Electronic Journal of Statistics
Vol. 8 (2014) 2619–2638
ISSN: 1935-7524
DOI: 10.1214/14-EJS964

Copula calibration

Johanna F. Ziegel

Institute of Mathematical Statistics and Actuarial Science, University of Bern
Sidlerstrasse 5, 3012 Bern, Switzerland
e-mail: johanna.ziegel@stat.unibe.ch

and

Tilmann Gneiting

Heidelberg Institute for Theoretical Studies and Karlsruhe Institute of Technology
HITS gGmbH, Schloß-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
e-mail: tilmann.gneiting@h-its.org

Abstract: We propose notions of calibration for probabilistic forecasts of general multivariate quantities. Probabilistic copula calibration is a natural analogue of probabilistic calibration in the univariate setting. It can be assessed empirically by checking for the uniformity of the copula probability integral transform (CopPIT), which is invariant under coordinate permutations and coordinatewise strictly monotone transformations of the predictive distribution and the outcome. The CopPIT histogram can be interpreted as a generalization and variant of the multivariate rank histogram, which has been used to check the calibration of ensemble forecasts. Kendall calibration is an analogue of marginal calibration in the univariate case. Methods and tools are illustrated in simulation studies and applied to compare raw numerical model and statistically postprocessed ensemble forecasts of bivariate wind vectors.

AMS 2000 subject classifications: 62.
Keywords and phrases: Copula, Kendall distribution, multivariate calibration, density forecast evaluation, ensemble prediction.

Received July 2013.

1. Introduction

The past two decades have witnessed major developments in the scientific approach to forecasting, in that probabilistic forecasts, which take the form of probability distributions over future quantities and events, have been replacing single-valued point forecasts in a wealth of applications (Gneiting and Katzfuss, 2014). The goal in probabilistic forecasting is to maximize the sharpness of the predictive probability distributions subject to calibration (Gneiting, Balabdaoui and Raftery, 2007). Calibration concerns the statistical compatibility between the predictive distributions and the realizing observations; in a nutshell, the observations are supposed to be indistinguishable from random numbers drawn from the predictive distributions.

For probabilistic forecasts of univariate quantities various types of calibration have been established (Gneiting and Ranjan, 2013). In particular, a forecast is

probabilistically calibrated if its probability integral transform (PIT), i.e., the value of the predictive cumulative distribution function at the realizing observation, is uniformly distributed. Accordingly, empirical checks for the uniformity of histograms of PIT values have formed a cornerstone of density forecast evaluation (Dawid, 1984; Diebold, Gunther and Tay, 1998; Gneiting, Balabdaoui and Raftery, 2007).

In this paper we introduce notions of calibration for probabilistic forecasts of multivariate quantities and propose tools for empirical calibration checks in such settings, as recently called for in hydrologic and meteorological applications (Schaake et al., 2010; Pinson, 2013; Schefzik, Thorarinsdottir and Gneiting, 2013). In Section 2 we study a natural multivariate extension of the univariate PIT that is invariant under coordinate permutations and coordinatewise strictly monotone transformations of the predictive distribution and the realizing observation, namely, the copula probability integral transform (CopPIT). Probabilistic copula calibration can be assessed empirically by checking the uniformity of the CopPIT histogram, which can be viewed as a generalization and variant of the multivariate rank histogram proposed by Gneiting et al. (2008). Furthermore, we introduce the notion of Kendall calibration, which is an analogue of marginal calibration in the univariate case. The strengths of these notions and tools include their ease of interpretability and their applicability to both density and ensemble forecasts.

In Section 3 we employ CopPIT histograms in a number of simulation studies, and in Section 4 we use them to compare raw numerical model and statistically postprocessed ensemble forecasts of bivariate wind vectors over Germany. The paper ends with a discussion in Section 5.

2. Multivariate notions of calibration

We introduce the copula probability integral transform (CopPIT) and the notions of probabilistic copula calibration and Kendall calibration within the prediction space setting of Gneiting and Ranjan (2013). Throughout, we identify a probability measure on $\mathbb{R}^d$ with its cumulative distribution function (CDF).

The Kendall distribution function $\mathcal{K}_H$ of a probability measure or CDF H on $\mathbb{R}^d$ is defined as

$$\mathcal{K}_H(w) = \mathrm{pr}\{H(X) \le w\} \quad \text{for} \quad w \in [0,1],$$

where the random vector X has distribution H. It is well known that if $d = 1$ and H is continuous then $\mathcal{K}_H$ corresponds to a uniform distribution on $[0,1]$. In dimension $d > 1$, the Kendall distribution depends only on the copula of the probability measure H and generally it is not uniform (Barbe et al., 1996). In fact, for any CDF K on $[0,1]$ with $K(w) \ge w$ for $w \in [0,1]$ and any integer $d > 1$, there exists a probability measure H on $\mathbb{R}^d$ such that $\mathcal{K}_H = K$ (Nelsen et al., 2003; Genest, Nešlehová and Ziegel, 2011).

2.1. Probabilistic copula calibration and Kendall calibration

As noted, we work in the prediction space setting introduced by Gneiting and Ranjan (2013). Specifically, let $(\Omega, \mathcal{A}, \mathbb{Q})$ be a probability space. Let Y be an $\mathbb{R}^d$-valued random vector on Ω, and let H be a d-variate CDF-valued random quantity that is measurable with respect to some sub σ-algebra $\mathcal{A}_0 \subseteq \mathcal{A}$.[1] Furthermore, let the random variable V be uniformly distributed on the unit interval $[0, 1]$ and independent of Y and $\mathcal{A}_0$.

The CDF-valued random quantity H provides an $\mathcal{A}_0$-measurable predictive probability measure for the $\mathbb{R}^d$-valued outcome Y. It is said to be ideal relative to $\mathcal{A}_0$ if it equals the conditional law of Y given $\mathcal{A}_0$, which we denote by $H = \mathcal{L}(Y|\mathcal{A}_0)$. Thus, an ideal forecast honors the information in the sub σ-algebra $\mathcal{A}_0 \subseteq \mathcal{A}$ to the full extent possible. For a function f on the real line, we use the notation $f(y-) = \lim_{x \uparrow y} f(x)$ to denote the left-hand limit, if it exists.

Definition 2.1 (CopPIT)**.** In the prediction space setting, the random variable

$$U_H = \mathcal{K}_H\{H(Y)-\} + V\left[\mathcal{K}_H\{H(Y)\} - \mathcal{K}_H\{H(Y)-\}\right] \tag{1}$$

is the copula probability integral transform (CopPIT) of the CDF-valued random quantity H.

If T is a deterministic coordinatewise strictly monotone transformation on $\mathbb{R}^d$, i.e.,

$$T(x_1, \ldots, x_d) = (T_1(x_1), \ldots, T_d(x_d))$$

where the mappings $T_1, \ldots, T_d$ are real-valued and strictly increasing, the distribution of U_H for the probabilistic forecast H and the outcome Y is the same as that of $U_{H \circ T^{-1}}$ for the probabilistic forecast $H \circ T^{-1}$ and the outcome $T(Y)$. The distribution of U_H also is invariant under coordinate permutations. An interesting open question is for the largest class of transformations under which this invariance holds, with the class of the locally orientation preserving functions being a candidate.

Definition 2.2. The forecast H is probabilistically copula calibrated if its CopPIT is uniformly distributed on the unit interval.

Probabilistic copula calibration can be viewed as a multivariate generalization of the notion of probabilistic calibration in the univariate case. In the prediction space setting, let F be a univariate CDF-valued random quantity for the real-valued outcome Y. Gneiting and Ranjan (2013, Definition 2.6) define F to be probabilistically calibrated if

$$U_F = F(Y-) + V\{F(Y) - F(Y-)\} \tag{2}$$

is standard uniformly distributed. If the dimension is $d = 1$ then equation (1) is the same as equation (2).

[1] That is, $\{H(y_j) \in B_j \text{ for } j = 1, \ldots, n\} \in \mathcal{A}_0$ for all finite collections $y_1, \ldots, y_n \in \mathbb{R}^n$ and Borel sets $B_1, \ldots, B_n \subseteq \mathbb{R}$.

Definition 2.3. The forecast H is Kendall calibrated if

$$\mathbb{Q}\{H(Y) \leq w\} = \mathbb{E}_{\mathbb{Q}}\{\mathcal{K}_H(w)\} \quad \text{for} \quad w \in [0,1]. \tag{3}$$

The concept of Kendall calibration can be interpreted as marginal calibration of the Kendall distribution, where marginal calibration refers to the univariate prediction space setting, as follows (Gneiting and Ranjan, 2013, Definition 2.6). If F is a univariate CDF-valued random quantity for the real-valued outcome Y, then it is marginally calibrated if $\mathbb{Q}(Y \leq y) = \mathbb{E}_{\mathbb{Q}}\{F(y)\}$ for $y \in \mathbb{R}$. Hence, marginal calibration ensures that the forecast distribution is at least correct on average over many prediction cases. The multivariate analogue introduced in Definition 2.3 ensures that the dependence structure is correctly predicted on average. Here, the dependence structure is summarized by the Kendall distribution of the underlying copula.

The following result justifies the quest for probabilistically copula calibrated and Kendall calibrated predictive distributions in practical settings.

Theorem 2.1. *If the forecast H is ideal with respect to the σ-algebra $\mathcal{A}_0$, then it is both probabilistically copula calibrated and Kendall calibrated.*

Proof. Suppose that $H = \mathcal{L}(Y|\mathcal{A}_0)$ and let $w \in [0,1]$. Then

$$\mathbb{Q}\{H(Y) \leq w\} = \mathbb{E}_{\mathbb{Q}}\,\mathbb{E}_{\mathbb{Q}}\,[\mathbb{1}\{H(Y) \leq w\}|\mathcal{A}_0] = \mathbb{E}_{\mathbb{Q}}\{\mathcal{K}_H(w)\},$$

whence H is Kendall calibrated. Turning to probabilistic copula calibration, observe that $\mathcal{K}_H$ and H are $\mathcal{A}_0$-measurable, and that the conditional distribution of $\mathbf{Y}$ given $\mathcal{A}_0$ is H, hence $\mathbb{Q}\{U_H \leq w\} = w$ by the known results for non-random CDFs. □

Suppose that the probabilistic forecasts $F_1, \ldots, F_d$ for the marginals of the random vector $Y = (Y_1, \ldots, Y_d)$ are probabilistically calibrated. Then probabilistic copula calibration can be seen as a property that depends only on the copula C of the forecast H and the copula of the outcome vector Y, as follows. Probabilistic calibration of the marginals implies that the random vector $W = (U_{F_1}, \ldots, U_{F_d})$ has uniformly distributed marginals. Therefore, the problem of predicting Y by H can be reduced to predicting W by a copula C. Then

$$H = C \circ (F_1, \ldots, F_d)$$

yields a multivariate probabilistic forecast of Y with probabilistically calibrated marginals. For a related discussion in the context of ensemble forecasts, see Schefzik, Thorarinsdottir and Gneiting (2013).

2.2. Empirical assessment of copula calibration

In the practice of forecast evaluation, one observes a sample

$$(H_1, y_1), \ldots, (H_J, y_J)$$

from the joint distribution of the probabilistic forecast and the outcome.

To assess probabilistic copula calibration one can plot a histogram of the empirical CopPIT values

$$u_j = \mathcal{K}_{H_j}\{H_j(y_j)-\} + v_j\,[\mathcal{K}_{H_j}\{H_j(y_j)\} - \mathcal{K}_{H_j}\{H_j(y_j)-\}] \tag{4}$$

for $j = 1, \ldots, J$, where $v_1, \ldots, v_J$ are independent standard uniformly distributed random numbers. Based on ideas in Czado, Gneiting and Held (2009), one can also define a non-randomized version of the CopPIT, but we do not pursue this here. In most cases of practical interest, the Kendall distribution is continuous and then we can write

$$u_j = \mathcal{K}_{H_j}\{H_j(y_j)\}, \tag{5}$$

without any need to invoke v_j. If $d = 1$, the CopPIT histogram coincides with the PIT histogram, the key tool in checking the calibration of univariate probabilistic forecasts (Diebold, Gunther and Tay, 1998; Gneiting, Balabdaoui and Raftery, 2007; Czado, Gneiting and Held, 2009). If the forecasts are probabilistically copula calibrated, the CopPIT histogram is uniform up to random fluctuations, and deviations from uniformity can be interpreted diagnostically, as illustrated in Section 3.1.

For multivariate distributions with an Archimedean copula the Kendall distribution function $\mathcal{K}_H$ is available in closed form (McNeil and Nešlehová, 2009), and then we can readily evaluate (4) or (5). For other types of distributions, one can approximate $\mathcal{K}_H$ by the empirical CDF of $H(x_1), \ldots, H(x_n)$ for some large n, where $x_1, \ldots, x_n$ is a sample from a d-variate population with CDF H. An alternative approximation that does not require the potentially costly evaluation of H, uses the empirical Kendall distribution function $\mathcal{K}_n$, i.e., the empirical CDF of the pseudo-observations

$$w_k = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(x_j \preceq x_k) \qquad \text{for } k = 1, \ldots, n, \tag{6}$$

where $x_j = (x_{j1}, \ldots, x_{jd}) \preceq x_k = (x_{k1}, \ldots, x_{kd})$ if $x_{jl} \le x_{kl}$ for $l = 1, \ldots, d$. As Barbe et al. (1996) show, the empirical Kendall distribution function $\mathcal{K}_n$ generally converges to $\mathcal{K}_H$. Both types of approximation are further discussed in Section 3.4.

To assess Kendall calibration one can plot

$$\frac{1}{J}\sum_{j=1}^{J} \mathbb{1}\{H_j(y_j) \le w\} \quad \text{vs.} \quad \frac{1}{J}\sum_{j=1}^{J} \mathcal{K}_{H_j}(w)$$

for $w \in [0, 1]$, which are the empirical analogues of the left- and right-hand sides of (3). We call this type of display a Kendall calibration diagram. If the forecasts are calibrated the resulting plot ought to be close to the diagonal.

2.3. Comparison to the multivariate rank histogram

As noted, the CopPIT histogram generalizes the multivariate rank histogram introduced by Gneiting et al. (2008) in the context of ensemble forecasts. This

refers to the situation in which the probabilistic forecasts $H_1, \dots, H_J$ are empirical measures with a fixed size m.

For ease of exposition, we drop the indices and suppose that the forecast H places mass $1/m$ at each of $x_1, \dots, x_m \in \mathbb{R}^d$, while the outcome is $y \in \mathbb{R}^d$. The associated multivariate rank is obtained as follows. Define pre-ranks $\rho_0 = 1 + \sum_{i=1}^m \mathbb{1}(x_i \preceq y)$ and

$$\rho_k = \mathbb{1}(y \preceq x_k) + \sum_{i=1}^m \mathbb{1}(x_i \preceq x_k) \quad \text{for } k = 1, \dots, m.$$

The multivariate rank then is the rank of the observation pre-rank ρ_0 among $\rho_0, \rho_1, \dots, \rho_m$, with ties resolved at random. Conditional on H and y we thus get a multivariate rank with a discrete uniform distribution on the integers

$$1 + \sum_{k=1}^m \mathbb{1}(\rho_k < \rho_0), \ \dots, 1 + \sum_{k=1}^m \mathbb{1}(\rho_k \le \rho_0). \tag{7}$$

We now link the multivariate rank and the CopPIT. If H is the empirical measure with mass $1/m$ at $x_1, \dots, x_m \in \mathbb{R}^d$, its Kendall distribution function can be expressed in terms of the pseudo-observations at (6), in that

$$\mathcal{K}_H(w) = \frac{1}{m} \sum_{k=1}^m \mathbb{1}(w_k \le w) \quad \text{for} \quad w \in [0, 1].$$

Since $\rho_0 = mH(y) + 1$ and $\rho_k = mw_k + \mathbb{1}(y \preceq x_k)$ for $k = 1, \dots, m$, we can express the CopPIT value (4) in terms of the pseudo-ranks. Solving these equations in terms of $H(y)$ and w_k and plugging into (4) shows that conditional on H and y the CopPIT value has a uniform distribution on the interval

$$\left[\frac{1}{m} \sum_{k=1}^m \mathbb{1}\{\rho_k - \mathbb{1}(y \preceq x_k) < \rho_0 - 1\}, \frac{1}{m} \sum_{k=1}^m \mathbb{1}\{\rho_k - \mathbb{1}(y \preceq x_k) \le \rho_0 - 1\} \right]. \tag{8}$$

A comparison of (7) and (8) suggests that if the ensemble size m is large the CopPIT and the multivariate rank histogram tend to look nearly identical. If m is small this may not be the case, as we illustrate in Section 4.

The multivariate rank histogram has also been used to assess the calibration of probabilistic forecasts in the form of continuous multivariate distributions. For example, Schuhen, Thorarinsdottir and Gneiting (2012) transform predictive densities for bivariate wind vectors into ensemble forecasts, by drawing a simple random sample from each predictive distribution, where the particular choice of the sample size $m = 8$ allows for a better comparison with the underlying ensemble forecast. In such settings we prefer to work with the CopPIT histogram, as it makes better use of the structure of the predictive distributions and does not induce additional randomness into the evaluation procedure. We illustrate this latter aspect in a simulation setting in Section 3.3.

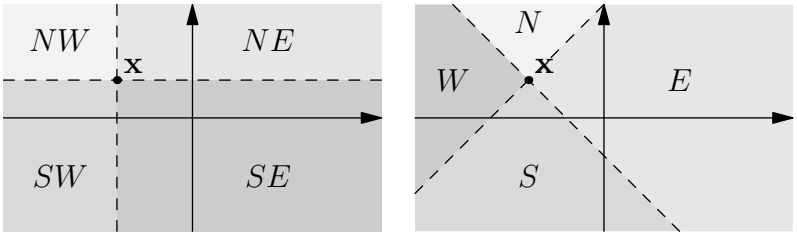

FIG 1. *Illustration of quadrants for directional CopPITs.*

2.4. Directional copula calibration

The CopPIT is a natural multivariate generalization of the PIT in the univariate setting. We now discuss a further generalization that allows for directional approaches. In doing so, we refer to the probabilistic forecast for the $\mathbb{R}^d$-valued outcome Y either by H or μ, with H denoting a CDF and μ the associated probability measure.

Let $e_1, \dots, e_d$ be an orthonormal basis of $\mathbb{R}^d$ and let $\mathcal{E}$ be the closed convex cone spanned by this basis. We define the $\mathcal{E}$-CDF of the probability measure μ as

$$H^{\mathcal{E}} : \mathbb{R}^d \to [0,1], \qquad x \mapsto \mu(x + \mathcal{E}).$$

Any function $H^{\mathcal{E}}$ characterizes the probability measure μ. The usual CDF is obtained by choosing $e_j = (e_{1j}, \dots, e_{dj})$ with $e_{ij} = -\mathbb{1}(i = j)$, whereas the survival function of μ is $H^{\mathcal{E}}$ with $e_{ij} = \mathbb{1}(i = j)$. The CopPIT depends on the particular CDF chosen, and distinct choices of $\mathcal{E}$ may reveal distinct facets of calibration or the lack thereof. In principle, one could envision a procedure in the style of a projection pursuit algorithm (Huber, 1985) that finds those $\mathcal{E}$ where the deviation of the CopPIT histogram from uniformity is the most pronounced. In the case of density forecasts a related idea was considered by Ishida (2005).

Certain choices of the cone $\mathcal{E}$ might be particularly useful. We illustrate this for $d = 2$, but the idea generalizes to higher dimensions. Let SW be the convex cone spanned by $(-1, 0)$ and $(0, -1)$, i.e., the south-west quadrant. Analogously we define the quadrants SE, NE, and NW, as illustrated in Figure 1. If the marginals are probabilistically calibrated, probabilistic copula calibration with respect to H^{SW}, which is the classical multivariate CDF, only depends on the forecast copula. This argument remain valid for H^{SE}, H^{NE}, and H^{NW}, with the latter being the multivariate survival function.

Similarly, we can assess directional Kendall calibration by plotting

$$\frac{1}{J}\sum_{j=1}^{J} \mathbb{1}\{H_j^{\mathcal{E}}(y_j) \le w\} \quad \text{vs.} \quad \frac{1}{J}\sum_{j=1}^{J} \mathcal{K}_{H_j^{\mathcal{E}}}(w)$$

for $w \in [0,1]$ and suitable choices of the cone $\mathcal{E}$.

TABLE 1
Parameters of forecast distributions in the simulation study

Forecast	First Margin F_1	Second Margin F_2	Copula C
T	correct $\mu_1 = 2 - B_1$	correct $\sigma_2^2 = 1/B_2$	correct $\tau = (B_1 + B_2)/2$
F	biased $\hat{\mu}_1 = 0.8(2 - B_1)$	underdispersed $\hat{\sigma}_2^2 = 0.8/B_2$	misspecified $\hat{\tau} = 0.6(B_1 + B_2)/2$

3. Simulation studies

We now illustrate the use of the CopPIT histogram and Kendall calibration diagram in simulation studies, all of which use R (R Core Team, 2013).

3.1. Interpretation of copula calibration

We consider the following simulation setting in dimension $d = 2$. Let B_1 and B_2 be independent beta variables with parameters $(\alpha_1, \beta_1) = (2, 5)$ and $(\alpha_2, \beta_2) = (5, 2)$, respectively. Conditional on (B_1, B_2) the outcome vector $Y = (Y_1, Y_2)$ has normal margins and a Gumbel copula with Kendall's τ equal to $(B_1 + B_2)/2$, as described by Nelsen (2006). The margin Y_1 has mean $\mu_1 = 2 - B_1$ and unit variance; the margin Y_2 has mean zero and variance $\sigma_2^2 = 1/B_2$.

We assess eight probabilistic forecasters with various types of forecast deficiencies. All forecasters have access to (B_1, B_2) and specify a Gumbel copula with Kendall's τ equal to $\hat{\tau}$ and normal marginals, where the first margin F_1 has mean $\hat{\mu}_1$ and unit variance, and the second margin F_2 has mean zero and variance $\hat{\sigma}_2^2$, with details provided in Table 1. We name each forecaster with a sequence of three letters, where T stands for true and F for false. For example, the forecaster TTF specifies the first and the second marginal distributions correctly, but misspecifies the copula. The forecaster TTT is ideal with respect to the σ-algebra generated by (B_1, B_2) in the sense defined in Section 2.1 and does not show any forecast deficiencies.

Figure 2 shows CopPIT histograms for the eight forecasters based on a sample of 4,000 forecast–observation pairs. It is interesting to observe that the standard CopPIT histogram detects misspecified marginals as well as misspecified copulas. Similar to the interpretation of univariate PIT histograms (Gneiting, Balabdaoui and Raftery, 2007), biases yield skewed histograms, underdispersed forecasts induce a U-shape, and overdispersed forecasts an inverse U-shape.

Figure 3 shows univariate PIT histograms along with directional CopPIT histograms based on another sample of 4,000 forecast–observation pairs. The joint consideration of the histograms can diagnose specific forecast deficiencies. As a rule of thumb, the CopPIT histograms mimic features seen in the univariate PIT histograms if the copula is well specified. For example, in the third row the first marginal distribution and the copula are specified correctly, whereas the second marginal distribution, F_2, is underdispersed. The underdispersion of F_2 is reflected by the U-shaped PIT histogram in the second column, and this

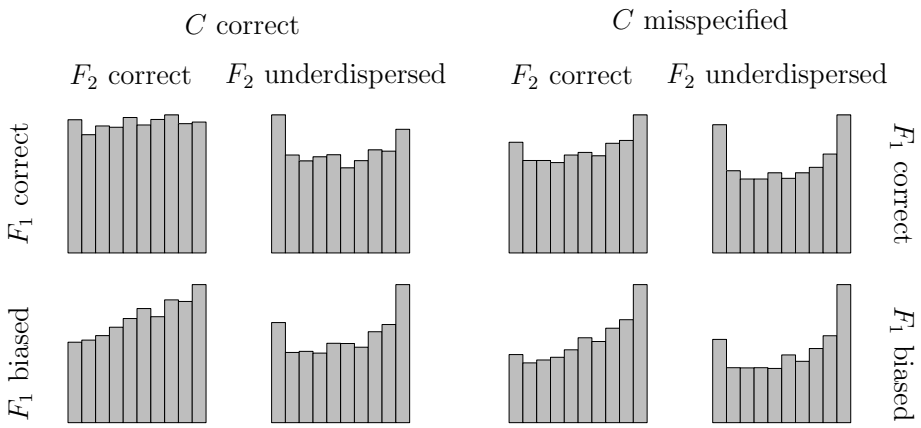

FIG 2. *CopPIT histograms for the forecasters in the simulation study in Section 3.1.*

U-shape carries over to the directional CopPIT histograms. In contrast, if the copula is ill specified, the CopPIT histograms show deviations from uniformity in shapes that are not necessarily reflected by the PIT histograms. For example, in the seventh row the first marginal distribution is specified correctly, and the second marginal distribution is underdispersed. However, the copula is now misspecified. Instead of only seeing U-shapes in the CopPIT histograms, as we do in row three, we now see hump shapes in columns four and six.

Finally, Figure 4 shows directional Kendall calibration diagrams. While misspecifications of the probabilistic forecasts are readily discernible, the Kendall calibration diagrams appear to be more difficult to interpret diagnostically than the CopPIT histograms.

3.2. Tawn copulas

Bivariate Archimedean copulas are characterized by their Kendall distribution function; the same is true in dimension $d = 3$ and has been conjectured for $d \geq 4$ (Genest, Nešlehová and Ziegel, 2011). This raises the question whether the CopPIT histogram is suited particularly well to this situation, such as in Section 3.1, and may fail otherwise. Therefore, we now consider another simulation example based on the Tawn copula family (Tawn, 1988). This is a three parameter family of extreme value copulas, which can be defined via their Pickands dependence function; see, for example, Gudendorf and Segers (2010). For all extreme value copulas, the Kendall distribution function has the remarkably simple form

$$\mathcal{K}_H(w) = w - (1-\tau)w \log w \quad \text{for} \quad w \in [0,1],$$

in terms of Kendall's τ, which derives from the Pickands dependence function A as

$$\tau = \int_0^1 \frac{t(1-t)}{A(t)} \, \mathrm{d}A'(t).$$

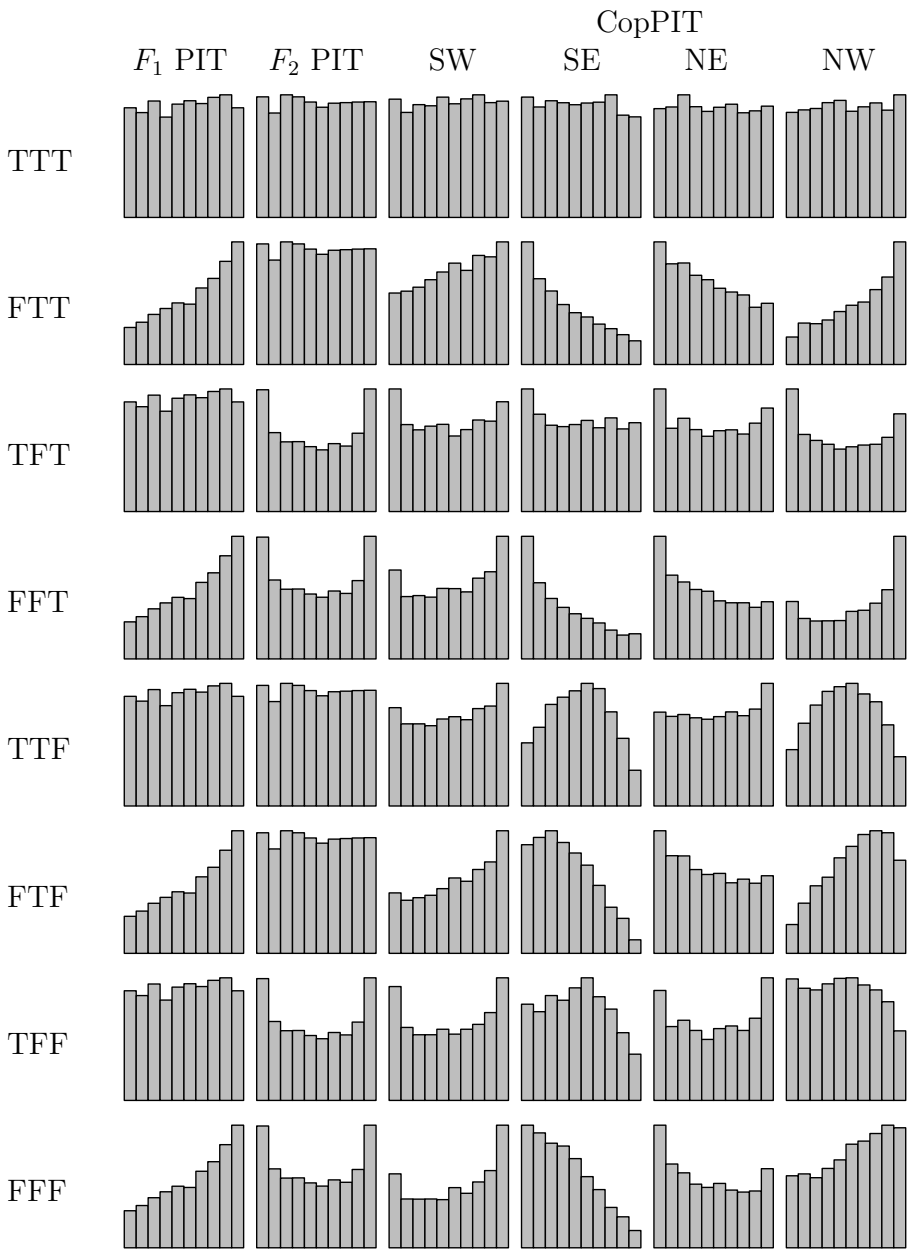

FIG 3. *Univariate PIT and directional CopPIT histograms for the forecasters in the simulation study in Section 3.1.*

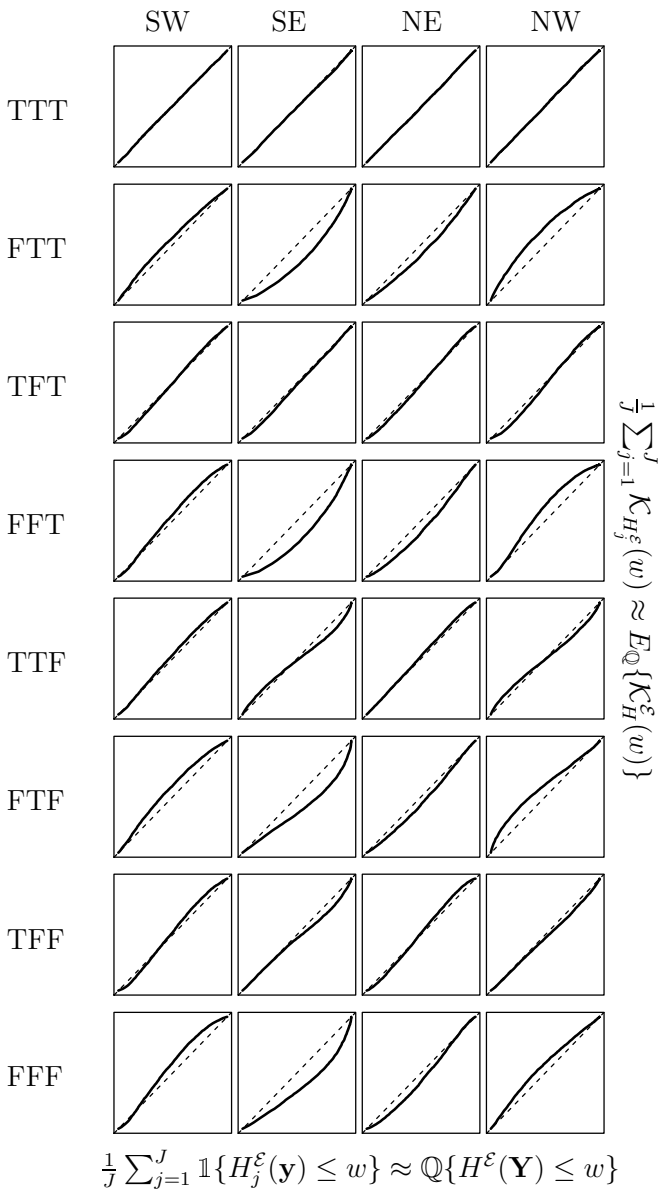

FIG 4. *Directional Kendall calibration diagrams for the forecasters in the simulation study in Section 3.1.*

It follows easily that, contrary to the Archimedean case, the Kendall distribution does not identify the copula uniquely, even in the bivariate case.

The Tawn copula family allows for left- as well as right-skewed Pickands dependence functions. We consider the left-skewed case only, which corresponds

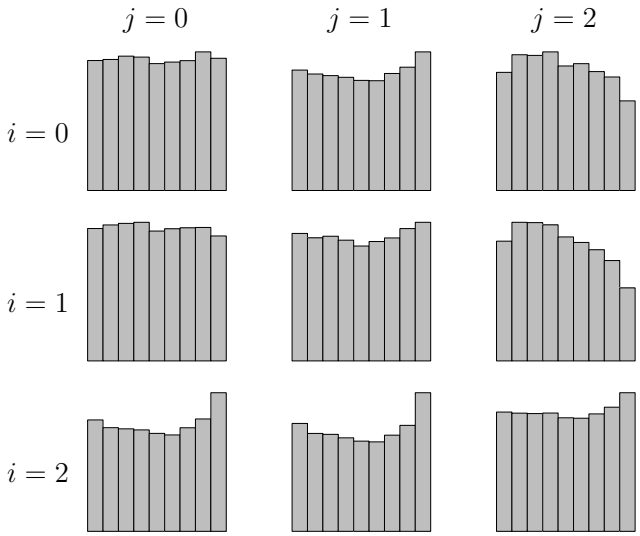

FIG 5. *CopPIT histograms for the forecasters F_{ij} in the simulation study in Section 3.2.*

to Pickands dependence functions of the form

$$A(t) = (1-\psi)t + \{(1-t)^\theta + (\psi t)^\theta\}^{1/\theta}, \qquad t \in [0,1],$$

where $\theta > 1$ and $\psi \in [0,1]$. Let the random variable γ be gamma distributed with shape and scale parameter equal to 6 and 0.3, respectively. Conditional on γ, the bivariate outcome is distributed as a Tawn copula with $\theta = \gamma + 1$ and $\psi = 1/(\gamma+1)$. We consider nine different forecasters F_{ij}, where $i, j = 0, 1, 2$. All nine forecasters predict the marginal distributions correctly. Forecaster F_{00} indeed predicts the full bivariate distribution correctly. Generally, forecaster F_{ij} predicts a Tawn copula with parameters $\theta = \theta_i$ and $\psi = \psi_j$, where

$$\theta_0 = \gamma + 1, \qquad \theta_1 = 2\gamma + 1, \qquad \theta_2 = \frac{\gamma}{5} + 1,$$

and $\psi_j = 1/\theta_j$ for $j = 0, 1, 2$. If we fix the value of γ at its expectation, 1.8, the outcome and forecaster F_{00} have a conditional Kendall's τ of 0.285. The respective values for the other forecasters vary from a minimum of 0.098 for F_{21}, 0.221 for F_{22}, and 0.320 for F_{10}, to a maximum of 0.605 for F_{12}.

Figure 5 shows the CopPIT histograms for 10,000 forecast–observation pairs for each forecaster. Overall, miscalibration is detected well by the CopPIT histograms. The type of deviation from uniformity varies depending on which parameters have been misspecified. Predicting θ_1 instead of θ_0 only has a small influence on the CopPIT histogram, whereas predicting θ_2 instead of θ_0 makes the CopPIT histogram attain a U-shape. Predicting ψ_1 instead of ψ_0 also yields U-shaped CopPIT histograms, while predicting ψ_2 instead of ψ_0 favors hump-shaped histograms. For forecaster F_{22}, the two effects nearly compensate each other, so the deviation from uniformity is minor only.

3.3. A higher dimensional example

As noted in Section 2.3, the multivariate rank histogram of Gneiting et al. (2008), which is tailored to discrete ensemble forecasts, can also be used to assess the calibration of probabilistic forecasts given in the form of continuous multivariate distributions. However, additional randomness is introduced in the evaluation procedure as a sample needs to be drawn from the predictive distribution in order to compute the multivariate rank histogram. In such situations, it seems advantageous to use the CopPIT histogram.

We illustrate this in a simulation setting in dimension $d = 20$, where we choose the sample size $m = 8$ to compute the multivariate rank histograms. In weather and climate forecasting, ensemble systems operate with small m and large d (Gneiting and Raftery, 2005; Leutbecher and Palmer, 2008), so this scenario is practically relevant. Specifically, let δ and λ be independent gamma variables with shape and scale parameters $(2, 2)$ and $(5, 1)$, respectively. Given λ, let $(p_{ij})_{i,j=1,\ldots,5}$ be independent Poisson variables with parameter λ. The outcome vector Y is given by

$$Y = \begin{pmatrix} I_5 & 0 & 0 \\ 0 & A & 0 \\ 0 & 0 & B \end{pmatrix} X,$$

where $X = (X_1, \ldots, X_{20})'$ is vector of independent normally distributed components with mean zero and variance δ^2. Furthermore, I_5 is the five-dimensional identity matrix, $A \in \mathbb{R}^{10\times 10}$ has entries $a_{i,i} = 1$, $a_{i,i+1} = 1/2$ and zeros otherwise, and $B \in \mathbb{R}^{5\times 5}$ has entries $b_{i,i} = (p_{i,i} + 1)/(\sum_{ij} p_{i,j} + 5)$ and $b_{i,j} = p_{i,j}/(\sum_{ij} p_{i,j} + 5)$ if $i \neq j$. In the context of weather forecasting, the dependence structure embodied in the matrix A might correspond to a given weather variable at a given location over ten subsequent prediction horizons, and that in B to five distinct weather variables at a given location and a given prediction horizon.

Forecaster F_1 predicts the distribution of a vector X_1 of independent normal components with variance δ^2. Forecaster F_2 predicts the distribution of

$$X_2 = \begin{pmatrix} I_5 & 0 \\ 0 & A' \end{pmatrix} X_1,$$

where $A' \in \mathbb{R}^{15\times 15}$ has entries $a'_{i,i} = 1$, $a'_{i,i+1} = 1/2$ and zeros otherwise. Hence, she correctly predicts the first 15 components of Y. Forecaster F_3 predicts the distribution of

$$X_3 = \begin{pmatrix} I_5 & 0 & 0 \\ 0 & A & 0 \\ 0 & 0 & B' \end{pmatrix} X_1,$$

where $B' \in \mathbb{R}^{5\times 5}$ has entries $b'_{i,i} = (\lambda + 1)/(25\lambda + 5)$ and $b'_{i,j} = \lambda/(25\lambda + 5)$ if $i \neq j$. This means that the last forecaster is almost predicting the correct dependence structure, but she does not have access to the $p_{i,j}$, so instead their mean, λ, is used.

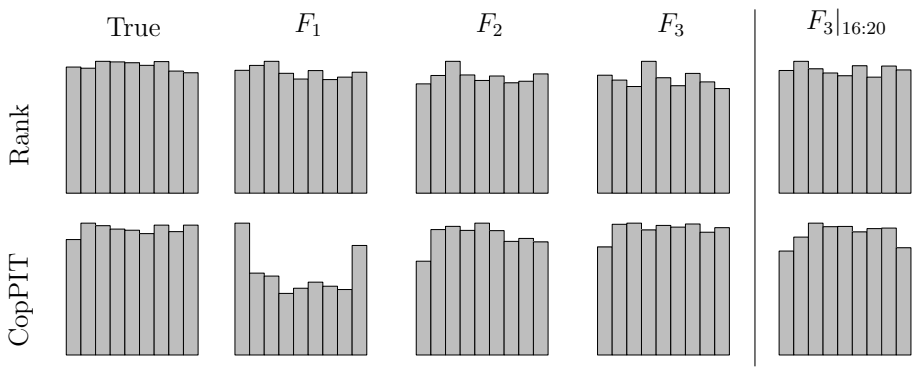

FIG 6. *Multivariate rank and CopPIT histograms in the high-dimensional simulation setting in Section 3.3.*

Figure 6 shows multivariate rank and CopPIT histograms in this setting, based on a sample of 4,000 forecast–observation pairs. As the Kendall distribution function of a multivariate normal law is not available in closed form, an approximate version is used, as discussed in the subsequent section. The rank histograms have difficulties in detecting the deficient probabilistic forecasts due to the discretization effect mentioned in Section 2.3. In contrast, the CopPIT histograms for the forecasts F_1 and F_2 with the misspecified copulas are non-uniform, as desired. For forecast F_3 it is debatable whether the CopPIT histogram deviates more from uniformity than the rank histogram. However, if we compare the rank histogram and the CopPIT histogram for the predicted distribution of the last five components of Y, denoted $F_3|_{16:20}$, the CopPIT histogram is able detect miscalibration, while the rank histogram remains essentially flat. When assessing calibration, it is not unnatural to check the calibration of this set of components separately. As described above, the dependence structure encoded in B, which is influencing the components 16 to 20 is of a different nature than the one of the first 15 components.

3.4. Approximating the Kendall distribution function

For many widely used multivariate distributions the Kendall distribution function is not available in closed form. In such cases the empirical Kendall distribution function (6) can be used to calculate the CopPIT if it is possible to draw samples from the distribution. We use this procedure with sample size $n = 5{,}000$ in Section 3.1 for the directional CopPIT histograms SE, NE, and NW, and throughout the bivariate real data study in Section 4.

However, in higher dimensions the sample size n may need to be very large in order to make the approximation sufficiently precise. Assuming that the predictive CDF, H, can be evaluated, it may then be preferable to approximate the Kendall distribution function by the empirical CDF of $H(x_1), \ldots, H(x_n)$, where

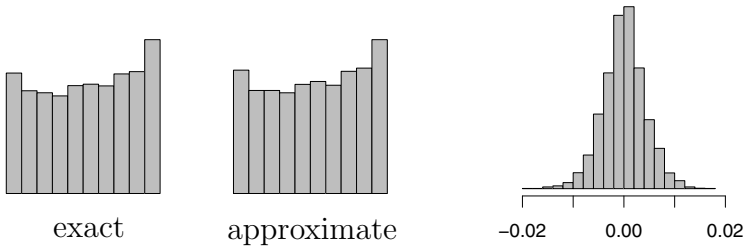

FIG 7. *CopPIT histograms for the forecaster TTF in Figure 3, computed using either the exact Kendall distribution function, or the approximation based on the empirical Kendall distribution function (6) with $n = 5{,}000$. The histogram at right shows the distribution of the error in the individual CopPIT values that is caused by using the approximation.*

$x_1, \dots, x_n$ is a sample drawn from H. In Section 3.3, we use this technique with sample size $n = 4{,}000$.

Here we demonstrate in two examples that the errors made in the two types of approximations typically are sufficiently small, so that they do not obstruct conclusions drawn from the CopPIT histogram.

Specifically, for the forecaster TTF in Section 3.1 we compute the CopPIT value using either the closed form expression of the Kendall distribution function derived by McNeil and Nešlehová (2009), to yield U_1, or based on the empirical Kendall distribution function(6) with $n = 5{,}000$, to yield U_2. Figure 7 displays the CopPIT histograms computed from U_1 and U_2, respectively, and shows the distribution of the approximation error, $U_1 - U_2$, for 4,000 forecast–observation pairs. The absolute value of the approximation error is below 0.01 in 97.5% of the cases. It is possible to spot tiny differences in the histograms. However, as CopPIT histograms are intended to be used as qualitative diagnostic tools, these small deviations are unimportant.

Figure 8 shows the results of an analogous study for forecaster F_1 in Section 3.3, who predicts a normal CDF, H, with $d = 20$ independent components. The

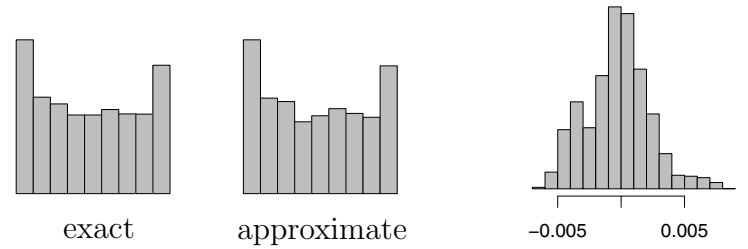

FIG 8. *CopPIT histograms for the forecaster F_1 in Figure 6 computed using either the exact Kendall distribution function, or the approximation based on the empirical CDF of $H(x_1), \dots, H(x_n)$, where $n = 4{,}000$ and $x_1, \dots, x_n$ is a sample from H. The histogram at right shows the distribution of the error in the individual CopPIT values that is caused by using the approximation.*

respective Kendall distribution function is given by

$$\mathcal{K}(w) = w \sum_{k=0}^{d-1} \frac{(-\log w)^k}{k!} \qquad \text{for} \qquad w \in [0,1].$$

We compare to the approximation via the empirical CDF of $H(x_1), \ldots, H(x_n)$, where $n = 4{,}000$ and $x_1, \ldots, x_n$ is a sample from H. Again, the approximation error is minor.

4. Case study: Probabilistic forecasts of wind vectors over the Pacific Northwest

In a recent change of paradigms, meteorologists have adopted probabilistic weather forecasting in the form of ensemble forecasts. An ensemble forecast is a collection of numerical weather prediction (NWP) model runs that are based on distinct initial conditions and/or model physics parameters (Gneiting and Raftery, 2005; Leutbecher and Palmer, 2008). Despite their undisputed success, ensemble forecasts tend to be biased and underdispersed, in the sense of the spread among the ensemble members being too small to be realistic. Therefore, methods for the statistical postprocessing of ensemble forecasts have been developed, such as the ensemble model output statistics (EMOS) approach of Gneiting et al. (2005), which generates Gaussian predictive distributions for univariate variables. In a more recent development, Schuhen, Thorarinsdottir and Gneiting (2012) developed a bivariate EMOS method that generates bivariate Gaussian predictive distributions for wind vectors. Parameter estimation is performed on a rolling 30-day training period, except for the model for the correlation coefficient, which is estimated on historical data; cf. Schuhen, Thorarinsdottir and Gneiting (2012, Sections 3.c and 3.d).

Here, we take up their work on probabilistic forecasts of surface wind vectors over the North American Pacific Northwest based on the University of Washington Mesoscale Ensemble (Eckel and Mass, 2005), which has $m = 8$ members. The test data comprise calendar year 2008 with a total of 19,282 forecast–observations pairs at 79 meteorological stations and a prediction horizon of 48 hours. We assess and compare the raw ensemble forecast, the statistically postprocessed regional bivariate EMOS forecast developed by Schuhen, Thorarinsdottir and Gneiting (2012), and an Independent EMOS forecast with the same bivariate Gaussian predictive distribution, except that the correlation coefficient is misspecified at zero.

Figure 9 shows univariate PIT histograms, the multivariate rank histogram, the CopPIT histogram, and the Kendall calibration diagram for the raw ensemble, Independent EMOS, and EMOS forecasts. The raw ensemble forecast shows U-shaped PIT, multivariate rank and CopPIT histograms, which attest to its underdispersion, and the Kendall calibration diagram points at severe forecast deficiencies. The univariate PIT histograms for the Independent EMOS and EMOS forecasts are identical and diagnose slight overdispersion. However,

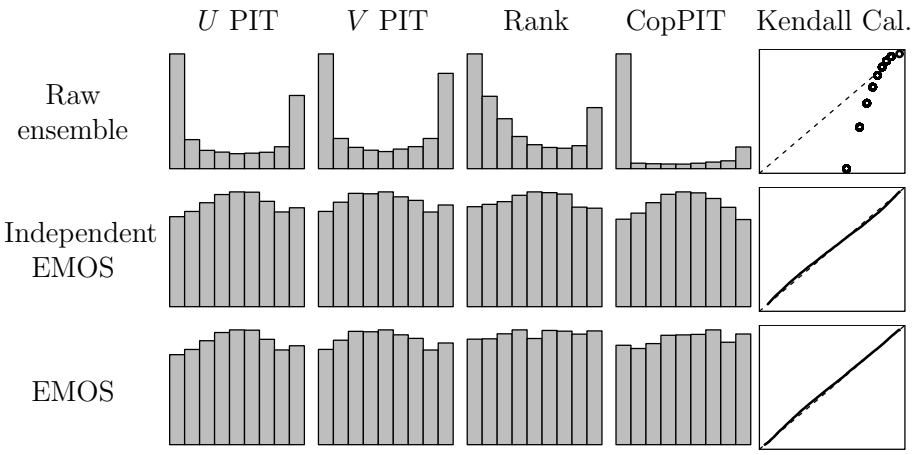

FIG 9. *Univariate PIT histograms, multivariate rank histogram, CopPIT histogram and Kendall calibration diagram for the raw ensemble, Independent EMOS, and EMOS forecasts of wind vectors. Following common practice, we label the wind vector components as u and v.*

the bivariate rank and CopPIT histograms for the EMOS forecast are more uniform than for the Independent EMOS forecast, as the Independent EMOS technique fails to take dependencies between the wind vector components into account, with the CopPIT histogram providing a much clearer diagnosis than the multivariate rank histogram.

5. Discussion

In this paper, we introduced the copula probability integral transform (CopPIT), and we proposed CopPIT histograms and Kendall calibration diagrams as diagnostic tools in the evaluation and comparison of probabilistic forecasts of multivariate quantities. These tools apply to non-parametric, semi-parametric and parametric approaches and thus can be employed to diagnose strengths and deficiencies of multivariate stochastic models in nearly any setting, be it predictive or not.

Extant methods for calibration checks for probabilistic forecasts of multivariate quantities apply either to ensemble forecasts only, such as rank histograms (Smith and Hansen, 2004; Wilks, 2004; Gneiting et al., 2008; Thorarinsdottir, Scheuerer and Heinz, 2014), or they apply to density forecasts only, such as the methods of Diebold, Hahn and Tay (1999), Ishida (2005), and González-Rivera and Yoldas (2012) that rely on the univariate PIT and the Rosenblatt transform (Rosenblatt, 1952; Rüschendorf, 2009) in one way or another. By way of contrast, CopPIT histograms and Kendall calibration diagrams apply to all types of probabilistic forecasts, including both ensemble forecasts and density forecasts.

In our case study, we assessed probabilistic forecasts of raw ensemble and statistically postprocessed density forecasts of bivariate wind vectors. However,

our methods also apply in higher dimensions and then it may be useful to plot CopPIT histograms and Kendall calibration diagrams for a range of subvectors of the outcome, too.

As noted, probabilistic forecasting strives to maximize the sharpness of the predictive probability distributions subject to calibration (Gneiting, Balabdaoui and Raftery, 2007), and the methods proposed here serve to evaluate calibration only. If probabilistic forecasters are to be ranked considering both calibration and sharpness, proper scoring rules can be employed (Gneiting and Raftery, 2007; Gneiting et al., 2008), with recent theoretical advances having been made by Ehm (2011). Diks, Panchenko and van Dijk (2010) and Röpnack et al. (2013) advocate the use of the logarithmic score to compare probabilistic forecasts of multivariate quantities. The event based approach of Pinson and Girard (2012) reduces a high-dimensional quantity to a binary event — essentially, the ultimate dimension reduction — and applies proper scoring rules to assess the induced probability forecasts for dichotomous events. While these techniques aim to rank probabilistic forecasters, CopPIT histograms and Kendall calibration diagrams are diagnostic tools that strive to inform model development and spur model improvement.

Acknowledgements

The authors thank Nina Schuhen and Thordis Thorarinsdottir for assistance with the data handling, and Fabian Krüger for support on some computational issues. Tilmann Gneiting acknowledges funding from the European Union Seventh Framework Programme under grant agreement no. 290976.

References

Barbe, P., Genest, C., Ghoudi, K. and Rémillard, B. (1996). On Kendall's process. *J. Multivariate Anal.* **58** 197–229. MR1405589

Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data. *Biometrics* **65** 1254–1261. MR2756513

Dawid, A. P. (1984). Statistical theory: The prequential approach. *J. R. Stat. Soc. A* **147** 278–290. MR0763811

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.* **39** 863–883.

Diebold, F. X., Hahn, J. and Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: High-frequency returns on foreign exchange. *Rev. Econ. Stat.* **81** 661–673.

Diks, C., Panchenko, V. and van Dijk, D. (2010). Out-of-sample comparisons of copula specifications in multivariate density forecasts. *J. Econ. Dyn. Contr.* **34** 1596–1609. MR2720290

Eckel, F. A. and Mass, C. F. (2005). Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting* **20** 328–350.

Ehm, W. (2011). Unbiased risk estimation and scoring rules. *C. R. Math.* **349** 699–702. MR2817395

Genest, C., Nešlehová, J. and Ziegel, J. (2011). Inference in multivariate Archimedean copula models. *Test* **20** 223–256. MR2834039

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. B* **69** 243–268. MR2325275

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Ann. Rev. Stat. Appl.* **1** 125–151.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science* **310** 248–249.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102** 359–378. MR2345548

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *El. J. Stat.* **7** 1747-1782. MR3080409

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.* **133** 1098–1118.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* **17** 211–235. MR2434318

González-Rivera, G. and Yoldas, E. (2012). Autocontour-based evaluation of multivariate predictive densities. *Int. J. Forecasting* **28** 328–342.

Gudendorf, G. and Segers, J. (2010). Extreme-value copulas. In *Copula Theory and Its Applications*, (P. Jaworski, F. Durante, W. K. Härdle and T. Rychlik, eds.). *Lecture Notes in Statistics – Proceedings* **198** 127–145. Springer Berlin Heidelberg. MR3051266

Huber, P. J. (1985). Projection pursuit. *Ann. Stat.* **13** 435–475. MR0790553

Ishida, I. (2005). Scanning multivariate conditional densities with probability integral transforms. Center for Advanced Research in Finance, University of Tokio, Working Paper F-045.

Leutbecher, M. and Palmer, T. N. (2008). Ensemble forecasting. *J. Computat. Phys.* **227** 3515–3539. MR2400226

McNeil, A. J. and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. *Ann. Stat.* **37** 3059–3097. MR2541455

Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed. Springer, New York. MR2197664

Nelsen, R. B., Quesada-Molina, J. J., Rodríguez-Lallena, J. A. and Úbeda Flores, M. (2003). Kendall distribution functions. *Stat. Prob. Lett.* **65** 263–268. MR2018039

Pinson, P. (2013). Wind energy: Forecasting challenges for its operational management. *Stat. Sci.* **28** 564–585. MR3161588

Pinson, P. and Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Appl. Energy* **96** 12–20.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0.

Röpnack, A., Hense, A., Gebhardt, C. and Majewski, D. (2013). Bayesian model verification of NWP ensemble forecasts. *Mon. Wea. Rev.* **141** 375–387.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Stat.* **23** 470–472. MR0049525

Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *J. Stat. Plann. Infer.* **139** 3921–3927. MR2553778

Schaake, J., Pailleux, J., Thielen, J., Arritt, R., Hamill, T., Luo, L., Martin, E., McCollor, D. and Pappenberger, F. (2010). Summary of recommendations of the first workshop on Postprocessing and Downscaling Atmospheric Forecasts for Hydrologic Applications held at Météo-France, Toulouse, France, 15–18 June 2009. *Atmos. Sci. Let.* **11** 59–63.

Schefzik, R., Thorarinsdottir, T. L. and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.* **28** 616–640. MR3161590

Schuhen, N., Thorarinsdottir, T. L. and Gneiting, T. (2012). Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.* **140** 3204–3219.

Smith, L. A. and Hansen, J. A. (2004). Extending the limits of ensemble forecast verification with the minimum spanning tree histogram. *Mon. Wea. Rev.* **132** 1522–1528.

Tawn, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika* **75** 397–415. MR0967580

Thorarinsdottir, T. L., Scheuerer, M. and Heinz, C. (2014). Assessing the calibration of high-dimensional ensemble forecasts using the rank histogram. *J. Comput. Graph. Stat.*, in press, DOI: 10.1080/10618600.2014.977447.

Wilks, D. S. (2004). The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Mon. Wea. Rev.* **132** 1329–1340.