

ABSTRACT

9 Proper scoring rules provide a theoretically principled framework for the
10 quantitative assessment of the predictive performance of probabilistic fore-
11 casts. While a wide selection of such scoring rules for univariate quantities
12 exists, there are only few scoring rules for multivariate quantities, and many
13 of them require that forecasts are given in the form of a probability density
14 function. The energy score, a multivariate generalization of the continuous
15 ranked probability score, is the only commonly used score that is applicable
16 in the important case of ensemble forecasts, where the multivariate predic-
17 tive distribution is represented by a finite sample. Unfortunately, its ability to
18 detect incorrectly specified correlations between the components of the mul-
19 tivariate quantity is somewhat limited. In this paper we present an alternative
20 class of proper scoring rules based on the geostatistical concept of variograms.
21 We study their sensitivity to incorrectly predicted means, variances, and cor-
22 relations in a number of examples with simulated observations and forecasts,
23 and show that the variogram-based scoring rules are distinctly more discrim-
24 inative with respect to the correlation structure. This conclusion is confirmed
25 in a case study with post-processed wind speed forecasts at five wind park
26 locations in Colorado, U.S.A.

27 **1. Introduction**

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

35 Along with the availability of probabilistic forecasts comes the need for both diagnostic and
36 quantitative methods to assess the quality of those forecasts and to compare the performance of
37 competing forecasters. A probabilistic forecast should be calibrated, i.e. statistically consistent
38 with the values that materialize, and sharp, i.e. very specific about the anticipated weather (Gneit-
39 ing et al. 2007). Sharpness can be assessed via numerical and graphical summaries of the width
40 of the prediction intervals that come with a predictive probability distribution. The notion of
41 calibration is more complex, and different types of calibration have been established. Marginal
42 calibration measures the similarity of the aggregated predictive distribution and the climatological
43 distribution of the predictand, and can be checked by comparing the average predictive cumulative
44 distribution function (CDF) with the empirical CDF of the observations (Gneiting et al. 2007).
45 Probabilistic calibration concerns the dynamical aspects of probabilistic forecasts and can be as-
46 sessed by studying verification rank histograms (Anderson 1996; Hamill and Colucci 1997; Hamill
47 2001).

48 In order to make a quantitative comparison of different forecast methods, summary measures
49 of their predictive performance are required. Those measures should take both calibration and

50 sharpness into account. To this end, scoring rules have been proposed which assign a numerical
51 score $S(F, y)$ to each pair (F, y) where F is the CDF of the predictive distribution and y is the
52 realized value. If we take scoring rules to be negatively oriented, $S(F, y)$ can be viewed as a
53 penalty that the forecasters wish to minimize. A crucial property that one should always require
54 from a scoring rule is that it is *proper*, which is formally defined by the requirement

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

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

62 The notions and methods mentioned above refer to probabilistic forecasts of univariate quanti-
63 ties. In some applications, however, multivariate quantities are of interest where multivariate can
64 either refer to several different weather variables, or to a single variable considered at different
65 locations in space or points in time simultaneously. River basin streamflow forecasts, for example,
66 rely heavily on the meteorological inputs, and the runoff of mountain streams in spring season de-
67 pends on both temperature (because of its impact on the amount of melt water) and precipitation
68 amounts. It is therefore important to know if an observed temperature above the predictive mean
69 is likely to be associated with observed precipitation amounts above the predictive mean. If there
70 is a positive or negative association between those two variables it should be reflected by the joint
71 probabilistic forecast. Moreover, simultaneous consideration of all locations in the river basin

72 and several lead times may be required. A recent article by Wilks (2014) considers probabilistic
73 forecasting of heat waves, which requires the simultaneous study of minimum temperature and
74 dewpoint temperature at two consecutive days, and Feldmann et al. (2014) study statistical post-
75 processing models that yield calibrated temperature forecasts simultaneously at several locations.
76 A number of multivariate generalizations of the verification rank histogram have been proposed
77 (Smith and Hansen 2004; Wilks 2004; Gneiting et al. 2008; Thorarinsdottir et al. 2014; Ziegel
78 and Gneiting 2014) that are sensitive to misrepresentations of both univariate characteristics and
79 correlations between the different components of the multivariate quantity under consideration.

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

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

83 where \mathbf{X} and \mathbf{X}' are independent random vectors that are distributed according to the multivariate
84 CDF F and $\|\cdot\|$ is the Euclidean norm. The energy score has the appealing property that it gen-
85 eralizes the univariate continuous ranked probability score (CRPS, Hersbach 2000) and is readily
86 applicable also to ensemble forecasts. It has been pointed out, however, that this score is often
87 not sufficiently sensitive to misspecifications of the correlations between the different components
88 (Pinson and Girard 2012; Pinson and Tastu 2013). This is a big drawback since unlike the means
89 and variances those correlations cannot be studied by applying univariate scores to the individual
90 components. On the other hand, there are scoring rules (e.g. the logarithmic score by Roulston
91 and Smith 2002, applied to a multivariate probability density function) that are more sensitive to
92 misspecified correlations, but require that the forecast is given in terms of a predictive density, and
93 are thus not applicable in the important case of ensemble forecasts. Dawid and Sebastiani (1999)
94 proposed some multivariate scoring rules that depend only on the mean vector μ_F and the covari-

95 ance matrix Σ_F of the predictive distribution F . A particularly appealing example is the scoring
96 rule (hereafter referred to as the Dawid-Sebastiani score or DSS)

$$S_{DS}(F, \mathbf{y}) = -\log \det \Sigma_F - (\mathbf{y} - \mu_F)' \Sigma_F^{-1} (\mathbf{y} - \mu_F).$$

97 It is equivalent to the logarithmic score for multivariate Gaussian predictive distributions and re-
98 mains a proper (though not strictly proper) score relative to the larger class probability distributions
99 for which the second moments of all components are finite (Gneiting and Raftery 2007). In prin-
100 ciple this score could be applied to empirical versions of μ_F and Σ_F that were estimated from
101 an ensemble, but unless the sample size is much larger than the dimension of the multivariate
102 quantity, sampling errors can have disastrous effects on the calculation of $\det \Sigma_F$ and Σ_F^{-1} , and
103 render this score useless in the context of ensemble forecasting (see e.g. Table 2 in Feldmann et al.
104 2014). Accordingly, in Section 2 we propose a new, proper, multivariate score that is based on
105 pairwise differences between all components of the multivariate quantity and that we hypothesize
106 is more readily usable for ensemble forecast diagnosis. Some simulation examples are presented
107 in Section 3. These will demonstrate that this new score is sensitive to misspecified correlations
108 between the different components, and that it is useful for ensemble forecast diagnosis even when
109 the number of ensemble members is moderate. An application of the new score in the context of
110 probabilistic wind speed forecasting at several locations in Colorado (U.S.A.) simultaneously is
111 presented in Section 4, before we conclude with a short discussion in Section 5.

112 2. A scoring rule based on pairwise differences

113 The basic idea of the class of multivariate scoring rules proposed in the following is to consider
114 pairwise differences of the components of the multivariate quantity of interest. This has already
115 been suggested in the context of rank histograms (e.g. Fig. 5 in Hamill 2001) and recently been

116 utilized by Feldmann et al. (2014) in a diagnostic plot to check the adequacy of a statistical model
 117 for spatial correlations. Denote by \mathbf{y} the vector of observations, by y_i its i -th component, and
 118 assume that \mathbf{y} is a realization of the random vector \mathbf{Y} . Adopting the concept of a *variogram* (also
 119 referred to as *structure function*) from geostatistics we study the quantity

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

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

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

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

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

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

134 *variogram score of order p (VS- p)*

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

135 where X_i and X_j are the i -th and the j -th component of a random vector \mathbf{X} that is distributed ac-
 136 cording to F , and w_{ij} are non-negative weights. The score S_{γ_p} measures the dissimilarity between
 137 approximations of the variograms of order p of observations and forecasts over all pairs of com-
 138 ponents of the quantity of interest. For the observations, our best guess of $E|Y_i - Y_j|^p$ is simply the
 139 powered absolute difference of y_i and y_j . When the forecast distribution is given in the form of an
 140 ensemble $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}$, the forecast variogram $E_F |X_i - X_j|^p$ can be approximated by

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

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

$$\begin{aligned} p = 1 &\Rightarrow E|Y_i - Y_j|^p = \sqrt{\frac{2}{\pi}} \sigma_{ij}, & \text{var}|Y_i - Y_j|^p &= \left(1 - \frac{2}{\pi}\right) \sigma_{ij}^2 \\ p = 2 &\Rightarrow E|Y_i - Y_j|^p = \sigma_{ij}^2, & \text{var}|Y_i - Y_j|^p &= 2\sigma_{ij}^4 \end{aligned}$$

149 This shows that in both cases, both magnitude and variability of pairs of weakly correlated com-
 150 ponents are higher than for strongly correlated components. The former would therefore dominate
 151 the VS- p on the one hand, and introduce more variability on the other hand, which implies that

152 down-weighting pairs that are expected to have relatively weak correlations can benefit the sig-
 153 nal to noise ratio. In situations where there is some notion of distance between the i -th and j -th
 154 component (e.g. time lag as in the examples in Section 3 or spatial distance as in Section 4), cor-
 155 relations at short distances are typically stronger than those at longer distances. As a pragmatic
 156 ad hoc choice of the weights we then suggest to let them be proportional to the inverse distances
 157 between the corresponding components. This idea of down-weighting certain pairs of components
 158 is conceptually similar to covariance localization in data assimilation (Houtekamer and Mitchell
 159 2001; Hamill et al. 2001), where elements in the empirical covariance matrix that correspond to
 160 conceivably weakly or uncorrelated components are tapered down towards zero to reduce the ef-
 161 fects of sampling error. When the multivariate quantity consists of variables of different type (e.g.
 162 temperature, pressure, and relative humidity), there is no obvious notion of distance and even the
 163 definition of S_{γ_p} seems doubtful as we would be subtracting quantities with potentially different
 164 units. In that situation, one could apply S_{γ_p} to standardized components

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

165 where $\mu_i^{(cl)}$ and $\sigma_i^{(cl)}$ are the climatological mean and variance of the respective variables. This
 166 approach has been suggested in multivariate geostatistics in the context of variance-based cross-
 167 variograms, which are the equivalent of our score in the situation where components can corre-
 168 spond to different variables. In the geostatistical context it can be justified by the fact that pre-
 169 dictors derived from variance-based cross-variograms do not depend on the particular unit, and
 170 so the user should work with standardized variables in order to minimize the effects of sampling
 171 error (Cressie and Wikle 1998). In some applications there might be better, more problem-specific
 172 meteorological concepts to transform weather variables of different type in a way that brings them

173 all to a scale in which they can be compared, one example being the total-energy norm (e.g. Hamill
174 et al. 2003).

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

$$E(Z - E(Z))^2 \leq E(Z - a)^2.$$

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

193 Before comparing it with the ES and DSS in simulations, we shall mention that the VS- p can
194 be viewed as a special case of a much larger class of scoring rules. Consider the mapping $g_{p, \tilde{w}} :$

195 $\mathbb{R}^d \rightarrow \mathbb{R}^{d^2}$ defined by

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

196 Choosing $\tilde{w}_{ij} = \sqrt{w_{ij}}$, we can rewrite the VS- p from eq. (3) as

$$S_{\gamma_p}(F, \mathbf{y}) = \sum_{i,j=1}^d \left((g_{p,\tilde{\mathbf{w}}}(\mathbf{y}))_{ij} - \mathbb{E}_F(g_{p,\tilde{\mathbf{w}}}(\mathbf{X}))_{ij} \right)^2,$$

197 which shows that the VS- p of a single, multivariate forecast is (up to the factor $1/d^2$) the same
 198 as the mean squared error (MSE) over the d^2 components of the transformed forecast vector. The
 199 generalization of the VS- p is now obvious: instead of the MSE, we can apply any other univariate
 200 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
 201 values as an alternative score for our multivariate quantity. Or, we can apply the ES to the d^2 -
 202 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
 203 also be studied in the subsequent section.

204 3. Simulation study

205 We compare the energy score, the Dawid-Sebastiani score, and the variogram score of order
 206 $p = 0.5, 1$, and 2, and inverse distance weights as described above. In all experiments we generate
 207 $n = 5000$ observation vectors of dimension d , and an m -member ensemble of forecast vectors
 208 of the same dimension with both correct and misspecified means, variances or correlations. To
 209 understand the impact of representing the predictive distribution by an ensemble on the different
 210 scores, we consider both small ($m = 20$) and medium-sized ($m = 100$) ensembles. While a formal
 211 definition of being *proper* exists and allows one to check this property mathematically, there does
 212 not seem to be a commonly accepted measure of a scoring rule's ability to discriminate between
 213 calibrated and uncalibrated forecasts. This is an important characteristic though, that determines
 214 its utility for forecast verification in practice. In this simulation study, we try to get some sense

215 of the discrimination ability of the various scores by repeating each experiment ten times and
 216 visualizing the respective outcomes by boxplots. Even though the scores are averaged over 5000
 217 cases, they still vary from one experiment to another. If the group of average scores obtained with
 218 calibrated forecasts is clearly separated from the one obtained with uncalibrated forecasts, we
 219 will interpret this as good discrimination ability of the scoring rule that was utilized. Conversely,
 220 if there is a strong overlap of the ranges of outcomes obtained with calibrated and uncalibrated
 221 forecasts, we will conclude that the scoring rule that produced these outcomes cannot reliably
 222 detect this particular type of miscalibration.

223 *Miscalibrated marginal distributions*

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

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

233 In this experiment we take $r = 3$. If we associate each component with a time point, Y can be
 234 viewed as a short, stationary AR(1) process. Note that the definitions of all scores studied here
 235 neither exploit nor rely on this property of stationarity. Moreover, since the scores are calculated
 236 separately for each of the 5000 cases and averaged only afterwards, they can also be applied in

237 situations where the distribution of the observation vector differs from one case to another. The
238 possibility of exploiting preliminary knowledge about the multivariate dependence structure is
239 further discussed in the second example below. To compare the sensitivity of the different scores
240 to misspecifications of means and variances, we generate forecasts with the same exponential
241 correlation function as above and

242 a) correct variances but biased means $\mu_F = (-0.5, -0.25, 0, 0.25, 0.5)'$

243 b) correct means and variances

244 c) correct means but too large variances $\sigma_i^2 = 1.5, i = 1, \dots, 5$

245 d) correct means but too small variances $\sigma_i^2 = 0.6667, i = 1, \dots, 5$

246 The corresponding boxplots are shown in Fig. 1. We note first of all that the influence of en-
247 semble size is rather different from one score to another. For the ES, there is hardly any difference
248 between $m = 20$ with $m = 100$. This can be an advantage if only an ensemble of very small size
249 is available, but it also suggests that the ES cannot distinguish a very good representation of the
250 predictive distribution F from a very sparse one. This is different for the VS- p 's, which consis-
251 tently improve with increasing ensemble size, thus showing that the finite sample representation
252 of F does have a noticeable effect on the score. This sampling effect, however, does not change
253 the qualitative conclusions about the predictive performance of the different forecasts (this is also
254 true for the examples considered below). A really substantial change of the scores due to the
255 different finite representations of the predictive distribution can be observed with the DSS (note
256 the different scales for $m = 20$ and $m = 100$). The approximation of μ_F and Σ_F by empirical
257 means and covariances estimated from the small ensemble is so poor that the resulting scores lead
258 to false conclusions about predictive performance, favoring the over-dispersive ensemble over the
259 calibrated one. For the larger ensemble, this score bias due to insufficient representation of F plays

260 a smaller role, and the DSS discriminates well between the correct and uncalibrated forecasts. The
261 ES is very effective in detecting the erroneous linear trend corresponding to the forecasts simulated
262 according to a), but the separation between the calibrated and over-/under-dispersive forecasts is
263 less distinct. Among the different VS- p studied here, the VS- p with $p = 0.5$ has clearly the best
264 discrimination ability. It identifies the miscalibration of the mean less clearly than the ES, but is
265 more effective in detecting over- and underdispersiveness. The VS- p with $p = 1$ still detects all
266 types of miscalibration reasonably well. It is noticeable, however, that with increasing p the ran-
267 dom variations between scores obtained with identical setups become larger and larger and blur
268 the systematic differences between calibrated and uncalibrated forecasts. Before we turn to the
269 genuinely multivariate aspects we would like to recall that the VS- p is not *strictly* proper. In the
270 present situation, for example, the effects of an erroneous trend and underdispersion can cancel out
271 (for $p = 2$ this can be seen directly from eq. (2)). We therefore emphasize again that an analysis of
272 the marginal distributions by means of univariate scores should precede the study of multivariate
273 properties.

274 *Misspecified correlation strength*

275 In our second experiment we focus on the correlation structure of the multivariate quantity un-
276 der consideration. We study the ability of the different scores to detect whether the correlations
277 between the different components of the forecast vectors are too weak, adequate, or too strong
278 compared to the corresponding correlations of the observation vectors. Moreover, we study the
279 effect of increasing the dimension from $d = 5$ to $d = 15$ on the different scores. In both cases, we
280 consider again a zero mean, unit variance AR(1) process with correlation function given in (5).
281 For the observation vectors we choose $r = 3$ as before and compare ensemble forecasts simulated
282 with the same correlation model but $r = 2, r = 3$, and $r = 4.5$. The boxplots in Fig. 2 for the

283 ES confirm the conclusion of Pinson and Tastu (2013) that the ES can hardly discriminate multi-
 284 variate forecasts that differ only with respect to their correlations between individual components.
 285 For the DSS the conclusion is as in the first experiment. It discriminates well between calibrated
 286 and uncalibrated forecasts if the ensemble that represents the predictive distribution is sufficiently
 287 large. A small ensemble, however, results in an inaccurate approximation of μ_F and Σ_F , and the
 288 corresponding DSS leads to misguided inference. This representation issue is much less severe
 289 for the VS- p , and for $p = 0.5$ and $p = 1$ it discriminates well between correct and incorrect corre-
 290 lation strengths. For $p = 2$ the discrimination ability is still better than for the ES but overall not
 291 very satisfactory with random differences between identical setups having the same magnitude as
 292 systematic score differences due to miscalibration. Increasing the dimension from $d = 5$ to $d = 15$
 293 has a slightly negative effect on the discrimination ability of the VS- p . This may be somewhat
 294 surprising since a larger dimension entails more data that are used for the calculation of S_{γ_p} . How-
 295 ever, since our definition of the VS- p in eq. (3) does not make any assumption (e.g. stationarity
 296 in a time series or spatial context) about the correlation structure of forecasts and observations,
 297 increasing the number of summands in (3) does *not* lead to an averaging of sampling error. If
 298 one was absolutely sure that some additional structural assumption is justified, i.e. that the set of
 299 all pairs (i, j) can be represented as a union of disjoint subsets I_1, \dots, I_N such that the component
 300 differences corresponding to the pairs in each subset have the same p -th absolute moment, one
 301 could replace definition (3) by

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

302 This way, additional structural information could be exploited and an increase of d would then
 303 likely reduce sampling error and improve the discrimination ability of the score. In the present
 304 example, the simulated AR(1) process is stationary and proceeding as described above with $I_k :=$

305 $\{(i, j) : |i - j| = k\}$ would be justified. In general, however, such information is not available, and
 306 while simplifying assumptions are common and appropriate in statistical modeling, we contend
 307 that verification methods should avoid unwarranted preliminary assumptions about forecasts and
 308 observations as far as possible. We therefore recommend retaining the definition in eq. (3), even
 309 though it is less favorable with respect to the VS- p 's discrimination ability. The fact that the
 310 discrimination ability in the present example even gets slightly worse from $d = 5$ to $d = 15$ can
 311 probably be explained by the fact that the fraction of pairs of components in $S_{\gamma_p}(F, \mathbf{y})$ with rather
 312 weak correlations increases, and thus more variability is introduced into the calculation of the
 313 score.

314 *Misspecified correlation model*

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

$$318 \quad \text{i) } \text{corr}(Y_i, Y_j) = \left(1 + \frac{|i-j|}{3}\right)^{-1},$$

$$319 \quad \text{ii) } \text{corr}(Y_i, Y_j) = \exp\left(-\frac{|i-j|}{4}\right) \cdot \left(0.75 + 0.25 \cos\left(\frac{|i-j|\pi}{2}\right)\right).$$

320 Both of them yield correlations at lag 1 that are very similar to the exponential model (5) with
 321 $r = 3$. Model i), however, has much stronger correlations at larger lags, and model ii) has a periodic
 322 component that makes it oscillate around this exponential reference model. Can the VS- p detect
 323 those difference between model (5) and model i) and ii), respectively, even though our proposed
 324 weighting scheme down-weights larger lags? Figure 3 confirms many of the conclusions from
 325 the preceding experiment. The ES again lacks sensitivity to misspecifications of the correlation
 326 structure while the VS- p 's distinguish much better between the correct and the incorrect correla-

327 tion model. Again however, the discrimination ability depends on p , with smaller values yielding
328 significantly better results. The DSS has similar issues in this example as in those discussed above.
329 Their magnitude drops dramatically when passing from 20 to 100 ensemble members, although
330 the underlying multivariate distribution is the same. In the case where the observations have long
331 range dependence, both ensemble sizes are insufficient to reduce this score's representation bias
332 enough to yield the proper ranking between correct and incorrect forecasts. In the example with
333 the oscillating correlation model, the DSS yields the correct ranking and separates the two cases
334 very well. However, it may well be that this is simply an example where the bias due to the finite
335 representation of the predictive distribution favors the correct ranking by chance.

336 *Misspecified generating process*

337 When we introduced the VS- p in Section 2, we emphasized that this family of scoring rules is
338 proper, but not strictly proper. It is based only on the p -th absolute moment of differences between
339 all pairs of components. It is clear that biases that are the same for all components cancel out.
340 It is also clear that certain combinations of misspecifications (e.g. overestimation of correlation
341 strength and overestimation of marginal variances) can partially or fully cancel out. But even if
342 it has been assured that the marginal distributions are calibrated, the p -th absolute moment of
343 component differences does in general not fully characterize the multivariate dependence. How
344 good is the VS- p in distinguishing forecasts that are entirely correct (i.e. have been generated
345 by the same process as the observations) from forecasts that have correct means, variances, *and*
346 correlations, but have been generated by a completely different mechanism? It can be expected
347 that the answer depends on the particular generating process, and we are careful to make general
348 claims as to this issue. Yet it is instructive to study at least one such example. We simulate
349 observations as follows:

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

$$y_t = \sqrt{\frac{15}{8}} \cdot \sum_{i=1}^v (1 - (t - t_i)^2)_+, \quad t = 1, \dots, 15 \quad (6)$$

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

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

358 We now compare forecasts that were generated in the same way as this shot noise observation
 359 process with forecasts that have the same means, variances, and correlations, but were simulated
 360 from a multivariate Gaussian distribution. An illustration of one sample path, respectively, on
 361 the full interval $[1, 15]$ is provided in the supplemental material to this paper. The results of this
 362 comparison are depicted in Figure 4. A few conclusions are very consistent with what we already
 363 observed before. The discrimination ability of the ES is rather poor, and the DSS favors the
 364 incorrect model as a result of insufficient approximation of μ_F and Σ_F , even in the case where
 365 $m = 100$. Recall that the DSS depends on the predictive distribution only through its component
 366 means and variances, and inter-component correlations, so for a perfect approximation of μ_F and
 367 Σ_F we would expect the DSS to be indifferent towards the particular forecast generation process.
 368 The same is true for the VS-2, while the effect of the generation process on the VS-1 and VS-
 369 0.5 is not quite as obvious. For the first time, we observe problems related to the finite sample

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

377 *Sensitivity of the variogram score of order p to the choice of weights*

378 So far, we have always chosen the weights in (3) proportional to the inverse distance between
379 the components. We have argued in Section 2 that such a choice is reasonable whenever there is
380 some natural notion of distance, and correlations between components are expected to decrease
381 with this distance. Yet, this choice is quite ad hoc, and it is natural to ask how sensitive the dis-
382 crimination ability of the VS- p is with regard to the choice of weights, and if other choices yield a
383 similar or even better performance. To answer this question, we repeat the first two experiments,
384 this time considering only the case where $d = 15$ and $m = 20$. We restrict our attention to the
385 VS-0.5 but study two alternative weighting schemes: no weighting at all (i.e. $w_{ij} \equiv 1$) and a kind
386 of localization scheme where $w_{ij} = \left(1 - \left(\frac{|i-j|}{3}\right)^2\right)_+$, i.e. pairs of components more than 3 units
387 apart are not considered at all. The results in Figure 5 are as one might have expected. Misspec-
388 ifying the range parameter in our exponential correlation model (5) affects correlations between
389 all pairs of components. As pointed out in Section 2, close-by, strongly correlated components
390 have a more favorable signal to noise ratio, and so it is not surprising that the localization weight-
391 ing scheme has the best, and the unweighted VS-0.5 has the worst discrimination ability. The
392 same conclusion holds in the experiment where the correlation function of the observations has

393 a periodic component. Even at short lags, this correlation functions differs quite strongly from
 394 the simple exponential model, and focusing on close-by component pairs therefore benefits the
 395 score’s discrimination ability. Differences between the long range correlation model and the expo-
 396 nential model, on the contrary, are more noticeable for pairs of components that are further apart,
 397 and hence the unweighted VS-0.5 performs best. Overall, we conclude that if prior knowledge
 398 about correlations is available, some sort of localization scheme with appropriately chosen cut-off
 399 radius should be used. In the absence of such knowledge, the inverse distance weighting scheme
 400 seems to be a good compromise. We finally note that even the unweighted score permits better
 401 identification of misspecified dependence structures than the ES.

402 *Generalizations of the variogram score of order p*

403 At the end of Section 2 we pointed out that the VS- p defined in eq. (3) can be viewed as a spe-
 404 cial case of a larger class of scoring rules which transforms both forecast and observation vectors
 405 to d^2 -dimensional vectors of weighted, powered, absolute differences between the components of
 406 the original vectors. Here, we fix $p = 0.5$ and define the weight vector $\tilde{\mathbf{w}}$ of the transformation
 407 $g_{0.5, \tilde{\mathbf{w}}}$ through $\tilde{w}_{ij} = 1/\sqrt{|i-j|}$. With these choices, the VS-0.5 with inverse distance weights is
 408 (up to a constant factor) the same as the mean squared error (MSE) of the componentwise means
 409 of the transformed forecasts with respect to the transformed observations. As alternative scores
 410 we consider the mean absolute error (MAE) of the componentwise medians of the transformed
 411 forecasts, the mean continuous ranked probability score (MCRPS) over all components of the
 412 transformed forecasts, and the ES of the vector of transformed forecast. Figure 6 shows results
 413 for the setting of our second experiment above with $d = 15$ and $m = 20$, where the observation
 414 is generated according to a correlation function with long range dependence, or a periodic com-
 415 ponent, respectively, and the scores are used to distinguish correct forecasts from those where an

416 exponential correlation model is used for the forecasts. The main point to note is that all scores
417 are able to distinguish the correct from the incorrect correlation model, showing that it is really
418 the transformation $g_{0.5, \tilde{w}}$, rather than the particular score applied to the transformed vectors, that
419 is crucial for detecting misspecified dependence structures. With the MAE and MCRPS being
420 particular discriminative in the example with long range dependence and the ES faring best in the
421 example with a periodic component, there is no clear ranking among the different scores. The
422 MSE, the score that corresponds to the VS-0.5, demonstrates good discrimination ability in both
423 examples. Its preference over the other options is by no means imperative, but it seems to be a
424 good compromise, and thus a reasonable standard choice.

425 **4. Evaluating multi-site wind speed forecasts**

426 We finally apply our score in a data example to evaluate and compare statistically calibrated,
427 probabilistic forecasts of wind speeds at five major wind park locations in the state of Colorado
428 (U.S.A.). Specifically, we consider the period from 1 January to 31 December 2013, use 80-m
429 wind-speed forecasts from the 2nd-generation GEFS reforecast data set (Hamill et al. 2013) and
430 the corresponding reanalyses for both calibration and verification. The reforecast ensemble has
431 11 members and was initialized once daily at 0000 UTC. We study 80-m wind-speed predictions
432 with lead times 24h, 48h, and 72h at the grid points that are closest to

- 433 • Cedar Point Wind Farm (250 MW, operational since 2011)
- 434 • Cedar Creek Wind Farms I and II (550 MW, operational since 2007/2010)
- 435 • Peetz Table Wind Energy Center (430 MW, operational since 2001/2007)
- 436 • Colorado Green Wind Farm (162 MW, operational since 2003)
- 437 • Cheyenne Ridge Wind Project (under development, project size 300-600 MW)

438 As explained above, the ensemble forecasts f_{1s}, \dots, f_{11s} , $s \in \mathcal{S}$, where \mathcal{S} denotes the set of
439 the five wind park locations, can be interpreted as a sample from the multivariate distribution
440 that describes the simultaneous predictions. The raw model output, however, often suffers from
441 systematic biases and typically fails to fully represent prediction uncertainty (Hamill and Colucci
442 1997). To calibrate the marginal predictive distributions, we follow Thorarinsdottir and Gneiting
443 (2010) and fit a heteroscedastic regression model to past forecast-observation pairs that turns the
444 ensemble mean \bar{f}_s and the ensemble variance S_s^2 at location s into a predictive truncated normal
445 distribution

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

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

456 The post-processing scheme just described only addresses the marginal distributions. In our
457 particular example, however, power network operators might be interested in whether low wind
458 speeds (and hence low wind power production) at one wind park will be compensated by higher
459 wind speeds at the other wind parks, or whether wind speeds will be low at all wind parks simul-
460 taneously. To account for this multivariate aspect of our prediction problem and address correla-

461 tions between the forecasts at the different locations, we use the ensemble copula coupling (ECC)
 462 technique (Scheffzik et al. 2013) which turns the 5 marginal predictive distributions back into an
 463 ensemble $\tilde{f}_{1s}, \dots, \tilde{f}_{11s}$, $s \in \mathcal{S}$ that has the same rank correlation structure as the original ensemble
 464 but calibrated margins. Specifically, if F_s denotes the predictive, truncated normal CDF at location
 465 s , calibrated ensemble forecasts are obtained via

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

466 where $\rho_s(k) = \text{rank}(f_{ks})$, $k = 1, \dots, 11$. With other words, the original forecasts are replaced by
 467 quantiles (this particular way of sampling is referred to as ECC-Q) of the calibrated marginal dis-
 468 tributions in such a way that the ordering of the ensemble member forecasts remains unchanged.
 469 In this way, the (flow-dependent) rank correlation information of the raw GEFS ensemble is pre-
 470 served.

471 Does this preservation of rank correlations really yield noticeably better multivariate forecasts
 472 than a sampling scheme in which ρ_s is a random perturbation of the set $\{1, \dots, 11\}$ (i.e. no spatial
 473 correlations) or one in which ρ_s is the identity (i.e. maximal spatial dependence)? We compute
 474 those alternative, marginally calibrated ensembles (“Random-Q”, “Ordered-Q”) as well, and use
 475 the ES, the VS-0.5, and the VS-1 to evaluate and compare the corresponding multivariate wind
 476 speed forecasts with those of the raw and ECC-Q ensemble. Again, we use inverse distance
 477 weights for the VS- p where distance is now the geographical distance (in km) between the wind
 478 farm locations. Since in Section 3 the empirical DSS turned out to be unreliable for small ensemble
 479 sizes and the VS-2 was always less discriminative than the VS-0.5 and VS-1, only the two latter
 480 are considered here as alternatives to the ES. In order to facilitate the comparison between the three
 481 different scores, we turn them into *skill scores* with respect to the raw ensemble. That is, instead of
 482 the energy score ES_* for method ‘*’ we state the energy skill score $ESS_* = 1 - ES_*/ES_{ens}$ which

483 measures the increase in predictive performance compared to the raw ensemble (likewise for the
484 variogram scores). All skill scores in Table 1 agree that ECC-Q yields the most skillful, multivari-
485 ate probabilistic forecasts. The Ordered-Q ensemble, for which wind speeds are simultaneously
486 low or high at all locations, is less skillful than the uncalibrated ensemble; the corresponding mul-
487 tivariate structure is clearly inappropriate. The comparison between ECC-Q and Random-Q is
488 more interesting, and confirms the above findings about the respective sensitivity of the ES and
489 the VS- p to miscalibration. The ESS yields a somewhat clearer distinction between the raw and
490 the ECC-Q ensemble, which differ in their marginal distribution but have the same rank correla-
491 tions. The Random-Q ensemble, however, scores almost as well as the ECC-Q ensemble, despite
492 its doubtful assumption of spatial independence. Under the VS-0.5 and VS-1, on the contrary,
493 the Random-Q ensemble fares distinctly worse than the ECC-Q ensemble, and has even negative
494 skill for lead times larger than 24h. Those two ensembles yield identical forecasts at each location
495 individually, but their components have different rank correlations. Again, the VS- p can detect
496 those differences more clearly.

497 **5. Discussion**

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

500 “There is a pressing need for the development of decision-theoretically principled
501 methods for the evaluation of probabilistic forecasts of multivariate variables.”

502 When the focus is on the correlation structure and the mean and covariance matrix of the predictive
503 distribution are given in closed form, the DSS is an excellent choice. The examples in Section 3
504 show, however, that the usage of this score can be problematic when the probabilistic forecasts are
505 represented by an ensemble of limited size, and empirical versions of the predictive mean vector

506 and covariance matrix have to be used. In spite of being proper, the DSS can then lead to entirely
507 wrong conclusions about predictive performance, which suggests that this scoring rule is far from
508 being *fair* in the sense of Fricker et al. (2013). In this paper, we have presented a new class of
509 multivariate scores based on powered differences between pairs of components of the multivariate
510 quantity, denoted as variogram scores of order p (VS- p). In our simulation studies the VS- p was
511 also negatively affected by the sampling error due to representing the predictive distribution by a
512 (possibly small) ensemble. In the majority of cases, however, it led to the correct conclusions about
513 predictive performance, which suggests that it is much closer to being *fair* than the DSS. Moreover,
514 it is more successful than the ES in distinguishing forecasts with different correlation structures.
515 Three different choices of powers p were studied for the VS- p , and it was found that the best results
516 are obtained with $p = 0.5$, while $p = 2$ was clearly suboptimal. Would a VS- p with $p < 0.5$ have
517 even better properties? At least for Gaussian predictive distributions, a square-root transformation
518 is likely already the best choice since the distribution of $|X_i - X_j|^{0.5}$ is almost perfectly symmetric
519 and thus has much better sampling properties than the strongly skewed distribution that comes with
520 the choice $p = 2$ (Cressie and Hawkins 1980). If the predictive distribution itself is already skewed,
521 however, then smaller powers may indeed be favorable to obtain a near symmetric distribution of
522 $|X_i - X_j|^p$.

523 In Section 4 we considered a data example with statistically post-processed wind speed forecasts.
524 Scoring rules in general, and the VS- p in particular, may however also be useful diagnostic tools
525 in the development process of ensemble prediction systems. In the context of data assimilation,
526 for example, it is important that the ensemble adequately represents the variances and covariances
527 between different variables at different locations. Comparing different ensembles via scoring rules
528 rather than empirical covariances (averaged over a certain time period) has the advantage that the
529 former evaluate every time point separately and average the scores rather than covariances. This

530 is more adequate if those covariances are flow-dependent. Moreover, if the scores are normalized
 531 in a reasonable way (for the VS- p this could be done by requiring that the weights sum to one
 532 on each day), even the space dimension may change over time, and averaging the corresponding
 533 scores would still make sense. If the distribution of the observation errors is known, those can be
 534 taken into account by simulating a sample $\varepsilon_{il}, i = 1, \dots, d, l = 1, \dots, M$ of such errors and adding
 535 them to the ensemble forecasts. The empirical version of the VS- p then becomes

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

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

540 We think that the class of VS- p proposed here is a useful contribution to address the above
 541 mentioned research issue of decision-theoretically principled methods for multivariate forecast
 542 evaluation. It has certain limitations, resulting from the fact that is not *strictly* proper as discussed
 543 in Section 2. Given the strong increase in the number of degrees of freedom with the dimen-
 544 sion of the quantity to be forecast it is unlikely, however, that there exists a single multi-variate
 545 score that serves all purposes. We strongly recommend to always consider several different scores
 546 before drawing conclusions. Some of the limitations of the VS- p can be addressed by studying
 547 the ES (which is more sensitive to misspecifications of the predictive mean and less affected by
 548 the finite representation of the predictive distribution) or univariate scores for the marginal distri-
 549 butions alongside with our VS- p . Focusing on differences between components is probably the
 550 most natural, but by no means the only possible transformation of the multivariate quantity that
 551 leads to a multivariate score that is sensitive to correlations between components. In some appli-

552 cations, studying composite quantities like minima, maxima, or averages over several locations or
553 lead times (Berrocal et al. 2007; Feldmann et al. 2014), or indexes that involve multiple quanti-
554 ties (Wilks 2014) is a natural way to turn multivariate quantities into univariate ones that can be
555 evaluated by standard univariate scores. This way, specific (and practically relevant) aspects of the
556 multivariate predictive distribution can be evaluated, and this sort of verification is another rec-
557 ommended supplement to general purpose multivariate scores like the ES or the VS- p presented
558 here.

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

564 **References**

565 Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensem-
566 ble model integrations. *J. Climate*, **9**, 1518–1530.

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

569 Bruno, R., and G. Raspa, 1989: Geostatistical characterization of fractal models of surfaces. *Geo-*
570 *statistics*, M. Armstrong, Ed., Quantitative Geology and Geostatistics, Vol. 4, Springer Nether-
571 lands, 77–89.

572 Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison
573 of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**,

574 1076–1097.

575 Cressie, N., and D. M. Hawkins, 1980: Robust estimation of the variogram I. *Math. Geol.*, **12**,
576 115–125.

577 Cressie, N., and C. K. Wikle, 1998: The variance-based cross-variogram: you can add apples and
578 oranges. *Math. Geol.*, **30** (7), 789–799.

579 Dawid, A. P., and P. Sebastiani, 1999: Coherent dispersion criteria for optimal experimental de-
580 sign. *Ann. Statist.*, **27**, 65–81.

581 Emery, X., 2005: Variograms of order ω : A tool to validate a bivariate distribution model. *Math.*
582 *Geol.*, **37**, 163–181.

583 Feldmann, K., M. Scheuerer, and T. L. Thorarinsdottir, 2014: Spatial postprocessing of ensemble
584 forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.*, doi:
585 10.1175/MWR-D-14-00210.1.

586 Fricker, T. E., C. A. T. Ferro, and D. B. Stephenson, 2013: Three recommendations for evaluating
587 climate predictions. *Meteor. Appl.*, **20**, 246–255.

588 Gneiting, T., 2011: Making and evaluating point forecasts. *J. Amer. Stat. Assoc.*, **106**, 746–762.

589 Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharp-
590 ness. *J. Roy. Stat. Soc. B*, **69**, 243–268.

591 Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Ann. Rev. Statist. Appl.*, **1**, 125–151.

592 Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J.*
593 *Amer. Stat. Assoc.*, **102**, 359–378.

- 594 Gneiting, T., L. I. Stanberry, E. P. Gritmit, L. Held, and N. A. Johnson, 2008: Assessing proba-
595 bilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface
596 winds (with discussion and rejoinder). *Test*, **17**, 211–264.
- 597 Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon.*
598 *Wea. Rev.*, **129**, 550–560.
- 599 Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. G. Jr., Y. Zhu, and
600 W. Lapenta, 2013: NOAA’s second-generation global medium-range ensemble reforecast data
601 set. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565.
- 602 Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts.
603 *Mon. Wea. Rev.*, **125**, 1312–1327.
- 604 Hamill, T. M., C. Snyder, and J. S. Whitaker, 2003: Ensemble forecasts and the properties of
605 flow-dependent analysis-error covariance. *Mon. Wea. Rev.*, **131**, 1741–1758.
- 606 Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background
607 error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790.
- 608 Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble pre-
609 diction systems. *Wea. Forecasting*, **15**, 559–570.
- 610 Houtekamer, P. L., and H. L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric
611 data assimilation. *Mon. Wea. Rev.*, **129**, 123–137.
- 612 Jung, T., and M. Leutbecher, 2008: Scale-dependent verification of ensemble forecasts. *Quart. J.*
613 *Roy. Meteor. Soc.*, **134**, 973–984.
- 614 Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539.

- 615 Lewis, J. M., 2005: Roots of ensemble forecasting. *Mon. Wea. Rev.*, **133**, 1865–1885.
- 616 Matérn, B., 1986: *Spatial variation*, Lecture Notes in Statistics, Vol. 36. 2nd ed., Springer-Verlag,
617 Berlin.
- 618 Pinson, P., and R. Girard, 2012: Evaluating the quality of scenarios of short-term wind power
619 generation. *Appl. Energ.*, **96**, 12–20.
- 620 Pinson, P., and J. Tastu, 2013: Discrimination ability of the energy score. Tech. rep., Technical
621 University of Denmark.
- 622 Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information
623 theory. *Mon. Wea. Rev.*, **130**, 1653–1660.
- 624 Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex
625 simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640.
- 626 Smith, L. A., and J. A. Hansen, 2004: Extending the limits of ensemble forecast verification with
627 the minimum spanning tree. *Mon. Wea. Rev.*, **132**, 1522–1528.
- 628 Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble
629 model output statistics using heteroskedastic censored regression. *J. Roy. Stat. Soc. A*, **173**,
630 371–388.
- 631 Thorarinsdottir, T. L., M. Scheuerer, and C. Heinz, 2014: Assessing the calibration of high-
632 dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Stat.*, doi:10.1080/
633 10618600.2014.977447.
- 634 Wilks, D. S., 2004: The minimum spanning tree histogram as verification tool for multidimen-
635 sional ensemble forecasts. *Mon. Wea. Rev.*, **132**, 1329–1340.

636 Wilks, D. S., 2014: Multivariate ensemble-MOS using empirical copulas. *Quart. J. Roy. Meteor.*
637 *Soc.*, doi:10.1002/qj.2414.

638 Ziegel, J. F., and T. Gneiting, 2014: Copula calibration. *Electron. J. Statist.*, **8 (2)**, 2619–2638.

639 **LIST OF TABLES**

640 **Table 1.** Skill scores of the ECC-Q, Random-Q, and Ordered-Q ensembles with respect
641 to the raw ensemble. 33

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

	lead time 24h			lead time 48h			lead time 72h		
	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

642 **LIST OF FIGURES**

643 **Fig. 1.** Scores for mean-biased, correct, overdispersive, and underdispersive forecasts (from left to
644 right) for ensemble size $m = 20$ and $m = 100$ 35

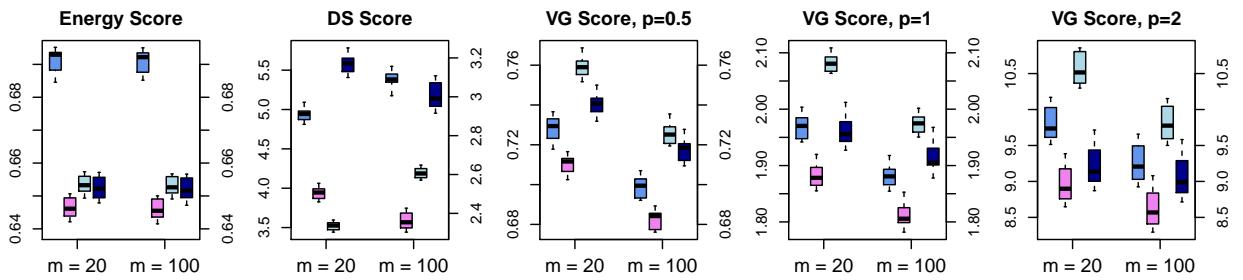
645 **Fig. 2.** Scores for forecasts with too weak, adequate, and too strong correlations (from left to right)
646 compared to the observations. The top row shows the results for $d = 5$, the bottom row
647 shows results for $d = 15$ 36

648 **Fig. 3.** Scores for forecasts with correct (left boxplot) and incorrect (right boxplot) correlation struc-
649 ture where the correct correlation function is that of model i) (top row) or model ii) (bottom
650 row), and the incorrect correlation function is in both cases the exponential model (5) with
651 $r = 3$ 37

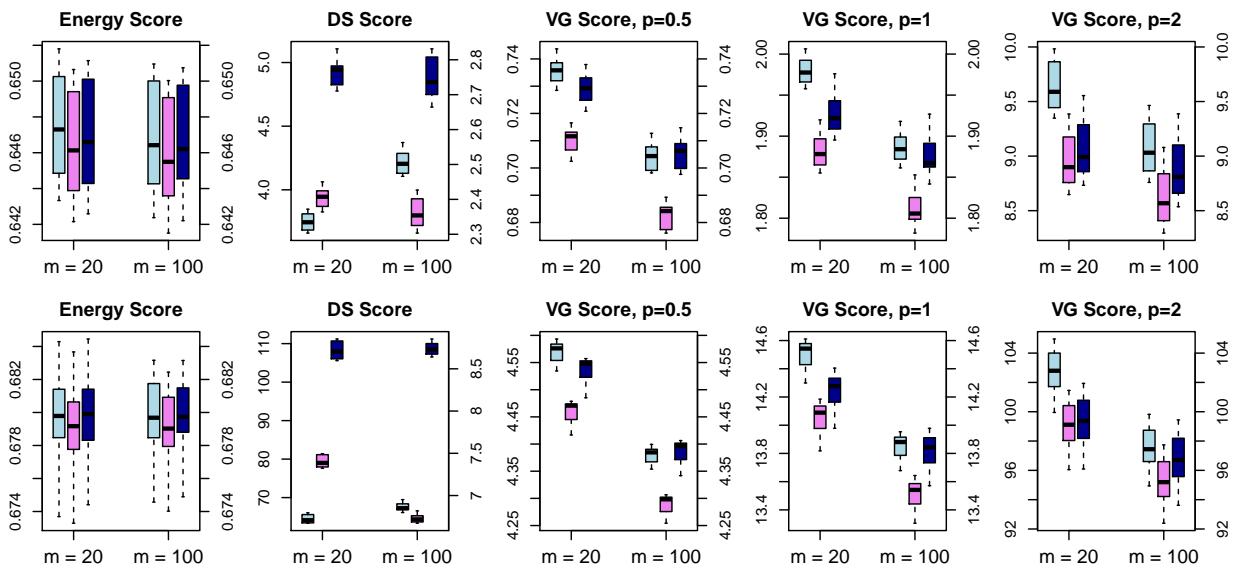
652 **Fig. 4.** Scores for forecasts of correct (shot noise) type (left boxplots) and incorrect (Gaussian)
653 process type (right boxplots). 38

654 **Fig. 5.** VS-0.5 for different component weights. The three plots on the left show results for the
655 correlation strength experiment, the three plots on the right show results for the correlation
656 model experiment. 39

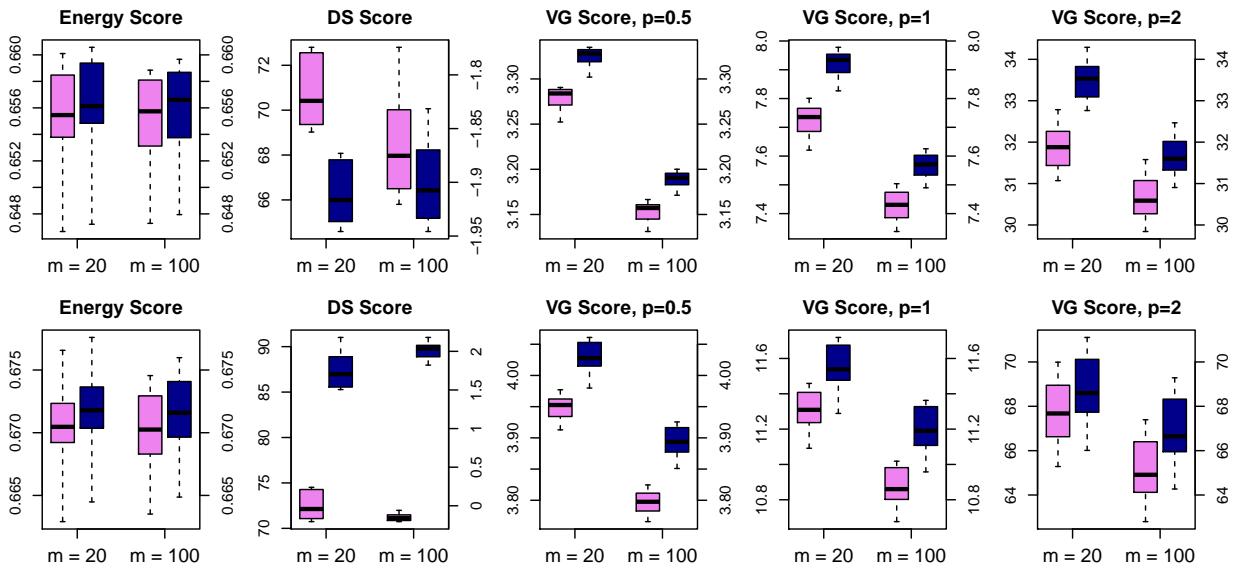
657 **Fig. 6.** Different scores applied to the $g_{0.5, \bar{w}}$ -transformed forecast and observation vectors. The
658 two left boxplots within each plot correspond to the experiments where the observation is
659 generated according model i). The two right boxplots correspond to the experiments where
660 the observation is generated according to model ii). 40



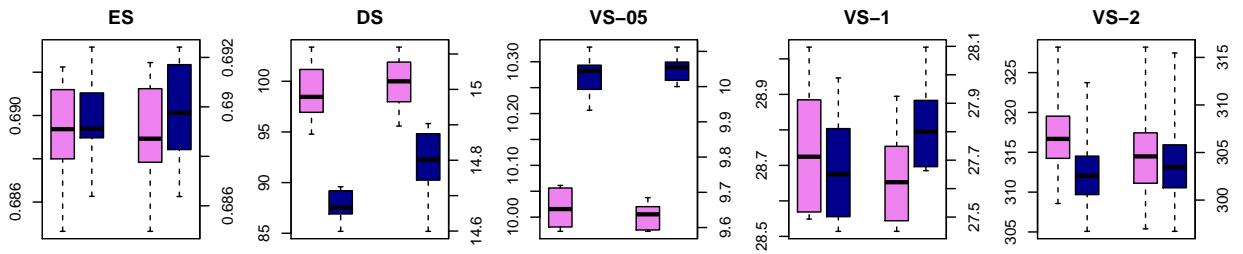
661 FIG. 1. Scores for mean-biased, correct, overdispersive, and underdispersive forecasts (from left to right) for
 662 ensemble size $m = 20$ and $m = 100$.



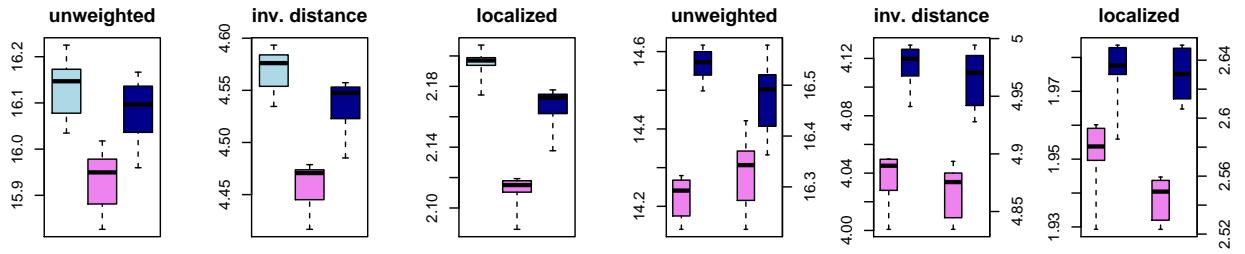
663 FIG. 2. Scores for forecasts with too weak, adequate, and too strong correlations (from left to right) compared
 664 to the observations. The top row shows the results for $d = 5$, the bottom row shows results for $d = 15$.



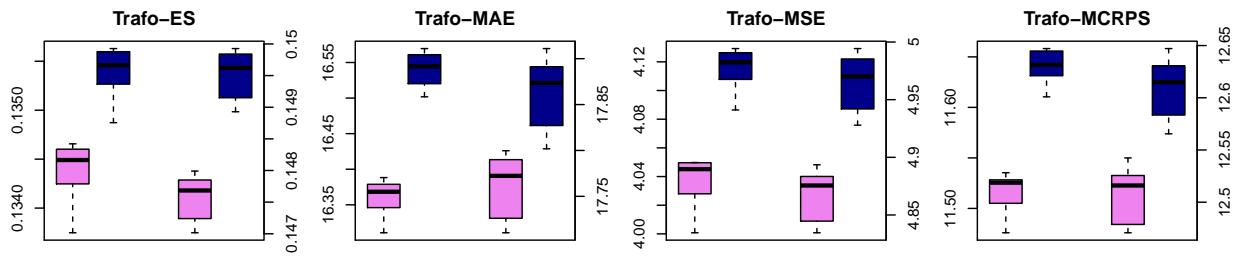
665 FIG. 3. Scores for forecasts with correct (left boxplot) and incorrect (right boxplot) correlation structure
 666 where the correct correlation function is that of model i) (top row) or model ii) (bottom row), and the incorrect
 667 correlation function is in both cases the exponential model (5) with $r = 3$.



668 FIG. 4. Scores for forecasts of correct (shot noise) type (left boxplots) and incorrect (Gaussian) process type
 669 (right boxplots).



670 FIG. 5. VS-0.5 for different component weights. The three plots on the left show results for the correlation
 671 strength experiment, the three plots on the right show results for the correlation model experiment.



672 FIG. 6. Different scores applied to the $g_{0.5, \tilde{w}}$ -transformed forecast and observation vectors. The two left
 673 boxplots within each plot correspond to the experiments where the observation is generated according model i).
 674 The two right boxplots correspond to the experiments where the observation is generated according to model ii).