Generation of scenarios from calibrated ensemble forecasts with a dynamic ensemble copula coupling approach

Ben Bouallègue, Zied; Heppelmann, Tobias; Theis, Susanne E.; Pinson, Pierre

Published in:
ArXiv

Publication date:
2015

Document Version
Peer reviewed version

Generation of scenarios from calibrated ensemble forecasts with a dynamic ensemble copula coupling approach

Zied Ben Bouallègue[a,b], Tobias Heppelmann[a], Susanne E. Theis[a] and Pierre Pinson[c]

[a]Deutscher Wetterdienst, Offenbach, Germany
[b]Meteorological Institute, University of Bonn, Germany
[c]Technical University of Denmark, Denmark

Abstract

Probabilistic forecasts in the form of ensemble of scenarios are required for complex decision making processes. Ensemble forecasting systems provide such products but the spatio-temporal structures of the forecast uncertainty is lost when statistical calibration of the ensemble forecasts is applied for each lead time and location independently. Non-parametric approaches allow the reconstruction of spatio-temporal joint probability distributions at a low computational cost.For example, the ensemble copula coupling (ECC) method consists in rebuilding the multivariate aspect of the forecast from the original ensemble forecasts. Based on the assumption of error stationarity, parametric methods aim to fully describe the forecast dependence structures. In this study, the concept of ECC is combined with past data statistics in order to account for the autocorrelation of the forecast error. The new approach which preserves the dynamical development of the ensemble members is called dynamic ensemble copula coupling (d-ECC). The ensemble based empirical copulas, ECC and d-ECC, are applied to wind forecasts from the high resolution ensemble system COSMO-DE-EPS run operationally at the German weather service. The generated scenarios are assessed in the form of time series by means of multivariate verification tools and in a product oriented framework. Verification results over a 3 month period show that the innovative method d-ECC outperforms or performs as well as ECC in all investigated aspects. In particular, the temporal variability of the time trajectories are better captured with d-ECC which preserves the information content of the original scenarios.

1 Introduction

Probabilistic information are essential for an optimal use of a forecast (Krzysztofowicz, 1983). Such information can be provided by an Ensemble Prediction System (EPS) which aims at

describing the flow-dependent forecast uncertainty (Leutbecher and Palmer, 2008). Several deterministic forecasts are run simultaneously accounting for uncertainties in the description of the initial state, the model parametrization and, for limited area models, the boundary conditions. Probabilistic products are derived from an ensemble, tailored to specific user's need. For example, wind forecasts in the form of quantile at selected probability levels are of particular interest for actors in the renewable energy sector (Pinson, 2013).

However, probabilistic products generally suffer from a lack of reliability, the system showing biases and failing to fully represent the forecast uncertainty. Statistical techniques allow to adjust the ensemble forecast correcting for systematic inconsistencies (Gneiting *et al.*, 2007). This step known as calibration is based on past data and usually focuses on a single or few aspects of the ensemble forecast. For example, calibration of wind forecast can be performed with univariate approaches (Bremnes, 2004; Sloughter *et al.*, 2010; Thorarinsdottir and Gneiting, 2010) or bivariate methods which account for correlation structures of the wind components (Pinson, 2012; Schuhen *et al.*, 2012). These calibration procedures provide reliable predictive probability distribution of wind speed or wind components for each forecast horizon and location independently. Decision making problems can however require information about the spatial and/or temporal structure of the forecast uncertainty. Examples of application in the renewable energy sector resemble the optimal operation of a wind-storage system in a market environment, the unit commitment over a control zone or the optimal maintenance planning (Pinson *et al.*, 2009). In other words, scenarios that describe spatio-temporal wind variability are relevant products for end-users of wind forecasts.

The generation of scenarios from calibrated ensemble forecasts is a step that can be performed with the use of empirical copulas. The empirical copula approaches are non-parametric and, in comparison with parametric approaches (Keune *et al.*, 2014; Feldmann *et al.*, 2015), simple to implement and computationally cheap. Empirical copulas can be based on climatological records (Schaake Shuffle (ScSh); Clark *et al.*, 2004) or on the original raw ensemble (ensemble copula coupling (ECC); Schefzik *et al.*, 2013). ECC, which consists in the conservation of the ensemble member rank structure from the original ensemble to the calibrated one, has the advantage to be applicable to any location of the model domain without restriction related to the availability of observations. But ScSh has been shown to outperform ECC (Wilks, 2014). ECC provides indeed not realistic scenarios when the post-processing indiscriminately increases the ensemble spread to a large extent. Non-representative correlation structures in the raw ensemble are magnified after calibration leading to unrealistic forecast variability. As a consequence, ECC can deteriorate the ensemble information content when applied to ensembles with relatively poor reliability as suggested, for example, by verification results in Flowerdew (2014).

In this paper, a new version of the ECC approach is proposed in order to overcome the

generation of unrealistic scenarios. Focusing on time series, a temporal component is introduced in the ECC scheme: it preserves the dynamical development of each member by accounting for the autocorrelation of the forecast error over consecutive forecast horizons. The assumption of forecast error stationarity, already adopted for the development of fully parametric approaches (Pinson *et al.*, 2009; Schölzel and Hense., 2011), is exploited in combination with the structure information of the original scenarios. The new approach, called dynamic ensemble copula coupling (d-ECC), generates scenarios which reflect the autocorrelation of the error in the original ensemble. Objective verification is performed in order to show the benefit of the proposed approach with regards to the standard ECC.

The manuscript is organized as follows: Section 2 describes the dataset used to illustrate the manuscript as well as the calibration method applied to derive calibrated quantile forecasts from the raw ensemble. Section 3 introduces the empirical copula approaches for the generation of scenarios and discuss in particular the ECC and d-ECC methods. Section 4 describes the verification process for the scenario assessment. Section 5 presents the results obtained by means of multivariate scores and in a product oriented verification framework.

2 Data

2.1 Ensemble forecasts and observations

COSMO-DE-EPS is the high resolution ensemble prediction system run operationally at DWD. It consists of 20 COSMO-DE forecasts with variations in the initial conditions, the boundary conditions and the model physics (Gebhardt *et al.*, 2011; Peralta *et al.*, 2012). COSMO-DE-EPS follows the multi-model ensemble approach, with 4 global models driving each 5 physically perturbed members. The ensemble configuration implies a clustering of the ensemble members as a function of the driving global model when large scale structures dominate the forecast uncertainty.

The focus is here on wind forecast at 100 meter height above ground. The post-processing methods are applied to forecasts of the 00UTC run with an hourly output interval and a forecast horizon of up to 21 hours. The observation dataset comprises quality controlled wind measurements from 7 stations: Risoe, FINO1, FINO2, FINO3, Karlsruhe, Hamburg and Lindenberg, as plotted in Figure 1. The verification period covers a 3 month period: March, April and May 2013.

Figure 2(a) shows an example of COSMO-DE-EPS wind forecast at hub-height. The forecast is valid on day March 2, 2013, at station FINO1 (see Figure 1). The ensemble members are drawn in grey while the corresponding observations are drawn in black. In Figure 2(b), the raw ensemble forecast is interpreted in the form of quantiles.

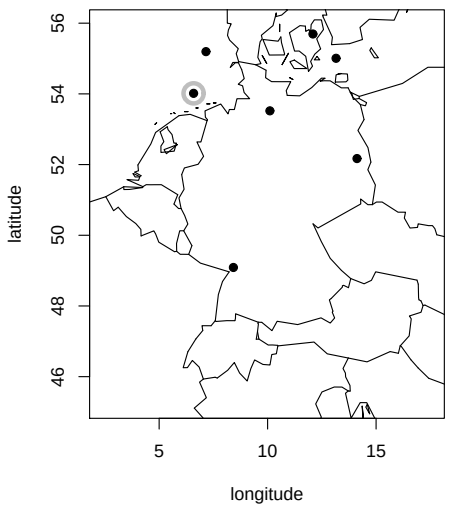

Figure 1: Map of Germany and neighboring areas (approximately the COSMO-DE domain) with latitude/longitude axes. Location of the 7 wind stations used in this study. The station FINO1 is highlighted with a grey mark.

Formally, a quantile forecast q_τ at probability level τ (with $0 \leq \tau \leq 1$) is defined as:

$$q_\tau := F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\} \tag{1}$$

where F is the predictive cumulative probability distribution of the random variable $Y \in \Re$ such:

$$F(y) = \mathbb{P}(Y \leq y). \tag{2}$$

In practice, at each forecast lead time, the member of rank n is interpreted as a quantile forecast at probability level τ_n such:

$$\tau_n = \frac{n - 0.5}{N_e} \tag{3}$$

where N_e is the number of ensemble members.

In the example of Figure 2, the raw ensemble is not able to capture the observation variability. Calibration aims to correct for this lack of reliability by adjusting the mean and enlarging the spread of the ensemble forecast.

2.2 Calibrated ensemble forecasts

Since COSMO-DE-EPS forecasts have shown to suffer from statistical inconsistencies (Ben Bouallègue, 2013; Ben Bouallègue, 2015), calibration has to be applied in order to provide reliable forecasts

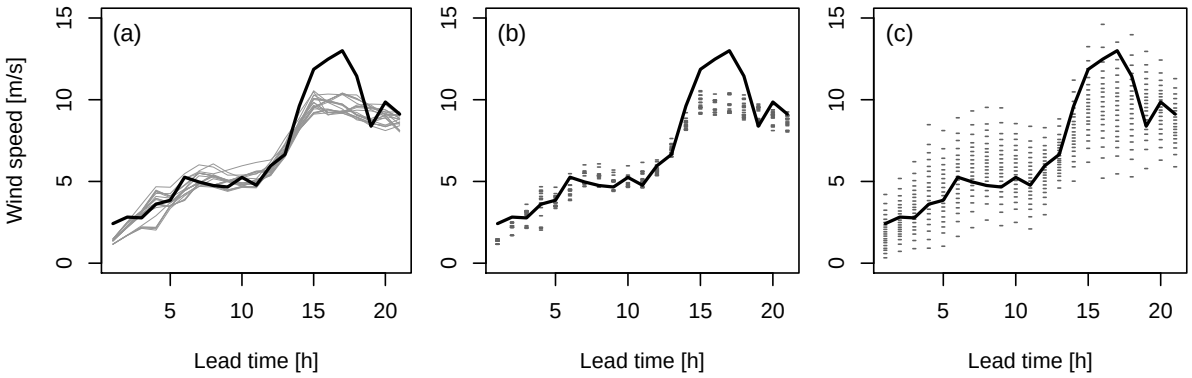

Figure 2: Wind speed at 100 meter height above ground, on March 2, 2013, at station FINO1: (a) COSMO-DE-EPS forecast (grey lines), (b) raw ensemble forecast in the form of quantiles (sorted members, see text), (c) calibrated quantile forecasts, and the corresponding observations (black lines).

to the users. The method applied in this study is the bivariate Non-homogeneous Gaussian Regression (EMOS, Schuhen *et al.*, 2012). The mean and variance of each wind component as well as the correlation between the two components characterize the predictive bivariate normal distribution. Correction applied to the raw ensemble mean and variance are optimized by minimizing the continuous ranked probability score ($CRPS$; Matheson and Winkler, 1976). The calibration coefficients are estimated for each station and each lead time separately (local version of EMOS), based on a training period being defined as a moving window of 45 days.

The final calibrated products considered here are N_e equidistant quantile forecasts of wind speed estimated for each location and each forecast horizon separately. The N_e probability levels associated to the forecast quantiles follow Eq. (3). Since N_e corresponds to the ensemble size, this definition of the probability levels matches the interpretation of an ensemble forecast in terms of quantiles which optimizes the $CRPS$ (Bröcker, 2012).

Calibrated quantile forecasts are shown in Figure 2(c). The spread of the ensemble is increased with respect to Figure 2(b) and thus the observation variability is now captured by the forecast. The calibration method provides from a statistical point of view reliable ensemble marginal distribution and reliable quantile forecasts as checked by means of rank histograms and quantile reliability plots (not shown). The performance of the applied calibration technique is similar to the one obtained with other methods such quantile regression (Koenker and Bassett, 1978; Bremnes, 2004).

Information about spatial and temporal dependence structures, which are crucial in many applications, are however no more available after this calibration step (see Figure 2(c)). The next post-processing step consists then in the generation of consistent scenarios based on the

calibrated samples.

3 Generation of scenarios

The generation of scenarios with empirical copulas is here briefly described. For a deeper insight into the methods, the reader is invited to refer to the original article of Schefzik *et al.* (2013), or to Wilks (2014) and references within.

First, consider the multivariate cumulative distribution function (*cdf*) G defined as:

$$G(y_1, ..., y_L) = \mathbb{P}[Y_1 \leq y_1, ..., Y_L \leq y_L] \tag{4}$$

of a random vector $(Y_1, ..., Y_L)$ with $y_1, ..., y_L \in \mathbb{R}$. As in Eq. (2), we note F_i the marginals defined :

$$F_i(y) = \mathbb{P}[Y_i \leq y]. \tag{5}$$

The Sklar's theorem (Sklar, 1959) states that G can be expressed as:

$$G(y_1, ..., y_L) = C(F_1(y_1), F_L(y_L)) \tag{6}$$

where C is a copula that links an L-variate cumulative distribution function G to its univariate marginal *cdf*s $F_1, ..., F_L$.

In Eq. (6), a joint distribution is represented as univariate margins plus copulas. The problem of estimating univariate distributions and the problem of estimating dependence can therefore be treated separately. Univariate calibration marginal *cdf*s $F_1, ..., F_L$ are provided by the univariate calibration step described in the previous Section. The choice of the copula C depends on the application and on the size L of the multivariate problem. We focus here on empirical copulas since they are suitable for problems with high dimensionality.

We note H the empirical copula. H is based on a multivariate dependence canvas, a specific discrete dataset $\boldsymbol{z}$ defined in $\mathbb{R}^L$. The chosen dataset is described formally as:

$$\boldsymbol{z} := \left\{(z_1^1, ..., z_1^N), ..., (z_L^1, ..., z_L^N)\right\} \tag{7}$$

consisting of L tuples of size N with entries in $\mathbb{R}$. In other words, L is the dimension of the multivariate variable and N is the number of scenarios. The rank of z_l^n for $n \in \{1, ..., N\}$ and $l \in \{1, ..., L\}$ is defined as:

$$R_l^n := \sum_{i=1}^{N} \mathbb{I}(z_l^i \leq z_l^n) \tag{8}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function taking value 1 if the condition in parenthesis is true

and zero otherwise. The empirical copula H induced by the dataset $\boldsymbol{z}$ is given by:

$$H(\frac{j_1}{N},...,\frac{j_L}{N}) := \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(R_1^i \leq j_1,...,R_L^i \leq j_L) \tag{9}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\prod_{l=1}^{L}\mathbb{I}(R_l^i \leq j_l) \tag{10}$$

for integers $0 \leq j_1,...,j_L \leq N$.

In practice, N equidistant quantiles of F_l with $l \in \{1,...,L\}$ are derived from the univariate calibration step:

$$\boldsymbol{q} := \left\{(q_1^1,...,q_1^N),...,(q_L^1,...,q_L^N)\right\} \tag{11}$$

with

$$q_l^n := F_l^{-1}(\tau_n)\,; \quad n \in \{1,..,N\} \tag{12}$$

where τ_n is defined in Eq. (3). The sample $\boldsymbol{q}$ is rearranged following the dependence structure of the reference canvas $\boldsymbol{z}$. The permutations $\pi_l(n) := R_l^n$ for $n \in \{1,..,N\}$ are derived from the univariate order statistics $z_l^{(1)} \leq ... \leq z_l^{(N)}$ for $l \in \{1,..,L\}$ and applied to the univariate calibrated sample $\boldsymbol{q}$. The post-processed ensemble $\tilde{x}_l^1,...,\tilde{x}_l^N$ for each margin l is expressed as:

$$\tilde{x}_l^1 := q_l^{\pi_l(1)},...,\tilde{x}_l^N := q_l^{\pi_l(N)} \tag{13}$$

The multivariate correlation structures are generated based on the rank correlation structures of a sample canvas $\boldsymbol{z}$. The empirical copulas presented here only differ in the way $\boldsymbol{z}$ is defined. In the following, let $k \in \{1,...,K\}$ be a location and $t \in \{1,...T\}$ a look-ahead time and let $L := K \times T$. For simplicity, we consider here a single weather variable.

3.1 Ensemble copula coupling

The rank structure of the ensemble is preserved after calibration when applying the standard ensemble copula coupling approach (ECC). The raw ensemble forecast is noted $\boldsymbol{x}$:

$$\boldsymbol{x} := \left\{(x_1^1,...,x_1^{N_e}),...,(x_L^1,...,x_L^{N_e})\right\} \tag{14}$$

where N_e is the ensemble size. ECC applies without restriction to any multivariate setting. The number of scenarios generated with ECC is however the same as the size of the original ensemble ($N = N_e$). The transfer of the rank structure from the raw ensemble forecast to the calibrated one consists then in taking $\boldsymbol{x}$ as the required canvas in Eq. (7).

Based on COSMO-DE-EPS forecasts of Figure 3(a) (identical to Figure 2(a)), an example

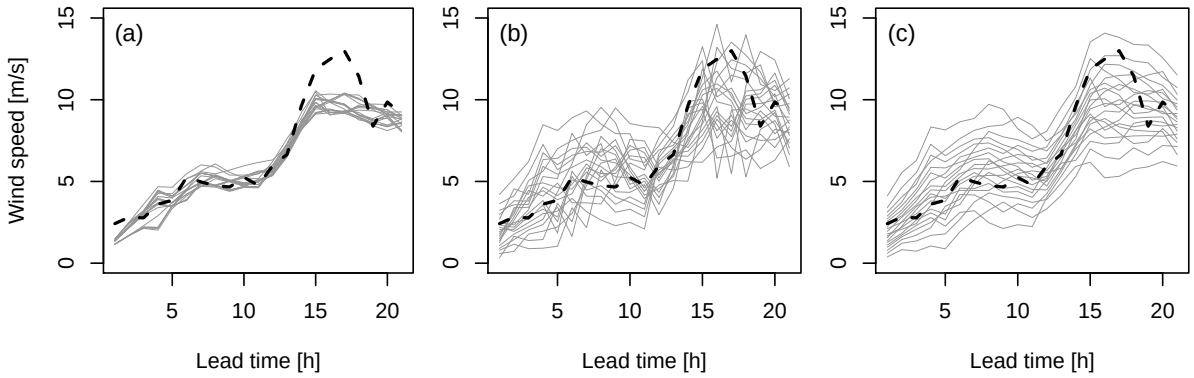

Figure 3: Same example as in Figure 2: (a) COSMO-DE-EPS scenarios, (b) ECC derived scenarios, (c) d-ECC derived scenarios, and the corresponding observations (black lines).

of scenarios derived with ECC is provided in Figure 3(b). The increase of spread after the calibration step implies a larger step-to-step variability in the time trajectories. Figure 4 focuses on a single scenario highlighting the difference between the original and post-processed scenarios.

3.2 Dynamic ensemble copula coupling

ECC assumes that the ensemble prediction system correctly describes the spatio-temporal dependence structures of the weather variable. This assumption is quite strong and cannot be valid in all cases. On the other side, based on an assumption of error stationarity, parametric methods have been developed focusing on covariance structures of the forecast error (Pinson *et al.*, 2009; Schölzel and Hense., 2011). We propose a new version of the ECC approach which is an attempt to combine both information: the structure of the original ensemble and the autocorrelation of the forecast error. The new scheme is called dynamic ensemble copula coupling (d-ECC) since it aims this way at preserving the dynamical development of the ensemble forecasts.

For this purpose, we note $\boldsymbol{e}$ the forecast error defined as the difference between deterministic forecasts and observations:

$$\boldsymbol{e} := \{e_1, ..., e_T\} \tag{15}$$

$$= \{x_1 - y_1, ..., x_T - y_T\} \tag{16}$$

where x_t and y_t with $t \in \{1, ..., T\}$ are a deterministic forecast and the corresponding observation, respectively. For simplicity, the dimensionality of the scenarios has been restricted to the temporal component. The temporal correlation of the error is described by a correlation matrix

$\boldsymbol{\Sigma_e}$ defined as:

$$\boldsymbol{\Sigma_e} = \begin{pmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,T} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{T,1} & \rho_{T,2} & \cdots & \rho_{T,T} \end{pmatrix} \tag{17}$$

where $\rho_{t_1 t_2}$ is the correlation coefficient between e_{t_1} and e_{t_2}, so the correlation between the forecast error at time step t_1 and the forecast error at time step t_2. An estimator of the correlation matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$ is performed by an ordinary least squares procedure based on the training samples used for the univariate calibration step at the different forecast horizons. In our setup, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$ is regularly updated on a daily basis from the moving windows defined as training datasets.

Again here, we aim at constructing a canvas (Eq. (7)) in order to establish the correlation structures within the calibrated ensemble $\boldsymbol{q} := \left\{(q_1^1, ..., q_1^N), ..., (q_L^1, ..., q_L^N)\right\}$. In the d-ECC approach, the canvas is built from an iterative process described by the following steps:

1. Define a reference sample canvas $\boldsymbol{z}$, choosing for example for the first iteration $\boldsymbol{z}$=$\boldsymbol{x}$,

2. Apply ECC in order to derive a post-processed ensemble of scenarios $\tilde{\boldsymbol{x}}$:

$$\tilde{\boldsymbol{x}} := \left\{(\tilde{x}_1^1, ..., \tilde{x}_1^{N_e}), ..., (\tilde{x}_L^1, ..., \tilde{x}_L^{N_e})\right\}, \tag{18}$$

3. Derive the "bias" correction $\boldsymbol{c}^i$ imposed to each scenario i ($i \in 1, ..., N_e$) of the reference canvas by the post-processing step:

$$\boldsymbol{c}^i := \left\{c_1^i, ..., c_L^i\right\} \tag{19}$$

$$= \left\{\tilde{x}_1^i - z_1^i, ..., \tilde{x}_L^i - z_L^i\right\}, \tag{20}$$

4. Adjust the bias correction of each scenario based on the estimate of the error autocorrelation $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$ and assuming $\boldsymbol{c}$ as representative of the forecast error:

$$\hat{\boldsymbol{c}}^i := \left\{\hat{c}_1^i, ..., \hat{c}_L^i\right\} \tag{21}$$

with

$$\hat{\boldsymbol{c}}^i = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} \boldsymbol{c}^{i^T}. \tag{22}$$

5. Derive the so-called adjusted ensemble $\hat{\boldsymbol{z}}$ such:

$$\hat{\boldsymbol{z}} := \left\{(\hat{z}_1^1, ..., \hat{z}_1^M), ..., (\hat{z}_L^1, ..., \hat{z}_L^M)\right\} \tag{23}$$

where a scenario $\hat{\boldsymbol{z}}^i = \left\{\hat{z}_1^i, ..., \hat{z}_L^i)\right\}$ of $\hat{\boldsymbol{z}}$ is defined as a combination of the reference member and the adjusted bias:

$$\hat{\boldsymbol{z}}^i = \boldsymbol{z}^i + \hat{\boldsymbol{c}}^i, \tag{24}$$

6. Set $\boldsymbol{z} = \hat{\boldsymbol{z}}$ defining $\hat{\boldsymbol{z}}$ as new reference canvas in point 1 and repeat the process until it converges. Convergence is checked computing the mean absolute difference between $\hat{\boldsymbol{z}}$ and $\boldsymbol{z}$.

The adjusted ensemble $\hat{\boldsymbol{z}}$ combines the raw ensemble structure and the autocorrelation of the forecast error for each member separately. The convergence is reached when the adjusted biases bring the reference canvas to the level of correction required by the calibration step.

The reasoning behind the d-ECC setup can be clarified as follows. Focusing on a single member, the impact of the calibration step at a given forecast horizon, *i.e.* the forecast difference after and before post-processing, is seen as a conditional bias correction. This bias correction is aimed to be consistent over the different forecast horizons. In Figure 4(b), we see that the bias correction associated to the ECC scenario can strongly vary from one forecast horizon to the next one (black line, Eq. (20) of step 3). The bias is smoothed based on the assumption of temporal autocorrelation of the error (dashed line, Eq. (22) of step 4). This bias adjustment is then superimposed to the reference forecast before to draw from again the correlation structure of the scenarios.

The new scheme reduces to the standard ECC in the case where $rank(z_{t,k}^i) = rank(\hat{z}_{t,k}^i)$ for all $i \in \{1, ..., M\}$ and $k \in \{1, ..., K\}$ after the first iteration, which means that the additional terms $\hat{\boldsymbol{c}}^i$ do not have any impact on the rank structure of the ensemble. This case occurs if:

- $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix, which means that there is no temporal correlation of the error in the original ensemble,
- $\boldsymbol{c} = \boldsymbol{0}$ where $\boldsymbol{0}$ is the null vector, which means that the calibration step does not impact the forecast, the forecast being already well calibrated.
- $\boldsymbol{c} = h \cdot \boldsymbol{J}$ where h is a constant and $\boldsymbol{J}$ an all-ones vector, which means that the calibration step corrects only for bias errors and the system is spread bias free.

So the d-ECC typically takes effect if calibration corrects the spread and if this correction is correlated in time at the member level.

The assumption of error autocorrelation is checked and adjusted by means of the correlation matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$ estimated from the training datasets used for calibration at the different time horizons. Based on the dataset described in Section 2, Figure 5 shows the lagged correlation of the forecast error derived from $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$. The correlation is decreasing as a function of the time lag,

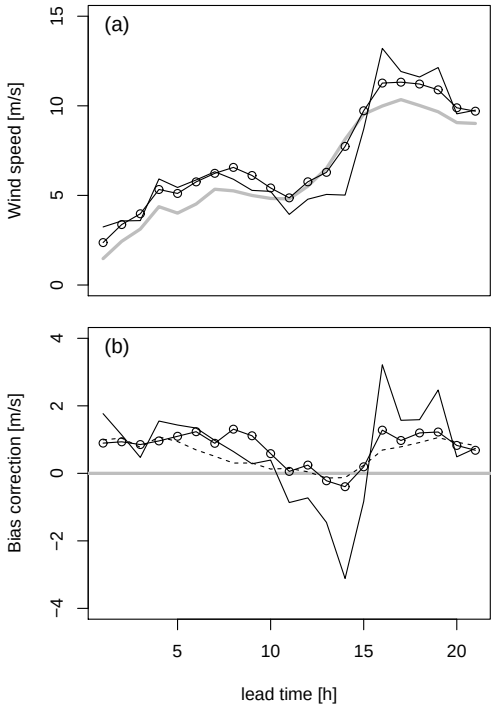

Figure 4: Illustration of the concept of d-ECC based on the example of Figure 3 showing (a) one among the 20 scenarios and (b) the bias correction applied to the original scenario after post-processing. The raw ensemble forecast (here the member 13) is represented in grey, the ECC scenario in black, and the d-ECC scenario in black with dots. The dashed line represents the adjusted bias correction suggested by d-ECC after the first iteration step (see text).

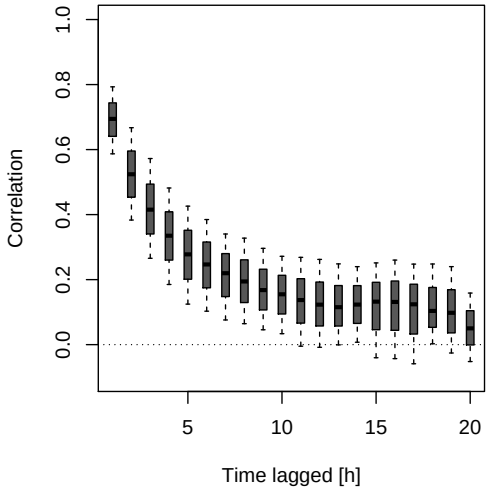

Figure 5: Temporal lagged correlation coefficients summarizing the error correlation matrix $\hat{\mathbf{\Sigma}}_e$ used in the d-ECC approach. The boxplots indicate the variability within the 3 month calibration period.

reaching near zero values for lags greater than 10 hours. However, for short and very short time lags, the correlation is high and stable over the rolling training datasets. In particular, focusing on a time lag of 1 hour, the correlation ranges between 60% and 80%. Its low variability over a 3 month period indicates that the temporal correlation of the forecast error is not flow dependent. As a consequence, d-ECC can be seen as a "universal" approach that does not suffer restriction related to the forecasted weather situation.

Considering again our case study, the scenarios generated with d-ECC based on the COSMO-DE-EPS forecasts are shown in Figure 3(c). The d-ECC derived scenarios are smoother and subjectively more realistic than the ones derived with ECC in Figure 3(b). In Figure 4, focusing on a single scenario, it is highlighted that the difference between the original and the d-ECC time trajectories varies gradually from one time interval to the next one while abrupt transitions occur in the case of the ECC scenario, as in this example between hours 15 and 17.

The time series variability is statistically explored by means of a spectral analysis, an analysis of the time series in the frequency domain. Such an analysis is useful in order to describe statistical properties of the scenarios but has also direct implications for user's applications (see Section 5; Vincent *et al.*, 2010). A Fourier transformation is applied to each forecasted and observed scenario and the contributions of the oscillations at various frequencies to the scenario variance examined (Wilks, 2006). In Figure 6, the amplitude of the forecast and observation time series is plotted as a function of their frequency components.

As already suggested by the case study, this analysis confirms that the ECC considerably

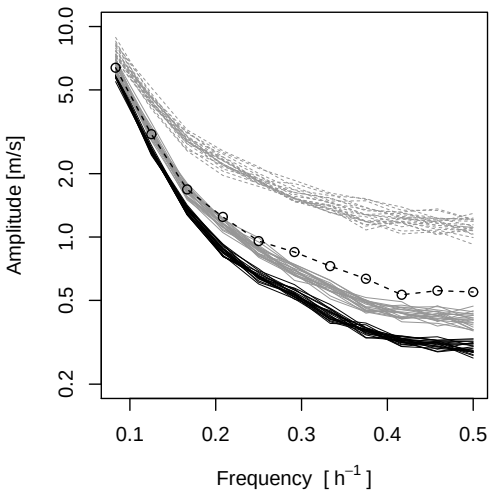

Figure 6: Spectral analysis of the scenarios from the raw ensemble (black lines), of the scenarios derived with ECC (dashed grey lines) and with d-ECC (grey lines). Each line corresponds to one scenario among the 20. The spectrum of the observed time series is represented by the dashed dotted line.

increases the variability of the time trajectories with respect to the original ensemble, in particular at high frequencies. ECC scenario fluctuations are also much larger than the observed ones. Indeed, the amplitude is on average more than two times larger at high frequencies in ECC time series than in the observed ones which explains the visual impression that ECC scenarios are unrealistic. Conversely, scenarios derived with the new copula approach do not exhibit such features. While the original ensemble shows a deficit of variability with respect to the observations, the d-ECC approach allows improving this aspect of the forecast. This first result, showing that d-ECC scenarios have a similar spectrum as the observation one, is completed with an objective assessment of the forecasted scenarios based on probabilistic verification measures.

4 Verification methods

4.1 Multivariate scores

Verification of scenarios is first performed assessing the multivariate aspect of the forecast by means of adequate scores. The scores are applied focusing on scenarios in the form of time series. Considering an ensemble with N_e scenarios $\boldsymbol{x}^{(n)}$ with $n \in \{1, ..., N_e\}$ and an observed scenario

$\boldsymbol{y}$, the energy score (ES; Gneiting *et al.*, 2008) is defined as:

$$ES = \frac{1}{N_e}\sum_{n=1}^{N_e}\|\boldsymbol{y} - \boldsymbol{x}^{(n)}\| - \frac{1}{2N_e^2}\sum_{m=1}^{N_e}\sum_{p=1}^{N_e}\|\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(p)}\| \tag{25}$$

where $\|.\|$ represents the Euclidean norm. ES is a generalization of the $CRPS$ to the multivariate case.

ES suffers from a lack of sensitivity to misrepresentation of correlation structures (Pinson and Tastu, 2013). We consider therefore additionally the p-variogram score (pVS; Scheuerer and Hamill, 2015), which has better discriminative property with this respect. Based on the geostatistical concept of variogram, pVS is defined as:

$$pVS = \sum_{i\neq j}\omega_{ij}\left(\mid y_i - y_j \mid^p - \frac{1}{N}\sum_{n=1}^{N}\mid x_i^{(n)} - x_j^{(n)} \mid^p\right)^2 \tag{26}$$

with p the order of the variogram and where ω_{ij} are weights and the indices i and j indicate the i-th and the j-th components of the marked vectors, respectively. As suggested by Scheuerer and Hamill (2015), the weights ω_{ij} are chosen proportional to the inverse distance in time such:

$$\omega_{ij} = \frac{1}{\mid i - j \mid}, \quad i \neq j, \tag{27}$$

since i and j are here forecast horizon indices.

4.2 Multivariate rank histograms

The multivariate aspect of the forecast is in a second step assessed by means of rank histograms applied to multi-dimensional fields (Thorarinsdottir *et al.*, 2014). Two variants of the multivariate rank histogram are applied: the averaged rank histogram (ARH) and the band depth rank histogram ($BDRH$). The difference of the two approaches lies in the way to defined pre-ranks from multivariate forecasts. ARH considers the averaged rank over the multivariate aspect while $BDRH$ assesses the centrality of the observation within the ensemble based on the concept of functional band depth.

The interpretation of ARH is the same as the interpretation of a univariate rank histogram: $\cup$-shaped, $\cap$-shaped, and flat rank histograms are interpreted as underdispersiveness, overdispersiveness, and calibration of the underlying ensemble forecasts, respectively. The interpretation of $BDRH$ is different: a $\cup$-shape is associated to a lack of correlation, a $\cap$-shape to a too high correlation in the ensemble, a skewed rank histogram to bias or dispersion errors and a flat rank histogram to calibrated forecasts.

4.3 Product oriented verification

Besides multivariate verification of time series scenarios, the forecasts are assessed in a product oriented framework. This type of scenario verification follows the spirit of the event oriented verification framework proposed by Pinson and Girard (2012). Probabilistic forecasts that require time trajectories are provided and assessed by means of well-established univariate probabilistic scores.

Two types of products derived from forecasted scenarios are here under focus. The first one is defined as the mean wind speed over a day (here, a day is limited to the 21 hour forecast horizon). The second product is defined as the maximal upward wind ramp over a day, a wind ramp being defined as the difference between two consecutive forecast intervals. For both products, 20 forecasts are derived from the 20 scenarios at each station and each verification day.

The performances of the ensemble forecasts for the two types of products are evaluated by means of the $CRPS$. The $CRPS$ is the generalization of the mean absolute error to predicitve distributions (Gneiting *et al.*, 2008), and can be seen as the integral of the Brier score (BS; Brier, 1950) over all thresholds or the integral of the quantile score (QS; Koenker and Bassett, 1978) over all probability levels. Considering an ensemble forecast, the $CRPS$ can be estimated as a weighted sum of QS applied to the sorted ensemble members (Bröcker, 2012). For a deeper insight in the forecast performance in terms of attributes, the $CRPS$ is decomposed following the same approach (Ben Bouallègue, 2015): the $CRPS$ reliability and resolution components are estimated as weighted sums of the reliability and resolution components of the QS at the probability levels defined by the ensemble size (see Eq. (3)), respectively. Formally, we write:

$$CRPS_{reliability} = \frac{2}{N_e} \sum_{n=1}^{N_e} QS_{reliability}^{(\tau_n)} \tag{28}$$

$$CRPS_{resolution} = \frac{2}{N_e} \sum_{n=1}^{N_e} QS_{resolution}^{(\tau_n)} \tag{29}$$

where $QS_{reliability}^{(\tau_n)}$ and $QS_{resolution}^{(\tau_n)}$ are the reliability and resolution components of the QS applied to the quantile forecasts at probability level τ_n, respectively. The QS decomposition is performed following Bentzien and Friederichs (2014). The $CRPS_{reliability}$ is negatively oriented (the lower the better) while the $CRPS_{resolution}$ is positively oriented (the higher the better).

4.4 Bootstrapping

The statistical significance of the results are tested applying a block-bootstrap approach. Bootstrapping is a resampling technique which provides an estimation of the statistical consistency (Efron and Tibshirani, 1986) and is commonly applied to meteorological datasets (e.g. Hamill, 1999).

A block-bootstrap approach is applied in the following which consists in defining a block as a single day of the verification period. Each day is considered as a separate block of fully independent data. The verification process is repeated 500 times using each time a random sample with replacement of the 92 verification days (March, April, May, 2013). The derived score distributions illustrate consequently the variability of the performance measures over the verification period and not between locations. Boxplots are used to represent the distributions of the performance measures, where the quantile of the distributions at probability levels 5%, 25%, 50%, 75 % and 95% are highlighted.

5 Results and discussion

Figure 7 shows the performance of the forecasted time trajectories by means of multivariate scores. The post-processed scenarios perform significantly better than the raw members in terms of ES (Figure 7(a)). In terms of pVS, the d-ECC scenarios are better than the ECC ones and significantly better than the raw ones when $p = 0.5$ (Figure 7(b)). For higher orders of the variogram (here $p = 1$, Figure 7(c)), the forecast improvement after post-processing is still clear when using d-ECC while the ECC results are slightly worse than the ones of the original forecasts.

Figure 8 depicts the results in terms of multivariate rank histograms, ARH (upper panel) and $BDRH$ (lower panel). The raw ensemble shows clear reliability deficiencies (Figures 8(a) and 8(d)) which motivated the use of post-processing techniques. Forecasts derived with ECC show still underdispersiveness but also too little correlation (Figures 8(b) and 8(e)) while forecasts derived with d-ECC are better calibrated according to the rank histograms in Figures 8(c) and 8(f). We note however a light tendency to overdispersiveness and too high correlation in the d-ECC case.

Figure 9 focuses on two products drawn from the time series forecasts: the daily mean wind speed (upper panel) and the daily maximal upward ramp (lower panel). The performances are assessed in terms of $CRPS$, $CRPS$ reliability and $CRPS$ resolution, from left to right, respectively. Looking at the results in terms of $CRPS$, we note the high similarity of Figures 9(a) and 9(d) with Figures 7(a) and 7(b), respectively. As for the ES, post-processing significantly improves the forecasts of the daily mean product. As for pVS with $p = 1$, d-ECC improves

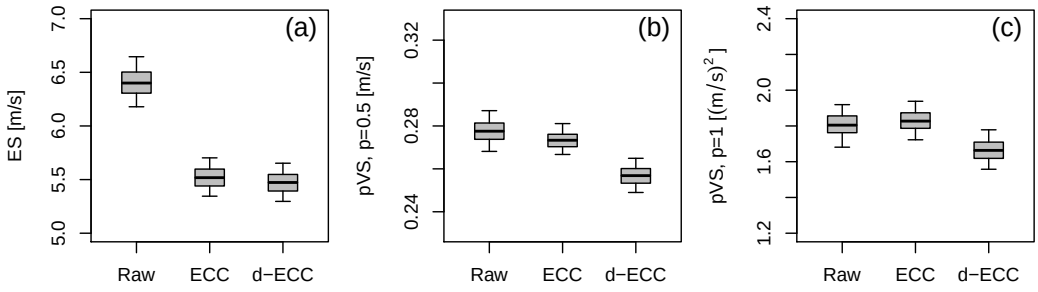

Figure 7: Multivariate scores of time series: energy score (a) and p-variogram score for $p = 0.5$ (b) and $p = 1$ (c) in the form of box plots drawn from the application of a 500-block bootstrapping.

the ramp product with respect to the original one while ECC deteriorates it. The $CRPS$ decomposition allows detailing the origin of these performances. We see in Figures 9(b) and 9(e) that the $CRPS$ results are mainly explained by the impact of the post-processing on the $CRPS$ reliability components. However, focusing on the results in terms of $CRPS$ resolution in Figures 9(c) and 9(f), we note that the resolution of the original and d-ECC products are similar while ECC deteriorates the resolution of the ramp product with respect to the original one.

Those verification results are interpreted as follows. Calibration corrects for the mean of the ensemble forecast and this is reflected, after the derivation of scenarios, by an improvement of the ES and daily mean product skill. Calibration also corrects for spread deficiencies increasing the variability of the ensemble forecasts. This increase of spread associated with a preservation of the rank structure of the original ensemble, as it is the case in the ECC approach, enlarges indiscriminately the temporal variability of the forecasts and leads to a deterioration of the pVS and ramp product results.

The d-ECC approach provides scenarios with a temporal variability comparable to the one of the observation. In that case, the benefit of the calibration step in terms of reliability (at single forecast horizons) persists at the multivariate level (looking at time trajectories) after the reconstruction of scenarios with d-ECC. The multivariate reliability, or the reliability of derived products, is improved but not perfect after postprocessing, exhibiting a tendency of overdispersiveness and too high correlation that could be related to an overestimation of the time correlation. However, unlike ECC, d-ECC is able to preserve the information content of the original ensemble forecasts which means that the value of end-products derived from d-ECC scenarios is not compromised with the application of a post-processing including d-ECC.

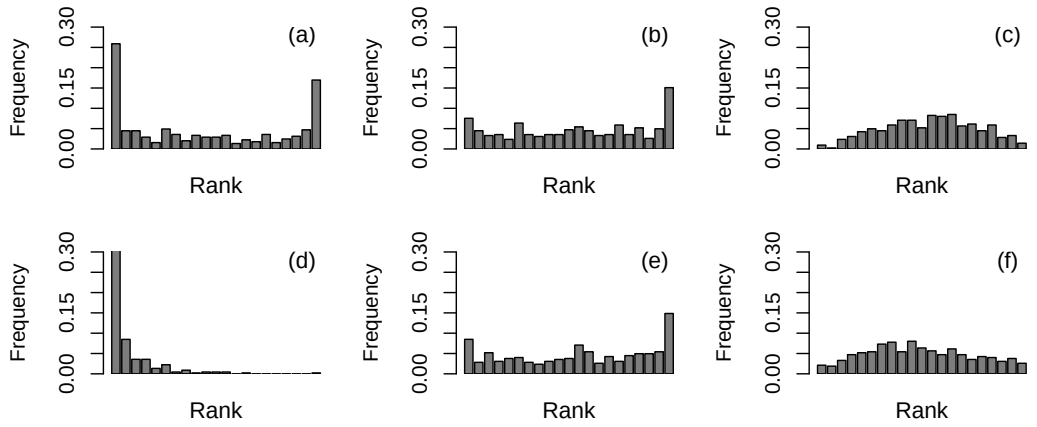

Figure 8: Multivariate rank histograms: (a,b,c) average rank histograms and (d,e,f) band depth rank histograms for time series from the raw ensemble (a,d) and derived with ECC (b,e) and d-ECC (c,f).

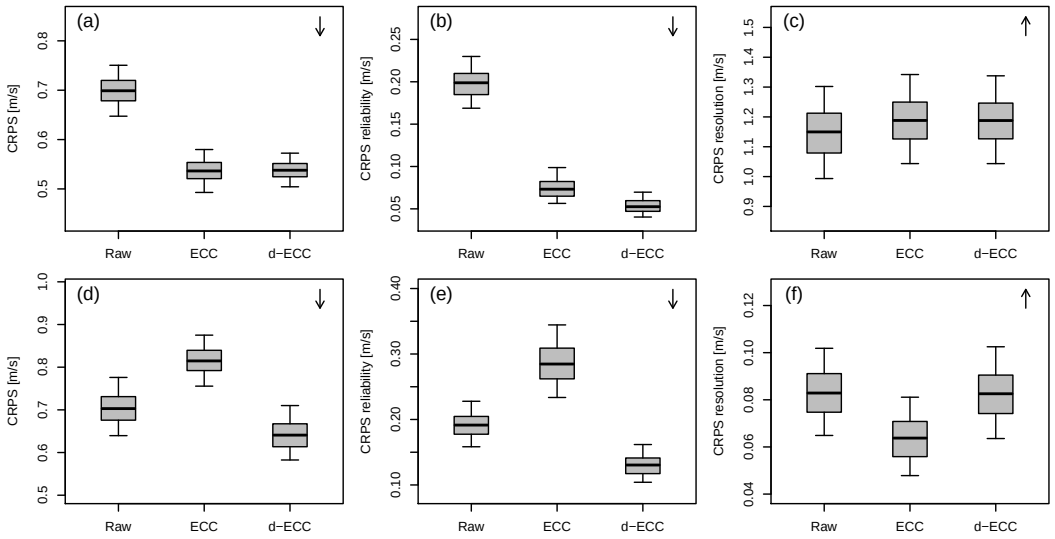

Figure 9: Product oriented verification of scenarios: (a,b,c) daily means at station, (d,e,f) maximal upward ramps within a day at station. Results are shown in terms of $CRPS$ (a,d), $CRPS$ reliability component (b,e) and $CRPS$ resolution component (c,f). The box plots indicate confidence intervals estimated with block bootstrapping. The arrows in the right corners indicate whether the performance measure is positively or negatively oriented.

6 Conclusion and outlook

A new empirical copula approach is proposed for the post-processing of calibrated ensemble forecasts. The so-called dynamic ensemble copula coupling approach is introduced with a focus on temporal structures of wind forecasts. The new scheme includes a temporal component in the ECC approach accounting for the bias autocorrelation associated to each member. The estimation of the correlation in the bias structure allows adjusting this assumption based on past data.

The scenarios derived with d-ECC show to be qualitatively realistic and quantitatively performant. Post-processing of wind speed based on EMOS and d-ECC improves the forecasts in many aspects. In comparison to ECC, d-ECC drastically improves the performance of the derived scenarios. Applications that require temporal trajectories will fully benefit of the dynamic approach. Though only the temporal aspect has been investigated in this study, the dynamic ensemble copula approach can be straightforwardly generalized to any multivariate setting. Further research is however required for the application of d-ECC at scales that are unresolved by the observations. For example, geostatistical tools could be applied for the description of the autocorrelation error structure at the model grid level.

d-ECC accounts for deficiencies or limitations in the generation and/or calibration of the ensemble forecast. As for any post-processing technique, the benefit of the dynamic copula can be weakened by improving the representation of the forecast uncertainty with more efficient member generation techniques and/or by improving the calibration procedure correcting for conditional biases. Meanwhile, at low additional complexity and computational costs, d-ECC can be considered as a valuable alternative to the standard ECC for the generation of consistent scenarios.

Acknowledgments

This work has been done within the framework of the EWeLiNE project (*Erstellung innovativer Wetter- und Leistungsprognosemodelle für die Netzintegration wetterabhängiger Energieträger*) funded by the German Federal Ministry for Economic Affairs and Energy. The authors acknowledge the department of wind energy of the Technical University of Denmark (DTU), the German Wind Energy Institute (DEWI GmbH), DNV GL, the Meteorological Institute (MI) of University of Hamburg and the Karlsruhe Institute of Technology (KIT) for providing wind measurements at stations Risoe, FINO1 and FINO3, FINO2, Hamburg and Lindenberg, and Karlsruhe, respectively.

References

Ben Bouallègue Z. 2013. Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Wea. Forecasting* **28**: 515–524, doi: 10.1175/WAF-D-12-00062.1.

Ben Bouallègue Z. 2015. Assessment and added value estimation of an ensemble approach with a focus on global radiation forecasts. *Mausam* **66**: 541–550.

Bentzien S, Friederichs P. 2014. Decomposition and graphical portrayal of the quantile score. *Q.J.R. Meteorol. Soc.* **140**: 1924–1934, doi:10.1002/qj.2284.

Bremnes JB. 2004. Probabilistic wind power forecasts using local quantile regression. *Wind Energ.* **7**: 47–54, doi:10.1002/we.107.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.* **78**: 1–3.

Bröcker J. 2012. Evaluating raw ensembles with the continuous ranked probability score. *Q.J.R. Meteorol. Soc.* **138**: 161–1617, doi:10.1002/qj.1891.

Clark M, Gangopadhyay S, Hay L, Rajagopalan B, Wilby R. 2004. The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology* **5**: 243–262, doi:10.1175/1525-7541(2004)005⟨0243:TSSAMF⟩2.0.CO;2.

Efron B, Tibshirani R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* **1**: 54–75.

Feldmann K, Scheuerer M, Thorarinsdottir T. 2015. Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous gaussian regression. *Mon. Wea. Rev.* **143**: 955–971, doi:10.1175/MWR-D-14-00210.1.

Flowerdew J. 2014. Calibrating ensemble reliability whilst preserving spatial structure. *Tellus A* **66**, doi:10.3402/tellusa.v66.22662.

Gebhardt C, Theis SE, Paulat M, Ben Bouallègue Z. 2011. Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.* **100**: 168–177, doi:10.1016/j.atmosres.2010.12.008.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration, and sharpness. *J. Roy. Stat. Soc.* **69B**: 243–268, doi:10.1111/j.1467-9868.2007.00587.x.

Gneiting T, Stanberry L, Grimit E, Held L, Johnson N. 2008. Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test* **17**: 211–235.

Hamill TM. 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Weather Forecasting* **14**: 155–167, doi:10.1175/1520-0434(1999)014⟨0155:HTFENP⟩2.0.CO;2.

Keune J, Ohlwein C, Hense A. 2014. Multivariate probabilistic analysis and predictability of medium-range ensemble weather forecasts. *Mon. Wea. Rev.* **142**: 4074–4090, doi:10.1175/MWR-D-14-00015.1.

Koenker R, Bassett G. 1978. Regression quantiles. *Econometrica* **46**: 33–50.

Krzysztofowicz R. 1983. Why should a forecaster and a decision maker use Bayes theorem. *Water Resour. Res.* **19**: 327–336, doi:10.1029/WR019i002p00327.

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *Journal of Computational Physics* **227**: 3515–3539, doi:10.1016/j.jcp.2007.02.014.

Matheson J, Winkler R. 1976. Scoring rules for continuous probability distributions. *Manage. Sci.* **22**: 1087–1096.

Peralta C, Ben Bouallègue Z, Theis SE, Gebhardt C. 2012. Accounting for initial condition uncertainties in COSMO-DE-EPS. *Journal of Geophysical Research* **117**, doi:10.1029/2011JD016581.

Pinson P. 2012. Adaptive calibration of (u,v)-wind ensemble forecasts. *Quart. J. Roy. Meteor. Soc.* **138**: 1273–1284, doi:10.1002/qj.1873.

Pinson P. 2013. Wind energy: Forecasting challenges for its operational management. *Statistical Science* **28**: 564–585, doi:10.1214/13-STS445.

Pinson P, Girard R. 2012. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy* **96**: 12–20, doi:10.1016/j.apenergy.2011.11.004.

Pinson P, Papaefthymiou G, Klockl B, Nielsen H, Madsen H. 2009. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy* **12**: 51–62, doi:10.1002/we.284.

Pinson P, Tastu J. 2013. Discrimination ability of the energy score. *Technical report, Technical University of Denmark*.

Schefzik R, Thorarinsdottir T, Gneiting T. 2013. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statist. Sci.* **28**: 616–640, doi:10.1214/13-STS443.

Scheuerer M, Hamill T. 2015. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Mon. Wea. Rev.* **143**: 1321–1334, doi:10.1175/MWR-D-14-00269.1.

Schölzel C, Hense A. 2011. Probabilistic assessment of regional climate change in Southwest Germany by ensemble dressing. *Climate Dyn.* **36**: 2003–2014, doi:10.1007/s00382-010-0815-1.

Schuhen N, Thorarinsdottir T, Gneiting T. 2012. Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.* **140**: 3204–3219, doi:10.1175/MWR-D-12-00028.1.

Sklar M. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8**: 229–231.

Sloughter J, Gneiting T, Raftery AE. 2010. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.* **105**: 25–35, doi:10.1175/MWR-D-12-00002.1.

Thorarinsdottir T, Gneiting T. 2010. Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Statist. Soc. Ser. A* **173**: 371–388, doi:10.1111/j.1467-985X.2009.00616.x.

Thorarinsdottir T, Scheuerer M, Heinz C. 2014. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics* **in press**.

Vincent C, Giebel G, Pinson P, Madsen H. 2010. Resolving nonstationary spectral information in wind speed time series using the hilbert–huang transform. *J. Appl. Meteor. Climatol.* **49**: 253–267, doi:10.1175/2009JAMC2058.1.

Wilks D. 2014. Multivariate ensemble model output statistics using empirical copulas. *Quart. J. Roy. Meteor. Soc.* **141**: 945–952, doi:10.1002/qj.2414.

Wilks DS. 2006. *Statistical methods in the atmospheric sciences*. 2nd Edn. Academic Press, New York, 627pp.