Evaluating the quality of scenarios of short-term wind power generation

P. Pinson[a,*], R. Girard[b]

[a] *Technical University of Denmark, Dpt. of Informatics and Mathematical Modelling, Denmark*
[b] *Mines ParisTech, Centre for Energy and Processes, Sophia-Antipolis, France*

Abstract

Scenarios of short-term wind power generation are becoming increasingly popular as input to multi-stage decision-making problems e.g. multivariate stochastic optimization and stochastic programming. The quality of these scenarios is intuitively expected to substantially impact the benefits from their use in decision-making. So far however, their verification is almost always focused on their marginal distributions for each individual lead time only, thus overlooking their temporal interdependence structure. The shortcomings of such an approach are discussed. Multivariate verification tools, as well as diagnostic approaches based on event-based verification are then presented. Their application to the evaluation of various sets of scenarios of short-term wind power generation demonstrates them as valuable discrimination tools.

Keywords: Renewable energy, forecasting, time trajectories, multivariate verification, diagnostic tools

1. Introduction

Numerous countries have set ambitious goals for the integration of renewable energies into their power systems, generally in a liberalized market environment, and with wind energy often being a primary choice. Taking the example case of Denmark, objectives were set to reach 50% of the energy consumption met by renewables in 2025-2030, while scenarios with 100% targets for 2050 are seriously investigated [1]. Such levels of renewable energy penetration call for a shift in the operation paradigm by relying more on (inherently uncertain) forecasts of energy production and consumption, as well as stochastic approaches to decision-making. More generally it may require

*Corresponding author
Email address: pp@imm.dtu.dk (P. Pinson)
URL: www2.imm.dtu.dk/~pp/ (P. Pinson)

the development of integrated management methodologies, potentially involving district heating [2], transport [3] and water [4] among others.

The development of forecasting methodologies has been the focus of intensive research over the last decade with an ever-increasing number of contributions appearing in international conferences and peer-reviewed scientific journals [5–7]. As for other applications, both theoretical and practical developments are more and more going towards various forms of probabilistic forecasting. For a large class of decision-making problems optimal decisions will always directly relate to quantiles of conditional predictive distributions, and this whatever the level of forecast uncertainty. This point has been discussed from a more theoretical point of view by a number of authors, e.g. [8, 9]. The case for the use of probabilistic forecasts (and demonstration of resulting benefits) has been made for the reserve quantification problem [10], unit commitment [11], overall system operation planning [12], and for the design of optimal trading strategies [13, 14].

Probabilistic forecasts for several successive lead times are often generated and presented as marginal distributions for each lead time individually, or as some of their characteristics i.e. quantiles or central prediction intervals. They can be sufficient if the decisions to be made for any given lead time are independent of the others. In a general manner, however, making optimal decisions requires knowledge of the interdependence of the stochastic process over successive lead times — possibly also for a number of locations and many different variables. This then calls for the forecasting of the characteristics of the joint distribution of the process for the set of lead times of interest. Since modelling, estimation and communication of complex multivariate densities may be intractable, this type of forecasts ideally takes the form of scenarios [15, 16], also referred to as time trajectories, or as ensemble forecasts in the meteorological community [17]. They can easily be used as input to stochastic optimization problems e.g. with a Monte-Carlo approach.

Frameworks for the assessment of wind power probabilistic forecasts in the form of predictive densities (for every lead time individually) have been introduced [18, 19], based on the idea that reliability and sharpness are the components of the skill of these forecasts. Very few articles discuss, however, the question of the evaluation of time trajectories or more generally of multivariate probabilistic forecasts [20–23]. The increased interest for such a type of forecasts in decision-making calls for the proposal and discussion of suitable verification frameworks. They should ideally be of practical nature so that not only forecasters, but also forecast users, may readily

appraise what the quality of the forecasts is. The main objective of the present paper is to describe a framework based on existing multivariate verification tools and on a diagnostic approach to the evaluation of the quality of short-term scenarios of wind power generation. Quality focuses on the degree of correspondence between predictions and related observations, hence relating the inherent statistical characteristics of forecasts and measurements. This comes in contrast with the value of the forecasts, which represents the benefits (economical or not) resulting from the use of forecasts in decision-making processes.

An issue with this type of forecast evaluation relates to the dimensionality of forecasts and observations. Reducing the effective dimensionality of the verification problem can be done by formulating structural assumptions about the underlying stochastic processes or with a geometric approach [24]. Alternatively, we propose here a diagnostic approach with an event-based view of the verification problem.

Time trajectories are formally defined in Section 2. The dataset used as a basis for the argumentation of the paper is introduced in Section 3, including various sets of time trajectories for a wind farm in the South of France. The limitations of existing univariate verification frameworks for the assessment of wind power probabilistic forecasts are illustrated in Section 4. The multivariate extension of probabilistic forecast verification is presented in Section 5, also showing its advantages and limitations. Subsequently, a diagnostic approach to the assessment of scenarios of short-term wind power generation is described in Section 6 based on an event-based view of scenario evaluation. The interest of this diagnostic framework is demonstrated in Section 7. Finally the paper ends with concluding remarks in Section 8.

2. Definition of time trajectories

Let us focus on a stochastic process $\{Y_t\}$ for $t \in \mathcal{N}^+$, with $\{y_t\}$ the related time-series of successive observations. No particular assumption is made about the stochastic process $\{Y_t\}$, since in practice wind power generation may be seen as nonstationary, as well as nonlinear and non-Gaussian owing to its bounded nature.

In a point forecasting set-up, a forecaster issues at time t his best estimate $\hat{y}_{t+k|t}$ of a characteristic of the random variable Y_{t+k} at lead time k. Depending upon the scores to be minimised, $\hat{y}_{t+k|t}$ may relate to the conditional expectation of Y_{t+k}, i.e. if the loss function to be minimized is quadratic, or of some of its quantiles for more general asymmetric linear loss functions [9].

In parallel in a probabilistic density forecasting set-up, a forecaster issues his predictive density $\hat{f}_{t+k|t}$ (with corresponding cumulative distribution function $\hat{F}_{t+k|t}$) being his best guess estimate of the conditional distributional characteristics of the random variable Y_{t+k} given the information available at time t.

We now place ourselves at given time point t, and look at a set of K lead times in the future. For simplicity, these K lead times are for the times $t+1, t+2, \ldots, t+K$, though they do not necessarily need to be sampled regularly. The multivariate random variable $\mathbf{Z}_t$ composed by $Y_{t+k}, k = 1, \ldots, K$, is the variable for which the joint distributional characteristics aimed at being predicted. Even in the most simple case where $\mathbf{Z}_t$ is assumed multivariate Gaussian as in [23], communicating a forecast distribution for $\mathbf{Z}_t$ is complex since consisting of a set of conditional expectations for the successive lead times, associated with a conditional covariance matrix summarising the second-order characteristics of $\mathbf{Z}_t$. In a more general case, it may not even be possible to fully characterize the K-dimensional distribution $\mathbf{Z}_t$. A solution is instead to generate time trajectories which can be seen as equally likely samples of the predictive distribution of $\mathbf{Z}_t$. We denote by $\hat{\mathbf{z}}_t^{(j)} = [\hat{y}_{t+1|t}^{(j)}, \hat{y}_{t+2|t}^{(j)}, \ldots, \hat{y}_{t+K|t}^{(j)}]^\top$ the j^{th} time trajectory ($j = 1, \ldots, J$, with J the number of trajectories). The corresponding observation is $\mathbf{z}_t = [y_{t+1}, \ldots, y_{t+K}]^\top$.

Time trajectories as defined in the above can be generalised to space-time or even multivariate space-time trajectories. This is for instance the case if looking at wind power generation for several locations and lead times simultaneously, or if considering the joint forecasting of the power output of wind and wave energy devices at some offshore location.

3. Datasets with scenarios of short-term wind power generation

A few methods for the generation of scenarios of short-term wind power generation have recently appeared in the literature e.g. [15, 16, 25]. We have selected the ensemble-based method of [25] as well as the Gaussian copula approach of [15]. The interest of selecting these two is that, by construction, we can ensure that in the existing framework for probabilistic forecast verification on a per-horizon basis, the quality of the resulting trajectories cannot be differentiated. It then serves as a basis for our argument regarding the need for more advanced approaches to the evaluation of scenarios.

Focus is given to a wind farm located at Oupia in the South of France, with a nominal capacity of 8.1MW. All power forecasts and corresponding measurements are normalized, hence taking values

in $[0, 1]$. The data available includes meteorological ensemble forecasts and power measurements covering a period of 18 months between July 2004 and December 2005. Power measurements have an hourly temporal resolution. The characteristics of the meteorological ensemble forecasts are discussed below when introducing the ensemble-based trajectories. The first year of data is used for building and training the statistical models, while the last 6 months of data (360 forecast series and corresponding measurements) is employed for the verification exercise. Further description of the test case is available in [25]. It is fairly typical for verification exercises (related to wind power forecasts) to be based on datasets covering a few months to a year.

3.1. Ensemble-based time-trajectories

Ensemble forecasts of 10-metre winds are some of the operational products of the European Centre for Medium-range Weather Forecasts (ECMWF). They are issued twice a day at 00UTC and 12UTC. Their original temporal resolution is 3-hourly up to 6 days ahead, and then 6-hourly up to 15 days ahead. Emphasis is placed on lead times up to 2-3 days ahead, being in line with current requirements for wind power management and trading. The spatial resolution of the ensemble forecasts at the time was of 50 kms. They are downscaled to the level of the wind farm after bi-linear interpolation of the gridded model output, i.e. as a distance-based weighted combination of model outputs at the 4 closest grid points. They also are linearly interpolated in time so that their temporal resolution matches that of observations.

The methodology employed for the generation of the ECMWF ensemble forecasts is well documented in the literature. A good overview is given in [26]. These ensemble predictions aim at representing uncertainties in both the knowledge of the initial state of the atmosphere and in the physical parametrization of the numerical model used for integrating these initial conditions. The former type of uncertainties is addressed by employing singular vectors to sample initial uncertainties with the largest growth [17]. The issue of uncertainty in physical model parametrization is in turn dealt with based on a stochastic physics approach, see [27] among others.

Ensemble forecasts consist of 51 time trajectories: the control (unperturbed) run, plus 50 others resulting from the perturbation of initial conditions and model parameters. The ensemble forecasts of 10-metre winds need to be transformed to power ensemble forecasts. This nonlinear transfer function is modelled as the relationship between the mean ensemble forecasts of wind speed and direction and actual wind power observations, following [25] and [28]. The random forest approach

of [25] is employed here, with a different transfer function model set up and estimated for every lead time. Since it is known that the resulting ensemble forecasts of wind power may not be probabilistically correct, they are subsequently recalibrated. This is done by inflating the variance of the ensemble forecasts based on a mean-variance model as presented in [28]. It is not our aim here to describe in detail how these ensemble forecasts of wind power are obtained since focus is not on these methods, but on the subsequent verification exercise.

3.2. Gaussian copula approach

For a given point in time t, a set of time trajectories for the coming period is available from the ensemble-based approach described above. For every lead time k, predictive densities $\hat{f}_{t+k|t}$ of wind power generation can be derived, as well as related cumulative distribution functions $\hat{F}_{t+k|t}$, e.g. by linear or cubic spline interpolation through the set of ensemble members. In the case of the ensemble-based time trajectories, the interdependence structure of the multivariate stochastic process originates from the physics of the numerical model. In the present case, we assume instead that this interdependence can be modelled with a multivariate Gaussian copula, following [15]. This means that ensemble-based trajectories and those based on a Gaussian approach will have the same marginal distributions for every lead time, as given by the predictive densities derived from the ensemble forecasts, though having different temporal dynamics.

By employing a multivariate Gaussian random number generator, one can issue at time t a number J of realizations $\{x_{t+1}^{(j)}, x_{t+2}^{(j)}, \ldots, x_{t+K}^{(j)}\}$, $j = 1, \ldots, J$ from a multivariate Gaussian variable, for a chosen covariance structure, modelled or estimated. Using the inverse probit function Φ, as well as the predictive cumulative distribution functions $\hat{F}_{t+k|t}$ for every lead time, these multivariate Gaussian realizations are transformed into trajectories of wind power generation having the same marginal distributions as the ensemble-based ones,

$$\hat{y}_{t+k|t}^{(j)} = \hat{F}_{t+k|t}{}^{-1}\left(\Phi(x_{t+k}^{(j)})\right), \quad j = 1, \ldots, J, \quad k = 0, \ldots, K \tag{1}$$

To be consistent with the ensemble forecasts of wind power from Section 3.1, at each time ensemble forecasts are issued, we also issue 51 time trajectories of wind power generation based on the Gaussian copula approach.

Two types of covariance structures are considered for comparison purposes. The first one is based on an exponential covariance function, which actually proved realistic in view of the empirical

correlations observed. Denoting by X_{t+k} the Gaussian random variable for lead time $t+k$, this writes

$$\mathrm{cov}(X_{t+k_1}, X_{t+k_2}) = \exp\left(-\frac{|k_1 - k_2|}{\nu}\right), \quad 0 \leq k_1, k_2 \leq K \tag{2}$$

where ν is the range parameter controlling the strength of the correlation of random variables among the set of lead times. Our analysis of the forecasts and measurements data at the wind farm of this study, and more precisely of the empirical temporal covariance structure, indicated that a value of $\nu = 7$ would be most appropriate. The value of ν could be optimized based on a proper fitting of the model of (2) to the data.

In parallel the second type of interdependence structure employed is based on the empirical correlation observed themselves. This alternative approach extensively described by [15] involves tracking this empirical covariance structure based on an exponential smoothing scheme. At a given time t, this structure is summarized by the covariance matrix $\mathbf{R}_t$, $\mathbf{R}_t \in \mathbb{R}^{K\times K}$, recursively updated with

$$\mathbf{R}_t = \lambda \mathbf{R}_{t-1} + (1-\lambda)\tilde{\mathbf{x}}_{t-K}\tilde{\mathbf{x}}_{t-K}^{\top} \tag{3}$$

where

$$\tilde{\mathbf{x}}_{t-K} = \left[\Phi^{-1}\left(\hat{F}_{t-K+1|t-K}(y_{t-K+1})\right), \Phi^{-1}\left(\hat{F}_{t-K+2|t-K}(y_{t-K+2})\right), \ldots, \Phi^{-1}\left(\hat{F}_{t|t-K}(y_t)\right)\right]^{\top} \tag{4}$$

is the vector of past observations transformed through the probabilistic forecasts series issued at time $t-K$, and then through the probit function Φ^{-1}. The rationale behind this transformation is that if probabilistic forecasts are probabilistically calibrated, the $\tilde{\mathbf{z}}_t$-vectors are distributed mutivariate Gaussian. Their interdependence structure is hence fully determined by the covariance matrix $\mathbf{R}_t$. It is initialized with $\mathbf{R}_0 = \mathbf{I}$, while an optimal value for λ was found to be 0.99. Note that if aiming at optimizing the quality of time trajectories, more advanced covariance structures for the Gaussian copula may be envisaged, for instance combining exponential decay with horizon-dependent range parameter and kernels for representing potential seasonalities.

3.3. Illustrative example scenarios

Three sets of scenarios are issued for the test case. The first one uses the ensemble-based method, while the other two are based on the Gaussian copula method with either the exponential covariance or the empirical covariance. Fig. 1 depicts an episode with these three sets of scenarios issued at the same date for the Oupia wind farm, for a forecast length of 72 hours and with a

hourly temporal resolution. Each set of time trajectories should be understood as 51 equally likely scenarios of power production for the coming period. In comparison with predictive densities that would be displayed for each lead time (see [29] for instance), these trajectories provide additional information about the interdependence structure of the process over this period e.g. timing of sudden level changes or variability clustering. Visually it is nearly impossible to make a difference between the various sets of trajectories.

4. The limitations of existing probabilistic verification frameworks

The classical approach to the verification of probabilistic forecasts involves marginal predictive densities for each lead time only. We follow the paradigm expressed by [18] of sharpness maximization under the constraint of calibration and place ourselves in the framework described by [19]. This exercise is carried out here to insist on the fact that, even though it is often used in practice in the energy community, this approach cannot differentiate trajectories that have similar marginal densities. It still comprises a necessary first step however, which can allow drawing first conclusions e.g. related to further calibration of the predictive densities.

4.1. Calibration of predictive densities

Calibration refers to the correspondence between forecast and observed probabilities. It is commonly evaluated with rank histograms that summarize the frequencies with which the observations fall between ordered time trajectories on a per-lead-time basis [30]. A flat rank histogram is a necessary condition for probabilistic calibration. In the present case of having 51 time trajectories, one is left with 52 bins in which observations may fall. Rank histograms may be looked at for each lead time individually, or for all lead times indifferently. Owing to the limited sample size for verification, the former approach is chosen here.

Fig. 2 depicts rank histograms for ensemble-based scenarios and for those based on the Gaussian copula approach. Even for perfectly calibrated forecasts, the limited sample size makes that deviations from the flat rank histogram case, i.e. the line with ordinate 1/52, are to be expected. The magnitude of potential deviations are illustrated by 95% intervals determined from making a Binomial assumption on the actual observation falling or not within a given bin. These histograms inform of a slight systematic overestimation of the quantiles of observed wind power generation. They also tell about a slight under-dispersion of the overall envelope of time trajectories since

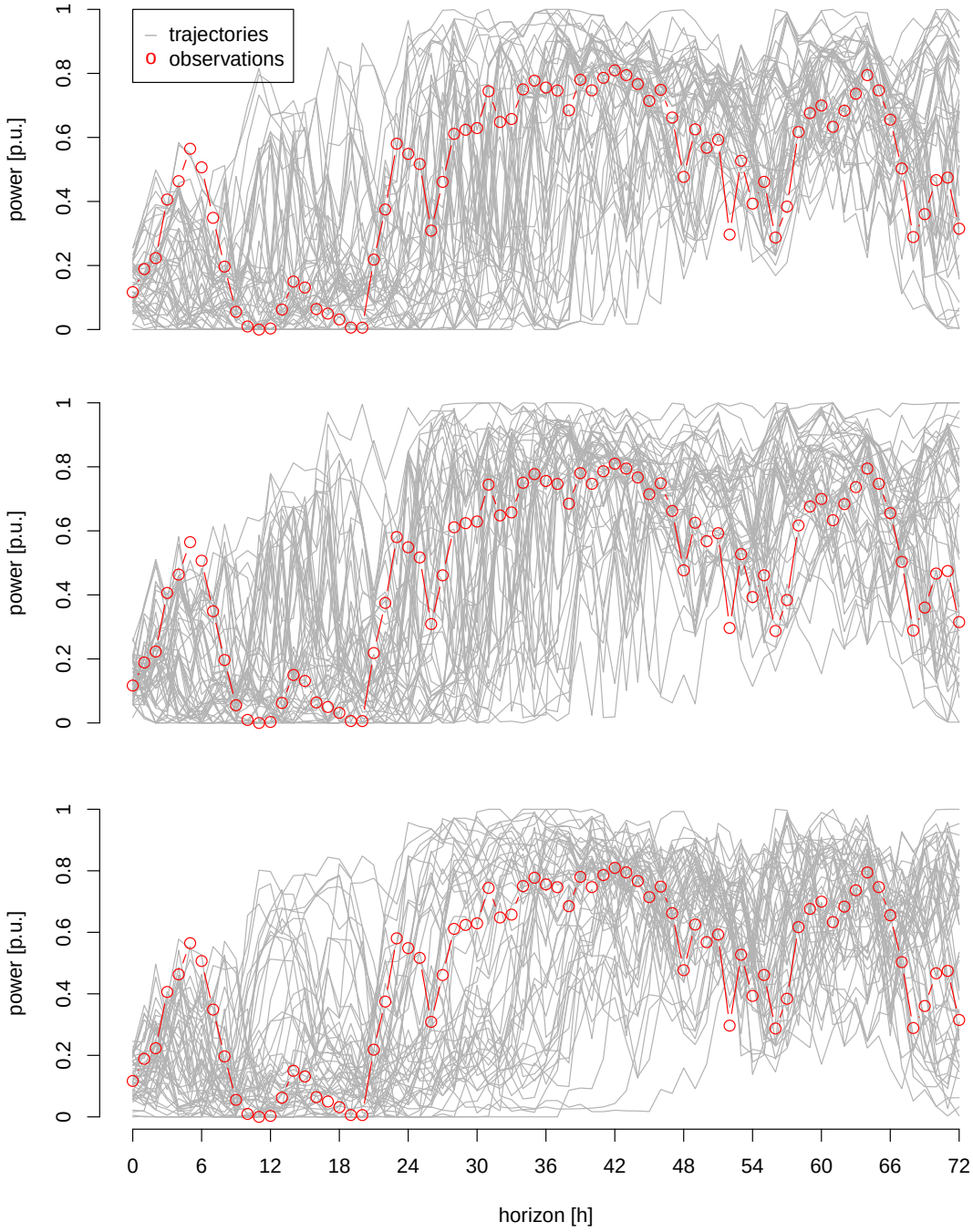


Figure 1: Example sets of time trajectories (51) of wind power production, based on (i) the ensemble-based method (bottom), (ii) the Gaussian copula method with exponential covariance (middle) and the empirical covariance (top). All three sets have the same marginal predictive distributions.

observations appear to fall more often than expected out of the range of this envelope (i.e. in bins 1 and 52). Globally these deviations are minimal. The marginals of the time trajectories hence cannot be deemed as not probabilistically calibrated. Most importantly it is not possible to make a difference between the various sets of scenarios. A similar exercise could be performed on a per-lead-time basis, though in that case sampling effects would translate to an increase in variability of frequencies among bins, which should be accounted for.

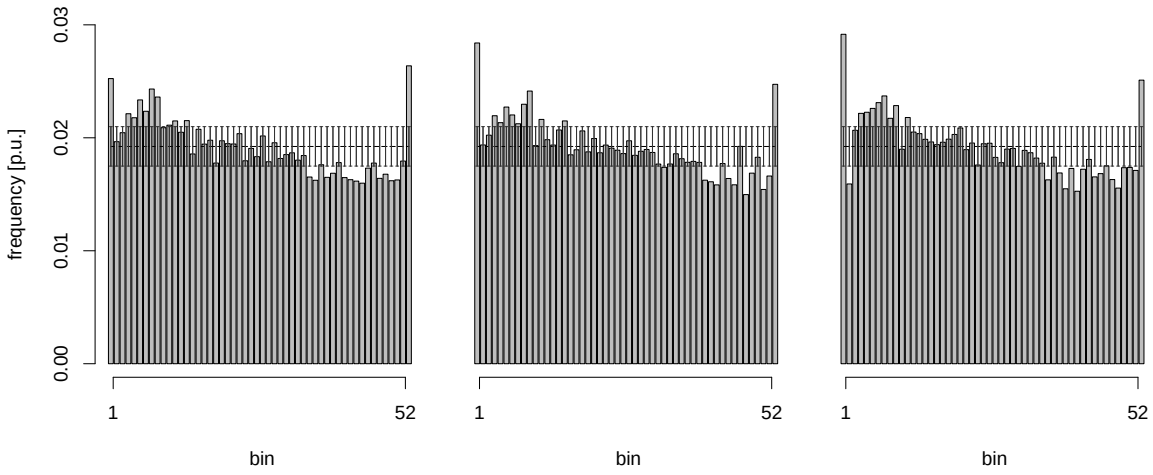


Figure 2: Probabilistic reliability of the three sets of short-term scenarios of wind power generation as evaluated by rank histograms. These results are for (i) the ensemble-based method (left), (ii) the Gaussian copula with exponential covariance (middle) and with empirical covariance (right).

4.2. Skill of predictive densities

The overall skill of the marginals of the time trajectories is evaluated based on the Continuous Rank Probability Score (CRPS), calculated as a function of the lead time. Overall skill encompasses both calibration and sharpness. The CRPS is a proper, negatively-oriented score, ensuring that lower CRPS values directly translate to actual higher skill of the predictive densities. In brief, the CRPS is a measure of the dissimilarity between the predicted cumulative distribution function $\hat{F}_{t+k|t}$ and that of the corresponding observation [18]. Since power values are normalized, the score is expressed in percentage of the wind farm nominal capacity. The value of the CRPS at time $t+k$, i.e. for a predictive cumulative distribution function $\hat{F}_{t+k|t}$ with corresponding measurement y_{t+k}, is calculated as

$$\mathrm{CRPS}_k = \frac{1}{T}\sum_{t=1}^{T}\left(\int_0^1 \left(\hat{F}_{t+k|t}(y) - \mathbf{1}(y - y_{t+k})\right)^2 dy\right) \tag{5}$$

where $\mathbf{1}(y - y_{t+k})$ represents the cumulative distribution function of the observation y_{t+k}, and with T the number of forecast series in the evaluation set.

Fig. 3 depicts the CRPS as a function of the lead time for the three sets of trajectories. It increases sharply for the first forecast horizons (1 to 10 hours ahead) then slowly augmenting from 9 to 12%. These are typical skill assessment results for probabilistic wind power forecasts, with skill deteriorating with forecast horizon [19]. The CRPS curves for the three sets of trajectories lay one of top of the other, confirming that there should not be any difference in skill between these three. Similar results would be obtained if considering alternative skill scores e.g. Ignorance, since also focusing on the properties of marginal predictive densities only. Overall the results presented show that the classical framework for probabilistic forecast evaluation focusing on the marginal predictive densities for individual lead times is not appropriate for discriminating between sets of trajectories with similar marginals but different temporal structures. Obviously still, it comprises an interesting framework for differentiating sets of scenarios with different marginal predictive densities.

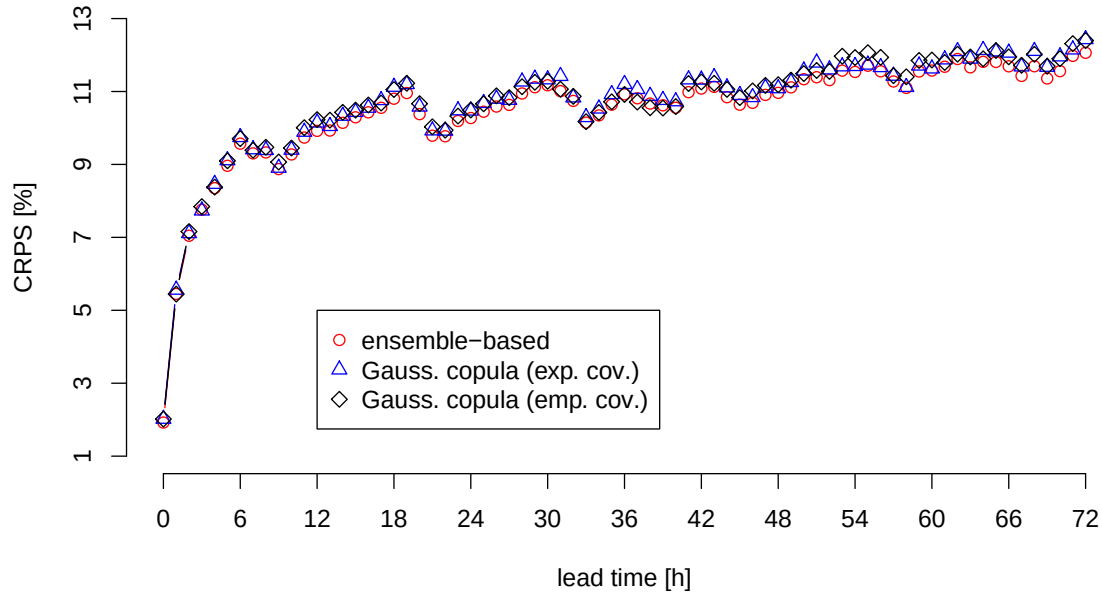


Figure 3: Evaluation of probabilistic forecast skill of the various sets of scenarios of wind power generation as a function of lead time with the CRPS.

5. The multivariate verification of scenarios

5.1. The Energy score as a multivariate skill score

A multivariate generalization of the CRPS has been proposed in order to cope with the aforementioned problem of skill scores informing on the forecast skill of predictive densities for individual lead times only [21]. The resulting score is referred to as the Energy score. For a given set of time trajectories $\hat{\mathbf{z}}_t^{(j)}$ issued at time t, it is given by

$$\mathrm{Es}_t = \frac{1}{J}\sum_{j=1}^{J} \|\mathbf{z}_t - \hat{\mathbf{z}}_t^{(j)}\|_2 - \frac{1}{2J^2}\sum_{i=1}^{J}\sum_{j=1}^{J} \|\hat{\mathbf{z}}_t^{(i)} - \hat{\mathbf{z}}_t^{(j)}\|_2 \tag{6}$$

where $\|.\|_2$ is the K-dimensional Euclidean norm (also called l^2 norm).

Similarly to the CRPS, it is averaged over the T forecast series of the evaluation set (here 360 forecast series). Es is a proper score, i.e. minimal when the true distribution is used for generating trajectories. It is a negatively-oriented score — the lower the better, while having the same unit than the variable of interest. The Energy score values for the ensemble and Gaussian copula time trajectories are collated in Table 1. For reference, we have also added the value for the non-recalibrated ensemble forecasts of wind power.

Table 1: Energy score for the various types of time trajectories. The standard deviation of the mean Energy score estimator is also given.

Method	Energy score Es (st. dev.)
Ensemble-based (non-recalibrated)	1.165 (0.014)
Gauss. copula (exp. cov.)	1.146 (0.014)
Gauss. copula (emp. cov.)	1.141 (0.014)
Ensemble-based	**1.130** (0.014)

The recalibration of the ensemble forecasts to obtain the ensemble-based scenarios allows for skill improvements. Note that the benefits of recalibration could also be observed from the rank histograms and skill scores discussed in Section 4 since recalibration is focused on the marginal distributions of the trajectories anyway. Using the Gaussian copula approach (with the exponential covariance structure) while relying on the marginals of the ensemble forecasts yields scores values that are better than for the non-recalibrated ensemble forecasts. Improvements appear more substantial when tracking the empirical covariance structure. Note that in view of the limited sample used for calculating these score values, the uncertainty of the Energy score (as given by the standard deviation of the estimator) is non-negligible. The recalibrated ensemble trajectories still have

an advantage though, which we expect to come from the physics embedded in the meteorological prediction models. These single score values do not inform though about *(i)* how the various sets of time trajectories differ in terms of their interdependence structures, *(ii)* how significant these differences are, and *(iii)* their ability to inform about specific events decision-makers are interested in, e.g. rapid changes in power generation levels.

5.2. Calibration based on multivariate rank histograms

Multivariate generalizations of the rank histograms of Section 4 were proposed in the literature for the evaluation of joint probabilistic forecasts for 2 or 3 variables [21, 31]. In the case of the trajectories considered here, these multivariate rank histograms should be produced for probabilistic forecasts of dimension the number of lead times. As an example, the Minimum Spanning Tree (MST) histogram extensively discussed in [31] is used. At a given time t, the lengths of the MSTs for all scenarios $\hat{\mathbf{z}}_t^{(j)}$ $(j = 1, \ldots, J)$ and for the observation $\mathbf{z}_t$ are determined and ordered. The MST histogram then depicts the empirical distribution over the evaluation set of the ranks of the MST lengths for the observed trajectories among the MST lengths of the predicted trajectories. The resulting histograms are gathered in Fig. 4 for the various sets of trajectories. As for Fig. 2, the magnitude of potential deviations from the flat-histogram case are illustrated by 95% intervals determined from making a Binomial assumption on the actual observation falling or not within a given bin. These intervals are necessarily wider here, since the rank histograms are based on 360 cases only, compared to the 26280 cases of Fig. 2.

Analysing the MST histograms for the three sets of trajectories, it is not possible to conclude that the scenarios are not probabilistically reliable. One can only notice a downward trend in the MST histogram for ensemble-based trajectories, and upward trends in those based on the Gaussian copula approach. This tells that the former set of scenarios tends to overestimate the temporal dependence structure in the observations, and inversely for the latter ones. Even if these MST histograms appear to allow some level of discrimination among the sets of trajectories, it still does not tell whether if, and why, a set of scenarios would be better than the other. Somehow these multivariate verification tools allow to confirm the first suspicion expressed after visual inspection of the trajectories. For scenarios with a more plausible temporal structures, differentiation is difficult even within such multivariate verification framework.

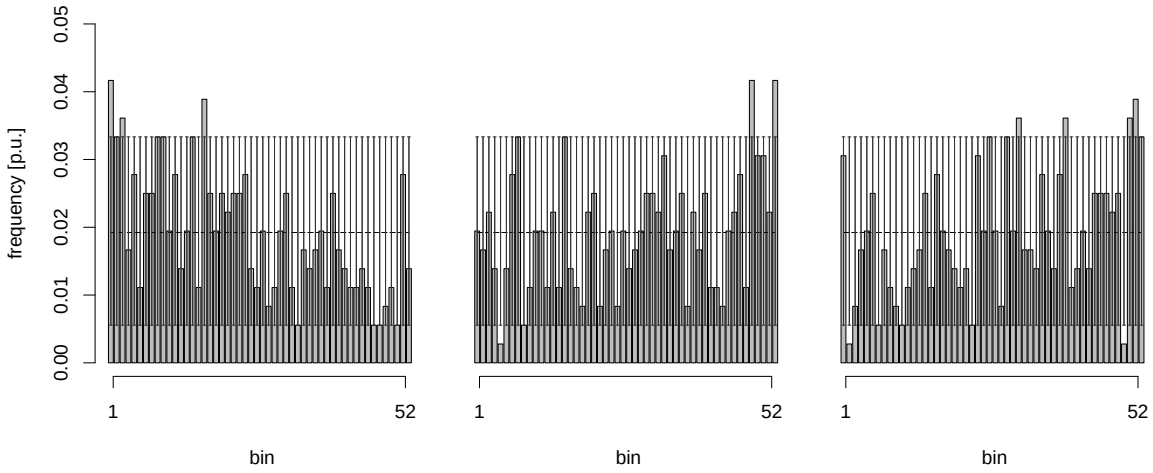


Figure 4: Multivariate probabilistic reliability of the three sets of short-term scenarios of wind power generation as evaluated by MST histograms. These results are for (i) the ensemble-based method (left), (ii) the Gaussian copula with exponential covariance (middle) and with empirical covariance (right).

6. Diagnostic event-based approach to scenario evaluation

The multivariate approach to forecast verification may permit to discriminate among various sets of trajectories, though it does not inform of their ability to mimic specific characteristics of the stochastic process. For that purpose the evaluation exercise should be focused on specific characteristics of the underlying processes. This motivates our proposal for an event-based approach to the evaluation of scenarios of short-term wind power generation. Event-based forecast verification has its roots in meteorology and climate science [33, 34]. Events are defined by setting a threshold on the value of a continuous variable, e.g. "wind power generation being greater than 50% of a wind farm's nominal capacity". The particularity of an event is that observations take values in $\{0, 1\}$ only, depending upon the event realizing or not. Related probability forecasts take values in $[0, 1]$. They are evaluated for each observation time, potentially as a function of the lead time of the forecasts.

Events can be generalized to the case of time trajectories. An event could then be formulated as "wind power generation being greater than 50% of a wind farm's nominal capacity for a period of 6 hours". In more mathematical terms, this generalization relies on functionals with input a time trajectory (forecast or realized) or a subset of this trajectory and whose output takes values

in $\{0,1\}$. For instance being at time t, one may introduce the functional g as

$$g(\mathbf{z}_t; k, h, \xi) = \prod_{i=k-h/2}^{i=k+h/2} \mathbf{1}\{y_{t+i} \geq \xi\} \tag{7}$$

where y_{t+i} is the i^{th} component of $\mathbf{z}_t$. In the above, $\mathbf{1}\{.\}$ is an indicator variable, being equal to 1 if the condition expressed within brackets realizes, and to 0 otherwise. This functional defines long-lasting events, i.e. with the process values being continuously above the threshold ξ over a number h of time steps around time step k. k can hence be seen as a form of lead time, by marking the centre of the window of interest in the future. Similarly, the functional

$$g(\mathbf{z}_t; k, h, \xi) = \mathbf{1}\left\{\left(\max_{i\in\{k-h/2,\ldots,k+h/2\}} y_{t+i} \; - \; \min_{i\in\{k-h/2,\ldots,k+h/2\}} y_{t+i}\right) \geq \xi\right\} \tag{8}$$

defines the significant gradient event, with the maximum absolute variation in the process over a window of length h, centred on time step k, being (or not) greater than ξ. In a generic manner, we write $g(\mathbf{z}_t; \boldsymbol{\theta})$ an event defined based on a time trajectory $\mathbf{z}_t$ and a parameter set $\boldsymbol{\theta}$. Other functionals could easily be defined for events relevant to forecast users interested in short-term wind power generation. For instance more general definitions for the gradient/ramp events are considered in [35].

The functionals introduced above define events from an observed time trajectory. In a similar fashion probability forecasts $\mathrm{P}_t\left[g(\mathbf{z}_t; \boldsymbol{\theta})\right]$ for these events can be obtained by applying the same functional g to the predicted set of time trajectories,

$$\mathrm{P}_t\left[g(\mathbf{z}_t; \boldsymbol{\theta})\right] = \frac{1}{J}\sum_{j=1}^{J} g(\hat{\mathbf{z}}_t^{(j)}; \boldsymbol{\theta}) \tag{9}$$

i.e. as the share of time trajectories predicting this event.

Evaluating time trajectories based on this event definition relates to an application-oriented approach, since concentrating on the type of events decision-makers want to extract from the forecasts they are provided with. For example the capacity of trajectories to inform about potential timing and durations of long-lasting period of power generation at nominal capacity level could be of crucial interest. Same would go for the duration of calm periods with no wind power generation. Events in that case would be defined using the functional in Eq. (7). If jointly looking at wind speed and power, one could also define events related to the cut-off of wind turbines. In parallel, a growing concern of forecast users is about the ability to predict significant gradients in power production

over windows of a few hours, as in Eq. (8), or about episodes with significant variability in the power output [36]. Somehow decision-makers already filter the forecasts they are provided with based on similar functionals, depending upon the decisions to be made. Verification procedures could be designed in a similar fashion.

Our proposal permits to rely on the existing frameworks for probability forecast verification for binary events, for which a wealth of methodological and applied developments exists. Here we concentrate on employing simple criteria like the Brier score and its decomposition [37, 38]. One could rely on a multitude of other approaches and criteria, considering more general scoring rules for instance [39], or diagnostic tools like the Relative Operating Characteristics (ROC) diagram, which explicitly account for hit and false alarm rates depending on probability thresholds for decisions. An extensive discussion of the relative merits of the Brier score and ROC diagrams is available in [40]. Let us remind here that the Brier score (Bs) is a proper score based on the quadratic deviation between the probability forecast of an event and its observation [41]. It does not explicitly account for hit and false alarm rates for various probability thresholds, though its propriety makes that these aspects are implicitly considered. The score value is calculated as the sample mean over the evaluation dataset of the quadratic distances between probability forecasts and corresponding observations,

$$\mathrm{Bs} = \frac{1}{T}\sum_{t=1}^{T}\left(\mathrm{P}_t\left[g(\mathbf{z}_t;\boldsymbol{\theta})\right] - g(\mathbf{z}_t;\boldsymbol{\theta})\right)^2 \tag{10}$$

where T is the length of the evaluation set. The Brier score can here be made horizon-dependent by evaluating it as a function of the lead time k.

The decomposition of the Brier score originally proposed by [37] will be employed in the application case-study. By splitting the range of possible predictions into 10 bins, this writes

$$\begin{aligned}\mathrm{Bs} \;=\; & \frac{1}{T}\sum_{i=1}^{10} n_i \left(\bar{\mathrm{P}}_t^i\left[g(\mathbf{z}_t;\boldsymbol{\theta})\right] - \bar{g}^i(\mathbf{z}_t;\boldsymbol{\theta})\right)^2 - \frac{1}{T}\sum_{i=1}^{10} n_i \left(\bar{g}^i(\mathbf{z}_t;\boldsymbol{\theta}) - \bar{\bar{g}}(\mathbf{z}_t;\boldsymbol{\theta})\right)^2 \\ & + \bar{\bar{g}}(\mathbf{z}_t;\boldsymbol{\theta})(1-\bar{\bar{g}}(\mathbf{z}_t;\boldsymbol{\theta}))\end{aligned} \tag{11}$$

where $\bar{\mathrm{P}}_t^i$ and $\bar{g}^i$ are the average predicted probabilities and observed frequencies of the event $g(\mathbf{z}_t;\boldsymbol{\theta})$ for the n_i forecasts in the i^{th} bin, while $\bar{\bar{g}}$ is the overall climatological frequency of the event. The three terms in Eq. (11) are the reliability, resolution and uncertainty components, respectively. The first two are joint attributes of forecasts and observations, while the last one

relates to observations only. Reliability was already introduced as the correspondence of predicted probabilities and observed frequencies of events. In turn resolution stands for the ability to resolve among situations with various levels of uncertainty. In the worst case a forecasting method that always predict the same probability of a given event happening has no resolution.

7. Example application of the event-based verification approach

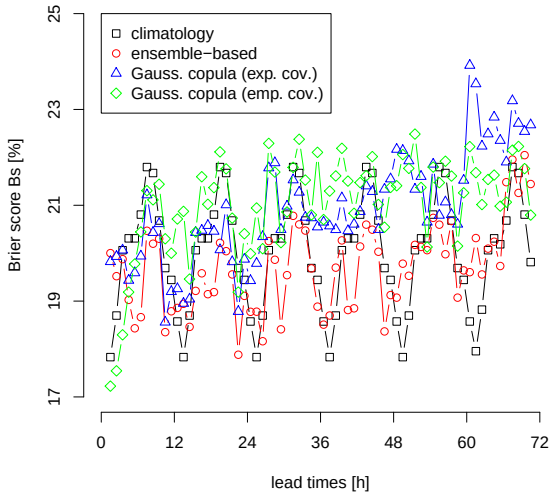
In view of the local wind climatology in this region of France, sudden changes in wind power output are to be expected. They are of great concern since possibly translating to high balancing costs if badly predicted. The energy score values in Table 1 seem to tell that the ensemble-based trajectories have a slightly higher skill overall, though the MST histograms of Fig. 4 did not show they were more probabilistically reliable. If concentrating on the issue of rapid changes in power generation, which sets of trajectories may be better, and to which extent?

To answer these questions, we set up an event-based verification exercise in the spirit of Section 6, where both predicted and observed time trajectories are filtered based on the functional of Eq. (8). Various time scales and magnitudes of level change may be relevant, hence yielding different values for the window length h and the threshold ξ. The minimum temporal resolution at which one expects dynamic information in the forecasts is of 3 hours owing to the resolution of the input meteorological forecasts. We therefore look at this time scale first, then followed by windows of 6 and 12 hours. Different thresholds are arbitrarily chosen for these various time scales, with higher thresholds for wider time windows. The parameters of this event-based verification exercise are collated in Table 2, along with the climatological frequencies of the defined events. The first two events concentrate on the short-term variability of power generation while the last two correspond more to regime changes.

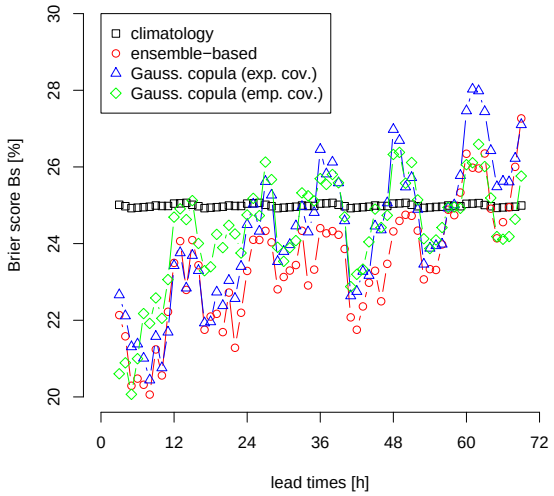
Table 2: Set-up of the event-based verification exercise, along with the climatological frequencies of these events.

Event n°	h	ξ	clim. frequency [%]
1	3	0.2	27.68
2	6	0.2	51.13
3	12	0.4	40.44
4	12	0.5	27.01

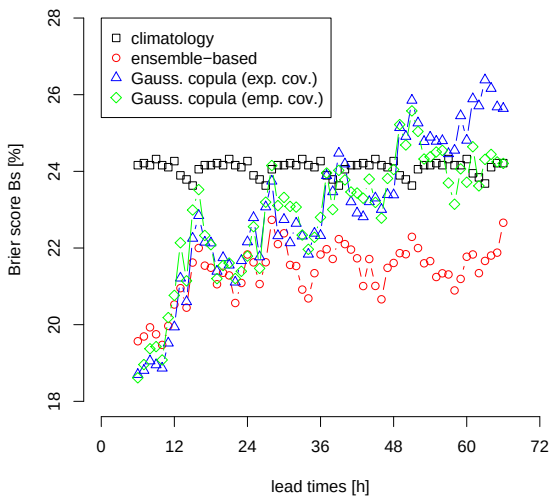
The verification results based on the horizon-dependent Brier score are gathered in Fig. 5, with climatology used as a benchmark. This benchmark unconditionally forecasts that the probability



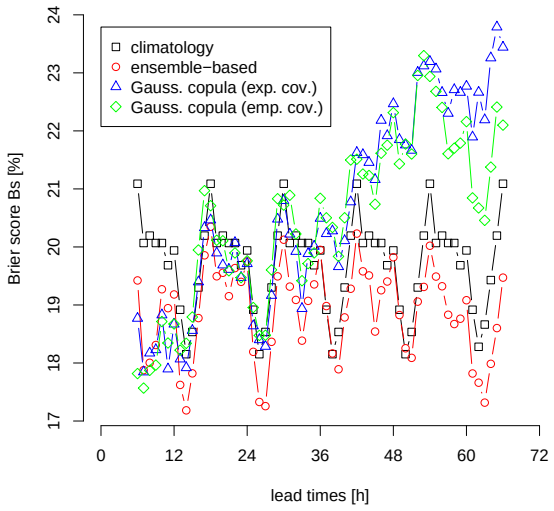
(a) Event n°1 - $h = 3$, $\xi = 0.2$



(b) Event n°2 - $h = 6$, $\xi = 0.2$



(c) Event n°3 - $h = 12$, $\xi = 0.4$



(d) Event n°4 - $h = 12$, $\xi = 0.5$

Figure 5: Event-based verification of time trajectories, for the maximum-gradient type of events. Different values of the window length h and of the threshold ξ are considered.

of the event realizing is given by the climatological frequencies gathered in Table 2. The lead time k is the centre of the window in the future. The periodicity in the skill of climatology are due to diurnal effects in wind power generation gradients. Over these four cases, the skill of the various

sets of time trajectories with respect to climatology is variable, sometimes being worse than that of this benchmark. Mainly for the most extreme events n° 1 and 4, climatology is already very competitive. For the case of event n° 1, such a low skill of the time trajectories is certainly due to phase errors i.e. errors in the timing of the events.

The point of this exercise is not to perform a comparison with climatology, but to compare the sets of time trajectories instead. For all types of events, the skill of the sets of trajectories is very similar for short lead times. The skill of Gaussian copula trajectories degrades faster than that of the ensemble-based ones, especially for the larger windows of 12 hours. This is a sensible result for the case of the exponential covariance structure, since short-term dependencies in wind power generation may be well captured, but not the changes of regimes at longer time scales. Using the empirical covariance for the dependence structure of the Gaussian copula appears to improve the event-based verification results for the first and final few lead times. This advantage certainly comes from overcoming the simplicity of the exponential covariance, which cannot represent complex temporal dependence structures. Finally however, the physics behind the ensemble-based trajectories still gives them an advantage.

To further investigate that aspect, the Brier score is decomposed following Eq. (11), though only looking at the reliability and resolution components. The uncertainty component is not looked at since being an attribute of the process itself and not of the forecasting methods. To illustrate this study, Fig. 6 depicts the result of such a decomposition for event n° 4, for which the largest deviation between the skill of the sets of time trajectories was observed. The reliability component should be minimized, and the resolution one maximized. Fig. 6 confirms the similarity in skill of the sets of trajectories for short lead times and not for longer ones. While the resolution of the various sets of trajectories stays at the same level, it is their reliability that diverges after a lead time of 20 hours. Employing the empirical covariance yields better results, though the evaluation of both reliability and resolution curves are qualitatively similar. The same behaviour of the reliability and resolution components was observed for the other defined events. This means that the Gaussian copula approach has the ability to resolve among situations with lower and higher uncertainties, though the predicted probabilities are not well calibrated. One could then conjecture that this similar resolution comes from the two sets of trajectories having the same marginals, while the reliability difference originates from the different modelling of the temporal

interdependence structure of the process.

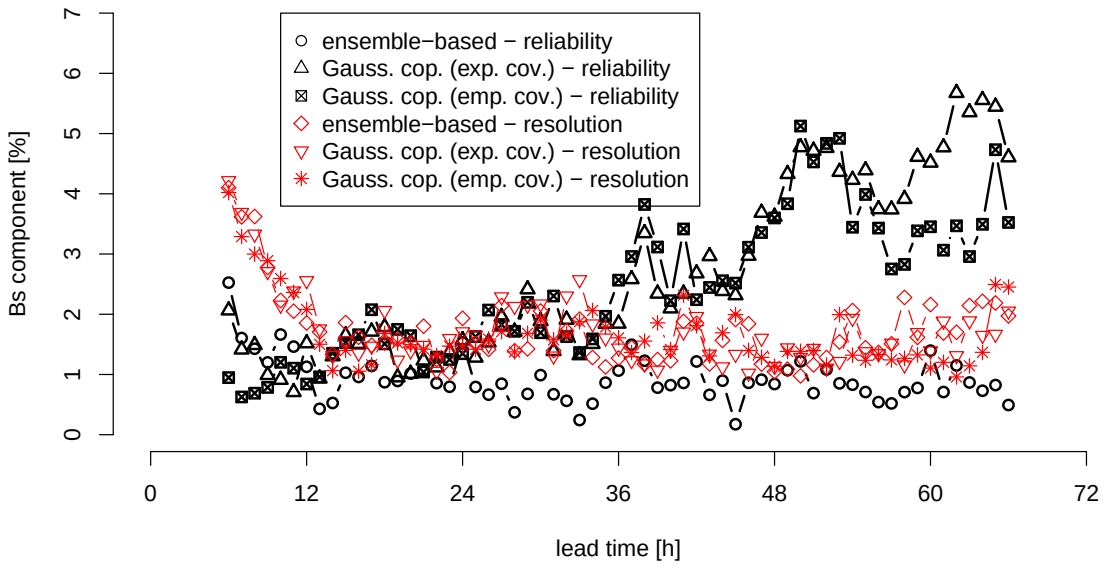


Figure 6: Decomposition of the horizon dependent Brier score into its reliability and resolution components for the two sets of time trajectories. Focus is on the maximum-gradient type of events, with $h = 12$ and $\xi = 0.5$.

While the energy score values only informed of a higher skill of the ensemble-based trajectories overall, this event-based verification exercise allowed focusing on specific attributes of the trajectories which are of interest to the decision-maker. Based on this exercise, the decision-maker would better know why to prefer ensemble-based trajectories, i.e. for their better ability to capture the interdependence structure at longer time scales. These results also are of interest to the statistical forecaster employing the Gaussian copula approach, since revealing that more advanced covariance structures for the Gaussian copula should be proposed and estimated.

8. Conclusions

Scenarios of short-term wind power generation are becoming an increasingly popular input to stochastic optimization problems. The question of the evaluation of their quality is seldom discussed or even considered. We explained and illustrated that the existing evaluation framework focusing on the marginal predictive densities of these trajectories does not allow to discriminate among

competing sets of scenarios, even if they had clearly unrealistic temporal structures. Some of the scores and diagnostic tools from the multivariate framework to probabilistic forecast evaluation may allow discarding scenarios with unrealistic temporal structures (especially the MST histograms). The Energy score is to be seen as a lead score for the evaluation of rival sets of trajectories, preferably based on datasets of sufficient length (at least a year).

Our main point has been that one clearly benefits from having a diagnostic approach to the evaluation of trajectories. In parallel in view of the high-dimensionality and difficulty of evaluating such complex forecast products, we showed how an event-based verification approach (relying on the definition of events that we would like the scenarios to mimic) certainly appears relevant. The event-defining functionals we used in this study allowed us to look at how rival sets of trajectories would represent short-term changes in power level as well as regimes changes at longer time scales. Other functionals could be proposed depending upon the focus of the forecaster aiming at evaluating and improving his scenario-generation method, or upon that of the practitioner's decision-making problems. For instance here the evaluation results hint at the fact that more advanced structures for the Gaussian copula employed should be thought of in order to improve the calibration of probability forecasts of level-change events over time scales of 6 hours or more and for lead times further than 20 hours. As of now, the physics of the ensemble-based trajectories appears to give them a slight advantage. It could certainly be reduced by using more advanced interdependence structures for generating trajectories based on statistical models. for From an applied point of view we aim at extensively using the concepts presented in the present paper for the thorough evaluation of rival approaches for the generation of scenarios of short-term wind power generation, and for various wind farms. Even if the evaluation results may be qualitatively and quantitatively different, the known properties of the scores and diagnostic approaches covered here ensure that they will allow to discriminate among rival sets of trajectories.

Methodological work will be necessary in the future for the proposal of new approaches to the evaluation of trajectories. In order to cope with the dimensionality issue, the idea of projecting them on empirical orthogonal functions may be appealing. If aiming at assessing some specific properties, more advanced transformations should be thought of by generalizing the idea of event-defining functionals. This may for instance allow focusing on ramp events and on the ability of rival approaches to inform about the uncertainty in their timing. In parallel, this framework should be

generalized to the case of multivariate variables and spatio-temporal processes, since in the future scenarios will be jointly issued and used for various renewable energy sources, along with load, and at various locations simultaneously. Actually, a more intuitive approach to the evaluation of sets of scenarios may be to concentrate on their value instead, i.e. on the comparative benefits from their use as input to various decision-making problems.

Acknowledgments

The work presented has been partly supported by the European Commission under the SafeWind project (ENK7-CT2008-213740) and the Anemos.plus project (ENK6-CT2006-038692), as well as by the Danish Council for Strategic Research, Technology and Production through the Ensymora project (10-093904/DSF). The authors are grateful to Electricité de France (EDF) for providing the power measurements, and to the European Centre for Medium-range Weather Forecasts (ECMWF) for the wind ensemble predictions. Henrik Madsen (Technical University of Denmark) and three anonymous reviewers are finally acknowledged for their comments and suggestions on an earlier version of this manuscript.

Bibliography

[1] Lund H, Mathiesen BV. Energy systems analysis of 100% renewable energy systems - The case of Denmark in years 2030 and 2050. Energy 2009;34:524–531.
[2] Lund H, Moller B, Mathiesen BV, Dyrelund A. The role of district heating in future renewable energy systems. Energy 2010;35:1381–1390.
[3] Juul N, Meibom P. Optimal configuration of an integrated power and transport system. Energy 2011;36:323–330.
[4] Flowers L, Miner-Nordstrom L. Wind energy applications for municipal water services: opportunities, situation analyses and case-studies. AWWA/WEF Joint Management Conference, Salt Lake City, Utah, 2006 (ref. NREL/CP-500-39178).
[5] Costa A, Crespo A, Navarro J, Lizcano G, Madsen H, Feitosa E. A review on the young history of the wind power short-term prediction. Renew Sust Energ Rev 2008;12:1725–1744.
[6] Monteiro C, Bessa R, Miranda V, Botterud A, Wang J, Conzelmann G. Wind power forecasting: state of the art 2009. Technical report, Argonne National Laboratory, ANL/DIS-10-1, 2009.
[7] Giebel G, Brownsword R, Kariniotakis G, Denhardt M, Draxl C. The state of the art in short-term prediction of wind power - A literature overview, 2nd edition. Technical report, EU project ANEMOS.plus, 2011.
[8] Granger CWJ. Prediction with a generalized cost of error function. Oper Res Quar 1969;20:199–207.
[9] Gneiting T. Quantiles as optimal point predictors. Int J Forecasting 2011;27:197–207.
[10] Matos M, Bessa R. Setting the operating reserve using probabilistic wind power forecasts. IEEE T Power Syst 2011;26:594–603.
[11] Wang J, Botterud A, Bessa R, keko H, Carvalho L, Issicaba D, Sumaili J, Miranda V. Wind power forecast uncertainty and unit commitment. Appl Energ 2011, available online.
[12] Meibom P, Barth R, Hasche B, Brand H, weber C, O'Malley M. Stochastic optimization model to study the operational impact of high wind power penetrations in ireland. IEEE T Power Syst 2011, available online.
[13] Pinson P, Chevallier C, Kariniotakis G. Trading wind energy from short-term probabilistic forecasts of wind power. IEEE T Power Syst 2007;22:1148–1156.
[14] Morales JM, Conejo AJ, Perez-Ruiz J. Short-term trading for a wind power producer. IEEE T Power Syst 2010;25:554–564.

[15] Pinson P, Papaefthymiou G, Klockl B, Nielsen HAa, Madsen H. From probabilistic forecasts to statistical scenarios of short-term wind power production. Wind Energy 2009;12:51–62.
[16] Morales JM, Minguez R, Conejo AJ. A methodology to generate statistically dependent wind speed scenarios. Appl Energ 2010;87,:843–855.
[17] Leutbecher M, Palmer TN. Ensemble forecasting. J Comput Phys 2008;227:3515–3539.
[18] Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J Royal Stat Soc B 2007;69:243–268.
[19] Pinson P, Nielsen HAa, Møller JK, Madsen H, Kariniotakis G. Nonparametric probabilistic forecasts of wind power: required properties and evaluation. Wind Energy 2007;10:497–516.
[20] Clements MP, Smith J. Evaluating multivariate forecast densities: a comparison of two approaches. Int J Forecasting 2002;18:397–407.
[21] Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 2008;17:211–235.
[22] Kaut M, Wallace SW. Evaluation of scenario-generation methods for stochastic programming. Pacif J Optim 2004;3:257–271.
[23] Jordà Ò, Marcellino M. Path forecast evaluation. J Appl Econom 2010;25:635–662.
[24] D'Urso P. Dissimilarity measures for time trajectories. J Italian Stat Soc 2000;9:53–83.
[25] Holmgren E. Risk indices for the estimation of uncertainty in wind power predictions based on ensembles of numerical weather predictions. M.Sc. Thesis, Chalmers University of Technology, Göteborg, Sweden.
[26] Palmer TN. Predicting uncertainty in forecasts of weather and climate. Rep Prog Phys 2000;63:71–116.
[27] Buizza R, Miller M, Palmer TN. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Q J Roy Meteor Soc 1999;131:2887–2908.
[28] Pinson P, Madsen H. Ensemble-based probabilistic forecasting at Horns Rev. Wind energy 2009;12:137–155.
[29] Pinson P, Kariniotakis G. Conditional prediction intervals of wind power generation. IEEE T Power Syst 2010;25:1845-1856.
[30] Hamill TM. Interpretation of rank histograms for verifying ensemble forecasts. Mon Weather Rev 2001;129:550–560.
[31] Wilks DS. The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. Mon Weather Rev 2004;132:1329–1340.
[32] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007;102:359-378.
[33] Jolliffe IT, Stephenson DB. Forecast verification - A practitioner's guide in atmospheric science. New York: Wiley; 2003.
[34] Wilks DS. Statistical methods in the atmospheric sciences. 2nd ed. London: Academic Press; 1995.
[35] Bossavy A, Girard R, Kariniotakis G. Forecasting ramps of wind power production with numerical weather prediction ensembles. Wind Energy 2011; in press.
[36] Cutler N, Kay M, Jacka K, Nielsen TS. Detecting, categorizing and forecasting large ramps in wind farm power output using meteorological observations and WPPT. Wind Energy 2007;10:453–470.
[37] Murphy AH. A new vector partition of the probability score. J App Meteor 1973;12:595–600.
[38] Stephenson DB, Coelho CAS, Jolliffe IT. Two extra components in the Brier score decomposition. Weather Forecast 2008;23:752–757.
[39] Benedetti R. Scoring rules for forecast verification. Mon Weather Rev 2010;138:201–211.
[40] Toth Z, Tallagrand O, Candille G, Zhu Y. Probability and ensemble forecasts. In: Jolliffe IT, Stephenson DB, editors. Forecast verification - A practitioner's guide in Atmospheric Science, London: John Wiley & Sons; 2006, p. 137–164.
[41] Brier R. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78:1–3.