimation are expected to produce variable-size intervals. This property is commonly named *sharpness* or *resolution* of the intervals [1, 18]. Note that here, we will introduce a nuance between sharpness and resolution: the former will relate to the average size of intervals while the latter is associated to their size variability.

Actually, the traditional view of interval forecast evaluation, which mainly comes from the econometric forecasting community, is based on the testing of correct conditional coverage. This means intervals have to be unconditionally reliable, and independent (see for instance [7], or [8] (ch. 3)). In the case of wind power forecasting, we know there exists a correlation among forecasting errors (at least for short time-lags). Thus, we do not expect prediction intervals to be independent. Then, it appears preferable to develop an evaluation framework that is based on an alternative paradigm. We propose to consider reliability as a primary requirement and then sharpness and resolution as an added value. It should be noted here that reliability can be increased by using some recalibration methods (e.g. conditional parametric models [22] or smoothed bootstrap [14]), while sharpness/resolution cannot be enhanced with post-processing procedures. This second aspect is the inherent (and invariant) ability of a probabilistic forecasting method to distinctly resolve future events [31].

3 Discussion on the nominal coverage rate

An important question concerning the intervals arises: how to choose an optimal nominal coverage rate ? When this pre-assigned probability is higher than 90%, intervals can be embarrassingly wide, because they will contain extreme prediction errors (or even outliers). In addition, working with high-coverage intervals means that we are aiming at modeling the very tails of distributions. Thus, the robustness of the uncertainty estimation methods becomes a critical aspect. However, if one defines lower pre-assigned probabilities (50% for instance), intervals will be much more narrow and more robust with respect to extreme prediction errors, but this will mean that actual future values are 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 [5].

Though, instead of focusing on a particular nominal coverage rate, it seems that producing a number of prediction intervals is a better solution. This will allow one to build the whole probability distribution of expected generation for each look-ahead time. For the example of the adapted resampling method presented in [26, 27], it is not more computationally expensive to estimate one or thirty quantiles. Either for energy management or trading, wind power forecast users will not rely on a single interval forecast only: usual decision-making methods need a complete approximation of the density function for providing an optimal management [10] or trading strategy [3, 11, 24].

4 Methods for the evaluation of probabilistic forecasts

In this Section, we only treat the aspects of evaluating the skill of pointwise quantiles or intervals, i.e. estimated on a per-horizon basis. Some methods are meant for producing intervals with a nominal coverage rate that holds for the whole forecast length [28], but this not the case for the methods dedicated to wind generation. In the following, all criteria will be evaluated as a function of the look-ahead time, or as an average over the forecast length. If a considered evaluation set is large enough, it may also be interesting to estimate the skill of probabilistic methods as a function of some other parameters (e.g. level of power).

The following methodology introduces a unique skill score for the quality evaluation of probabilistic forecasts of wind generation. Consequently, focus is given to the various aspects composing the overall skill, in a hierarchical manner: reliability has to be assessed first, followed by a study of shapness, and then resolution.

4.1 The indicator variable

Before going further with the evaluation of interval forecasts, it is necessary to introduce the *indicator variable* $I_{t,k}^{(\alpha)}$ (following the definition by Christoffersen [7]), which is defined for a prediction made at time t and for horizon k as follows

$$I_{t,k}^{(\alpha)} = \begin{cases} 1, & \text{if} \quad p_{t+k} \in \hat{\mathcal{I}}_{t+k/t}^{(\alpha)} \\ 0, & \text{if} \quad p_{t+k} \notin \hat{\mathcal{I}}_{t+k/t}^{(\alpha)} \end{cases} . \tag{2}$$

This indicator variable tells if the actual outcome p_{t+k} at time $t+k$ lies ("*hit*") or not ("*miss*") in the prediction interval estimated for that lead time.

We would like to mention that this definition of the indicator variable can be easily adapted when working with quantiles of a probabilistic distribution. Indeed, the value of p_{t+k} being inside or not of the interval is replaced by p_{t+k} being below or above the estimated quantile $\hat{r}_{t+k/t}^{(\alpha)}$.

Consequently, let us define as $n_{k,1}^{(\alpha)}$ the sum of hits and $n_{k,0}^{(\alpha)}$ the sum of misses (for a given horizon k) over the N realizations:

$$n_{k,1}^{(\alpha)} = \#\{I_{t,k}^{(\alpha)} = 1\} = \sum_{t=1}^{N} I_{t,k}^{(\alpha)}, \tag{3}$$

$$n_{k,0}^{(\alpha)} = \#\{I_{t,k}^{(\alpha)} = 0\} = N - n_{k,1}^{(\alpha)}. \tag{4}$$

It is by studying the time-series of indicator variables $\{I_{t,k}^{(\alpha)}\}_{t=1,\ldots,N}$ over the test set that we will evaluate the reliability and the overall skill of probabilistic forecasts.

4.2 Defining a unique skill score

As for point-forecast evaluation, it is often demanded that a unique skill score would give the whole information on a given method performance. Such measure would be given by scoring rules that associate a single numerical value $S(\hat{q}, p)$ to a predictive distribution $\hat{q}$ if the event p materializes. Then, we can define as

$$S(\hat{q}', \hat{q}) = \int S(\hat{q}', p)\hat{q}(p)dp \tag{5}$$

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

A scoring rule should reward a forecaster that expresses his true beliefs. It is said to be *proper* if it does so. Murphy [20] referred to that aspect as the forecast *constistency* and reminds that a forecast (probabilistic or not) should correspond to the forecaster's judgment. If we assume that a forecaster wish to maximize his skill score over an evaluation set, then a scoring rule is said to be proper if for any two predictive distributions $\hat{q}$ and $\hat{q}'$ we have

$$S(\hat{q}', \hat{q}) \leq S(\hat{q}, \hat{q}), \quad \forall \hat{q}, \hat{q}'. \tag{6}$$

Hence, if $\hat{q}$ corresponds to the forecaster's judgment, it is by quoting this particular predictive distribution that he will maximize his skill score.

If we consider that a predictive distribution $\hat{q}$ is characterized by its quantiles $\hat{\mathbf{r}} = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_l\}$ at levels $\alpha_1, \alpha_2, \ldots, \alpha_l$, Gneiting et al. [13] recently showed that any scoring rule of the form

$$S(\hat{\mathbf{r}}, p) = \sum_{i=1}^{l} \big(\alpha_i s_i(r_i) + (s_i(p) - s_i(r_i))I^{(\alpha_i)} + f(p)\big), \tag{7}$$

with $I^{(\alpha_i)}$ the indicator variable (for the quantile with proportion α_i) introduced above, s_i non-decreasing functions and f arbitrary, was proper for evaluating this set of quantiles. Here $S(\hat{\mathbf{r}}, p)$ is a positively rewarding score: a higher score value stands for an higher skill. The specific case of central prediction intervals corresponds to the case where only two quantiles are quoted (cf. Equation (1)). Note that for a unique quantile, the scoring rule given by Equation (7) generalizes the loss functions considered in quantile regression [22] and local quantile regression [4].

Using a single proper score may allow one to compare the overall skill of rival approaches. It can also be utilized as a criterion for optimizing the parameters of a given quantile estimation method. However, a unique score does not tell what are the contributions of reliability or sharpness to the skill (or to the lack of skill)[1]. This is why we focus on both of these two aspects in the following Paragraphs.

[1]This has already been stated by Roulston et al. [29] 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.

4.3 Reliability

The easiest way to check the calibration of interval forecasts is to compare their empirical coverage to the nominal one (i.e. the required probability (1-α)). An estimation $\hat{a}_k^{(\alpha)}$ of the actual coverage $a_k^{(\alpha)}$, for a given horizon k, is obtained by calculating the mean of the $\{I_{t,k}^{(\alpha)}\}_{t=1,\ldots,N}$ time-series over the test set:

$$\hat{a}_k^{(\alpha)} = E[I_{t,k}^{(\alpha)}] = \frac{1}{N}\sum_{t=1}^{t=N} I_{t,k}^{(\alpha)} = \frac{n_{k,1}^{(\alpha)}}{n_{k,0}^{(\alpha)} + n_{k,1}^{(\alpha)}}. \tag{8}$$

This standard approach for evaluating prediction intervals was proposed by Ballie et al. [2] and by McNees [19]. This is the idea used in *reliability diagrams* which give the empirical coverage versus the nominal coverage for various nominal coverage values. The closer to the diagonal the better. They can alternatively be depicted as the deviation from the 'perfect reliability' case for which empirical coverage would equal the nominal one (calculated as the difference between these two quantities). This idea is similar to the use of Probability Integral Transform histograms as proposed by Gneiting et al. [12] except that reliability diagrams directly provide that additional information about the magnitude of the deviation from the 'perfect reliability' case.

Reliability diagrams allow one to summarize the calibration assessment of several quantiles or intervals and thus to see at one glance if a given method tends to systematically underestimate (or overestimate) the uncertainty. Figure 1 shows an example of a reliability diagram for the evaluation of a given quantile estimation method. Deviations from the 'perfect reliability' case are given as a function of the quantile nominal proportions, as an average over the forecast length. Here, one notices a rather good calibration of the method since 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 a bit too narrow.

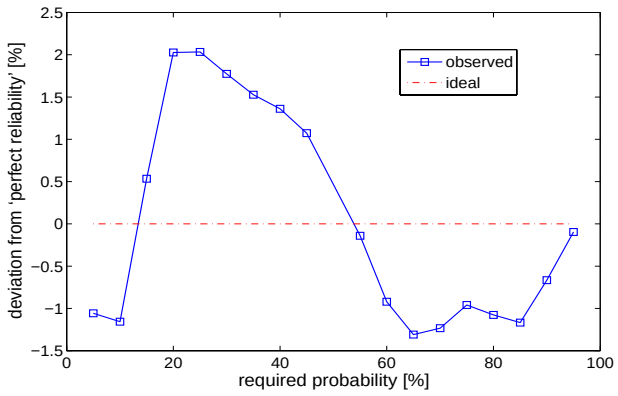

FIGURE 1: Example of a reliability diagram depicting deviations as a function of the nominal coverage rate, for the reliability evaluation of a method providing probabilistic forecasts of wind generation.

Using that kind of comparison between the nominal and

empirical coverages introduces subjectivity in the evaluation: the decision of whether the intervals have correct coverage or not is left to the analyst. This is why a more objective framework based on hypothesis testing has been introduced in the forecasting literature (mainly in econometric forecasting). For instance, Christoffersen [7] proposed a likelihood ratio χ^2-test for evaluating the unconditional coverage of interval forecasts of economic variables, accompanied by another test of independence. In the area of wind generation forecasting, Bremnes [4] recently used 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. Power forecasting errors are correlated (at least for short time-lags), and it is expected that series of interval hits and misses can come clustered together in a time-dependant fashion. This actually means that independence of the indicator variable sequence cannot be assumed in our case (except if independence is proven in a prior analysis). In general, it is known that statistical hypothesis tests cannot be directly applied for assessing the reliability of probabilistic forecasts due to the serial (or spatial) correlation stucture [15]. In the Appendix, this is illustrated 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. Finally, if hypothesis testing is envisaged for evaluating the unconditional coverage of quantiles or intervals, we would alternatively propose to use a non-parametric binomial test. This test is indeed meant for evaluating proportions when studying binary time-series.

4.4 Sharpness and Resolution

When dealing with sharpness or resolution, focus is given to the size of the prediction intervals. Let us define

$$\delta_{t,k}^{(\alpha)} \;=\; \hat{r}_{t+k/t}^{(1-\alpha/2)} - \hat{r}_{t+k/t}^{(\alpha/2)} \tag{9}$$

the size of the central interval forecast (with associated probability $(1-\alpha)$) estimated at time t for lead time $t+k$.

If two uncertainty estimation methods provide intervals at an acceptable level of reliability, it is the method that yields the narrowest intervals that is to be preferred. Here, we relate the sharpness aspect to the average size $\bar{\delta}_k^{(\alpha)}$ of the prediction intervals for a given horizon k:

$$\bar{\delta}_k^{(\alpha)} \;=\; \frac{1}{N}\sum_{t=1}^{N}\delta_{t,k}^{(\alpha)} \;=\; \frac{1}{N}\sum_{t=1}^{N}\left(\hat{r}_{t+k/t}^{(1-\alpha/2)} - \hat{r}_{t+k/t}^{(\alpha/2)}\right). \tag{10}$$

Both Bremnes [4] and Nielsen et al. [22] used such a measure for evaluating the sharpness of the their probabilistic forecasts as a function of the horizon. When focusing on the distance between the quantiles for the proportions 0.25 and 0.75 (the quartiles), this measure is commonly known as the inter quartile range. This measure has been found more useful (and also more robust) than the standard deviation for studying asymmetric distributions. However, it may be interesting not to focus only on these two particular quantiles but also to look at $\bar{\delta}_k^{(0.8)}$ and $\bar{\delta}_k^{(0.2)}$ that are the average size of repectively the 20%- and 80%-confidence central intervals.

The resolution concept is standing for the ability of providing a situation-dependant assessment of the uncertainty. If two approaches have similar sharpness, then a higher resolution translates to a higher quality of related interval forecasts. It is not possible to directly verify that property, though we may study the variation in size of the intervals. The standard deviation $\sigma_k^{(\alpha)}$ of the interval size (for a given horizon k and nominal coverage rate $(1-\alpha)$) provides that information. Because of the non-linear and conditionally heteroskedastic nature of the wind generation process, the forecast uncertainty is highly variable and it is thus expected that the interval size also greatly varies.

Finally, δ-diagrams and σ-diagrams, which give respectively $\bar{\delta}_k^{(\alpha)}$ and $\sigma_k^{(\alpha)}$ as a function of the nominal coverage rate for a given look-ahead time k (or over the forecast length), permit to better visualize the shape (and shape variations) of predictive distributions. We will underline the interest of such diagnostic tools in the following Sections.

5 Comparing the quality of various interval forecasting methods

In this Section, we apply the methodology introduced above for highlighting the properties of some of the approaches currently in use for providing interval forecast. The case-study is the Tunø Knob wind farm, which is located 6km off the cost of Jutland, in Denmark and has an installed capacity P_n of 5MW. The set of probabilistic predictions to be evaluated consists in central prediction intervals with a 50% degree of confidence. Forecasts are issued on a daily basis (at noon), with an hourly resolution, over a 153-days period. The evaluation set is thus composed by 153 forecast series only. We focus here on look-ahead times between 12 and 30-hour ahead, which are the relevant horizons for application to most of the European electricity markets.

Three different approaches are considered for producing the interval forecasts. The first one is based on the assumption that wind generation can be modelled with a Gaussian distribution whose mean is the estimated point forecast (given by a state-of-the-art approach) and whose standard deviation is estimated from adaptive statistics on the past performance of the predictor. It is known that this Gaussian assumption does not hold (see [16, 26] for instance), but still, we consider that approach as a baseline method. The second approach relies on ensemble forecasts of wind power. They are derived from the 51-member meteorological ensembles provided by the European Center for Medium-Range Weather Forecasts (ECMWF) following the method described in [22]. However, wind power ensembles have not been recalibrated in the present work, due to the small number of forecast series. Finally, the third approach is the quantile regression approach developed in [21] that uses power point predictions, as well as forecasts of explanatory variables e.g. wind speed and direction, for estimating prediction

intervals of wind power.

We mentioned that the first requirement for the intervals is that they have to be reliable. Thus we first check the empirical coverages of estimated quartiles and corresponding intervals (cf. Table 1). One notes a significant deviation from nominal coverage for the quartiles whatever the method, though quantile regression seems to produce the more reliable intervals. Knowing that indicator variable sequences are correlated and because we have only a small set of forecasts for evaluation, it is not possible to draw strict conclusions from these figures. Still, the deviation of the Gaussian lower quartile from the nominal value appears very large. The fact we cannot conclude on reliability prevent us from giving a rigorous picture on the various aspects of the methods' quality. However, we carry on with that study for highlighting some of the methods' properties concerning sharpness and resolution.

Nominal coverage	25	75	50
Gaussian	46.21	83.89	37.68
Ensemble-based	15.89	79.57	63.68
Quantile regression	28.76	84.34	55.58

TABLE 1: Empirical coverages of the estimated quartiles and empirical coverage of the related central prediction intervals (in %, averaged over the whole range of prediction horizons).

We look at sharpness first by plotting the average size of the central prediction intervals (Figure 2), as a function of the prediction horizon. The interval size tends to increase with lead time for the two statistical methods while that increase is really noteworthy for the intervals produced from ensembles. However, since their coverage is higher than for the two other methods, it seems normal to have larger intervals. Also, an interesting point is that although Gaussian and quantile-regression intervals have a similar size on average, the reliability of the Gaussian intervals is a lot lower. This would be an argument for preferring the more advanced approach.

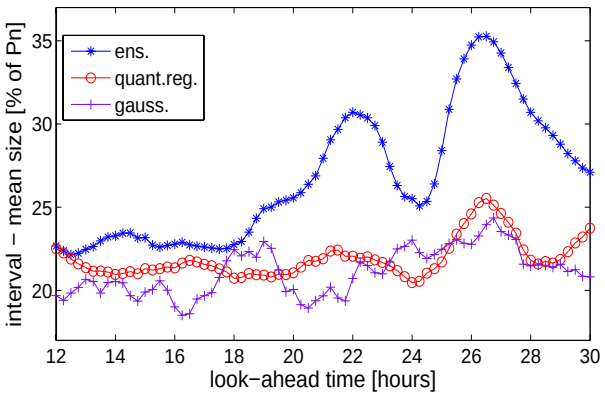

FIGURE 2: Sharpness evaluation - Average size of the central 50% prediction intervals as a function of the look-ahead time.

A crucial difference between the three methods can be seen from the plot that depicts the standard deviation of the interval size (Figure 3). Such a plot actually tells how probabilistic methods resolve future events (that is exactly what the resolution concept stands for). If two methods are considered as reliable (and similarly sharp), it is the one that proposes the most variable uncertainty assessment which is to be preferred. From that plot, one notes that Gaussian intervals have a very low variation of their size: they have no-resolution. Resolution is a lot higher for quantile-regression-based intervals and even higher for the ensemble-based ones. This is what we expect from these methods: statistical procedures may permit to estimate good quantiles, but ensembles are expected to provide an additional information about uncertainty and its variability.

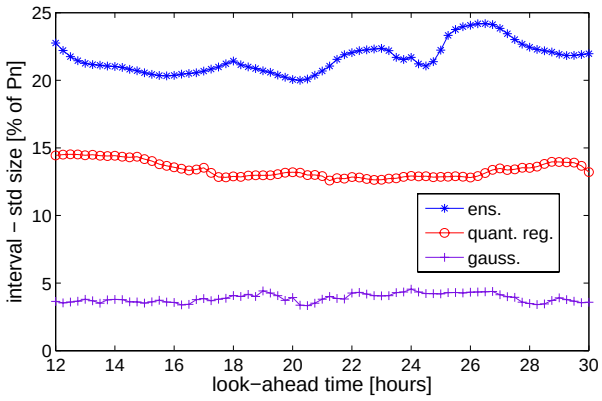

FIGURE 3: Resolution evaluation - Standard deviation of the size of the central 50% prediction intervals as a function of the look-ahead time.

Both sharpness and resolution can be assessed as a function of some other explanatory variables, such as predicted power (or wind speed). This permit to describe how the methods behave in various conditions. However, one then has to verify first the conditional reliability of the methods as a function of the considered explanatory variable.

Finally, we consider the skill score that we previously introduced (cf. Equation (7)). The set of quantiles consists of the two quartiles (i.e. $\alpha_1 = 0.25$ and $\alpha_2 = 0.75$). We put $s_i(p) = 4p, (i = 1, 2)$, and $f(p) = -2p$ following Gneiting [13] for the intuitive appeal of the resulting score, which rewards the forecaster for narrow intervals and gives a penalty if they do not cover the related observations. The best approach is the one that reaches the highest score. Figure 4 depicts the evolution of the skill score as a function of the look-ahead time for the three approaches.

There is a trend that the score value decreases for longer horizons. This meets the general statement that it is harder to forecast events that are further in the future. Though, what is surprising about Figure 4 is that the three different approaches have a similar skill, and they all have a range of horizons for which they perform better than the two others. This would actually mean that using a basic Gaussian assumption, developing an advanced approach based on statistical methods, or investing in the use of ensemble forecasting would lead to probabilistic forecasts of the same quality. This cannot be envisaged and we have shown above that even

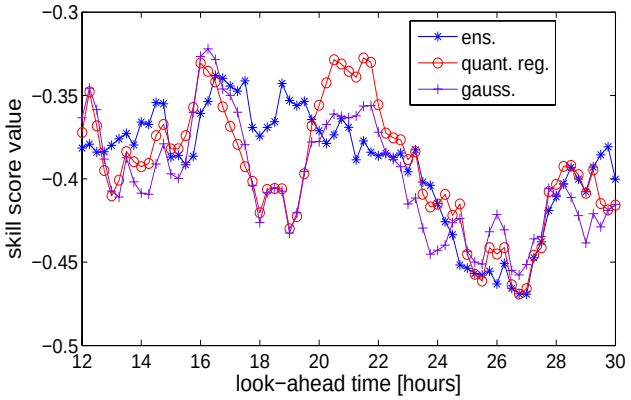

FIGURE 4: Skill score value as a function of the prediction horizon for the three interval forecasting approaches.

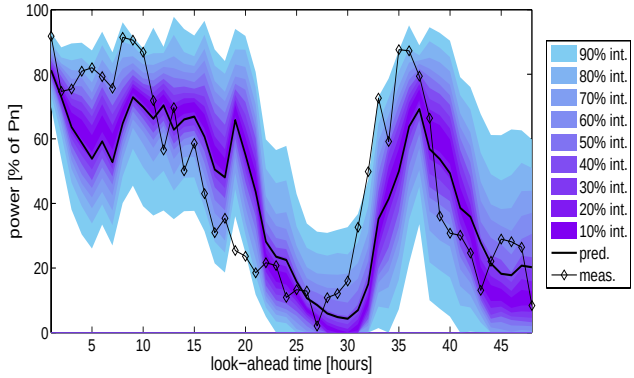

FIGURE 5: Example of wind power point prediction associated with a set of interval forecasts. The point predictions are given by WPPT and interval forecasts are estimated consequently with the adapted resampling method.

if we could not conclude on the reliability aspect, the two advanced approaches have nicer properties, i.e. distribution-free quantile estimation and variability of uncertainty estimations. Thus we show here that the use of that single score only can lead to misleading conclusions regarding the quality of evaluated approaches.

As a conclusion, we advise not to use a unique skill score for evaluating or comparing methods, unless the reliability aspect has been validated on a prior analysis. Then, maximizing the skill score value would be equivalent to maximizing sharpness.

6 Sensitivity analysis on the skill of an uncertainty estimation method

Adapted resampling is a distribution-free approach for the estimation of quantiles and interval forecasts suitable for non-linear processes. This approach has recently been applied to the wind power forecasting problem [25, 27] and evaluated on a variety of case-studies consisting of periods ranging from several months to several years for several European wind farms. This method provides probabilistic forecasts of wind generation based on the past performance of the point prediction method it is associated to. Here, adapted resampling is applied for post-processing WPPT forecasts (WPPT standing for Wind Power Prediction Tool [23]) for the Tunø Knob wind farm. The evaluation set is composed by 3200 series of 48-hour ahead hourly probabilistic forecasts (4.5 months) given by sets of estimated quantiles for proportion 0.05, 0.1,..., 0.95. This corresponds to central prediction intervals for nominal coverage rates 10, 20,..., and 90%.

Figure 5 depicts an episode with the wind power point predictions for the following 48 hours compared to the measured power values. The prediction intervals are associated to the point prediction and represented in the form of a *fan chart*. Since adapted resampling is a non-parametric method (i.e. quantiles are estimated without assuming a specified distribution), the resulting intervals are asymmetric.

Here, our aim is to show how the introduced evaluation criteria may be used for assessing the influence of a given method parameters on its quality. The degrees of freedom of the adapted resampling method are the number of fuzzy sets for mapping the range of possible predicted power values, the number of elements in error samples, and the number of bootstrap replications. In the following Paragraphs, we will focus on the first two degrees of freedom and leave aside the effect of the number of bootstrap replications.

6.1 Influence of the power curve mapping

The idea of adapted resampling is to propose a situation-dependent assessment of the forecast uncertainty: fuzzy logic is used for mapping several zones with different characteristics of the prediction uncertainty. Regarding wind generation, we know that the level of predicted power greatly affects the related level of forecast uncertainty [16]. Also, the level of predicted wind speed has an effect on the error distribution characteristics, since for very low wind speeds there is no risk of cut-off event, while for wind speed higher than 20 m.s^{-1} the cut-off risk is present. Figure 6 depicts an example of the mapping of the power curve for considering both of these effects: three fuzzy sets are used on the power range and two on the wind speed range. Due to the few occurences of cut-off events in the present dataset, we concentrate the analysis on the impact of the number of fuzzy sets used to map the predicted power range both on reliability and resolution. It is expected that increasing the number of fuzzy sets will mainly have an effect on the resolution aspect since it is its primary aim.

Three possibilities are envisaged: using only one fuzzy set on the power range (which is equivalent to using a classical resampling method), and a mapping with alternatively 3 or 5 fuzzy sets. We set the sample size to 300 elements and the number of bootstrap replications to 50.

For assessing the reliability of the probabilistic forecasts produced with these three settings, we use a reliability diagram which gives the deviation from 'perfect reliability', as a function of the nominal coverage of the quantiles (Figure 7). These deviations are shown as the average deviations over the whole range of look-ahead times. The figures given in the legend are the average absolute deviations from 'perfect

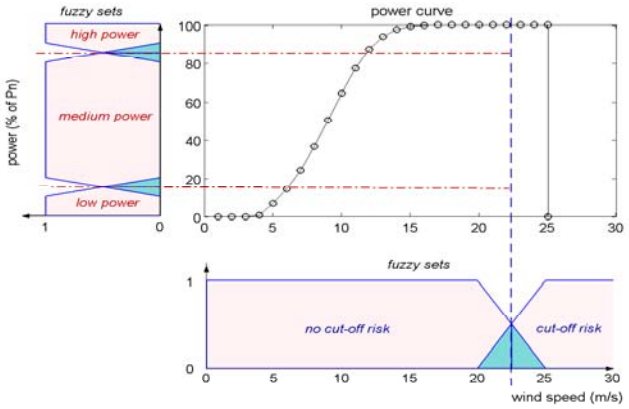

FIGURE 6: Situation dependence is achived through the mapping of the power curve with three power fuzzy sets and two cut-off risk zones.

reliability' (over the various nominal coverages and forecast horizons).

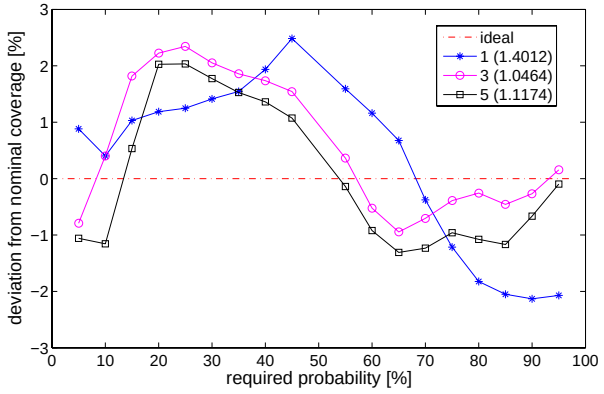

FIGURE 7: Reliability diagram for evaluating the influence of the power curve mapping on the resulting probabilistic forecasts reliability.

One sees from Figure 7 that the deviations from 'perfect reliability' are of the same order for the various settings: it is less than $\pm 2.5\%$. Even if it is not the primary aim of the mapping, it seems that using several fuzzy sets permit to increase the overall reliability of estimated quantiles. In this example, the average absolute deviation is 1.40 % for classical reampling intervals while it is 1.05 and 1.1% for the two other settings.

Then, we turn our attention to sharpness and resolution. Since sharpness proves to be similar for the three settings and that the power curve mapping mainly has an effect on the resolution aspect, we only focus here on the latter. Figure 8 gives the standard deviation of the interval size as a function of the interval nominal coverage rate. As an example, we focus on the σ-diagram of 24-hour ahead probabilistic forecasts. When going from classical to adapted resampling, the resolution is significantly augmented, whatever the level of confidence. For the example of the 50%-confidence prediction interval, the standard deviation of the interval size is actually multiplied by a factor 3. Also, one sees that by using more fuzzy sets for mapping the power curve, the resolution can be increased even more, mostly for high degrees of confidence. This means that the method has a better ability to differenciate the tail ends of predictive distributions. In a general manner, increasing the resolution of probabilistic forecasting methods is expected to give them more value for the management or the trading of wind generation.

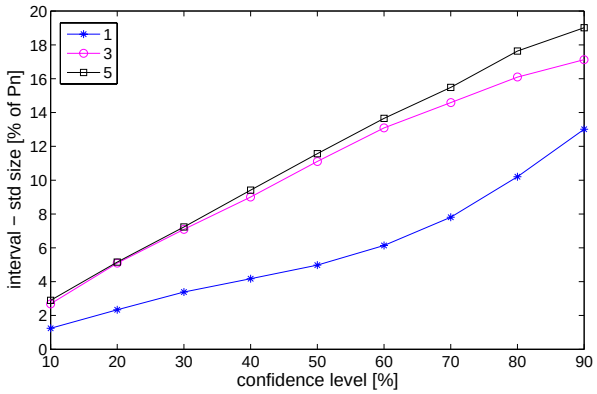

FIGURE 8: σ-diagram for 24-hour ahead forecasts for evaluating the influence of the power curve mapping on the resulting probabilistic forecasts resolution.

6.2 Influence of the sample size

The second part of the study concerns the influence of the sample size, i.e. the number of past prediction errors, on the skill of the estimated intervals provided by the adapted resampling method. Intuitively, considering more past errors should permit to better understand the uncertainty of the process and thus to augment the reliability of estimated predictive distributions. However, considering very large error samples would make the approach less dynamic. Here, the number of fuzzy sets is set to five and the number of bootstrap replications to 50. We produce probabilistic forecasts with error samples containing 50, 100, 200 and 300 elements.

The reliability diagram given by Figure 9 shows how the sample size affects the predictive distributions reliability. The absolute average deviation from 'perfect reliability' greatly diminishes as we use more elements in the resampling procedure. This absolute deviation is divided by 2 if considering the last 300 errors instead of dealing with the last 50 only. One notes from the reliability diagram that quantiles for proportions below 0.5 are overestimated while quantiles over the median are underestimated. This tells that intervals are too narrow on average. Augmenting the sample size diminishes that trend. It should be understood here that having too narrow intervals is more likely than having too large intervals: methods for estimating future uncertainty usually rely on past experience of a given model performance and then do not integrate the additional uncertainty of predicting new data [5].

For evaluating the sharpness of the interval forecasts, we superpose δ-diagrams for the various method settings (Figure 10). This Figure is for 24-hour ahead probabilistic forecasts. In a general manner, the average interval size

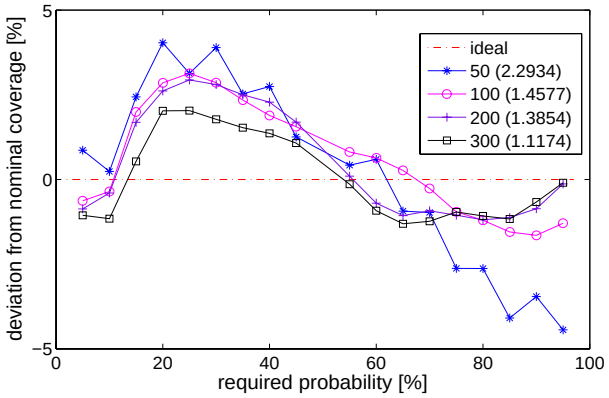

FIGURE 9: Reliability diagram for evaluating the influence of error sample size on the resulting probabilistic forecasts reliability.

ranges from ∼5% of nominal power for intervals at a 10%-confidence to ∼50% for those associated with a 90% degree of confidence. Whatever the nominal coverage rate, the average size decreases when considering more past prediction errors for estimating predictive distributions. This diminution in the mean size is up to 10% when going from 50 to 300 sample elements. Therefore, by increasing the sample size (up to 300 items here), we improve both the forecasts reliability and their sharpness. Note that these conclusions may be generalized for larger samples. Though, using too large samples may not be desirable from an operational point of view, due to the data storage requirements if considering e.g. a high-resolution mapping of the power curve or several applications at the same time.

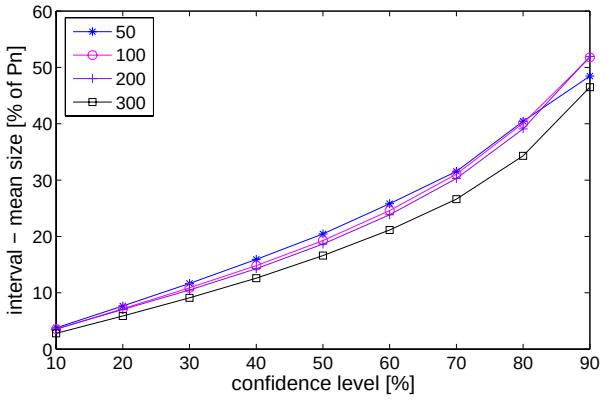

FIGURE 10: δ-diagram for 24-hour ahead forecasts for evaluating the influence of the error sample size on the resulting probabilistic forecasts sharpness.

7 Conclusions

In this paper, we have adressed the issue of evaluating probabilistic forecasts of wind power. The concepts of overall skill, reliability, sharpness and resolution have been introduced. Reliability, which stands for the forecasts' ability of respecting nominal probabilistic coverages (e.g. 50% interval forecasts should actually cover approximately 50% of observations over the evaluation set), is a primary requirement. Then, sharpness and resolution, which relates to the method's ability to provide a situation-dependent assessment of the forecast uncertainty, represents the added value. By describing various criteria for evaluating either quantile or interval forecasts, it has been emphasized that a unique skill score cannot give the whole information on what contributes to a given prediction method performance. It is thus necessary to first have a look at the reliability assessment and to focus on sharpness and resolution in a second stage. In addition, we have underlined the fact that when assessing the reliability of probabilistic forecast by the mean of hypothesis tests one should be careful regarding the underlying independence assumption. Indeed, hit-miss sequences are correlated and thus usual tests for unconditional coverage cannot be applied directly.

An appropriate set of measures and diagrams have been considered for comparing the current methods for estimating probabilistic wind power forecasts. These methods are based either on quantile prediction from advanced statistical methods or on non-linear transformation of ensemble forecasts of meteorological variables. The results drawn on reliability are sensitive on the size of the evaluation set that proves to be rather small here. However if one considers average reliability for all time steps then the quantile regression method seems to outperform the two others. Were the ensemble-based intervals recalibrated that they would certainly be more reliable. Consequently, a study of the prediction intervals' resolution revealed that both advanced methods can provide a situation-dependent assessment of the uncertainty. It will be of particular interest to further study these rival approaches for better highlighting their (maybe complementary) qualities.

In a final part of the paper, we have focused on a particular statistical quantile estimation method (adapted resampling), with the aim of assessing the influence of its degrees of freedom on its resulting performance. We showed that the mapping of the power curve was mainly improving the method resolution, while augmenting the sample size was increasing both the reliability and sharpness of estimated predictive distributions.

This paper concentrates on the *quality* of probabilistic forecasts of wind power. This means that we have considered the evaluation of their performance from a statistical point of view. However, as these forecasts are meant for the management or trading of wind generation, it would also be worth evaluating their *value*. The value corresponds to the increased benefits (economical or not) resulting from the daily use of these forecasts in an operational context.

Acknowledgements

This work was performed in the frame of the ANEMOS Project (ENK5-CT2002-00665) funded in part by the European Commission. The authors gratefully acknowledge ELSAM and the European Centre for Medium-Range Weather Forecasts for providing data for the realization of

the study.

References

[1] F. Atger. The skill of ensemble prediction systems. *Monthly Weather Review*, 127:1941–1957, September 1999.

[2] R. T. Baillie and T. Bollerslev. Prediction in dynamic models with time-dependant conditional variances. *Journal of Econometrics*, 51:91–113, 1992.

[3] G. N. Bathurst, J. Weatherhill, and G. Strbac. Trading wind generation in short-term energy markets. *IEEE Trans. on Power Syst.*, 17(3):782–789, August 2002.

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

[5] C. Chatfield. *Time-Series Forecasting*. Chapman & Hall/CRC, 2000.

[6] C. Chatfield. *The Analysis of Time Series: An Introduction*. Chapman & Hall/CRC, 6th edition, 2003.

[7] P. F. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39(4):841–862, 1998.

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

[9] W. J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, 1999.

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

[11] A. Fabbri, T. G. Gómez San Román, J. R. RivierAbbad, and V. H. Méndez Quezada. Assessment of the cost associated with wind generation prediction errors in a liberalized electricity market. *IEEE Trans. on Power Syst.*, 20(3):1440–1446, 2005.

[12] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. Technical report, University of Washington, Department of Statistics, May 2005. Technical report no. 483.

[13] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. Technical report, University of Washington, Department of Statistics, September 2004. Technical report no. 463.

[14] P. Hall and A. Rieck. Improving coverage accuracy of nonparametric prediction intervals. *J. Royal Stat. Soc.*, 63(4):717–725, 2001.

[15] T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 2000. Notes and Correspondance.

[16] M. Lange. On the uncertainty of wind power predictions - Analysis of the forecast accuracy and statistical distribution of errors. *Trans. ASME, J. Solar Energy Eng.*, 127(2):177–184, May 2005.

[17] H. Madsen, H. Aa. Nielsen, T. S. Nielsen, G. Kariniotakis, and P. Pinson. A protocol for standardizing the performance evaluation of short-term wind power prediction models. In *CD-Proceedings of the 2004 Global Windpower Conference, Chicago, U.S.*, March 2004.

[18] J. Mason. Definition of technical terms in forecast verification and examples of forecast verification scores. In *Proc. of the IRI Workshop on Forecast Quality, New York*, October 2000.

[19] S. McNees. Forecast uncertainty: can it be measured? mimeo, Federal Reserve Bank of Boston, Boston, 1995.

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

[21] H. Aa. Nielsen, H. Madsen, and T. S. Nielsen. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. In *Proc. of the 2004 European Wind Energy Association Conference, EWEC'04, London, United Kingdom*, pages 34–38, November 2004. Scientific Track.

[22] H. Aa. Nielsen, T. S. Nielsen, H. Madsen, J. Badger, G. Giebel, L. Landberg, K. Sattler, and H. Feddersen. Wind power ensemble forecasting. In *CD-Proceedings of the 2004 Global Windpower Conference, Chicago, U.S.*, March 2004.

[23] T. S. Nielsen and H. Madsen. ZEPHYR - The prediction models. In *Proc. of the 2001 European Wind Energy Association Conference, EWEC'01, Copenhagen, Denmark*, pages 868–871, June 2001.

[24] P. Pinson, C. Chevallier, and G. Kariniotakis. Optimizing benefits from wind power participation in electricity markets using advanced tools for wind power forecasting and uncertainty assessment. In *Proc. of the 2004 European Wind Energy Association Conference, EWEC'04, London, United Kingdom*, November 2004.

[25] P. Pinson and G. Kariniotakis. On-line adaptation of confidence intervals based on weather stability for wind power forecasting. In *CD-Proceedings of the 2004 Global Windpower Conference, Chicago, U.S.*, March 2004.

[26] P. Pinson and G. Kariniotakis. On-line assessment of prediction risk for wind power production forecasts. *Wind Energy*, 7(2):119–132, May-June 2004.

[27] P. Pinson, T. Ranchin, and G. Kariniotakis. Short-term wind power prediction for offshore wind farms - Evaluation of fuzzy-neural network based models. In *CD-Proceedings of the 2004 Global Windpower Conference, Chicago, U.S.*, March 2004.

[28] N. Ravishankar, L. Shiao-Yen Wu, and J. Glaz. Multiple prediction intervals for time-series: comparison of simultaneous and marginal intervals. *Journal of Forecasting*, 10:445–463, 1991.

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

[30] S.-E. Thor and P. Weis-Taylor. Long-term research and development needs for wind energy for the time frame 2000-2020. *Wind Energy*, 5:73–75, April-June 2003.

[31] Z. Toth, O. Tallagrand, G. Candille, and Y. Zhu. Probability and ensemble forecasts. In I.T. Jolliffe and D.B. Stephenson, editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, Ltd, 2003.

[32] A. Zervos. Developing wind energy to meet the Kyoto targets in the European Union. *Wind Energy*, 6(3):309–319, July-September 2003.

ϕ	Level of	Horizon (steps)											
	significance	1	2	3	4	5	6	7	8	9	10	11	12
0.0	0.05	45	45	45	45	45	45	45	45	45	45	45	45
	0.10	93	93	93	93	93	93	93	93	93	93	93	93
0.9	0.05	46	109	181	243	288	345	359	381	407	440	449	482
	0.10	91	189	270	332	396	432	454	472	495	514	532	567

TABLE 2: Number of cases out of 1000 in which the exact binomial test rejected the hypothesis of reliability.

Appendix: The influence of correlation on tests and estimates of reliability

The purpose of this appendix is to demonstrate that correlation inherent in the forecasts influences the precision by which e.g. actual coverage of an interval forecast is estimated and that tests are invalidated by the correlation.

Let x_{t+k} be the variable at time $t+k$ for which forecasts are generated at time t. In the simulation study it is assumed that the stochastic process $\{x_t\}$ is known and therefore it is possible to generate interval forecasts which are *assured to be reliable*. Specifically, an $AR(1)$ process is used:

$$x_t = \phi x_{t-1} + e_t \,, \tag{11}$$

where $\phi \in (-1, 1)$ is the pole of the process and $\{e_t\}$ is zero-mean Gaussian white noise with variance $1-\phi^2$, whereby $\{x_t\}$ has unit variance.

The k-step ahead forecast, i.e. the conditional mean, can be expressed as (see [6], p. 104)

$$\hat{x}_{t+k/t} = \phi^k x_t \,. \tag{12}$$

Since $\{x_t\}$ is a Gaussian process, $\hat{x}_{t+k/t}$ can also be interpreted as a 50% quantile forecast. For this quantile forecast $I_{t,k}^{(0.5)}$, $n_{k,1}^{(0.5)}$, $n_{k,0}^{(0.5)}$, and $\hat{a}_k^{(0.5)}$ are determined. Furthermore, a test for reliability is performed. Since the forecasts are known to be reliable the test should reject the hypothesis of reliability with a probability corresponding to the level of significance used. To avoid issues regarding approximative tests, like e.g. the χ^2-test, an exact binomial test is used (see [9], Section 3.1).

In the simulations 1- to 12-step ahead forecasts are considered. For each of $\phi = 0$ and $\phi = 0.9$, 1000 series of length 500 are simulated. For $\phi = 0$ the process $\{x_t\}$ is white noise and the exact binomial test should reject the hypothesis of reliability in approximately 50 of the 1000 cases for a 0.05 significance level (and respectively 100 out of 1000 for the 0.1 significance level). Table 2 displays the number of times the test were rejected for two levels of significance. From the table it is seen that when the process is correlated the test does not perform according to specifications for 2- to 12-step ahead forecasts. For 1-step ahead forecasts and for $\phi = 0$ (all horizons) the test seems to perform according to the nominal level of significance.

To further understand these aspects consider the histograms of the 1000 realizations of the estimate $\hat{a}_k^{(0.5)}$ for horizons 1, 3, 6, and 12 shown in Figure 11. It is seen that in all cases the distribution is centered around the nominal value of 0.5, corresponding to the 50% quantile forecast. However, the spread of the distribution increases with the horizon. In fact for 1-step ahead forecasts the central 95% of the values are between 0.45 and 0.54 and for 12-step ahead forecasts the corresponding limits are 0.38 and 0.62.

The implication of these findings is that due to correlation the uncertainty of the estimate $\hat{a}_k^{(\alpha)}$ can be much larger than expected from standard theory based on an assumption of independence. In turn this often results in a deviation from the nominal value which is much larger than expected and therefore a test based on the assumption of independent samples rejects the hypothesis of reliability far too often.

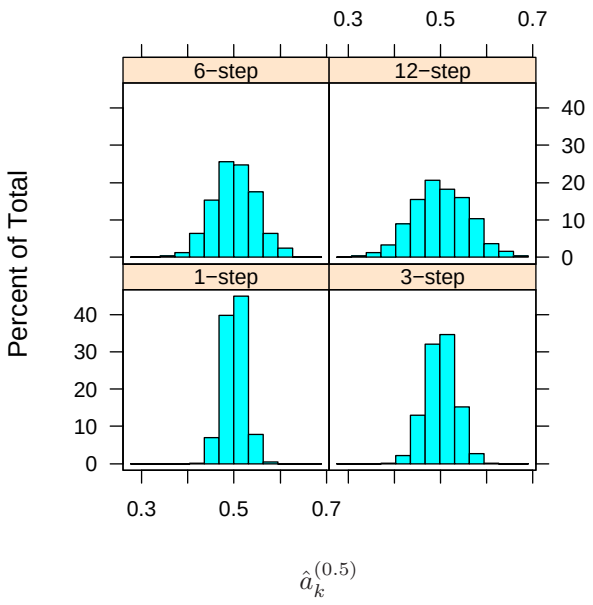

FIGURE 11: Histograms of the 1000 realizations of $\hat{a}_k^{(0.5)}$ for horizons 1, 3, 6, and 12.