Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities*

MICHAEL SCHEUERER AND THOMAS M. HAMILL

NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 19 August 2014, in final form 2 December 2014)

ABSTRACT

Proper scoring rules provide a theoretically principled framework for the quantitative assessment of the predictive performance of probabilistic forecasts. While a wide selection of such scoring rules for univariate quantities exists, there are only few scoring rules for multivariate quantities, and many of them require that forecasts are given in the form of a probability density function. The energy score, a multivariate generalization of the continuous ranked probability score, is the only commonly used score that is applicable in the important case of ensemble forecasts, where the multivariate predictive distribution is represented by a finite sample. Unfortunately, its ability to detect incorrectly specified correlations between the components of the multivariate quantity is somewhat limited. In this paper the authors present an alternative class of proper scoring rules based on the geostatistical concept of variograms. The sensitivity of these variogram-based scoring rules to incorrectly predicted means, variances, and correlations is studied in a number of examples with simulated observations and forecasts; they are shown to be distinctly more discriminative with respect to the correlation structure. This conclusion is confirmed in a case study with postprocessed wind speed forecasts at five wind park locations in Colorado.

1. Introduction

During the last two decades a paradigm shift has occurred in the practice of numerical weather prediction (NWP). To account for the various sources of uncertainty in the NWP model output, ensemble prediction systems were developed and have now become state of the art in meteorological forecasting (Buizza et al. 2005; Lewis 2005; Leutbecher and Palmer 2008). Those ensemble forecasts aim to represent the range of possible outcomes, and probabilistic statements like the probability of exceeding a certain amount of precipitation can be derived from them and help making informed decisions.

Along with the availability of probabilistic forecasts comes the need for both diagnostic and quantitative methods to assess the quality of those forecasts and to compare the performance of competing forecasters. A probabilistic forecast should be calibrated (i.e., statistically consistent with the values that materialize) and sharp (i.e., very specific about the anticipated weather; Gneiting et al. 2007). Sharpness can be assessed via numerical and graphical summaries of the width of the prediction intervals that come with a predictive probability distribution. The notion of calibration is more complex, and different types of calibration have been established. Marginal calibration measures the similarity of the aggregated predictive distribution and the climatological distribution of the predictand, and can be checked by comparing the average predictive cumulative distribution function (CDF) with the empirical CDF of the observations (Gneiting et al. 2007). Probabilistic calibration concerns the dynamical aspects of probabilistic forecasts and can be assessed by studying verification rank histograms (Anderson 1996; Hamill and Colucci 1997; Hamill 2001).

To make a quantitative comparison of different forecast methods, summary measures of their predictive performance are required. Those measures should take both calibration and sharpness into account. To this end, scoring rules have been proposed which assign a numerical score $S(F, y)$ to each pair (F, y) where F is the CDF of the predictive distribution and y is the

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/MWR-D-14-00269.s1.

Corresponding author address: Michael Scheuerer, NOAA/ESRL, Physical Sciences Division 325 Broadway, R/PSD1, Boulder, CO 80305.
E-mail: michael.scheuerer@noaa.gov

DOI: 10.1175/MWR-D-14-00269.1

realized value. If we take scoring rules to be negatively oriented, $S(F, y)$ can be viewed as a penalty that the forecasters wish to minimize. A crucial property that one should always require from a scoring rule is that it is *proper*, which is formally defined by the requirement

$$E_G S(G, Y) \leq E_G S(F, Y) \quad \forall \quad F, G, \tag{1}$$

where $E_G S(F, Y)$ denotes the expected score of the forecast CDF F when the verifying observations y are realizations of a random variable Y with CDF G, and $\forall$ means "for all." The score is *strictly proper* if the equality holds only if $F = G$ (Gneiting and Raftery 2007). Using only proper scoring rules is important in practice because the above inequality implies that a forecaster who knows the true distribution G has no incentive to predict any $F \neq G$, and is encouraged to quote her true belief. It has been demonstrated that the use of improper scores can lead to misguided inferences about predictive performance (Gneiting 2011).

The notions and methods mentioned above refer to probabilistic forecasts of univariate quantities. In some applications, however, multivariate quantities are of interest where multivariate can either refer to several different weather variables, or to a single variable considered at different locations in space or points in time simultaneously. River basin streamflow forecasts, for example, rely heavily on the meteorological inputs, and the runoff of mountain streams in spring season depends on both temperature (because of its impact on the amount of meltwater) and precipitation amounts. It is therefore important to know if an observed temperature above the predictive mean is likely to be associated with observed precipitation amounts above the predictive mean. If there is a positive or negative association between those two variables it should be reflected by the joint probabilistic forecast. Moreover, simultaneous consideration of all locations in the river basin and several lead times may be required. A recent article by Wilks (2014) considers probabilistic forecasting of heat waves, which requires the simultaneous study of minimum temperature and dewpoint temperature at two consecutive days, and Feldmann et al. (2015) study statistical postprocessing methods that yield calibrated temperature forecasts simultaneously at several locations. A number of multivariate generalizations of the verification rank histogram have been proposed (Smith and Hansen 2004; Wilks 2004; Gneiting et al. 2008; Thorarinsdottir et al. 2015; Ziegel and Gneiting 2014) that are sensitive to misrepresentations of both univariate characteristics and correlations between the different components of the multivariate quantity under consideration.

As far as proper scoring rules are concerned, the forecast verification toolbox is still rather limited. On the one hand there is the energy score (ES) and generalizations of it (Gneiting and Raftery 2007):

$$S_{\mathrm{en}}(F, \mathbf{y}) = E_F\|\mathbf{X} - \mathbf{y}\| - \frac{1}{2} E_F\|\mathbf{X} - \mathbf{X}'\|,$$

where $\mathbf{X}$ and $\mathbf{X}'$ are independent random vectors that are distributed according to the multivariate CDF F and $\|\cdot\|$ is the Euclidean norm. The energy score has the appealing property that it generalizes the univariate continuous ranked probability score (CRPS; Hersbach 2000) and is readily applicable also to ensemble forecasts. It has been pointed out, however, that this score is often not sufficiently sensitive to misspecifications of the correlations between the different components (Pinson and Girard 2012; Pinson and Tastu 2013). This is a big drawback since unlike the means and variances those correlations cannot be studied by applying univariate scores to the individual components. On the other hand, there are scoring rules [e.g., the logarithmic score by Roulston and Smith (2002), applied to a multivariate probability density function] that are more sensitive to misspecified correlations, but require that the forecast is given in terms of a predictive density, and are thus not applicable in the important case of ensemble forecasts. Dawid and Sebastiani (1999) proposed some multivariate scoring rules that depend only on the mean vector $\boldsymbol{\mu}_F$ and the covariance matrix $\boldsymbol{\Sigma}_F$ of the predictive distribution F. A particularly appealing example is the scoring rule [hereafter referred to as the Dawid–Sebastiani score (DSS)]:

$$S_{\mathrm{DS}}(F, \mathbf{y}) = \log \det \boldsymbol{\Sigma}_F + (\mathbf{y} - \boldsymbol{\mu}_F)' \boldsymbol{\Sigma}_F^{-1} (\mathbf{y} - \boldsymbol{\mu}_F).$$

It is equivalent to the logarithmic score for multivariate Gaussian predictive distributions and remains a proper (though not strictly proper) score relative to the larger class probability distributions for which the second moments of all components are finite (Gneiting and Raftery 2007). In principle this score could be applied to empirical versions of $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$ that were estimated from an ensemble, but unless the sample size is much larger than the dimension of the multivariate quantity, sampling errors can have disastrous effects on the calculation of $\det \boldsymbol{\Sigma}_F$ and $\boldsymbol{\Sigma}_F^{-1}$, and render this score useless in the context of ensemble forecasting (see e.g., Table 2 in Feldmann et al. 2015). Accordingly, in section 2 we propose a new, proper, multivariate score that is based on pairwise differences between all components of the multivariate quantity and that we hypothesize is more readily usable for ensemble forecast diagnosis. Some simulation examples are presented in section 3. These

will demonstrate that this new score is sensitive to misspecified correlations between the different components, and that it is useful for ensemble forecast diagnosis even when the number of ensemble members is moderate. An application of the new score in the context of probabilistic wind speed forecasting at several locations in Colorado simultaneously is presented in section 4, before we conclude with a short discussion in section 5.

2. A scoring rule based on pairwise differences

The basic idea of the class of multivariate scoring rules proposed in the following is to consider pairwise differences of the components of the multivariate quantity of interest. This has already been suggested in the context of rank histograms (e.g., Fig. 5 in Hamill 2001) and recently been utilized by Feldmann et al. (2015) in a diagnostic plot to check the adequacy of a statistical model for spatial correlations. Denote by $\mathbf{y}$ the vector of observations, by y_i its ith component, and assume that $\mathbf{y}$ is a realization of the random vector $\mathbf{Y}$. Adopting the concept of a *variogram* (also referred to as *structure function*) from geostatistics we study the quantity

$$\gamma_2(i, j) = \frac{1}{2} E|Y_i - Y_j|^2,$$

where E denotes the expectation under the (multivariate) distribution of $\mathbf{Y}$, which is assumed to have finite second moments. Denoting $\mu_i := E(Y_i)$, $\sigma_i^2 := \mathrm{var}(Y_i)$ and $\rho_{ij} := \mathrm{corr}(Y_i, Y_j)$ we have

$$E|Y_i - Y_j|^2 = (\mu_i - \mu_j)^2 + (\sigma_i^2 - 2\sigma_i\sigma_j\rho_{ij} + \sigma_j^2), \qquad (2)$$

which shows that γ_2 depends not only on the first and second moments of the individual components, but also on their correlations. More generally, one can consider variograms of order $p > 0$:

$$\gamma_p(i, j) = \frac{1}{2} E|Y_i - Y_j|^p.$$

The special cases $p = 1$ and $p = 0.5$ are known as *madogram* and *rodogram*, respectively (Bruno and Raspa 1989; Emery 2005). Variograms of order p can be defined for any multivariate distribution for which the pth absolute moments exist. For $p \neq 2$ and non-Gaussian distributions they can usually not be expressed as simple functions of the means, variances, and correlations of Y_i and Y_j, but they still depend on all of those quantities, and are therefore potentially useful for comparing the multivariate dependence structure of forecasts and observations. While condensing the information about the dependence of Y_i and Y_j into a single number $\gamma_p(i, j)$ implies a certain loss of information, we shall see that utilizing these quantities in the framework of scoring rules results in a performance measure that is sensitive to various types of miscalibration of multivariate forecasts. For a given d-variate observation vector $\mathbf{y}$ and forecast distribution F we define the *variogram score of order p* (VS-p):

$$S_{\gamma_p}(F, \mathbf{y}) = \sum_{i,j=1}^{d} w_{ij}(|y_i - y_j|^p - E_F|X_i - X_j|^p)^2, \qquad (3)$$

where X_i and X_j are the ith and the jth component of a random vector $\mathbf{X}$ that is distributed according to F, and w_{ij} are nonnegative weights. The score S_{γ_p} measures the dissimilarity between approximations of the variograms of order p of observations and forecasts over all pairs of components of the quantity of interest. For the observations, our best guess of $E|Y_i - Y_j|^p$ is simply the powered absolute difference of y_i and y_j. When the forecast distribution is given in the form of an ensemble $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, the forecast variogram $E_F|X_i - X_j|^p$ can be approximated by

$$E_F|X_i - X_j|^p \approx \frac{1}{m}\sum_{k=1}^{m} |x_i^{(k)} - x_j^{(k)}|^p, \quad i, j = 1, \ldots, d. \qquad (4)$$

Pairs of squared variogram differences can be emphasized or downweighted through the choice of the weights. This might be motivated by a subjective decision of an expert to put focus on certain component combinations. In a spatial context, for example, the possibility of emphasizing differences corresponding to pairs of locations that are either close by or a certain distance apart is related to the idea of scale-dependent verification (e.g., Jung and Leutbecher 2008). Downweighting certain pairs can also help mitigating the effects of sampling error. To see this, assume for simplicity that the random vector $\mathbf{Y}$ follows a multivariate Gaussian distribution with identical mean in all components. Defining $\sigma_{ij}^2 := \sigma_i^2 - 2\sigma_i\sigma_j\rho_{ij} + \sigma_j^2$ we then have

$$E|Y_i - Y_j|^1 = \sqrt{\frac{2}{\pi}}\sigma_{ij}, \quad \mathrm{var}|Y_i - Y_j|^1 = \left(1 - \frac{2}{\pi}\right)\sigma_{ij}^2$$

$$E|Y_i - Y_j|^2 = \sigma_{ij}^2, \qquad \mathrm{var}|Y_i - Y_j|^2 = 2\sigma_{ij}^4.$$

This shows that in both cases, both magnitude and variability of pairs of weakly correlated components are higher than for strongly correlated components. The former would therefore dominate the VS-p on the one hand, and introduce more variability on the other hand, which implies that downweighting pairs that are expected to have relatively weak correlations can benefit

the signal-to-noise ratio. In situations where there is some notion of distance between the ith and jth component (e.g., time lag as in the examples in section 3 or spatial distance as in section 4), correlations at short distances are typically stronger than those at longer distances. As a pragmatic ad hoc choice of the weights we then suggest to let them be proportional to the inverse distances between the corresponding components. This idea of downweighting certain pairs of components is conceptually similar to covariance localization in data assimilation (Houtekamer and Mitchell 2001; Hamill et al. 2001), where elements in the empirical covariance matrix that correspond to conceivably weakly or uncorrelated components are tapered down toward zero to reduce the effects of sampling error. When the multivariate quantity consists of variables of different type (e.g., temperature, pressure, and relative humidity), there is no obvious notion of distance and even the definition of S_{γ_p} seems doubtful as we would be subtracting quantities with potentially different units. In that situation, one could apply S_{γ_p} to standardized components:

$$\tilde{y}_i := \frac{y_i - \mu_i^{(\mathrm{cl})}}{\sigma_i^{(\mathrm{cl})}}, \quad \tilde{X}_i := \frac{X_i - \mu_i^{(\mathrm{cl})}}{\sigma_i^{(\mathrm{cl})}}, \quad i = 1, \ldots, d,$$

where $\mu_i^{(\mathrm{cl})}$ and $\sigma_i^{(\mathrm{cl})}$ are the climatological mean and variance of the variables, respectively. This approach has been suggested in multivariate geostatistics in the context of variance-based cross variograms, which are the equivalent of our score in the situation where components can correspond to different variables. In the geostatistical context it can be justified by the fact that predictors derived from variance-based cross variograms do not depend on the particular unit, and so the user should work with standardized variables in order to minimize the effects of sampling error (Cressie and Wikle 1998). In some applications there might be better, more problem-specific meteorological concepts to transform weather variables of different type in a way that brings them all to a scale in which they can be compared, one example being the total-energy norm (e.g., Hamill et al. 2003).

We now show that S_{γ_p} is proper relative to the class of the probability distributions for which the $(2p)$th moments of all components are finite. To see this, consider first a single pair (i, j). For any such pair, the mean of the random variable $Z := |Y_i - Y_j|^p$ minimizes the expected squared deviation of Z from any fixed number $a \in \mathbb{R}$, that is,

$$E[Z - E(Z)]^2 \le E(Z - a)^2.$$

This means that the inequality in (1) holds separately for any pair (i, j), but then it also holds for the weighted sum over all pairs, for any choice of nonnegative weights. Note, however, that the VS-p is not strictly proper because it only depends on the pth absolute moment of the distribution of component differences, and can therefore not distinguish between distributions of Z that have the same pth absolute moment but different higher moments. Moreover, large-scale random errors that are the same for all components cancel out when differences are considered; likewise, a bias that is the same for all components will go undetected. The simulation study in section 3 shows, however, that for suitable choices of p the VS-p is quite sensitive to misspecifications of the correlation structure of $\mathbf{Y}$. More importantly, this is still true when $E_F|X_i - X_j|^p$ has to be estimated as in (4) from an ensemble that represents the predictive distribution F. This approximation introduces quite a bit of additional sampling error, but the effects on the score's propriety and discrimination ability will be shown to be much less severe as for the Dawid–Sebastiani score. This makes the VS-p a favorable score in the context of ensemble forecasting, on which we focus in the rest of this paper.

Before comparing it with the ES and DSS in simulations, we shall mention that the VS-p can be viewed as a special case of a much larger class of scoring rules. Consider the mapping $g_{p,\tilde{\mathbf{w}}}: \mathbb{R}^d \to \mathbb{R}^{d^2}$ defined by

$$[g_{p,\tilde{\mathbf{w}}}(\mathbf{y})]_{ij} = \tilde{w}_{ij}|y_i - y_j|^p, \quad i, j = 1, \ldots, d,$$

where $\tilde{\mathbf{w}}$ is the weight vector of the transformation $g_{p,\tilde{\mathbf{w}}}$. Choosing $\tilde{w}_{ij} = \sqrt{w_{ij}}$, we can rewrite the VS-$p$ from (3) as

$$S_{\gamma_p}(F, \mathbf{y}) = \sum_{i,j=1}^{d} \{[g_{p,\tilde{\mathbf{w}}}(\mathbf{y})]_{ij} - E_F[g_{p,\tilde{\mathbf{w}}}(\mathbf{X})]_{ij}\}^2,$$

which shows that the VS-p of a single, multivariate forecast is (up to the factor $1/d^2$) the same as the mean squared error (MSE) over the d^2 components of the transformed forecast vector. The generalization of the VS-p is now obvious: instead of the MSE, we can apply any other univariate scoring rule to the components of $g_{p,\tilde{\mathbf{w}}}(\mathbf{X})$ and $g_{p,\tilde{\mathbf{w}}}(\mathbf{y})$, and take the mean over the resulting d^2 values as an alternative score for our multivariate quantity. Or, we can apply the ES to the d^2-variate vectors $g_{p,\tilde{\mathbf{w}}}(\mathbf{X})$ and $g_{p,\tilde{\mathbf{w}}}(\mathbf{y})$, rather than to $\mathbf{X}$ and $\mathbf{y}$ directly. These generalizations will also be studied in the subsequent section.

3. Simulation study

We compare the energy score, the Dawid–Sebastiani score, and the variogram score of order $p = 0.5$, 1, and 2, using inverse distance weights as described above. In all

experiments we generate $n = 5000$ observation vectors of dimension d, and an m-member ensemble of forecast vectors of the same dimension with both correct and misspecified means, variances, or correlations. To understand the impact of representing the predictive distribution by an ensemble on the different scores, we consider both small ($m = 20$) and medium-sized ($m = 100$) ensembles. While a formal definition of being *proper* exists and allows one to check this property mathematically, there does not seem to be a commonly accepted measure of a scoring rule's ability to discriminate between calibrated and uncalibrated forecasts. This is an important characteristic though that determines its utility for forecast verification in practice. In this simulation study, we try to get some sense of the discrimination ability of the various scores by repeating each experiment 10 times and visualizing the respective outcomes by boxplots. Even though the scores are averaged over 5000 cases, they still vary from one experiment to another. If the group of average scores obtained with calibrated forecasts is clearly separated from the one obtained with uncalibrated forecasts, we will interpret this as good discrimination ability of the scoring rule that was utilized. Conversely, if there is a strong overlap of the ranges of outcomes obtained with calibrated and uncalibrated forecasts, we will conclude that the scoring rule that produced these outcomes cannot reliably detect this particular type of miscalibration.

a. Miscalibrated marginal distributions

Although we contend that multivariate verification should focus on the correlations between the different components (predictive means and variances can be compared in a first step with univariate verification techniques), we shall start with a first experiment that compares the different scores with respect to their ability to detect biases and over- or underdispersion of the forecasts. We already noted that the VS-p is unable to detect a bias that is the same for all components, but we can consider a situation where this simple type of bias has been removed while an erroneous trend is present in the forecast means. Specifically, let the observation vectors be realizations of a Gaussian random vector $\mathbf{Y}$ of dimension $d = 5$ with zero mean, unit variance, and correlation function:

$$\mathrm{corr}(Y_i, Y_j) = \exp\left(-\frac{|i-j|}{r}\right), \quad i, j = 1, \ldots, d. \quad (5)$$

In this experiment we take $r = 3$. If we associate each component with a time point, Y can be viewed as a short, stationary autoregressive AR(1) process. Note that the definitions of all scores studied here neither exploit nor rely on this property of stationarity. Moreover, since the scores are calculated separately for each of the 5000 cases and averaged only afterward, they can also be applied in situations where the distribution of the observation vector differs from one case to another. The possibility of exploiting preliminary knowledge about the multivariate dependence structure is further discussed in the second example below. To compare the sensitivity of the different scores to misspecifications of means and variances, we generate forecasts with the same exponential correlation function as above and

1) correct variances but biased means: $\mu_F = (-0.5, -0.25, 0, 0.25, 0.5)'$;
2) correct means and variances;
3) correct means but too large variances: $\sigma_i^2 = 1.5$, $i = 1, \ldots, 5$; and
4) correct means but too small variances: $\sigma_i^2 = 0.6667$, $i = 1, \ldots, 5$.

The corresponding boxplots are shown in Fig. 1. We note first of all that the influence of ensemble size is rather different from one score to another. For the ES, there is hardly any difference between $m = 20$ with $m = 100$. This can be an advantage if only an ensemble of very small size is available, but it also suggests that the ES cannot distinguish a very good representation of the predictive distribution F from a very sparse one. This is different for the VS-p values, which consistently improve with increasing ensemble size, thus showing that the finite sample representation of F does have a noticeable effect on the score. This sampling effect, however, does not change the qualitative conclusions about the predictive performance of the different forecasts (this is also true for the examples considered below). A really substantial change of the scores due to the different finite representations of the predictive distribution can be observed with the DSS (note the different scales for $m = 20$ and $m = 100$). The approximation of $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$ by empirical means and covariances estimated from the small ensemble is so poor that the resulting scores lead to false conclusions about predictive performance, favoring the overdispersive ensemble over the calibrated one. For the larger ensemble, this score bias due to insufficient representation of F plays a smaller role, and the DSS discriminates well between the correct and uncalibrated forecasts. The ES is very effective in detecting the erroneous linear trend corresponding to the forecasts simulated according to 1), but the separation between the calibrated and over/underdispersive forecasts is less distinct. Among the different VS-p studied here, the VS-p with $p = 0.5$ has clearly the best discrimination ability. It identifies the miscalibration of the mean less clearly than the ES, but is more effective in detecting over and underdispersiveness. The VS-p with $p = 1$ still detects all types of miscalibration reasonably well. It is

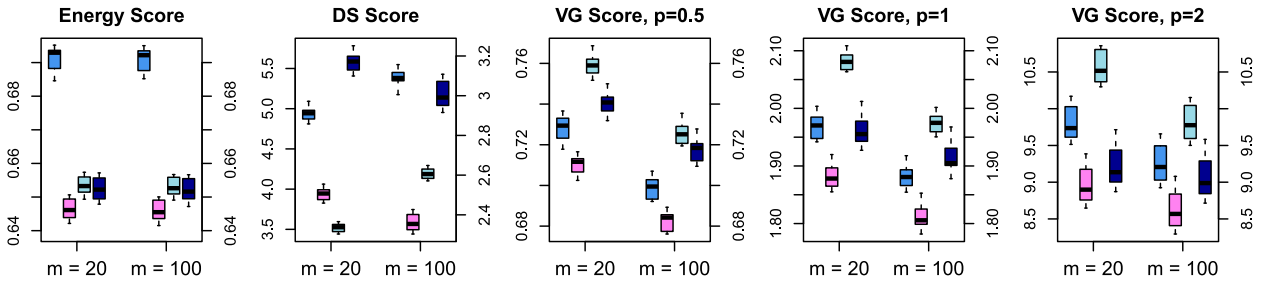

FIG. 1. (from left to right) Energy, Dawid–Sebastiani, and variogram scores of order $p = 0.5$, 1, and 2 for ensemble size $m = 20$ and $m = 100$. The boxplots corresponding to mean-biased, correct, over-, and underdispersive forecasts are cornflower blue, violet, light blue, and dark blue, respectively. The boxes cover the first to third quartile of the 10 outcomes, the black line shows the median, and the whiskers extend to the data extremes.

noticeable, however, that with increasing p the random variations between scores obtained with identical setups become larger and larger and blur the systematic differences between calibrated and uncalibrated forecasts. Before we turn to the genuinely multivariate aspects we would like to recall that the VS-p is not *strictly* proper. In the present situation, for example, the effects of an erroneous trend and underdispersion can cancel out [for $p = 2$ this can be seen directly from (2)]. We therefore emphasize again that an analysis of the marginal distributions by means of univariate scores should precede the study of multivariate properties.

b. *Misspecified correlation strength*

In our second experiment we focus on the correlation structure of the multivariate quantity under consideration. We study the ability of the different scores to detect whether the correlations between the different components of the forecast vectors are too weak, adequate, or too strong compared to the corresponding correlations of the observation vectors. Moreover, we study the effect of increasing the dimension from $d = 5$ to $d = 15$ on the different scores. In both cases, we consider again a zero mean, unit variance AR(1) process with correlation function given in (5). For the observation vectors, we choose $r = 3$ as before and compare ensemble forecasts simulated with the same correlation model, but $r = 2$, $r = 3$, and $r = 4.5$. The boxplots in Fig. 2 for the ES confirm the conclusion of Pinson and Tastu (2013) that the ES can hardly discriminate multivariate forecasts that differ only with respect to their correlations between individual components. For the DSS the conclusion is as in the first experiment. It discriminates well between calibrated and uncalibrated forecasts if the ensemble that represents the predictive distribution is sufficiently large. A small ensemble, however, results in an inaccurate approximation of $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$, and the corresponding DSS leads to misguided inference. This representation issue is much less severe for the VS-p, and for $p = 0.5$ and $p = 1$ it discriminates well between correct and incorrect correlation strengths. For $p = 2$ the discrimination ability is still better than for the ES but overall not very satisfactory with random differences between identical setups having the same magnitude as systematic score differences due to miscalibration. Increasing the dimension from $d = 5$ to $d = 15$ has a slightly negative effect on the discrimination ability of the VS-p. This may be somewhat surprising since a larger dimension entails more data that are used for the calculation of S_{γ_p}. However, since our definition of the VS-p in (3) does not make any assumption (e.g., stationarity in a time series or spatial context) about the correlation structure of forecasts and observations, increasing the number of summands in (3) does *not* lead to an averaging of sampling error. If one was absolutely sure that some additional structural assumption is justified [i.e., that the set of all pairs (i, j) can be represented as a union of disjoint subsets $I_1, \ldots, I_N$ such that the component differences corresponding to the pairs in each subset have the same pth absolute moment], one could replace definition (3) by

$$S_{\gamma_p}(F, \mathbf{y}) := \sum_{k=1}^{N} w_k \left[\sum_{(i,j)\in I_k} |y_i - y_j|^p - \sum_{(i,j)\in I_k} E_F |X_i - X_j|^p \right]^2 .$$

This way, additional structural information could be exploited and an increase of d would then likely reduce sampling error and improve the discrimination ability of the score. In the present example, the simulated AR(1) process is stationary and proceeding as described above with $I_k := [(i, j): |i - j| = k]$ would be justified. In general, however, such information is not available, and while simplifying assumptions are common and appropriate in

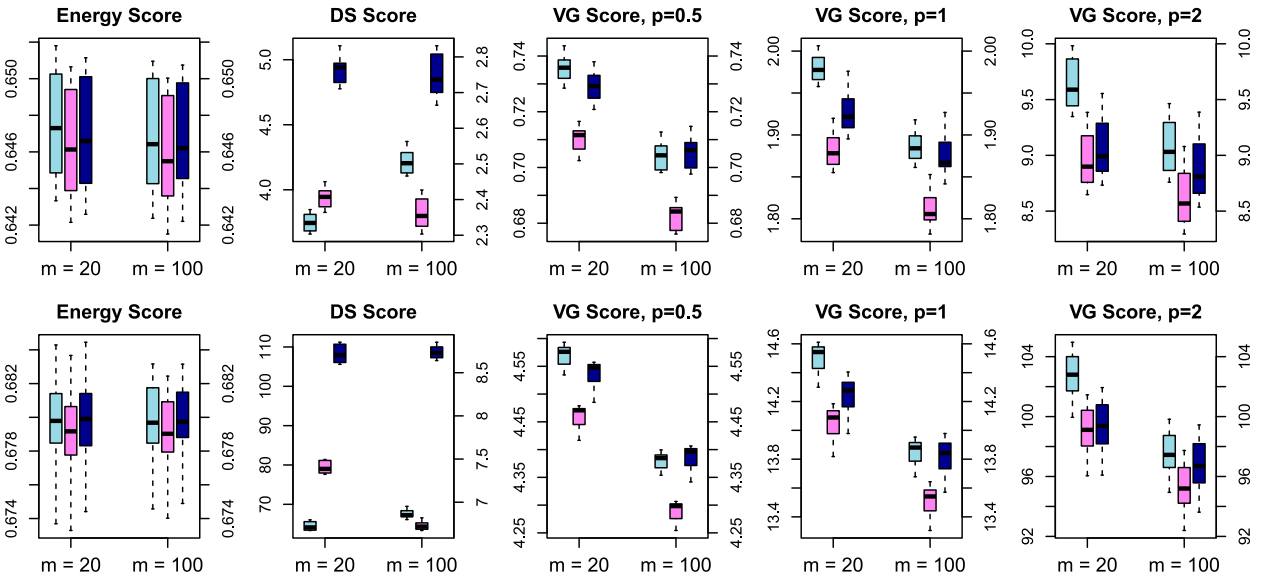

FIG. 2. As in Fig. 1, but for forecasts with too weak (light blue), adequate (violet), and too strong (dark blue) correlations compared to the observations for (top) $d = 5$ and (bottom) $d = 15$.

statistical modeling, we contend that verification methods should avoid unwarranted preliminary assumptions about forecasts and observations as far as possible. We therefore recommend retaining the definition in (3), even though it is less favorable with respect to the VS-p's discrimination ability. The fact that the discrimination ability in the present example even gets slightly worse from $d = 5$ to $d = 15$ can probably be explained by the fact that the fraction of pairs of components in $S_{\gamma_p}(F, \mathbf{y})$ with rather weak correlations increases, and thus more variability is introduced into the calculation of the score.

c. *Misspecified correlation model*

In the third experiment, we vary the entire correlation model rather than just the correlation strength. We now consider only the case $d = 15$ and simulate observations with zero mean, unit variance, and correlation function:

$$\text{(i)}\quad \operatorname{corr}(Y_i, Y_j) = \left(1 + \frac{|i-j|}{3}\right)^{-1}, \quad \text{and}$$

$$\text{(ii)}\quad \operatorname{corr}(Y_i, Y_j) = \exp\left(-\frac{|i-j|}{4}\right)\left[0.75 + 0.25\cos\left(\frac{|i-j|\pi}{2}\right)\right].$$

Both of them yield correlations at lag 1 that are very similar to the exponential model in (5) with $r = 3$. Model (i), however, has much stronger correlations at larger lags, and model (ii) has a periodic component that makes it oscillate around this exponential reference model. Can the VS-p detect those differences between the model in (5) and models (i) and (ii), respectively, even though our proposed weighting scheme down-weights larger lags? Figure 3 confirms many of the conclusions from the preceding experiment. The ES again lacks sensitivity to misspecifications of the correlation structure while the VS-p distinguishes much better between the correct and the incorrect correlation model. Again, however, the discrimination ability depends on p, with smaller values yielding significantly better results. The DSS has similar issues in this example as in those discussed above. Their magnitude drops dramatically when passing from 20 to 100 ensemble members, although the underlying multivariate distribution is the same. In the case where the observations have long-range dependence, both ensemble sizes are insufficient to reduce this score's representation bias enough to yield the proper ranking between correct and incorrect forecasts. In the example with the oscillating correlation model, the DSS yields the correct ranking and separates the two cases very well. However, it may well be that this is simply an example where the bias due to the finite representation of the predictive distribution favors the correct ranking by chance.

d. *Misspecified generating process*

When we introduced the VS-p in section 2, we emphasized that this family of scoring rules is proper, but not strictly proper. It is based only on the pth absolute moment of differences between all pairs of components. It is clear that biases that are the same for all components cancel out. It is also clear that certain combinations of

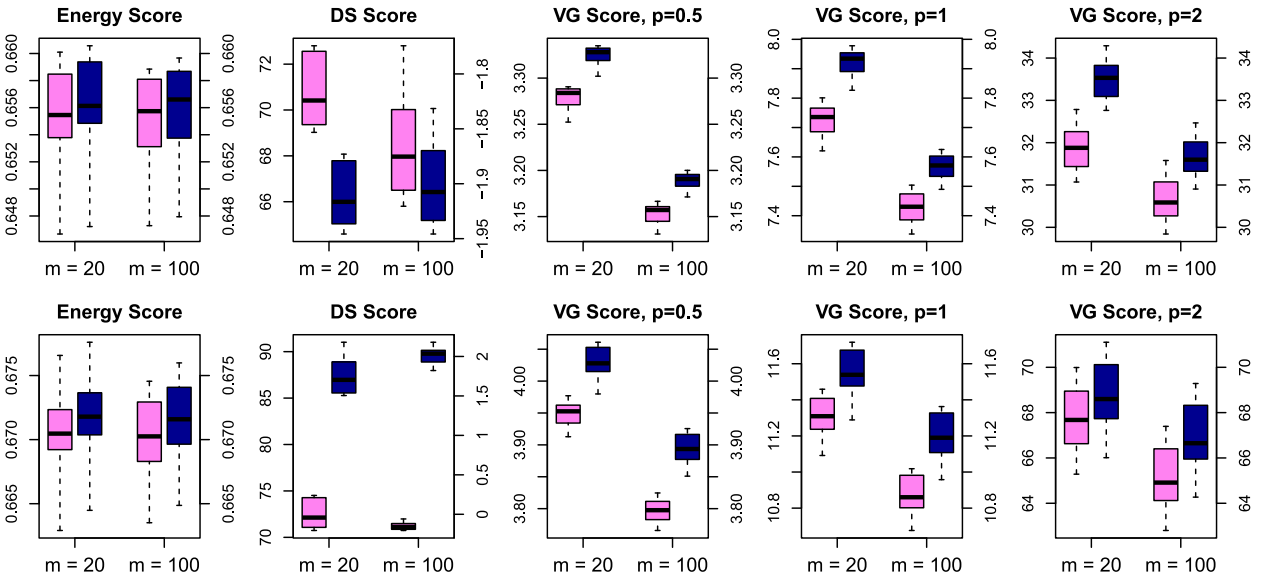

FIG. 3. As in Fig. 1, but for forecasts with (left violet boxplots) correct and (right dark blue boxplots) incorrect correlation structure where the correct correlation function is that using (top) model (i) or (bottom) model (ii) in section 3c ; and the incorrect correlation function is, in both cases, the exponential model in (5) with $r = 3$.

misspecifications (e.g., overestimation of correlation strength and overestimation of marginal variances) can partially or fully cancel out. But even if it has been assured that the marginal distributions are calibrated, the pth absolute moment of component differences does in general not fully characterize the multivariate dependence. How good is the VS-p in distinguishing forecasts that are entirely correct (i.e., have been generated by the same process as the observations) from forecasts that have correct means, variances, *and* correlations, but have been generated by a completely different mechanism? It can be expected that the answer depends on the particular generating process, and we are careful to make general claims as to this issue. Yet it is instructive to study at least one such example. We simulate observations as follows:

1) Draw a random number v from a Poisson distribution with parameter $\lambda = 8$.
2) Draw v locations $t_1, \ldots, t_v$ from a uniform distribution on the interval [0, 16].
3) Denoting by $(\cdot)_+$ the maximum of 0 and the function in brackets, define

$$y_t = \sqrt{\frac{15}{8}} \sum_{i=1}^{v} [1 - (t - t_i)^2]_+, \quad t = 1, \ldots, 15. \qquad (6)$$

One can think of $t_1, \ldots, t_v$ as storm centers that have an influence on all locations within a radius of one unit, expressed by the influence function $(1 - x^2)_+$. The different local storms are then added up to the final outcome. This process is a special case of a so-called *shot noise process*. Using results from Matérn (1986, chapter 3.3), one can show that with the specific choices made above $\mathbf{y}$ is a sample of a stationary time series with mean $\sqrt{5/3}$, variance 1, and correlation function:

$$\mathrm{corr}(Y_i, Y_j) = \left(1 + \frac{3|i-j|}{2} + \frac{|i-j|^2}{4}\right)\left(1 - \frac{|i-j|}{2}\right)_+^3.$$

We now compare forecasts that were generated in the same way as this shot noise observation process with forecasts that have the same means, variances, and correlations, but were simulated from a multivariate Gaussian distribution. An illustration of one sample path, respectively, on the full interval [1, 15] is provided in the supplemental material to this paper. The results of this comparison are depicted in Fig. 4. A few conclusions are very consistent with what we already observed before. The discrimination ability of the ES is rather poor, and the DSS favors the incorrect model as a result of insufficient approximation of $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$, even in the case where $m = 100$. Recall that the DSS depends on the predictive distribution only through its component means and variances, and intercomponent correlations, so for a perfect approximation of $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$ we would expect the DSS to be indifferent toward the particular forecast generation process. The same is true for the VS-2, while the effect of the generation process on the VS-1 and VS-0.5 is not quite as obvious. For the first time, we observe problems related to the finite sample representation of the predictive distribution also with

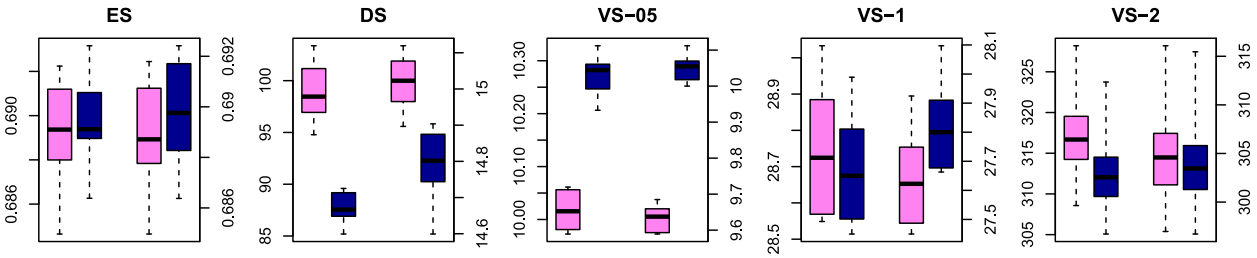

FIG. 4. As in Fig. 1, but for forecasts of (left violet boxplots) correct (shot noise) type and (right dark blue boxplots) incorrect (Gaussian) process type.

the VS-2 and VS-1. The good discrimination ability of the VS-0.5 may be based on several factors. On the one hand, the 0.5th absolute moment of differences seems to be very informative about the generating process. It is not clear though, whether this is specific to the present example or whether this is true in general. On the other hand, we have already observed that the choice $p = 0.5$ entails less sampling variability compared to larger values, and this likely contributes to the favorable performance of the VS-0.5 in the present example as well.

e. *Sensitivity of the variogram score of order p to the choice of weights*

So far, we have always chosen the weights in (3) proportional to the inverse distance between the components. We have argued in section 2 that such a choice is reasonable whenever there is some natural notion of distance, and correlations between components are expected to decrease with this distance. Yet, this choice is quite ad hoc, and it is natural to ask how sensitive the discrimination ability of the VS-p is with regard to the choice of weights, and if other choices yield a similar or even better performance. To answer this question, we repeat the first two experiments, this time considering only the case where $d = 15$ and $m = 20$. We restrict our attention to the VS-0.5, but study two alternative weighting schemes: no weighting at all (i.e., $w_{ij} \equiv 1$) and a kind of localization scheme where $w_{ij} = [1 - (|i - j|/3)^2]_+$ (i.e., pairs of components more than three units apart are not considered at all). The results in Fig. 5 are as one might have expected. Misspecifying the range parameter in our exponential correlation model in (5) affects correlations between all pairs of components. As pointed out in section 2, close by, strongly correlated components have a more favorable signal to noise ratio, and so it is not surprising that the localization weighting scheme has the best, and the unweighted VS-0.5 has the worst discrimination ability. The same conclusion holds in the experiment where the correlation function of the observations has a periodic component. Even at short lags, this correlation functions differs quite strongly from the simple exponential model, and focusing on close-by component pairs, therefore, benefits the score's discrimination ability. Differences between the long-range correlation model and the exponential model, on the contrary, are more noticeable for pairs of components that are farther apart, and hence the unweighted VS-0.5 performs best. Overall, we conclude that if prior knowledge about correlations is available, some sort of localization scheme with appropriately chosen cutoff radius should be used. In the absence of such knowledge, the inverse distance weighting scheme seems to be a good compromise. We

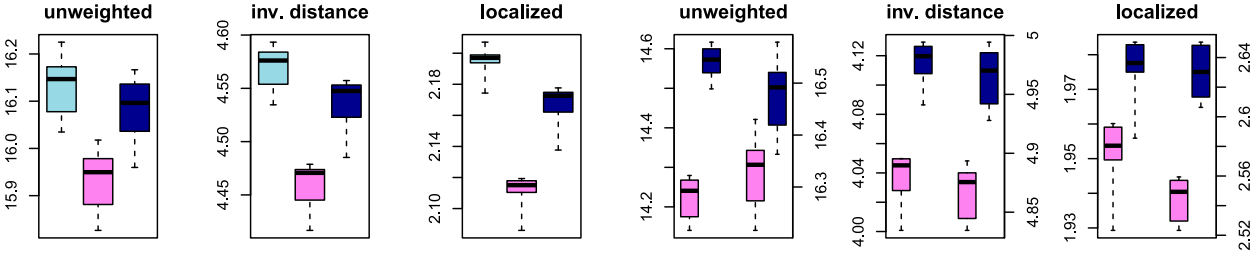

FIG. 5. VS-0.5 for three different component weights—unweighted, inverse distance, and localized—for the case where $d = 15$ and $m = 20$. The three plots on the left are for the correlation strength experiment with too weak, adequate, and too strong correlations being represented by light blue, violet, and dark blue boxplots, respectively, and those on the right are for the correlation model experiment with correct forecasts being represented by violet and incorrect forecasts being represented by dark blue boxplots.

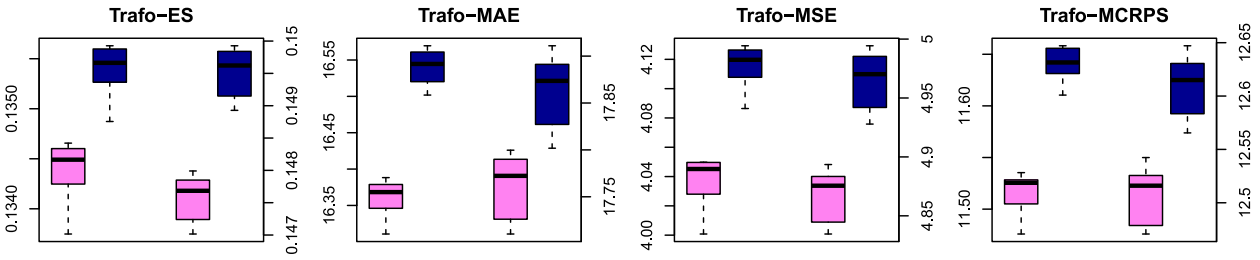

FIG. 6. Different scores that result from applying the ES, MAE, MSE, and MCRPS to the $g_{0.5,\tilde{\mathbf{w}}}$-transformed forecast and observation vectors. The two left and right boxplots within each panel correspond to the experiments where the observation is generated according to section 3c models (i) and (ii), respectively, with correct forecasts colored in violet and incorrect forecasts colored in dark blue.

finally note that even the unweighted score permits better identification of misspecified dependence structures than the ES.

f. Generalizations of the variogram score of order p

At the end of section 2, we pointed out that the VS-p defined in (3) can be viewed as a special case of a larger class of scoring rules which transforms both forecast and observation vectors to d^2-dimensional vectors of weighted, powered, absolute differences between the components of the original vectors. Here, we fix $p = 0.5$ and define the weight vector $\tilde{\mathbf{w}}$ of the transformation $g_{0.5,\tilde{\mathbf{w}}}$ through $\tilde{w}_{ij} = 1/\sqrt{|i-j|}$. With these choices, the VS-0.5 with inverse distance weights is (up to a constant factor) the same as the MSE of the componentwise means of the transformed forecasts with respect to the transformed observations. As alternative scores, we consider the mean absolute error (MAE) of the componentwise medians of the transformed forecasts, the mean continuous ranked probability score (MCRPS) over all components of the transformed forecasts, and the ES of the vector of transformed forecast. Figure 6 shows results for the setting of our correlation model experiment in section 3c with $d = 15$ and $m = 20$, where the observation is generated according to section 3c model (i) and (ii), respectively, and the scores are used to distinguish correct forecasts from those that erroneously use an exponential correlation model. The main point to note is that all scores are able to distinguish the correct from the incorrect correlation model, showing that it is really the transformation $g_{0.5,\tilde{\mathbf{w}}}$, rather than the particular score applied to the transformed vectors, that is crucial for detecting misspecified dependence structures. With the MAE and MCRPS being particular discriminative in the example with long-range dependence and the ES faring best in the example with a periodic component, there is no clear ranking among the different scores. The MSE, the score that corresponds to the VS-0.5, demonstrates good discrimination ability in both examples. Its preference over the other options is by no means imperative, but it seems to be a good compromise, and thus a reasonable standard choice.

4. Evaluating multisite wind speed forecasts

We finally apply our score in a data example to evaluate and compare statistically calibrated, probabilistic forecasts of wind speeds at five major wind park locations in the state of Colorado. Specifically, we consider the period from 1 January to 31 December 2013, use 80-m wind speed forecasts from the second-generation GEFS reforecast dataset (Hamill et al. 2013) and the corresponding reanalyses for both calibration and verification. The reforecast ensemble has 11 members and was initialized once daily at 0000 UTC. We study 80-m wind speed predictions with lead times of 24, 48, and 72 h at the grid points that are closest to

- Cedar Point Wind Farm (250 MW, operational since 2011);
- Cedar Creek Wind Farms I and II (550 MW, operational since 2007/10);
- Peetz Table Wind Energy Center (430 MW, operational since 2001/07);
- Colorado Green Wind Farm (162 MW, operational since 2003); and
- Cheyenne Ridge Wind Project (under development, project size 300–600 MW).

As explained above, the ensemble forecasts $f_{1s}, \ldots, f_{11s}$, $s \in \mathscr{S}$, where $\mathscr{S}$ denotes the set of the five wind park locations, can be interpreted as a sample from the multivariate distribution that describes the simultaneous predictions. The raw model output, however, often suffers from systematic biases and typically fails to fully represent prediction uncertainty (Hamill and Colucci 1997). To calibrate the marginal predictive

TABLE 1. Skill scores of the ECC-Q, random-Q, and ordered-Q ensembles with respect to the raw ensemble.

	Lead time 24 h			Lead time 48 h			Lead time 72 h		
	ESS	VSS-0.5	VSS-1	ESS	VSS-0.5	VSS-1	ESS	VSS-0.5	VSS-1
ECC-Q	0.184	0.171	0.151	0.119	0.119	0.096	0.063	0.036	0.027
Random-Q	0.175	0.047	0.088	0.108	−0.020	−0.017	0.051	−0.087	−0.063
Ordered-Q	−0.284	−0.147	−0.062	−0.420	−0.231	−0.145	−0.493	−0.461	−0.299

distributions, we follow Thorarinsdottir and Gneiting (2010) and fit a heteroscedastic regression model to past forecast–observation pairs that turns the ensemble mean $\overline{f}_s$ and the ensemble variance S_s^2 at location s into a predictive truncated normal distribution,

$$Y_s \,|\, f_{1s}, \ldots, f_{11s} \sim \mathcal{N}_0(a_s + b_s\overline{f}_s, c_s + d_s S_s^2), \quad (7)$$

for the observed wind speed Y_s at s. A separate model is fitted for each location, each forecast lead time, and each month of the verification period from 1 January to 31 December 2013. For each month, forecasts and observations from the same, the preceding, and the subsequent month in the years 2010, 2011, and 2012 are used as training data for the model fitting procedure [for details about that procedure we refer to Thorarinsdottir and Gneiting (2010)]. Once the parameters a_s, b_s, c_s, and d_s for each month, location, and lead time are determined, a predictive distribution for the day under consideration can be obtained by plugging the corresponding ensemble mean and variance into (7). Diagnostic plots (not shown here) confirm that the univariate probabilistic forecasts obtained in this way are calibrated (i.e., they are unbiased and represent the prediction uncertainty adequately).

The postprocessing scheme just described only addresses the marginal distributions. In our particular example, however, power network operators might be interested in whether low wind speeds (and hence low wind power production) at one wind park will be compensated by higher wind speeds at the other wind parks, or whether wind speeds will be low at all wind parks simultaneously. To account for this multivariate aspect of our prediction problem and address correlations between the forecasts at the different locations, we use the ensemble copula coupling (ECC) technique (Schefzik et al. 2013), which turns the five marginal predictive distributions back into an ensemble $\tilde{f}_{1s}, \ldots, \tilde{f}_{11s}$, $s \in \mathcal{S}$ that has the same rank correlation structure as the original ensemble but calibrated margins. Specifically, if F_s denotes the predictive, truncated normal CDF at location s, calibrated ensemble forecasts are obtained via

$$\tilde{f}_{1s} = F_s^{-1}\left(\frac{\rho_s(1)}{12}\right), \ldots, \tilde{f}_{11s} = F_s^{-1}\left(\frac{\rho_s(11)}{12}\right), \quad s \in \mathcal{S}, \quad (8)$$

where F_s^{-1} is the predictive quantile function at s and $\rho_s(k) = \text{rank}(f_{ks})$, $k = 1, \ldots, 11$. With other words, the original forecasts are replaced by quantiles (this particular way of sampling is referred to as ECC-Q) of the calibrated marginal distributions in such a way that the ordering of the ensemble member forecasts remains unchanged. In this way, the (flow dependent) rank correlation information of the raw GEFS ensemble is preserved.

Does this preservation of rank correlations really yield noticeably better multivariate forecasts than a sampling scheme in which ρ_s is a random perturbation of the set $\{1, \ldots, 11\}$ (i.e., no spatial correlations) or one in which ρ_s is the identity (i.e., maximal spatial dependence)? We compute those alternative, marginally calibrated ensembles (''random-Q,'' ''ordered-Q'') as well and use the ES, the VS-0.5, and the VS-1 to evaluate and compare the corresponding multivariate wind speed forecasts with those of the raw and ECC-Q ensemble. Again, we use inverse distance weights for the VS-p where distance is now the geographical distance (in kilometers) between the wind farm locations. Since in section 3 the empirical DSS turned out to be unreliable for small ensemble sizes and the VS-2 was always less discriminative than the VS-0.5 and VS-1, only the two latter are considered here as alternatives to the ES. In order to facilitate the comparison between the three different scores, we turn them into *skill* scores with respect to the raw ensemble. That is, instead of the energy score ES_* for method ''*'' we state the energy skill score $\text{ESS}_* = 1 - \text{ES}_*/\text{ES}_{\text{ens}}$, which measures the increase in predictive performance compared to the raw ensemble (likewise for the variogram scores). All skill scores in Table 1 agree that ECC-Q yields the most skillful, multivariate probabilistic forecasts. The ordered-Q ensemble, for which wind speeds are simultaneously low or high at all locations, is less skillful than the uncalibrated ensemble; the corresponding multivariate structure is clearly inappropriate. The comparison between ECC-Q and random-Q is more interesting, and confirms the above findings about the respective sensitivity of the ES and the VS-p to miscalibration. The ESS yields a somewhat clearer distinction between the raw and the ECC-Q ensemble, which differ in their marginal distribution, but have the same rank correlations. The random-Q

ensemble, however, scores almost as well as the ECC-Q ensemble, despite its doubtful assumption of spatial independence. Under the VS-0.5 and VS-1, on the contrary, the random-Q ensemble fares distinctly worse than the ECC-Q ensemble, and has even negative skill for lead times larger than 24 h. Those two ensembles yield identical forecasts at each location individually, but their components have different rank correlations. Again, the VS-p can detect those differences more clearly.

5. Discussion

In their recent review on probabilistic forecasting, Gneiting and Katzfuss (2014, p. 146) note as one out of eight key issues for future research that

> "There is a pressing need for the development of decision-theoretically principled methods for the evaluation of probabilistic forecasts of multivariate variables."

When the focus is on the correlation structure and the mean and covariance matrix of the predictive distribution are given in closed form, the DSS is an excellent choice. The examples in section 3 show, however, that the usage of this score can be problematic when the probabilistic forecasts are represented by an ensemble of limited size, and empirical versions of the predictive mean vector and covariance matrix have to be used. In spite of being proper, the DSS can then lead to entirely wrong conclusions about predictive performance, which suggests that this scoring rule is far from being *fair* in the sense of Fricker et al. (2013). In this paper, we have presented a new class of multivariate scores based on powered differences between pairs of components of the multivariate quantity, denoted as variogram scores of order p (VS-p). In our simulation studies the VS-p was also negatively affected by the sampling error due to representing the predictive distribution by a (possibly small) ensemble. In the majority of cases, however, it led to the correct conclusions about predictive performance, which suggests that it is much closer to being *fair* than the DSS. Moreover, it is more successful than the ES in distinguishing forecasts with different correlation structures. Three different choices of powers p were studied for the VS-p, and it was found that the best results are obtained with $p = 0.5$, while $p = 2$ was clearly suboptimal. Would a VS-p with $p < 0.5$ have even better properties? At least for Gaussian predictive distributions, a square root transformation is likely already the best choice since the distribution of $|X_i - X_j|^{0.5}$ is almost perfectly symmetric and thus has much better sampling properties than the strongly skewed distribution that comes with the choice $p = 2$ (Cressie and Hawkins 1980). If the predictive distribution itself is already skewed, however, then smaller powers may indeed be favorable to obtain a near-symmetric distribution of $|X_i - X_j|^p$.

In section 4, we considered a data example with statistically postprocessed wind speed forecasts. Scoring rules in general, and the VS-p in particular, may however also be useful diagnostic tools in the development process of ensemble prediction systems. In the context of data assimilation, for example, it is important that the ensemble adequately represents the variances and covariances between different variables at different locations. Comparing different ensembles via scoring rules rather than empirical covariances (averaged over a certain time period) has the advantage that the former evaluate every time point separately and average the scores rather than covariances. This is more adequate if those covariances are flow dependent. Moreover, if the scores are normalized in a reasonable way (for the VS-p this could be done by requiring that the weights sum to one on each day), even the space dimension may change over time, and averaging the corresponding scores would still make sense. If the distribution of the observation errors is known, those can be taken into account by simulating a sample $\epsilon_{il}, i = 1, \ldots, d,\ l = 1, \ldots, M$ of such errors and adding them to the ensemble forecasts. The empirical version of the VS-p then becomes

$$S_{\gamma_p}(F, \mathbf{y}) = \sum_{i,j=1}^{d} w_{ij} \left(|y_i - y_j|^p - \frac{1}{mM} \sum_{k=1}^{m} \sum_{l=1}^{M} |f_i^{(k)} + \epsilon_{il} - f_j^{(k)} - \epsilon_{jl}|^p \right)^2,$$

and by choosing M—the number of simulated observation error vectors—large enough, one can reduce at least part of the additional variability that is introduced into the score. It remains to be seen if the signal-to-noise ratio in those applications is large enough for this score to be still sufficiently discriminative.

We think that the class of VS-p proposed here is a useful contribution to address the above-mentioned research issue of decision theoretically principled methods for multivariate forecast evaluation. It has certain limitations, resulting from the fact that is not *strictly* proper as discussed in section 2. Given the

strong increase in the number of degrees of freedom with the dimension of the quantity to be forecast, it is unlikely, however, that there exists a single multivariate score that serves all purposes. We strongly recommend that several different scores be always considered before drawing conclusions. Some of the limitations of the VS-*p* can be addressed by studying the ES (which is more sensitive to misspecifications of the predictive mean and less affected by the finite representation of the predictive distribution) or univariate scores for the marginal distributions alongside with our VS-*p*. Focusing on differences between components is probably the most natural, but by no means the only possible transformation of the multivariate quantity that leads to a multivariate score that is sensitive to correlations between components. In some applications, studying composite quantities like minima, maxima, or averages over several locations or lead times (Berrocal et al. 2007; Feldmann et al. 2015), or indexes that involve multiple quantities (Wilks 2014) is a natural way to turn multivariate quantities into univariate ones that can be evaluated by standard univariate scores. This way, specific (and practically relevant) aspects of the multivariate predictive distribution can be evaluated, and this sort of verification is another recommended supplement to general purpose multivariate scores like the ES or the VS-*p* presented here.

Acknowledgments. The authors thank Tilmann Gneiting, Martin Leutbecher, and two anonymous reviewers for useful discussions and comments on the manuscript. This research was performed while the first author held a National Research Council Research Associateship Award at NOAA's Earth System Research Laboratory. The publication was partially supported by a NOAA/Office of Weather and Air Quality (OWAQ) USWRP grant.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2007: Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Wea. Rev.,* **135,** 1386–1402, doi:10.1175/MWR3341.1.

Bruno, R., and G. Raspa, 1989: Geostatistical characterization of fractal models of surfaces. *Geostatistics,* M. Armstrong, Ed., Quantitative Geology and Geostatistics, Vol. 4, Springer, 77–89.

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.,* **133,** 1076–1097, doi:10.1175/MWR2905.1.

Cressie, N., and D. M. Hawkins, 1980: Robust estimation of the variogram I. *Math. Geol.,* **12,** 115–125, doi:10.1007/BF01035243.

——, and C. K. Wikle, 1998: The variance-based cross-variogram: You can add apples and oranges. *Math. Geol.,* **30,** 789–799, doi:10.1023/A:1021770324434.

Dawid, A. P., and P. Sebastiani, 1999: Coherent dispersion criteria for optimal experimental design. *Ann. Stat.,* **27,** 65–81, doi:10.1214/aos/1018031101.

Emery, X., 2005: Variograms of order ω: A tool to validate a bivariate distribution model. *Math. Geol.,* **37,** 163–181, doi:10.1007/s11004-005-1307-4.

Feldmann, K., M. Scheuerer, and T. L. Thorarinsdottir, 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.

Fricker, T. E., C. A. T. Ferro, and D. B. Stephenson, 2013: Three recommendations for evaluating climate predictions. *Meteor. Appl.,* **20,** 246–255, doi:10.1002/met.1409.

Gneiting, T., 2011: Making and evaluating point forecasts. *J. Amer. Stat. Assoc.,* **106,** 746–762, doi:10.1198/jasa.2011.r10138.

——, and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.,* **102,** 359–378, doi:10.1198/016214506000001437.

——, and M. Katzfuss, 2014: Probabilistic forecasting. *Ann. Rev. Stat. Appl.,* **1,** 125–151, doi:10.1146/annurev-statistics-062713-085831.

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

——, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson, 2008: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST,* **17,** 211–235, doi:10.1007/s11749-008-0114-x.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167, doi:10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327, doi:10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

——, J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.,* **129,** 2776–2790, doi:10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.

——, C. Snyder, and J. S. Whitaker, 2003: Ensemble forecasts and the properties of flow-dependent analysis-error covariance. *Mon. Wea. Rev.,* **131,** 1741–1758, doi:10.1175/2559.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.,* **94,** 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Houtekamer, P. L., and H. L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.,* **129,** 123–137, doi:10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.

Jung, T., and M. Leutbecher, 2008: Scale-dependent verification of ensemble forecasts. *Quart. J. Roy. Meteor. Soc.,* **134,** 973–984, doi:10.1002/qj.255.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.,* **227,** 3515–3539, doi:10.1016/j.jcp.2007.02.014.

Lewis, J. M., 2005: Roots of ensemble forecasting. *Mon. Wea. Rev.,* **133,** 1865–1885, doi:10.1175/MWR2949.1.

Matérn, B., 1986: *Spatial Variation.* Lecture Notes in Statistics, Vol. 36, 2nd ed. Springer-Verlag, 151 pp.

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

——, and J. Tastu, 2013: Discrimination ability of the energy score. Tech. Rep., Technical University of Denmark, 16 pp.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660, doi:10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

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

Smith, L. A., and J. A. Hansen, 2004: Extending the limits of ensemble forecast verification with the minimum spanning tree. *Mon. Wea. Rev.,* **132,** 1522–1528, doi:10.1175/1520-0493(2004)132<1522:ETLOEF>2.0.CO;2.

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

——, M. Scheuerer, and C. Heinz, 2015: Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Stat.,* doi:10.1080/10618600.2014.977447, in press.

Wilks, D. S., 2004: The minimum spanning tree histogram as verification tool for multidimensional ensemble forecasts. *Mon. Wea. Rev.,* **132,** 1329–1340, doi:10.1175/1520-0493(2004)132<1329:TMSTHA>2.0.CO;2.

——, 2014: Multivariate ensemble Model Output Statistics using empirical copulas. *Quart. J. Roy. Meteor. Soc.,* doi:10.1002/qj.2414, in press.

Ziegel, J. F., and T. Gneiting, 2014: Copula calibration. *Electron. J. Stat.,* **8,** 2619–2638, doi:10.1214/14-EJS964.