Downloaded from orbit.dtu.dk on: Nov 07, 2026

DTU Library

Reliability diagrams for non-parametric density forecasts of continuous variables

Accounting for serial correlation

Pinson, Pierre; McSharry, Patrick; Madsen, Henrik

Published in:
Quarterly Journal of the Royal Meteorological Society

Link to article, DOI:
10.1002/qj.559

Publication date:
2010

Document Version
Early version, also known as pre-print

Link back to DTU Orbit

Citation (APA):
Pinson, P., McSharry, P., & Madsen, H. (2010). Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. *Quarterly Journal of the Royal Meteorological Society*, *136*(646), 77-90. https://doi.org/10.1002/qj.559

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **00**: 1–16 (2009)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.000

Reliability diagrams for nonparametric density forecasts of continuous variables: accounting for serial correlation

Pierre Pinson[1*], Patrick McSharry[2,3,4], Henrik Madsen[1]
[1] *DTU Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark*
[2] *Saïd Business School, Oxford, UK*
[3] *Mathematical Institute, University of Oxford, Oxford, UK*
[4] *Smith School of Enterprise and the Environment, Oxford, UK*

Abstract: Reliability is seen as a primary requirement when verifying probabilistic forecasts since a lack of reliability would introduce systematic bias in subsequent decision-making. Reliability diagrams comprise popular and practical diagnostic tools for the reliability evaluation of density forecasts of continuous variables. Such diagrams relate to the assessment of the unconditional calibration of probabilistic forecasts. A reason for their appeal is that deviations from perfect reliability can be visually assessed based on deviations from the diagonal. Deviations from the diagonal may however be caused by both sampling effects and serial correlation in the forecast-verification pairs. We build on a recent proposal consisting of associating reliability diagrams with consistency bars that would reflect the deviations from the diagonal potentially observable even if density forecasts were perfectly reliable. Our consistency bars however reflect potential deviations originating from the combined effects of limited counting statistics and of serial correlation in the forecast-verification pairs. They are generated based on an original surrogate consistency resampling method. Its ability to provide consistency bars with a significantly better coverage against the i.i.d. resampling alternative is shown from simulations. A practical example of the reliability assessment of nonparametric density forecasts of short-term wind power generation is finally given. Copyright © 2009 Royal Meteorological Society

KEY WORDS probabilistic forecasting; verification; calibration; surrogate; consistency resampling; wind power

Received ; Revised ; Accepted

1 Introduction

Over the past few decades, one of the major breakthroughs in forecasting meteorological variables for applications such as weather derivatives and renewable energy generation, comes from the transition from point to probabilistic forecasting (Gneiting , 2008a). One has to acknowledge the significant contribution of some of the leading meteorological centres e.g. NCEP (National Centres for Environmental Prediction) and ECMWF (European Centre for Medium-Range Weather Forecasts) in developing ensemble forecasting systems, as well as a probabilistic view of meteorological forecasting. For a detailed overview of ensemble forecasting and the underlying probabilistic forecasting philosophy developed in the meteorological community, the reader is referred to Leutbecher and Palmer (2008) and references therein. From a decision-making perspective, it has been shown that pricing weather derivatives based on density forecasts would bring significant benefits (Taylor and Buizza , 2006). In parallel for the example case of renewable energy, the production of which is a direct function of meteorological conditions, it is argued that optimal management and trading of generated energy should be based on probabilistic forecasts, see e.g. Pinson *et al.* (2007a). The transition from point to probabilistic forecasts is not only observed in the meteorological literature, as probabilistic forecasts are also becoming customary products in economics and finance (Abramson and Clemen , 1995; Tay and Wallis , 2000; Timmermann , 2000) or more generally in management sciences. When considering continuous variables such as wind speed, the most complete information about expected realization for a given lead time takes the form of a density forecast (equivalently referred to as predictive distribution), giving the probability density function of the corresponding random variable. Having a broader view of decision-making for real-world problems, Gneiting (2008b) argues that for a large class of cost functions of forecast users, optimal decisions directly relate to given quantiles of predictive densities of the variable of interest.

In parallel to these developments towards probabilistic forecasting and subsequent optimal decision-making, the issue of probabilistic forecast verification has attracted significant attention. For a recent overview of verification methods for probabilistic forecasts of categorical and continuous variables in atmospheric sciences, see e.g. Jolliffe and Stephenson (2003). A primary requirement for probabilistic forecasts relates to their calibration (equivalently referred to as their reliability) which corresponds to their probabilistic correctness. We hereby follow the paradigm introduced by Gneiting *et al.* (2007), i.e. based on maximizing the sharpness of predictive distributions subject to calibration. Note that this framework

*Correspondence to: DTU Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark. Tel: +45 4525 3428, fax: +45 4588 2673, email: pp@imm.dtu.dk

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls *[Version: 2007/01/05 v1.00]*

for probabilistic forecast evaluation is different from that based on testing for correct conditional coverage of density forecasts, as described in e.g. (Christoffersen , 1998; Diebold *et al.*, 1998), where focus is given to one-step ahead forecast only, in a time-series framework. The calibration requirement mentioned above calls for a thorough reliability assessment prior to proceeding with sharpness. Even if considering the use of proper scoring rules, see Bröcker and Smith (2007b) among others, calculating scores alone does not permit to see whether better score values come from higher reliability or increased sharpness. A decomposition of these proper scores into their reliability and sharpness components should then be performed. This is since a reliability assessment is necessary in order to make sure that systematic bias would not be introduced in further decision-making.

A popular and straightforward way of assessing calibration of probabilistic forecasts is via the use of reliability diagrams (Atger , 1998, 2004). The reliability diagrams we consider here for the reliability assessment of nonparametric density forecasts of continuous variables consist of an equivalent cumulative version of the popular probability integral transform (abbreviated PIT) histograms – otherwise called Talagrand diagrams, used for the verification of ensemble forecasts, see e.g. Hamill (2001). An extensive presentation of reliability diagrams for density forecasts of continuous variables (and equivalently for ensembles) is given in Section 2.

Recently for the case of probability forecasts of binary events, Bröcker and Smith (2007a) have explained how reliability diagrams may be misinterpreted, since even for perfectly reliable probabilistic forecasts there will always be deviations from the diagonal originating from sampling effects. They have proposed an elegant framework permitting one to easily integrate information about the impact of sampling effects directly in reliability diagrams. While their point is highly relevant, it is argued here that it is not only counting statistics but also serial correlation in forecast-verification pairs that affect interpretation of reliability diagrams produced from finite-size datasets. Somehow, the methodology proposed by Bröcker and Smith (2007a) is based on the fallacy that for reliable forecasts the random variables whose realizations are given by the probability integral transforms should be independent and identically distributed (i.i.d.) $\mathrm{U}[0,1]$. It is true that by definition, a necessary condition for density forecasts to be reliable is that such random variables are distributed $\mathrm{U}[0,1]$. It is not a necessary condition however for successive realizations to be independent. In the case of reliability diagrams for probabilistic forecasts of binary events, it might be acceptable to assume independence as forecast-verification pairs corresponding to a given probability class may be sparsely and randomly distributed over the verification period. It cannot be the case when assessing reliability of probabilistic forecasts of multi-categorical variables or of density forecasts of continuous variables. This will be illustrated by a simple practical example based on climatology density forecasts in Section 3. The role of serial correlation and sampling effects has already been hinted by e.g. Pinson *et al.* (2007b) for the case of the verification of wind power density forecasts, or by Hamill (2001) when considering verification of ensemble forecasts of meteorological variables. Section 4 introduces a methodology to provide consistency bars in the spirit of Bröcker and Smith (2007a), but accounting for serial correlation effects, using an original surrogate consistency resampling method. Simulations in Section 5 show how serial correlation adds to the sampling effect, and demonstrates the pitfall stemming from inference on the confidence one may have in reliability diagrams based on an i.i.d. assumption for probability integral transforms of reliable density forecasts. It also evaluates the validity and accuracy of the methodology introduced. The calibration assessment of short-term forecasts of wind power generation serves as an illustrative application in Section 6. Concluding remarks end the paper in Section 7.

2 Reliability diagrams for nonparametric density forecasts of continuous variables

Using t as the time index, we denote by $\{Y_t\}$ the stochastic process of interest and by $\{y_t\}$ the time-series of observed realizations. Each random variable Y_t can take values in $\mathcal{C} \subset \mathbb{R}$, with for instance $\mathcal{C} = \mathbb{R}^+$ for the case of wind speed. For simplicity, it is assumed that random variables and corresponding realizations are equally spaced in time. Also, it is considered that focus is on this stochastic process at a single location only, as reliability assessment of probabilistic forecasts for spatial fields would also be impacted by spatial sampling and correlation effects which are not treated here. Discussions on these spatial effects can be found in Wilks (1995) and in Jolliffe and Stephenson (2003).

Write $\hat{f}_{t+k|t}(y)$ a density forecast for the stochastic process of interest issued at time t for lead time $t+k$, and $\hat{F}_{t+k|t}(y)$ the related cumulative distribution function. $\hat{f}_{t+k|t}(y)$ is a forecast of the probability distribution function of Y_{t+k} given the information set available at time t, which may be derived from ensemble forecasts or from some nonparametric statistical forecasting techniques e.g. quantile regression. In a general manner, if no assumption is made about the shape of predictive distributions, a nonparametric density forecast $\hat{f}_{t+k|t}(y)$ can be summarized by a set of m quantile forecasts

$$\hat{f}_{t+k|t}(y) = \{\hat{q}^{(\alpha_i)}_{t+k|t} \,|\, 0 \leq \alpha_1 < \ldots < \alpha_i < \ldots < \alpha_m \leq 1\} \tag{1}$$

that is, with chosen nominal proportions α_i spread on the unit interval. In turn, a quantile forecast $\hat{q}^{(\alpha_i)}_{t+k|t}$ with nominal proportion α_i is defined by

$$\mathrm{P}_{\hat{F}_{t+k|t}}\left[y_{t+k} < \hat{q}^{(\alpha_i)}_{t+k|t}\right] = \alpha_i, \quad \alpha_i \in [0,1] \tag{2}$$

Both forecasts and observations are available for a limited time period $\mathcal{T} \subset \mathbb{N}^+$ used for forecast verification. Note that for ensemble forecasts with exchangeable members, nonparametric density forecasts can also be defined

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

as in (1). In such case, by ordering the J ensemble members in ascending order, the j^{th} member gives the quantile with nominal proportion $\alpha_j = j/(J+1)$. The (continuous) density forecast $\hat{f}_{t+k|t}(y)$ can then be built by interpolation through the set of quantiles.

For a given forecast horizon k, the core concept behind the use of reliability diagrams for evaluating density forecasts of continuous variables is that the series of predictive densities $\{\hat{f}_{t+k|t}(y)\}$ are reliable if and only if the random variable $Z_{t,k} = \hat{F}_{t+k|t}(Y_{t+k})$ is distributed U$[0,1]$. In practice this is performed by studying the realizations of $Z_{t,k}$ which are given by the sequence $\{z_{t,k}\}$ of probability integral transforms, with $z_{t,k} = \hat{F}_{t+k|t}(y_{t+k})$.

Such a definition of reliability actually corresponds to an unconditional calibration of the density forecasts, since no distinction is made between different time points in the data set, or other conditions that may affect density forecast reliability. This contrasts with the idea of conditional calibration introduced by Christoffersen (1998) and Diebold *et al.* (1998) for the case of density forecasts in a time-series context. It also contrasts with the idea of conditional reliability assessment presented by Pinson *et al.* (2007b) and with that of forecast stratum introduced by Bröcker (2009). Both Pinson *et al.* (2007b) and Bröcker (2009) base their argument on the fact that for the case of nonlinear processes, the reliability of probabilistic forecasts may be influenced by a set of external factors, or may even simply vary as a function of the forecasts themselves. What we refer to as unconditional calibration thus corresponds to the overall probabilistic bias of the density forecasts as discussed by Murphy (1993) and Taylor (1999) for instance, or to what Gneiting *et al.* (2007) refers to as probabilistic calibration.

A consequence of the fact that reliability diagrams relate to an unconditional calibration assessment is that one cannot assume that the sequence of probability integral transforms $\{z_{t,k}\}$ for a sequence of reliable forecast densities and corresponding verifications is to be i.i.d. This would be the case only if for each given time of the evaluation set density forecasts were equal to the true conditional densities of the stochastic data generating process (Diebold *et al.*, 1998; Gneiting *et al.*, 2007).

In practice, since nonparametric density forecasts as defined by (1) consist of a collection of quantile forecasts for which the nominal proportions are known, evaluating the reliability of density forecasts is achieved by verifying the reliability of each individual quantile forecast. Let us introduce in a first stage the indicator variable $\xi_{t,k}^{(\alpha_i)}$. Given a quantile forecast $\hat{q}_{t+k|t}^{(\alpha_i)}$ issued at time t for lead time $t+k$, and the verification y_{t+k}, $\xi_{t,k}^{(\alpha_i)}$ is given by

$$\begin{aligned} \xi_{t,k}^{(\alpha_i)} &= \mathbf{1}\{y_{t+k} < \hat{q}_{t+k|t}^{(\alpha_i)}\} = \mathbf{1}\{z_{t,k} < \alpha_i\} \\ &= \begin{cases} 1, & \text{if} \quad y_{t+k} < \hat{q}_{t+k|t}^{(\alpha_i)} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

It can be appealing to project into the standard Gaussian domain by using an inverse Normal transformation, i.e. based on the inverse probit function Φ^{-1},

$$\Phi^{-1} \; : \; p \rightarrow \sqrt{2}\text{erf}^{-1}(2p-1) \tag{4}$$

with 'erf^{-1}' the inverse error function. The resulting random variables and corresponding realizations are denoted $\tilde{Z}_{t,k}$ and $\tilde{z}_{t,k}$, respectively, with simply $\tilde{Z}_{t,k} = \Phi^{-1}(Z_{t,k})$ and $\tilde{z}_{t,k} = \Phi^{-1}(z_{t,k})$. Considering $\{\tilde{z}_{t,k}\}$ instead of $\{z_{t,k}\}$ for assessing calibration of density forecasts has some advantages discussed in the literature, see e.g. Berkowitz (2001), for hypothesis testing or for studying potential serial correlation. In the present case, it will permit us to apply classical tools from linear time-series analysis. Consequently, the indicator variable introduced in (3) can be equivalently defined as

$$\xi_{t,k}^{(\alpha_i)} \; = \; \mathbf{1}\{\tilde{z}_{t,k} < q_G^{(\alpha_i)}\} \tag{5}$$

with $q_G^{(\alpha_i)}$ the quantile with proportion α_i of a standard Gaussian distribution. In the following, we will equivalently consider the possibility of working with the indicator variable definitions of (3) and (5) in order to assess the reliability of density forecasts.

The time-series $\{\xi_{t,k}^{(\alpha_i)}\}$ is a binary sequence that corresponds to the series of hits (if the verification y_{t+k} lies below the quantile forecast) and misses (if otherwise) over the evaluation set. It is by studying $\{\xi_{t,k}^{(\alpha_i)}\}$ that one can assess the reliability of a time series of quantile forecasts. Indeed, an estimate $\hat{a}_{k,i}$ of the actual proportion $a_{k,i} = \mathbb{E}[\xi_{t,k}^{(\alpha_i)}]$, for a given horizon k, is obtained by calculating the mean of the $\{\xi_{t,k}^{(\alpha_i)}\}$ time-series over the test set

$$\hat{a}_{k,i} \; = \; \frac{1}{N}\sum_{t=1}^{N} \xi_{t,k}^{(\alpha_i)} \; = \; \frac{n_{k,1}^{(\alpha_i)}}{n_{k,0}^{(\alpha_i)} + n_{k,1}^{(\alpha_i)}} \tag{6}$$

where N is the number of time indices in $\mathcal{T}$, and where $n_{k,1}^{(\alpha_i)}$ and $n_{k,0}^{(\alpha_i)}$ correspond to the sum of hits and misses, respectively. They are calculated with:

$$n_{k,1}^{(\alpha_i)} \; = \; \#\{\xi_{t,k}^{(\alpha_i)} = 1\} \; = \; \sum_{t=1}^{N} \xi_{t,k}^{(\alpha)} \tag{7}$$

$$n_{k,0}^{(\alpha_i)} \; = \; \#\{\xi_{t,k}^{(\alpha_i)} = 0\} \; = \; N - n_{k,1}^{(\alpha)} \tag{8}$$

This measure of observed proportion serves as a basis for drawing reliability diagrams for density forecasts of continuous variables, which give the observed proportions $\hat{a}_{k,i}$ of the quantile forecasts against the nominal ones. They therefore consist of Q-Q plots (Q-Q standing for Quantile-Quantile), which are a classical diagnostic tool in the statistical literature. A particular feature of this definition of reliability diagrams is that each quantile is evaluated individually, which will allow us defining a consistency bar for each nominal proportion independently of the other quantiles. Our consistency bars then are pointwise consistency bars. Also, note that the reliability diagrams we consider here are for density forecasts of continuous variables and thus somewhat different from those considered for probabilistic forecasting

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

of binary/categorical variables e.g. (Bröcker and Smith , 2007a). The argument developed in the present paper regarding the fact that serial correlation effects should be accounted for is still valid though. In parallel, the methodology described in the following can be straightforwardly applied for the case of the verification of ensemble forecasts, owing to the definition employed for nonparametric density forecasts.

When visually inspecting calibration from reliability diagrams, a common intuitive thought is that the closer the observed proportions to the diagonal the better. This is because asymptotically as the number of forecast-verification pairs tends towards infinity, one wishes that

$$\lim_{N\to\infty} \hat{a}_{k,i} \;=\; \alpha_i, \quad \forall i \tag{9}$$

implying that observed proportions asymptotically equal nominal ones. In practice however, evaluation sets consisting of forecast-verification pairs are of finite (and often quite limited) size, and it is not expected that observed proportions exactly lie along the diagonal, even if the density forecasts are perfectly reliable. This issue is discussed in detail in Jolliffe and Stephenson (2003) and in Bröcker and Smith (2007a), while a more general discussion on uncertainty of verification measures can be found in Jolliffe (2007). Our contribution concerns the fact that not only sampling effects but also serial correlation in sequences of forecast-verification pairs may affect the observed reliability of even perfectly reliable density forecasts of continuous variables. A simple mathematical proof of that effect is given in Appendix A.

3 Example of serial correlation in probability integral transforms for reliable probabilistic forecasts

Consider here the issue of density forecasting of some continuous meteorological variable, say wind speed, at a forecast horizon k. It is of common knowledge that climatology comprises a benchmark density forecast, which has the nice property of being well calibrated, and the characteristic of having no resolution since consisting of unconditional density forecasts, see for instance the discussion in Pinson and Madsen (2009). Figure 1(a) depicts a time-series of mean hourly wind speed at the Horns Rev wind farm in Denmark over a period of almost 1,000 hours†. This offshore wind farm has been the first large-scale offshore wind farm worldwide, and has hence motivated a number of studies for e.g. the characterization of local wind characteristics e.g. Vincent *et al.* (2009), or the (probabilistic) forecasting of its power output (Pinson and Madsen , 2009). The time-series of wind speed is normalized by the maximum wind speed observed over the period, consequently taking values in $[0, 1]$.

†Wind speed measurements with a 10-minute resolution for the Horns Rev wind farm may be freely obtained at the website `www.winddata.com`

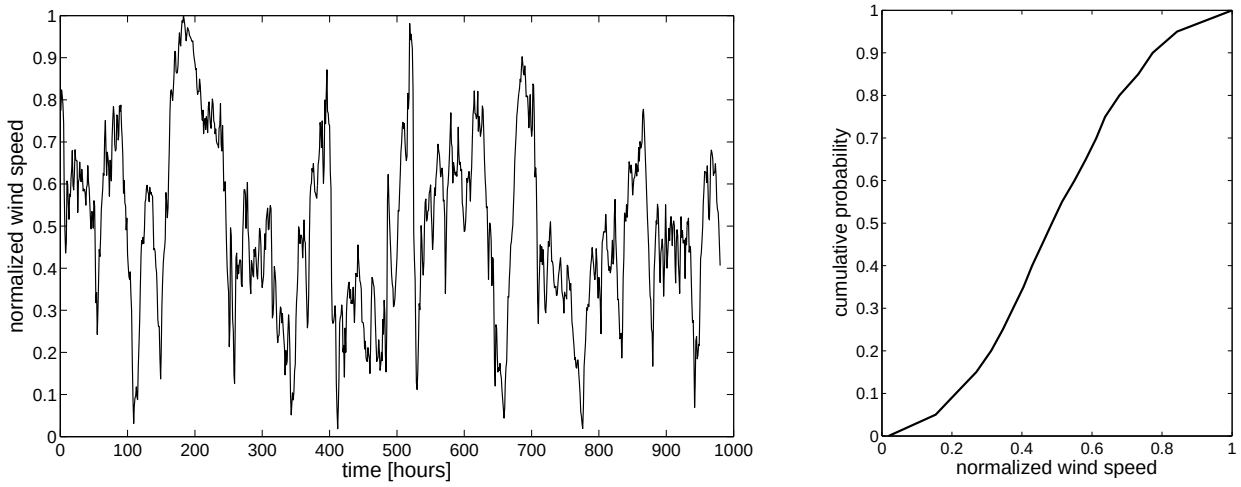
For any time t in this dataset, a perfectly reliable climatology density forecast $\hat{f}_{t+k|t}(y)$ for a given look-ahead time k can be obtained by defining it from the distribution of wind speed hourly averages over this period of 1,000 hours. The corresponding cumulative distribution function $\hat{F}_{t+k|t}$ is represented in Figure 1(b). This is obviously not the way an actual climatology density forecast would be built, since it would be based on a long record times-series of past measurements. For the purpose of our example however, the climatology density forecast we build is similar in essence to actual climatology forecasts based on long records of data since being a unique unconditional density based on recorded data, while we insure it is perfectly calibrated by relying on the measurements themselves. As a consequence, if producing a reliability diagram for evaluating the calibration of such density forecasts, the observed proportions would exactly lie on the diagonal.

Following the discussion of Section 2, the calibration assessment of density forecasts with reliability diagrams is based on studying the distribution of the random variables $Z_{t,k} = \hat{F}_{t+k|t}(Y_{t+k})$ whose realizations are given by the probability integral transforms $z_{t,k} = \hat{F}_{t+k|t}(y_{t+k})$. Intuitively, since it is expected that time-series of hourly wind speed averages will exhibit a significant autocorrelation pattern (see discussion in Vincent *et al.* (2009)), and since such time-series is transformed through a monotonic (strictly) increasing function, an autocorrelation pattern is also expected to be present in the time-series $\{z_{t,k}\}$ of probability integral transforms. This argument also applies to the corresponding time-series $\{\tilde{z}_{t,k}\}$.

The density forecasts defined by $\hat{F}_{t+k|t}(y)$ are by construction perfectly reliable: we thus have $Z_{t,k} \sim \mathrm{U}[0, 1]$. For random variables distributed $\mathrm{U}[0, 1]$, it appears more relevant to define autocorrelation in terms of rank correlation for various lags. It is depicted here in Figure 2(a). In parallel, Figure 2(b) shows the linear correlogram for the time-series $\{\tilde{z}_{t,k}\}$. One sees from Figure 2 that the rank and linear correlograms, in the Uniform and Gaussian domains, respectively, look very similar. This remark goes along a comment of e.g. De Oliveira (2003) such that, surprisingly, transformations may not significantly affect an observed (serial or spatial) correlation structure. Most important, one notices that autocorrelation values appear to be significantly positive (at a 5% significance level) for lags up to 24 hours, and significantly different from 0 for other lags. This confirms the statement made in Section 1 such that, even if a necessary condition for density forecasts to be reliable is that $Z_{t,k} \sim \mathrm{U}[0, 1]$, successive random variables (and corresponding probability integral transforms) do not have to be independent. A consequence of this result is that such correlations should be taken into account when performing hypothesis testing about reliability of density forecasts, or alternatively if issuing consistency or confidence bars when inferring on observed reliability. Note also that as mentioned by Diebold *et al.* (1998), one could look at the correlograms of centred probability integral transforms at the power

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

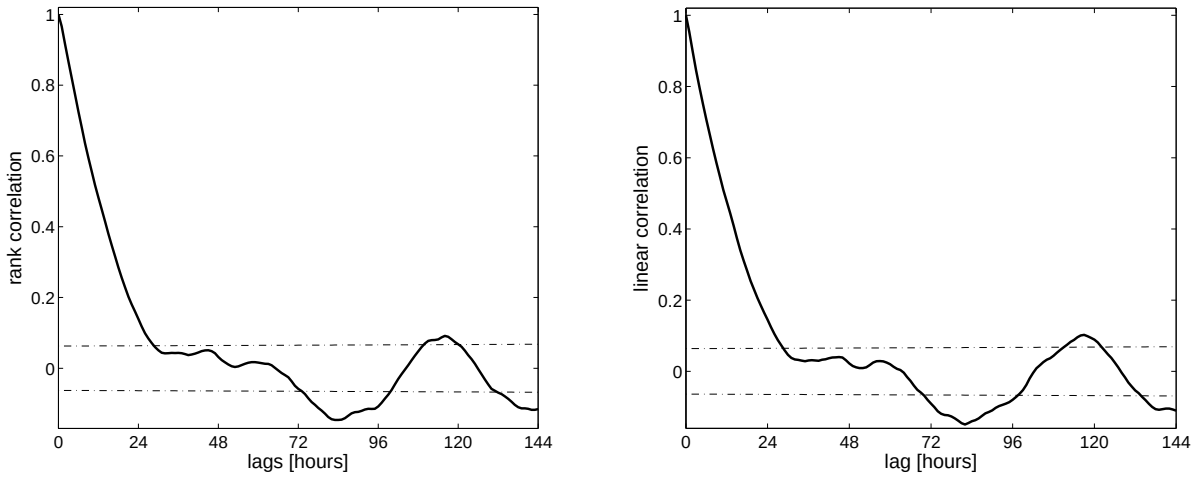
DOI: 10.1002/qj



(a) Time-series with wind speed hourly averages

(b) Cumulative distribution function

Figure 1. Episode with wind speed measurements at Horns Rev, and the corresponding cumulative distribution function corresponding to what would be an ideal climatology density forecast.



(a) Rank correlogram

(b) Linear correlogram

Figure 2. Rank correlogram for the sequence of probability integral transforms, and linear correlogram of their inverse Normal transformations. The dashed lines give the critical values at a 5% level of significance.

j, $j \geq 2$ in order to observe dependencies in higher order moments, i.e. mainly for variance, skewness and kurtosis.

4 Consistency bars using surrogate consistency resampling

Let us recall Bröcker and Smith (2007a), following Smith (1997), who introduced the proposal of generating consistency bars for reliability diagrams. The idea of consistency resampling is extended here to also account for a correlation pattern in the $\{z_{t,k}\}$ time-series. The aim of our consistency bars is then, for each of the quantile forecasts composing density forecasts, to reflect what can be a possible range (for a given confidence level $(1-\beta)$) of observed proportions for this quantile if it were indeed reliable, given the temporal dependencies observed in the forecast-verification pairs induced by the process and forecasting system of interest. From an hypothesis testing point of view, one could say that having the observed proportion of quantile forecasts (for a given nominal proportion) within the range of consistency bars implies that one cannot reject the hypothesis of these quantile forecasts being reliable. This paradigm translates to accepting that there may be a number of perfectly reliable forecast systems for the time-series of measurements of interest, and that the combination of forecasts and verifications may

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

induce different types of dependence structure in the time-series $\{z_{t,k}\}$ of probability integral transforms.

The surrogate consistency resampling method belongs to the more general class of resampling methods for dependent data. For an overview of those methods, see e.g. Lahiri (2003). An important advantage of this surrogate approach is that complete time-series are simulated based on their spectrum, instead of considering subsamples or blocks. An alternative to the use of surrogate consistency resampling would be to employ a model-based consistency resampling approach, where the general class of ARMA(p,q) (Auto-Regressive Moving Average) models could be envisaged for modelling the underlying process in the $\{\tilde{z}_{t,k}\}$ time-series. A subtle issue in this case would relate to the optimal selection of the autoregressive and moving average orders p and q. Model misspecification could have a dramatic impact on the quality of generated consistency bars. An additional advantage of the surrogate approach is that by construction it can be insured that surrogate time-series $\{\tilde{z}^{(.)}_{t,k}\}$ will be distributed $\mathcal{N}(0,1)$; this would not be the case if employing a model-based approach.

4.1 Basics of the surrogate method

The basis of our surrogate consistency resampling method is related to the Amplitude Adjusted Fourier Transform (AAFT) algorithm described by Theiler *et al.* (1992). This algorithm is based on the assumption that observed time-series are a monotonic transformation of realizations from a linear Gaussian process. It is indeed the case that for reliable density forecasts the time-series $\{z_{t,k}\}$ are a monotonic transformation of $\{\tilde{z}_{t,k}\}$, for which each of the realizations are generated from a $\mathcal{N}(0,1)$. The consecutive steps of the AAFT algorithm can be described as:

(i) the time-series $\{z_{t,k}\}$ is rescaled and transformed (while insuring that ranks in the data are conserved) to obtain a linear Gaussian process with time-series of realizations $\{\tilde{z}_{t,k}\}$

(ii) a Fourier-Transform based algorithm is employed to obtain a Gaussian surrogate time-series $\{\tilde{z}^{(.)}_{t,k}\}$, by randomizing the phases. The Gaussian surrogate time-series $\{\tilde{z}^{(.)}_{t,k}\}$ has the same marginal distribution and correlogram as the time-series $\{\tilde{z}_{t,k}\}$

(iii) $\{\tilde{z}^{(.)}_{t,k}\}$ is transformed back to the domain of the original time-series, leading to the surrogate time-series $\{z^{(.)}_{t,k}\}$, but with a slightly different autocorrelation function and power spectrum

A sound property of the AAFT algorithm is that the surrogate time-series have the same periodogram and marginal distribution as the original time-series $\{z_{t,k}\}$ of probability integral transforms, hence also insuring to have a similar correlogram. For a more detailed description of the surrogate data method and of the AAFT algorithm, the reader is referred to Theiler *et al.* (1992). Potential deviations in the marginal distribution or in the spectrum can be corrected by employing the iteratively refined surrogate method of Schreiber and Schmitz (2000). The core of the method i.e. the phase scrambling, is further discussed in Davison and Hinkley (1997).

One could think of directly employing this surrogate method for generating a set of surrogates being distributed U$[0,1]$ and having the same correlogram than the sequence of observed probability integral transform. This would be performed by first transforming the sequence $\{z_{t,k}\}$ to U$[0,1]$ (while preserving the ranks), and by then employing the AAFT algorithm for generating a set of surrogate Gaussian time-series. The last step, consisting of transforming back the Gaussian surrogates to U$[0,1]$ can actually be avoided, due to the equivalence in the definition of the indicator variable from (3) and (5). The counting necessary for building consistency bars can then be equivalently performed using the surrogate time-series $\{z^{(.)}_{t,k}\}$ or $\{\tilde{z}^{(.)}_{t,k}\}$.

If directly employing the above method for generating surrogate time-series however, the resulting surrogates will not be sampled from the true process behind $\{\tilde{z}_{t,k}\}$, but merely from the periodogram obtained from a limited sample. Consistency bars generated from this method would thus not be valid. Consequently, a proposal method for surrogate consistency resampling from the true process behind $\{\tilde{z}_{t,k}\}$ is described below. It consists of first identifying the true process behind $\{\tilde{z}_{t,k}\}$ with spectral analysis, and then of generating surrogates having periodograms sampled from the smooth spectrum of the process.

4.2 Smooth spectrum and periodogram sampling

Since the core of the surrogate data method relates to the phase scrambling of a linear Gaussian process, a spectral analysis framework appears relevant for characterizing the process $\{\tilde{Z}_{t,k}\}$. Write $g(\omega)$ the spectrum of the linear Gaussian process $\{\tilde{Z}_{t,k}\}$, with ω the angular frequency, and $I_N(\omega)$ the periodogram of the time-series $\{\tilde{z}_{t,k}\}$. The periodogram $I_N(\omega)$ corresponds to the sample spectrum observed from $\{\tilde{z}_{t,k}\}$ and is not a consistent (though unbiased) estimate of $g(\omega)$. A consistent estimate can be obtained instead as a smooth spectrum based on truncated periodograms and lag windows. While only the main equations and results are given here, the reader is referred to Madsen (2007) for more extended developments.

The smooth spectrum estimate $\hat{g}(\omega)$ of $g(\omega)$ based on the time-series $\{\tilde{z}_{t,k}\}$ is given as

$$\hat{g}(\omega) \;=\; \frac{1}{2\pi} \sum_{k=-(N-1)}^{k=N-1} \lambda_k C_k \exp(-i\omega k) \qquad (10)$$

where C_k is the sample autocorrelation for lag k, and λ_k is a lag window, allowing to reduce the influence of further lags on the estimate of the spectrum. While there exists a large number of potential lag windows, the class of general Tuckey windows is considered here, defined by

$$\lambda_k \;=\; \begin{cases} 1-2a+2a\cos\left(\frac{\pi k}{M}\right), & |k| \leq M \\ 0, & |k| > M \end{cases} \qquad (11)$$

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

where M is the truncation point, and a is the parameter controlling the shape of the lag window. In particular here, we employ the Tukey-Hanning window, defined by $a = 1/4$. Deciding on an appropriate truncation point M may be quite difficult as this relies on the expertise of the practitioner. It will be shown from the simulations below that the surrogate consistency resampling method proposed is not that sensitive to the choice of M. A limited expertise on expected serial correlation structure present in the $\{\tilde{z}_{t,k}\}$ time-series can provide a relevant guess.

When the smooth spectrum estimate $\hat{g}(\omega)$ is obtained, it can be used for generating a number of realistic periodograms for the surrogate time-series $\{z^{(.)}_{t,k}\}$. For that, let us recall here some properties of the periodogram for linear Gaussian processes:

(i) the periodogram values $I_N(\omega_p)$, with $\omega_p = \frac{2\pi}{N}p$, $p = 1, \ldots, N/2$, the so-called fundamental frequencies, are independent

(ii) $2\frac{I_N(\omega_p)}{g(\omega_p)} \sim \chi^2(2)$, $p \neq 1, N/2$

(iii) $\frac{I_N(\omega_p)}{g(\omega_p)} \sim \chi^2(1)$, $p = 1, N/2$

Property *(i)* relating to the independence of periodogram values for the fundamental frequencies ω_p (i.e. those defining the orthogonal basis of the Fourier series) is actually a crucial property behind the phase scrambling in the surrogate data method introduced above (Davison and Hinkley , 1997; Theiler *et al.*, 1992). In parallel, the smooth spectrum estimate $\hat{g}(\omega)$ can be plugged into properties *(ii)* and *(iii)* in order to simulate periodograms $I^{(.)}_N(\omega_p)$ for the surrogate time-series $\{z^{(.)}_{t,k}\}$, from independent random draws of $\chi^2(1)$ and $\chi^2(2)$ random variables.

4.3 Surrogate consistency resampling

A full description of the surrogate consistency resampling method is given here based on the elements introduced in the above Paragraphs. Again, consider the question of reliability assessment of a sequence $\{\hat{f}_{t+k|t}(y)\}$ of density forecasts (thus for a given forecast horizon k), for a corresponding time-series $\{y_{t+k}\}$ of observations. Following the paradigm introduced above, one accepts that even if this set of density forecasts were reliable, the forecast-verification pairs would induce a temporal correlation structure in the time-series $\{z_{t,k}\}$ of probability integral transforms.

The first step of the surrogate consistency resampling approach is to rescale the time-series $\{z_{t,k}\}$ so that they have a marginal distribution $\mathrm{U}[0,1]$ with a rank-preserving transformation, thus not altering its rank correlogram. The rescaled time-series are then projected into the standard Gaussian domain by using the inverse probit function. It is assumed that the obtained time-series $\{\tilde{z}_{t,k}\}$ is a sample of what would be the sequence of realizations coming from the forecast-verifications pairs if the forecasting system were indeed reliable. The smooth spectrum $\hat{g}(\omega)$ of the related linear Gaussian process $\{\tilde{Z}_{t,k}\}$ is estimated given the choice of a truncation point M. The j^{th} resampling cycle then consists of the following steps:

(i) generate a surrogate periodogram $I^{(j)}_N(\omega_p)$ from the smooth spectrum $\hat{g}(\omega)$

(ii) employ a Fourier-Transform based algorithm to generate a surrogate time-series $\{\tilde{z}^{(j)}_{t,k}\}$ by randomizing the phases of the surrogate periodogram $I^{(j)}_N(\omega_p)$.

(iii) calculate the observed proportions for each of the quantiles with nominal proportions $\alpha_i, i = 1, \ldots, m$ of the density forecasts using (6), based on the surrogate time-series $\{\tilde{z}^{(j)}_{t,k}\}$, yielding the surrogate observed proportions $\hat{a}^{(j)}_{i,k}, i = 1, \ldots, m$. Note that one can obtain surrogate time-series of probability integral transforms $\{z^{(j)}_{t,k}\}$ and of verifications $\{y^{(j)}_{t+k}\}$ by inverse transformation

This resampling cycle is repeated a number of times B and yields an empirical distribution of surrogate observed proportions for each quantile of the density forecasts. Let us define $\hat{G}_{i,k}$ as the cumulative version of this empirical distribution. $\hat{G}_{i,k}$ is a nonparametric estimate of what would be the distribution of proportions that could be observed for the dataset considered, for the quantile with nominal proportion α_i, if it were reliable and given the serial correlation structure induced by the forecast-verification pairs. Given the chosen confidence level $(1 - \beta)$, the lower and upper bounds of the consistency bars are given by

$$\underline{a}_{i,k} = \hat{G}^{-1}_{i,k}(\beta/2) \quad (12)$$

$$\bar{a}_{i,k} = \hat{G}^{-1}_{i,k}(1 - \beta/2) \quad (13)$$

The same argument as that developed by Bröcker and Smith (2007a) applies here, implying that, by construction, the surrogate time-series of probability integral transforms directly relate to a hypothetical sequence of forecast densities $\{\hat{f}^{(j)}_{t+k|t}(y)\}$ which would be reliable in view of the corresponding time-series $\{y^{(j)}_{t+k}\}$ of verifications. This is since it is imposed that surrogate time-series $\{\tilde{z}^{(j)}_{t,k}\}$ are drawn from a $\mathcal{N}(0,1)$ distribution. In addition, consistency bars can be generated for all quantile forecasts in parallel — they are pointwise consistency bars, since as explained in Section 2 the calibration assessment is individually performed for each each quantile with a given nominal proportion. There is finally no binning effect to be considered, as the calibration of all quantiles is verified against the same number of observations, corresponding to the number of time indices in $\mathcal{T}$.

5 Simulations

In this Section are performed simulations allowing us to demonstrate the pitfall stemming from inference on the confidence one may have in reliability diagrams based

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

on an i.i.d. assumption when issuing consistency bars. Simulations are also employed for demonstrating the validity of our approach before applying it to real-world test cases and data. The simulation set-up is described in a first stage, followed by simulations results and related comments.

5.1 Simulation set-up

The simulations are performed based on $\{\tilde{z}_{t,k}\}$ time-series only (i.e. based on linear Gaussian processes), since the conversions from $\mathcal{N}(0,1)$ to U$[0,1]$ (using the probit function Φ) and from U$[0,1]$ to the original domain of the observations are strictly monotonic, thus preserving ranks and counts. As a consequence, observed proportions of quantile forecasts are equivalent if calculated in the original, Uniform, or Gaussian domains.

Imagine generating a time-series of $\{\tilde{z}_{t,k}\}$ of length N for which each realization is drawn from a standard Gaussian distribution $\mathcal{N}(0,1)$ and with a linear correlogram $\rho(h)$, h being the difference between time indices. By definition, such time-series of probability integral transforms projected into a standard Gaussian domain corresponds to that for a reliable density forecasting system. Two types of correlograms are considered: on the one hand a dampened exponential correlogram $\rho_d(h)$, corresponding to a stationary first-order Markovian process, and on the other hand a correlogram $\rho_s(h)$ taking the form of a dampened exponential with oscillations, reflecting a seasonality in the sequence of probability integral transforms in the standard Gaussian domain. The damped exponential is simply given by

$$\rho_d \; : \; h \rightarrow \exp(-\tau h), \quad \tau > 0 \tag{14}$$

with τ the parameter controling the steepness of the exponential decay. In parallel, the dampened exponential with oscillations is defined by

$$\rho_s \; : \; h \rightarrow \frac{1}{2}\left\{\cos\left(\frac{2\pi h}{p}\right)+1\right\}\exp\left(-\frac{2\tau h}{p}\right), \quad \tau, p > 0 \tag{15}$$

with τ being the same type of parameter, while p controls the period of oscillations. For the case of $\rho_d(h)$, τ is set to $\tau = 0.3$, while $\rho_s(h)$ is parameterized with $\tau = 0.6$ and $p = 12$. The corresponding correlograms are depicted in Figure 3.

As explained in the above Section, if calculating observed proportions of quantiles composing density forecasts from a time-series $\{\tilde{z}_{t,k}\}$ of limited size, with correlograms ρ_d or ρ_s, there will clearly be deviations from the diagonal, even though $\{\tilde{z}_{t,k}\}$ relates to a reliable forecast system. This is illustrated in Figure 4 for time-series $\{\tilde{z}_{t,k}\}$ of 400 realizations having correlograms ρ_d or ρ_s (in Figures 4(a) and 4(b), respectively). The nominal proportions for the quantiles composing density forecasts are chosen as ranging from 0.05 to 0.95 with a 0.05 increment. The three different curves in Figures 4(a) and 4(b) correspond to the observed proportions for 3 different draws of $\{\tilde{z}_{t,k}\}$ for each of the two correlograms.

Note that we have chosen in the present paper to present reliability diagrams in the most classical manner, that is, by depicting observed against nominal probabilities. As argued by Bröcker and Smith (2007a), Bröcker (2009) or Pinson *et al.* (2007b), however, one may present such diagrams in a different manner in order to focus on the area around the ideal diagonal case. Especially, the proposal by Bröcker and Smith (2007a) and by Bröcker (2009) of plotting reliability diagrams on the probability paper may be seen as attractive, owing to simplicity of presentation and of interpretation of consistency bars. Considering some other presentation of reliability diagrams would not call for any change in the methodology described for the derivation of consistency bars.

For both correlograms, one notices that for certain draws the observed proportions may lie fairly close to the diagonal, while for some other draws they may be quite far from this same diagonal. This is while they all relate in the same way to reliable forecast systems. It can also be seen from Figure 4 that for a stronger correlation pattern like ρ_s, deviations from the diagonal may be larger. This suggests that serial correlation in forecast-verifications pair magnifies the sampling effects.

Given a chosen correlogram, a number m of linear Gaussian time-series $\{\tilde{z}_{t,k}\}$ (of length N) are generated. Consistency bars are produced based on i.i.d. consistency resampling, and based on the surrogate consistency resampling method described above. It is arbitrarily chosen to focus on 90% consistency bars (that is, for $(1-\beta) = 0.9$), partly because it is the nominal coverage rate considered by Bröcker and Smith (2007a) and partly because this is a quite common choice for a number of real-world applications. Note that similar simulations could be performed for other nominal coverage rates $(1-\beta)$ in order to verify the quality of generated consistency bars. Over the m time-series, one counts the number of times the observed proportions for each quantile forecasts composing reliable density forecasts (i.e. the quantile of a standard Gaussian distribution) lie below and above the consistency bars. This provides us with an approach to verify their actual coverage.

5.2 Simulation results

In the following we carry out simulations on time-series of length $N = 400$ time steps with correlograms ρ_d or ρ_s. Counts performed on such short time-series will clearly be affected by the correlation patterns, as illustrated in Figure 4. In order to verify the coverage rate of the consistency bars, $m = 1000$ different time-series are generated for each type of correlogram. Comparison is made between consistency bars generated from an i.i.d. consistency resampling method, and from the surrogate consistency resampling method introduced above. For this latter case, a benchmark consists of consistency bars generated from the true spectrum of the generated time-series. In parallel, consistency bars generated from the estimated smooth spectrum (with various values of the truncation point M) are also evaluated. A number of $B = 1000$ surrogate time-series are used for deriving consistency bars for

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

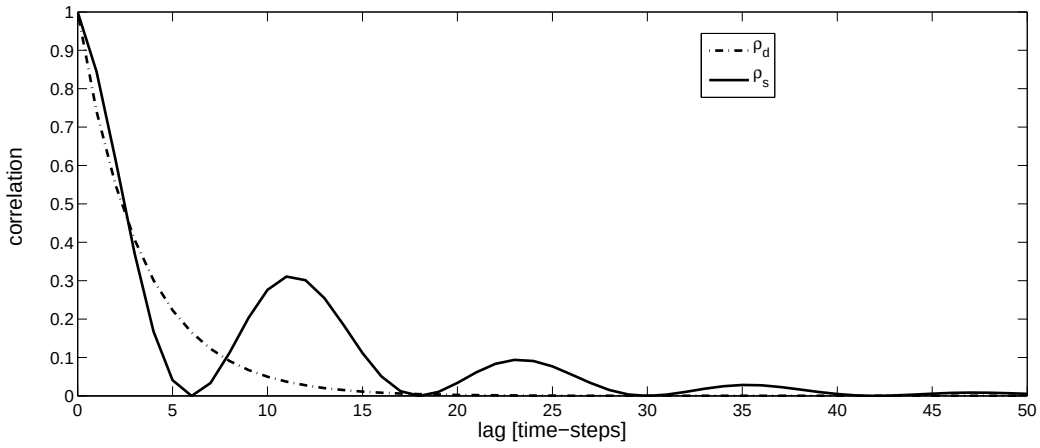
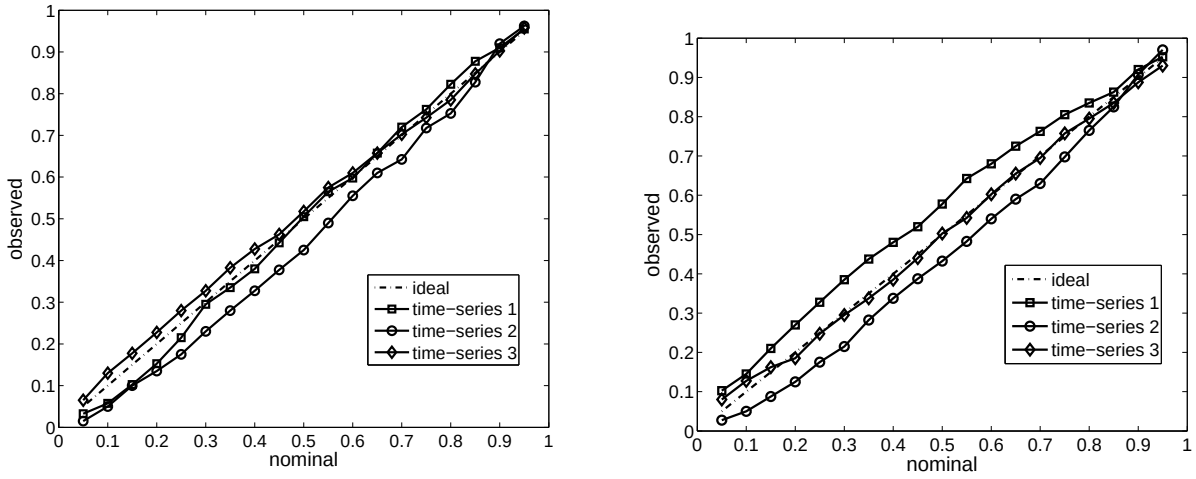


Figure 3. Dampened exponential (with and without oscillations) correlograms used in the simulations.



(a) Correlogram ρ_d (b) Correlogram ρ_s

Figure 4. Reliability diagrams giving examples of observed proportions of reliable forecasting systems for two different correlation structures in the time-series $\{\tilde{z}_{t,k}\}$ (consisting of 400 successive realizations).

each of the m time-series. As is the case for any computer-intensive resampling method, the number of replications B should be chosen sufficiently large in order to obtain realistic confidence bounds, though not too large in order to keep computational time reasonable. The values of the truncation points are chosen from an expert guess based on the analysis of the periodograms of the time-series $\{\tilde{z}_{t,k}\}$, as it would be done for real-world applications. For the case of the correlogram ρ_d, potential expert guesses could be $M = \{12, 24, 36\}$ while for the case of the ρ_s correlogram, they could consist of $M = \{24, 36, 48, 60\}$. It is often said that a reasonable choice for M is such that $M = 2\sqrt{N}$, see e.g. (Upton and Cook , 2002, pp. 324-325), which would translate to $M = 40$ here. The corresponding results, consisting of observed coverage of consistency bars, are gathered in Table I for the correlograms ρ_d or ρ_s, respectively.

In both cases, one clearly sees that there is a significant difference between i.i.d. and surrogate consistency resampling methods in terms of observed coverage of the generated consistency bars. As expected, the lack of coverage is more significant as the correlation pattern is stronger, i.e. for the correlogram ρ_s. This is an illustration of the pitfall stemming from assuming independence when serial correlation is indeed present in the sequence of forecast-verification pairs. Such lack of coverage could translate to concluding a lack of calibration of density forecasts over the period considered, while in fact the observed deviation from perfect reliability cannot be deemed significant.

When employing the surrogate consistency method with the true spectrum of the time-series $\{\tilde{z}_{t,k}\}$ in order to generate surrogate time-series, the observed coverage of the consistency bars is very close to the target 90%

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

Table I. Summary of the observed coverage rate of the 90% consistency bars generated with either i.i.d. or surrogate consistency resampling methods, for time-series with correlogram ρ_d and ρ_s. Consistency bars are for a set of quantiles defining nonparametric density forecasts. Surrogate consistency resampling based on the true spectrum of the time-series is used as a benchmark. For the surrogate consistency resampling method using estimated smooth spectrum, several expert guesses on the truncation point M are considered. Note that asymptotically as m and B tend towards infinity, the observed coverage rate for proportions α_i and $(1-\alpha_i)$ should be the same. The differences here are due to sampling effects.

(a) Correlogram ρ_d

α_i	i.i.d.	surrogate ($M=12$)	surrogate ($M=24$)	surrogate ($M=36$)	surrogate (true)
0.05	66.7	86.7	90.2	91.1	90.4
0.10	61.4	88.0	88.7	88.9	90.2
0.15	59.0	84.8	87.3	89.0	90.7
0.20	58.4	85.5	87.7	88.6	90.5
0.25	56.4	85.3	88.1	88.7	89.8
0.30	57.6	85.0	86.8	88.9	89.9
0.35	55.2	84.8	87.9	87.9	89.7
0.40	56.0	85.6	88.7	89.1	90.5
0.45	56.7	86.3	89.2	88.6	89.7
0.50	56.3	86.0	88.3	88.7	90.6
0.55	54.6	87.2	88.3	88.7	89.4
0.60	54.0	86.9	88.9	89.4	88.9
0.65	55.8	86.5	89.5	89.8	89.4
0.70	54.4	85.4	88.8	89.8	89.3
0.75	56.5	85.6	88.1	89.6	89.1
0.80	59.7	86.3	88.8	90.9	89.2
0.85	59.3	86.1	89.4	89.2	89.8
0.90	61.2	87.5	90.0	89.2	90.0
0.95	65.1	88.8	91.1	90.4	90.6

(b) Correlogram ρ_s

α_i	i.i.d.	surrogate ($M=24$)	surrogate ($M=36$)	surrogate ($M=48$)	surrogate ($M=60$)	surrogate (true)
0.05	57.7	86.8	88.4	88.2	91.1	90.9
0.10	53.7	84.4	86.6	87.9	90.4	90.8
0.15	52.0	83.3	85.9	87.5	90.4	90.0
0.20	50.8	82.6	85.7	87.0	90.2	90.7
0.25	51.1	83.1	86.1	86.7	90.5	90.7
0.30	48.7	83.5	86.9	87.5	90.0	90.5
0.35	46.6	83.2	86.7	87.6	91.0	89.3
0.40	45.8	83.4	86.5	87.3	90.3	90.1
0.45	46.3	83.2	86.7	87.1	90.7	90.6
0.50	45.4	83.3	85.8	87.6	90.1	90.6
0.55	46.4	82.9	86.2	88.4	90.5	89.8
0.60	45.5	83.0	85.9	88.0	89.8	89.4
0.65	47.8	83.2	86.7	88.7	89.7	89.8
0.70	47.7	83.0	87.3	88.9	89.2	90.0
0.75	48.5	83.2	86.2	89.6	88.6	89.8
0.80	51.1	83.9	87.3	90.1	89.1	90.3
0.85	51.8	84.1	87.5	88.8	89.3	90.0
0.90	54.8	85.5	87.2	88.2	88.2	88.9
0.95	59.9	86.4	88.8	89.7	89.8	91.7

nominal coverage. Furthermore when employing the estimated smooth spectrum instead, the coverage of generated consistency bars is also close to the target 90% nominal coverage. Choosing a truncation point which is too

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

small results in under-representing the correlation structure present in the sequence of forecast-verification pairs, leading to generation of consistency bars that are too narrow. By making a reasonable guess for the truncation point, i.e. $M = \{24, 36\}$ and $M = \{48, 60\}$ for the case of the correlograms ρ_d and ρ_s, respectively, the generated consistency bars appear to have an acceptable coverage. It is true that for practical real-world verification studies it might be difficult to pick an optimal truncation point (as would be the case for any spectral analysis study in any case), but a reasonable guess from the practitioner should provide a sufficiently accurate estimate of the spectrum, leading to appropriate consistency bars.

6 Application to the reliability assessment of density forecasts of wind power generation

Wind power is the renewable energy with the fastest growth over the last few years. It has a significant share in the electricity generation mix in a number of European countries, most notably in Denmark and Spain. The optimal integration of this renewable energy in the existing electricity system requires forecasts for various ranges of horizons depending on the decisions to be made, i.e. from a few minutes ahead for the control of wind farm output, to several days ahead for offshore maintenance planning. The forecasts that are used the most today have an hourly resolution up to 48-72 hours ahead, are employed for the trading and management of the wind power generation, and issued based on one or several forecasts of relevant meteorological variables for the site(s) of interest. If considering lead times from few minutes up to 2 hours ahead, forecasts are then generated from purely statistical methods relying on local measurements only, as for instance in (Gneiting *et al.*, 2006; Pinson and Madsen , 2009b). For an overview of motivations, techniques and practical experience with wind power forecasting, the reader is referred to Giebel *et al.* (2003) and Costa *et al.* (2008). Among the various types of forecasting products employed for wind energy management, maintenance planning and trading, nonparametric density forecasts are becoming more and more popular, since benefits from their use have been demonstrated, see e.g. (Matos and Bessa , 2009; Pinson *et al.*, 2007a). This is because the loss functions of forecast users commonly differ from the classical quadratic loss function for which point forecasts relating to the conditional expectation are optimal. Such loss functions may also vary in time due to the changes in the structure and dynamics of electricity markets. As wind power generation is a nonlinear and bounded process, predictive densities may be highly skewed and with heavy tails (Lange , 2005), hence being difficult to accurately model with known parametric families of density functions, see discussion by Pinson (2006). This has motivated the development of a large number of nonparametric methods for wind power density forecasting, based on statistical methods and/or ensemble forecasts, see (Bremnes , 2006; Møller *et al.*, 2008; Nielsen *et al.*, 2006; Pinson and Madsen , 2009) among others.

We consider here nonparametric density forecasts of wind power generation for the whole installed capacity in Western Denmark, which approximately represents P_n =2.5GW over the period considered. All forecasts and measurements are normalized by this nominal capacity, and therefore expressed in percentages of P_n. Forecasts are issued hourly, and have an hourly temporal resolution up to a forecast length of 43 hours. The point forecasts of wind power generation were provided by the Wind Power Prediction Tool (WPPT) as described in e.g. Nielsen *et al.* (2002), while the nonparametric density forecasts were generated based on the adapted resampling method initially described in Pinson (2006). The period for which both measurements and forecasts are available goes from the beginning of January 2006 until mid-November 2007. Figure 5 depicts an example with wind power point forecasts issued on the 8th January 2007 at noon, related nonparametric density forecasts, as well as the corresponding measurements. Density forecasts take the form of a set of central prediction intervals (i.e. centered in probability around the median) with increasing nominal proportions from 10% to 90%. They thus are defined by 18 quantile forecasts with nominal proportions from 5% to 95% with 5% increment, except for the median.

Three different sets of forecast series are arbitrarily selected from the two years which were available. These three periods are picked over Autumn 2006, Spring and Summer 2007. The first two sets consists of 600 forecast series, while the latter one consists of 800 forecast series. The reliability of nonparametric density forecasts is studied based on reliability diagrams in the form of those presented in Section 2. Each lead time is considered individually. Inspection of the periodograms and correlograms of the $\{\tilde{z}_{t,k}\}$ time-series suggests that truncation points between 36 and 60 hours could be relevant for the first two sets of forecasts series, while M may take values between 72 and 96 hours for the case of the last set, for all forecast horizons. A common value of 48 hours is selected for the first two sets, while a value of 84 hours is picked for the third one. As shown and discussed in Section 5, the width of the generated consistency bars is not highly sensitive to the choice for M. A sufficiently large number B of surrogate time-series is chosen as $B = 1000$.

Let us focus for instance on the calibration assessment of 42-hour ahead density forecasts of wind power generation, which is summarized in Figure 6. After the series of probability integral transforms are converted to having a Normal distribution, with a rank-preserving transformation, the smooth spectra of the obtained time-series related to the three sets of forecast series are estimated with the method described in Section 4.2. These smooth spectrum estimates are gathered in Figure 6(a). Significant differences can be observed among these smooth spectrum estimates. While the smooth spectrum estimated for set one could relate to a first-order Markovian process with moving average, that for set two is typical of a second-order Markovian process. Finally for set

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

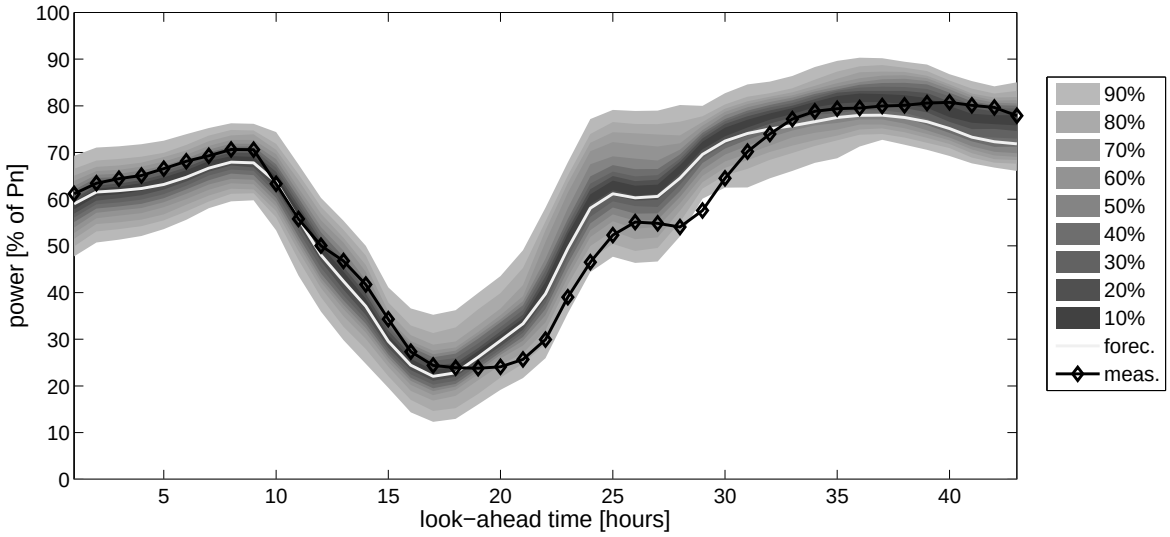


Figure 5. Example of nonparametric density forecasts of wind power generation for the whole Western Denmark (issued on the 8th January 2007 at noon) in the form of a river of blood fan chart. Density forecasts are represented as a set of central prediction intervals with increasing nominal proportions. Power values are normalized by the total wind capacity P_n for the region. Measurements and point forecasts are also depicted.

three the estimated smooth spectrum can be seen as that of a first-order Markovian process with seasonalities. This therefore makes us expect different widths for the consistency bars to be generated. The reliability diagrams for the three sets of forecats series are depicted in Figures 6(b), 6(c) and 6(d). Consistency bars, generated using the methods described in Section 4, are depicted as pointwise consistency bars informing, for each nominal proportion of the quantile forecasts composing nonparametric density forecasts, about consistent deviations that can be expected even if such forecasts were perfectly reliable.

The three sets of consistency bars indeed have different widths, thus reflecting the effect of the identified correlation structures on the potential range of observed proportions for perfectly reliable density forecasts. They are generally tighter for set two, i.e. for which the smooth spectrum takes the form of that for a simple second-order Markovian process. Notice that the larger consistency bars are for the set with the most forecast series (set three), illustrating the fact that it is not because more forecast series are available that one should expect smaller consistency bars — again due to the stronger correlation pattern present for that set.

Let us now interpret these reliability diagrams. If consistency bars were not available, one would subjectively appreciate the observed deviations from the diagonal and decide on acceptable reliability (or not) of the various quantile forecasts composing nonparametric density forecasts. One would then certainly accept all quantile forecasts to be reliable over set three. In contrast, quantile forecasts with nominal proportions between 0.45 and 0.8 for set one and with nominal proportions between 0.1 and 0.65 for set two would be deemed as non-reliable owing to an increased deviation between the ideal diagonal case and observed proportions. Now consider the sets of consistency bars. For sets one and three, the observed proportions of all quantiles composing density forecasts lie within the consistency bars, even though deviations from the diagonal are of different magnitudes. This does not tell us that the quantiles forecasts are reliable, but inversely that it cannot be concluded that they are not reliable (for a 10% level of significance). This goes against the subjective evaluation given before. In contrast for set two, the fact that observed proportions for quantile forecasts with nominal proportions between 0.1 and 0.55 lie outside of the consistency bars confirms that quantile forecasts for such nominal proportions should not be considered as reliable.

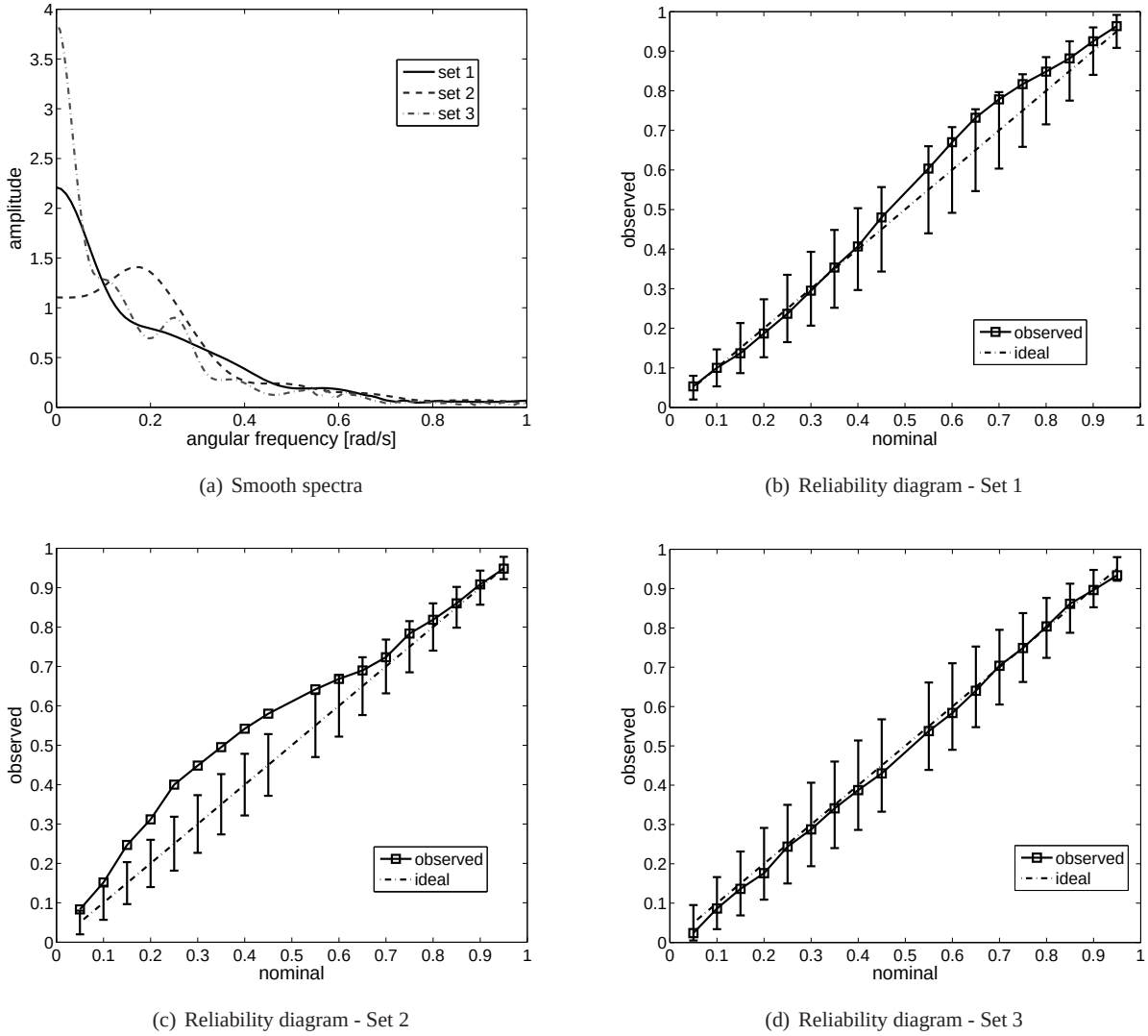
7 Concluding remarks

Focus has been given to the question of the calibration assessment of density forecasts of continuous variables, originating from ensemble forecasts or statistical methods, with reliability diagrams. It has been explained that employing such reliability diagrams relates to the evaluation of unconditional calibration. In many applications one should acknowledge the presence of serial correlation in the sequence of probability integral transforms even for reliable density forecasts, in turn induced by the sequence of forecast-verification pairs.

We have built on an interesting proposal by Bröcker and Smith (2007a) consisting of associating reliability diagrams with consistency bars. Consistency bars here reflect the potential impact of both limited counting statistics and serial correlation on what would be the observed reliability of a perfectly reliable forecast system

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj



(a) Smooth spectra

(b) Reliability diagram - Set 1

(c) Reliability diagram - Set 2

(d) Reliability diagram - Set 3

Figure 6. Example results from the reliability assessment of 42-hour ahead nonparametric density forecasts of wind power generation. The smooth spectrum estimates related to the 3 sets of forecast series are gathered in (a). Reliability diagrams with consistency bars for a 90% confidence level are depicted in (b), (c) and (d).

over the set of available observations. An original surrogate consistency resampling method has been introduced and evaluated for that purpose. The presence of serial correlation clearly increases the width of consistency bars. Even though the serial correlation pattern has to be estimated (here in the frequency domain), it has been demonstrated from simulations that the actual coverage of our consistency bars is close to their intended nominal coverage.

The consistency bars which have been considered in the present paper are pointwise consistency bars. This means that they relate to the individual reliability assessment of the quantile forecasts (for a given nominal proportion) composing nonparametric density forecasts. As a possible extension of the work presented here, one may consider the definition of consistency envelopes, which in contrast would relate to the simultaneous reliability assessment of quantile forecasts with various nominal proportions (thus for the full densities), in a multiple hypothesis testing framework.

Our most important message here is not that one should mandatorily use the approach introduced for generating consistency bars, but instead that one must consider the potential effect of serial correlation in reliability assessment. Indeed, it has been shown that assuming independence of the sequence of probability integral transforms would clearly lead to underestimating the range of potentially observed proportions for a perfectly reliable probabilistic forecasting system over the period of interest. One may decide on one's own way of modelling or accounting for serial correlation, potentially simulating different plausible serial correlation patterns and assessing their impact on the width of consistency bars. Note that the question of interdependence among forecast-verification pairs should also be considered if focusing on probability

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj

forecasts for binary events. It might be that interdependence is not an issue for a large number of cases, but one should still be aware of this potential issue before to apply i.i.d. resampling methods.

For the surrogate consistency resampling method developed, only one parameter, i.e. the truncation point M, has to be selected for the estimation of the smooth spectrum. This is somehow the price to pay for capturing the interdependence structure in the sequence of probability integral transforms. Even though the selection of M may call for some statistical (and/or signal processing) expertise from the practitioner, the actual coverage of consistency bars is not highly sensitive to the choice for M, especially if consistency bars are to be used for visual assessment of density forecasts calibration and not for thorough hypothesis testing. In parallel the number B of replications of the consistency resampling method, which corresponds to the number of surrogate time-series to be generated, should be chosen sufficiently large, e.g. $B \geq 1000$. Since computational power is rapidly increasing, picking a large value for B should not be a problem. For reference, only 4 minutes were necessary for generating the reliability diagrams with consistency bars of Section 6 with Matlab, i.e. for time-series of length 600 and 800 with the number of surrogates $B = 1000$.

Unconditional calibration is only one aspect of probabilistic forecast verification. It is a crucial aspect though as a probabilistic bias in density forecasts would directly translate to a bias in decisions to be made from such probabilistic forecasts. If being strict about forecast verification, density forecasts which cannot be deemed as reliable should not be considered further for decision-making. Fortunately, one can easily correct for a lack of unconditional calibration, with e.g. conditional parametric models (Nielsen *et al.*, 2006) or smoothed bootstrap (Hall and Rieck , 2001). For the case of the application considered, the use of consistency bars has permitted us to carry out a reliability assessment of nonparametric density forecasts of wind power generation, with results stronger than those obtained before, i.e. solely based on subjective evaluation of the deviation between observed proportions and the ideal diagonal case of reliability diagrams. We intend to promote the use of consistency bars as a generic feature of reliability diagrams for the evaluation of density forecasts of wind power generation.

Acknowledgments

The work presented has been partly supported by the European Commission under the SafeWind Project (ENK7-CT2008-213740), and by the Danish Research Council for Technology and Production Sciences (grant no. FTP-274-08-0573), which are hereby acknowledged. The authors would like to thank DONG Energy and Vattenfall Denmark for originally providing the wind speed measurements for the Horns Rev wind farm. The authors are also grateful to Energinet.dk, the Transmission System Operator in Denmark, for providing wind power measurements for Western Denmark, and to ENFOR A/S for the wind power (point) forecasts. Acknowledgments are due to two Reviewers and one Associate Editor, whose suggestions and comments permitted to enhance the paper. Acknowledgments are finally due to Henrik Aa. Nielsen, James W. Taylor, as well as Tilmann Gneiting, for fruitful discussion about reliability of ensemble and density forecasts.

Appendix A. A mathematical proof of the effect of serial correlation on the size of consistency bars

In this Appendix a simple mathematical proof is given of the effect of serial correlation on the size of consistency bars. More precisely, we show here that for any type of correlation in the sequence of forecast-verification pairs, the consistency bars are to be wider than in the i.i.d. case.

Let us focus, without loss of generality, on a given nominal proportion α_i. The forecast horizon k is omitted in the developments below in order to lighten the notations. In view of the definition of the indicator variable $\xi_t^{(\alpha_i)}$ in (3), it appears that $\xi_t^{(\alpha_i)}$ is the realization at time t of a Bernoulli random variable with parameter α_i. As a consequence, the observed proportion $\hat{a}_i$ is a realization of a random variable defined as the sum of Bernoulli trials, scaled by the number of trials N, being here the length of the evaluation period. Below, we will denote this sum by X_i or $\tilde{X}_i$, for the i.i.d. and correlated case, respectively.

In the case where there is no serial correlation present in the sequence of forecast-verification pairs, the corresponding Bernoulli trials are i.i.d. By definition, the sum X_i of N i.i.d. Bernoulli trials with chance of success α_i follows a binomial distribution, $X_i \sim B(N, \alpha_i)$. The first two moments of the distribution of the proportion X_i/N are then given by

$$\mathbb{E}\left[X_i/N\right] = \alpha_i \tag{16}$$

$$\text{var}\left[X_i/N\right] = \frac{\alpha_i(1-\alpha_i)}{N} \tag{17}$$

In contrast when serial correlation is present in the sequence of forecast-verification pairs, the corresponding Bernoulli trials cannot be independent. In such a case, it is known that the sum $\tilde{X}_i$ of N dependent Bernouilli trials can be modelled with a Beta-Binomial distribution, see e.g. Ahn and Chen (1995) or Tsai *et al.* (2003). This distribution is defined as

$$\tilde{X}_i \sim B(N, \theta) \tag{18}$$

with

$$\theta \sim \text{Beta}(\alpha_i, \sigma_\theta) \tag{19}$$

Note that for the sake of simplicity, the Beta distribution $\text{Beta}(\alpha_i, \sigma_\theta)$ in the above is characterized by its mean α_i and variance σ_θ, instead of its two shape parameters. Consequently, the first two moments of the distribution of the proportion $\tilde{X}_i/N$ are given by

$$\mathbb{E}\left[\tilde{X}_i/N\right] = \alpha_i \tag{20}$$

$$\text{var}\left[\tilde{X}_i/N\right] = \frac{\alpha_i(1-\alpha_i)}{N}\left(1 + \frac{N-1}{\sigma_\theta + 1}\right) \tag{21}$$

Copyright © 2009 Royal Meteorological Society

Prepared using qjrms3.cls
DOI: 10.1002/qj

Since necessarily $N > 1$ and $\sigma_\theta > 0$, one has

$$\text{var}\left[\tilde{X}_i/N\right] > \text{var}\left[X_i/N\right] \tag{22}$$

meaning that, whatever α_i, the distribution of proportions in the case of serial correlation will have a higher second-order moment that if there is no correlation. Such distributions will in any case be centred on α_i and symmetric around it. Therefore, for any confidence level $(1-\beta)$, consistency bars will be wider if serial correlation is present in the sequence of forecast-verification pairs.

References

Abramson B, Clemen R. 1995. Probability forecasting. *Int. J. Forecasting* **11**: 1–4.

Ahn H, Chen JJ. 1995. Generation of over-dispersed and under-dispersed Binomial variates. *J. Comput. Graph. Stat.* **4**: 55–64.

Atger F. 1998. The skill of ensemble prediction systems. *Mon. Wea. Rev.* **127**: 1941–1953.

Atger F. 2004. Estimation of the reliability of ensemble probabilistic forecasts. *Quat. J. Roy. Met. Soc.* **130**: 1509–1523.

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

Bremnes JB. 2006. A comparison of a few statistical models for making quantile wind power forecasts. *Wind Energ.* **9**: 3–11.

Bröcker J. 2009. On reliability analysis of multi-categorical forecasts. *Nonlin. Processes Geophys.* **15**: 661–673.

Bröcker J, Smith LA. 2007. Increasing the reliability of reliability diagrams. *Weather Forecast.* **22**: 651–661.

Bröcker J, Smith LA. 2007. Scoring probabilistic forecasts: The importance of being proper. *Weather Forecast.* **22**: 382–388.

Christoffersen PF. 1998. Evaluating interval forecasts. *Int. Econ. Rev.* **39**: 841–862.

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

Davison AC, Hinkley DV. 1997. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Methods, Cambridge University Press: Cambridge.

De Oliveira V. 2003. A note on the correlation structure of transformed Gaussian random fields. *Aust. N. Z. J. Stat.* **45**: 353–366.

Diebold FX, Gunther TA, Tay AS. 1998. Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.* **39**: 863–883.

Giebel G, Kariniotakis G, Brownsword R. 2003. The state of the art in short-term prediction of wind power - A literature overview. Technical report, ANEMOS EU project, deliverable report D1.1. [Available online: http://www.anemos-project.eu]

Gneiting T. 2008. Editorial: Probabilistic forecasting. *J. Roy. Stat. Soc. A* **171**: 319–321.

Gneiting T. 2008. Quantiles as optimal point predictors. Technical Report, University of Washington, Department of Statistics, Technical Report no. 538.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. B* **69**: 243–268.

Gneiting T, Larson K, Westrick K, Genton MG, Aldrich E. 2006. Calibrated probabilistic forecasting at the stateline wind energy center - The regime-switching space-time method. *J. Am. Stat. Assoc.* **101**: 968–979.

Hall P, Rieck A. 2001. Improving coverage accuracy of nonparametric prediction intervals. *J. Roy. Stat. Soc. B* **63**: 717–725.

Hamill T. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.* **129**: 550–560.

Jolliffe IT. 2007. Uncertainty and inference for verification measures. *Weather Forecast.* **22**: 137–150.

Jolliffe IT, Stephenson DB. 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley: New York.

Lange M. 2003. On the uncertainty of wind power predictions - Analysis of the forecast accuracy and statistical distribution of errors. *J. Sol. Energ. - Trans. ASME* **127**: 177–184.

Lahiri SN. 2003. *Resampling Methods for Dependent Data*. Springer Series in Statistics, Springer: Berlin.

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *J. Comput. Phys.* **227**: 3515–3539.

Luo X, Nakamura T, Small M. 2007. Surrogate data methods applied to nonlinear time series. *Actas Primer Encountro Sobre Análisis No Lineal De Series Temporales*, ed. Grupo de Investigación Interdisciplinar en Sistemas Dinámicos, pp. 1-16.

Madsen H. 2007. *Time Series Analysis*. Chapman & Hall/CRC: London.

Matos M, Bessa R. 2009. Decision support tools for power balance and reserve management. Technical report, ANEMOS.plus EU project, deliverable report D3.3.

McSharry P, Bouwman S, Bloemhof G. 2005. Probabilistic forecasts of the magnitude and timing of peak electricity demand. *IEEE T. Power Syst.* **20**: 1166–1172.

Murphy AH. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather Forecast.* **8**: 281–293.

Møller JK, Nielsen HAa, Madsen H. 2008. Time-adaptive quantile regression. *Comput. Stat. Data An.* **52**: 1292–1303.

Nielsen HAa, Nielsen TS, Madsen H, Badger J, Giebel G, Landberg L, Sattler K, Voulund L, Tøfting J. 2006. From wind ensembles to probabilistic information about future wind power production - Results from an actual application. *Proceedings IEEE PMAPS Conference, Probabilistic Methods Applied to Power Systems*, Stockholm, Sweden.

Nielsen TS, Madsen H, Nielsen HAa. 2002. Prediction of wind power using time-varying coefficient functions. *Proceedings 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. [Available online: http://pastel.paristech.org]

Pinson P, Chevallier C, Kariniotakis G. 2007. Trading wind generation with short-term probabilistic forecasts of wind power. *IEEE T. Power Syst.* **22**: 1148–1156.

Pinson P, Madsen H. 2009. Ensemble-based probabilistic forecasting at Horns Rev. *Wind Energy* **12**: 137–155.

Pinson P, Madsen H. 2009b. Adaptive modeling and forecasting of wind power fluctuations with Markov-switching autoregressive models. Technical Report, Technical University of Denmark, DTU Informatics, Denmark.

Pinson P, Nielsen HAa, Møller JK, Madsen H, Kariniotakis G. 2007. Nonparametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy* **10**: 497–516.

Schreiber T, Schmitz A. 2000. Surrogate time series. *Physica D* **142**: 346–382.

Smith LA. 1997. The maintenance of uncertainty. *Proceedings of the International School of Physics Enrico Fermi* **133**: 177–246.

Tay AS, Wallis KF. 2000. Density forecasting: a survey. *J. Forecasting* **19**: 235–254.

Taylor JW. 1999. Evaluating volatility and interval forecasts. *J. Forecasting* **18**: 111–128.

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

Taylor JW, Buizza R. 2006. Density forecasting for weather derivative pricing. *Int. J. Forecasting* **22**: 29–42.

Theiler J, Eubank S, Longtin A, Galdrikan B, Farmer JD. 1992. Testing for nonlinearity in time-series: the method of surrogate data. *Physica D* **58**: 77–94.

Timmermann A. 2000. Density forecasting in economics and finance. *J. Forecasting* **19**: 231–234.

Tsai CA, Hsueh HM, Chen JJ. 2003. Estimation of false discovery rates in multiple testing: Application to microarray data. *Biometrics* **59**: 1071–1081.

Upton G, Cook I. 2002. *A Dictionary of Statistics*. Oxford University Press: Oxford.

Vincent CL, Giebel G, Pinson P, Madsen H. 2009. Resolving non-stationary spectral signals in wind speed time-series using the Hilbert-Huang transform. *J. Appl. Meteorol. Clim.*, available online.

Wilks DS. 1995. *Statistical Methods in the Atmospheric Sciences, 2nd Ed*. International Geophysics Series, Vol. 59, Academic Press.

Copyright © 2009 Royal Meteorological Society
Prepared using qjrms3.cls

DOI: 10.1002/qj