Statistical Science
2013, Vol. 28, No. 4, 616–640
DOI: 10.1214/13-STS443

Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling

Roman Schefzik, Thordis L. Thorarinsdottir and Tilmann Gneiting

Abstract. Critical decisions frequently rely on high-dimensional output from complex computer simulation models that show intricate cross-variable, spatial and temporal dependence structures, with weather and climate predictions being key examples. There is a strongly increasing recognition of the need for uncertainty quantification in such settings, for which we propose and review a general multi-stage procedure called ensemble copula coupling (ECC), proceeding as follows:

1. Generate a raw ensemble, consisting of multiple runs of the computer model that differ in the inputs or model parameters in suitable ways.
2. Apply statistical postprocessing techniques, such as Bayesian model averaging or nonhomogeneous regression, to correct for systematic errors in the raw ensemble, to obtain calibrated and sharp predictive distributions for each univariate output variable individually.
3. Draw a sample from each postprocessed predictive distribution.
4. Rearrange the sampled values in the rank order structure of the raw ensemble to obtain the ECC postprocessed ensemble.

The use of ensembles and statistical postprocessing have become routine in weather forecasting over the past decade. We show that seemingly unrelated, recent advances can be interpreted, fused and consolidated within the framework of ECC, the common thread being the adoption of the empirical copula of the raw ensemble. Depending on the use of Quantiles, Random draws or Transformations at the sampling stage, we distinguish the ECC-Q, ECC-R and ECC-T variants, respectively. We also describe relations to the Schaake shuffle and extant copula-based techniques. In a case study, the ECC approach is applied to predictions of temperature, pressure, precipitation and wind over Germany, based on the 50-member European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble.

Key words and phrases: Bayesian model averaging, empirical copula, ensemble calibration, nonhomogeneous regression, numerical weather prediction, probabilistic forecast, Schaake shuffle, Sklar's theorem.

1. INTRODUCTION

In a vast range of applications, critical decisions depend on the output of complex computer simulation models, with examples including weather and climate

Roman Schefzik is Ph.D. Student and Tilmann Gneiting is Professor, Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany (e-mail: r.schefzik@uni-heidelberg.de; t.gneiting@uni-heidelberg.de). Thordis L. Thorarinsdottir is Senior Research Scientist, Norwegian Computing Center, P.O. Box 114, Blindern, 0314 Oslo, Norway (e-mail: thordis@nr.no).

predictions and the management of floods, wildfires, air quality and groundwater contaminations. There is a much increased recognition of the need for quantifying the uncertainty in the model output, as evidenced by the creation of pertinent American Statistical Association (ASA) and Society for Industrial and Applied Mathematics (SIAM) interest groups, and by the recent launch of the SIAM/ASA Journal on Uncertainty Quantification. As SIAM President Nick Trefethen (2012) notes succinctly,

> "An answer that used to be a single number may now be a statistical distribution."

Frequently, the goal is prediction, and we are witnessing a transdisciplinary change of paradigms in the transition from deterministic or point forecast to probabilistic or distributional forecasts (Gneiting, 2008). The goal is to obtain calibrated and sharp, joint predictive distributions of future quantities of interest, from which any desired functionals, such as event probabilities, moments, quantiles and prediction intervals can be extracted, for a full quantification of the predictive uncertainty. In this context, calibration refers to the statistical compatibility of the probabilistic forecasts and the observations, in that events predicted to occur with probability p ought to realize with empirical frequency p. Sharpness refers to the concentration of the predictive distributions and is a property of the probabilistic forecasts only (Gneiting, Balabdaoui and Raftery, 2007). While our data examples all concern weather forecasting, where the recognition of the need for uncertainty quantification can be traced at least to Cooke (1906), the methods and principles we discuss apply in much broader contexts, both predictive and in other settings, where one seeks to quantify the uncertainty in our incomplete knowledge of current or past quantities and events.

Focusing attention on the setting of our case study, accurate predictions of future weather are of considerable value for society. Medium-range weather forecasts, with lead times up to two weeks, are obtained by numerically solving the partial differential equations that describe the physics of the atmosphere, with initial conditions provided by estimates of the current state of the atmosphere (Kalnay, 2003). In order to account for the uncertainties in the forecast, national and international meteorological centers use ensembles of numerical weather prediction (NWP) model output, where the ensemble members differ in terms of the two major sources of uncertainty, namely, the initial conditions and the parameterization of the NWP model (Palmer, 2002; Gneiting and Raftery, 2005). To give an example, Figures 1 and 2 illustrate forecasts of surface temperature and six-hour precipitation accumulation over

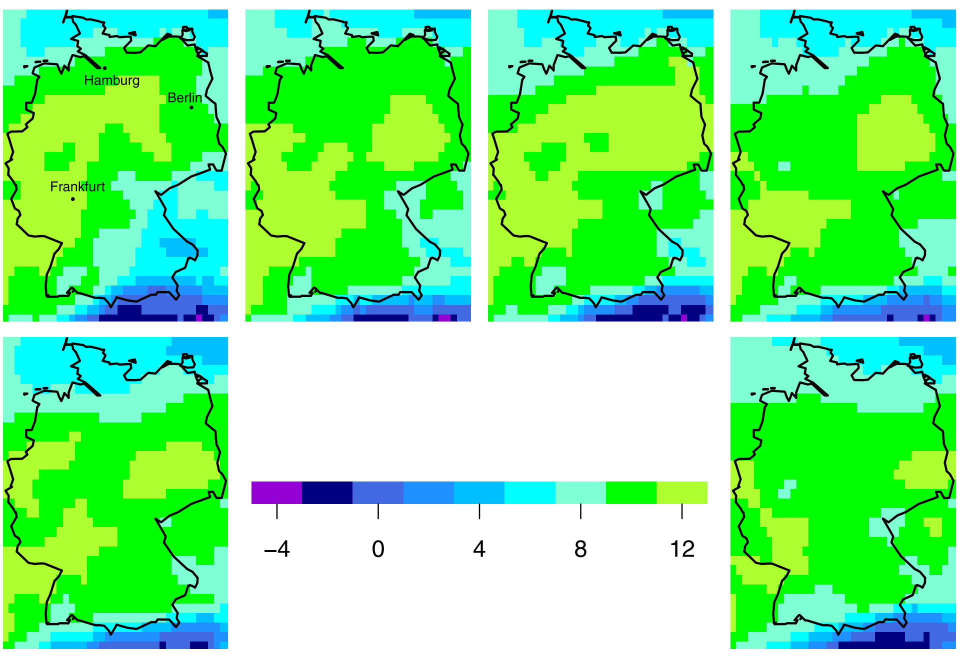


FIG. 1. 48-*hour ahead ECWMF ensemble forecast for temperature over Germany valid* 2:00 *am on April* 1, 2011, *in the unit of degrees Celsius. Six randomly selected members are shown. The top left panel shows the locations of the three stations used in the subsequent case study.*

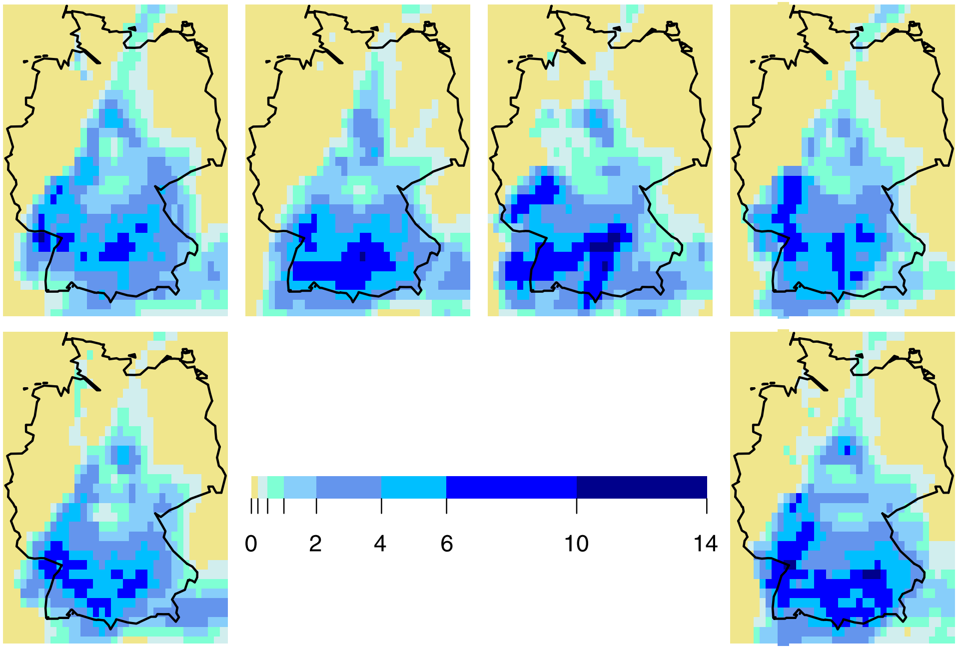


FIG. 2. *24-hour ahead ECWMF ensemble forecast for six-hour precipitation accumulation over Germany valid* 2:00 *am on May* 20, 2010, *in the unit of millimeters. Six randomly selected members are shown.*

Germany issued by the European Centre for Medium-Range Weather Forecasts (ECMWF) as a part of its 50-member real-time ensemble, which operates at a horizontal resolution of approximately 32 km and lead times up to ten days (Molteni et al., 1996; Leutbecher and Palmer, 2008). The valid time of these forecasts is 00:00 Universal Time Coordinated (UTC) in meteorological format, which we convert to local time in what follows.

While the goal of NWP ensemble systems is to capture the inherent uncertainty in the prediction, they are subject to systematic errors, such as biases and dispersion errors. It is therefore common practice to statistically postprocess the output of NWP ensemble forecasts, with state of the art techniques including the ensemble Bayesian model averaging (BMA) approach developed by Raftery et al. (2005) and the nonhomogeneous regression (NR) or ensemble model output statistics (EMOS) technique proposed by Gneiting et al. (2005).

To illustrate the idea, let y denote the weather quantity of interest, such as temperature at a specific location and look-ahead time, and write $x_1, \dots, x_M$ for the corresponding M ensemble member forecasts. The ensemble BMA approach employs mixture distributions of the general form

$$y|x_1, \dots, x_M \sim \sum_{m=1}^{M} w_m f(y|x_m),$$

where the left-hand side refers to the conditional distribution given the ensemble member forecasts. Here $f(y|x_m)$ denotes a parametric probability distribution or kernel that depends on the ensemble member forecast x_m in suitable ways, with the mixture weights $w_1, \dots, w_M$ reflecting the members' relative contributions to predictive skill over a training period. BMA postprocessed predictive distributions based on the 50-member ECMWF ensemble are illustrated in Figure 3 for temperature, where the kernel is normal and the postprocessing corrects for both a low bias and underdispersion, and in Figure 4 for precipitation, where the kernel comprises a point mass at zero along with a power transformed gamma distribution for positive accumulations.

In contrast, the NR predictive distribution is a single parametric distribution of the general form

$$y|x_1, \dots, x_M \sim g(y|x_1, \dots, x_M),$$

where g is a parametric distribution function with location, scale and shape parameters depending on the ensemble values in suitable ways. For example, g could be normal with the mean an affine function of the ensemble member forecasts and the variance an affine function of the ensemble variance.

Statistical postprocessing techniques such as ensemble BMA and NR have been shown to substantially improve the predictive skill of the NWP ensemble output

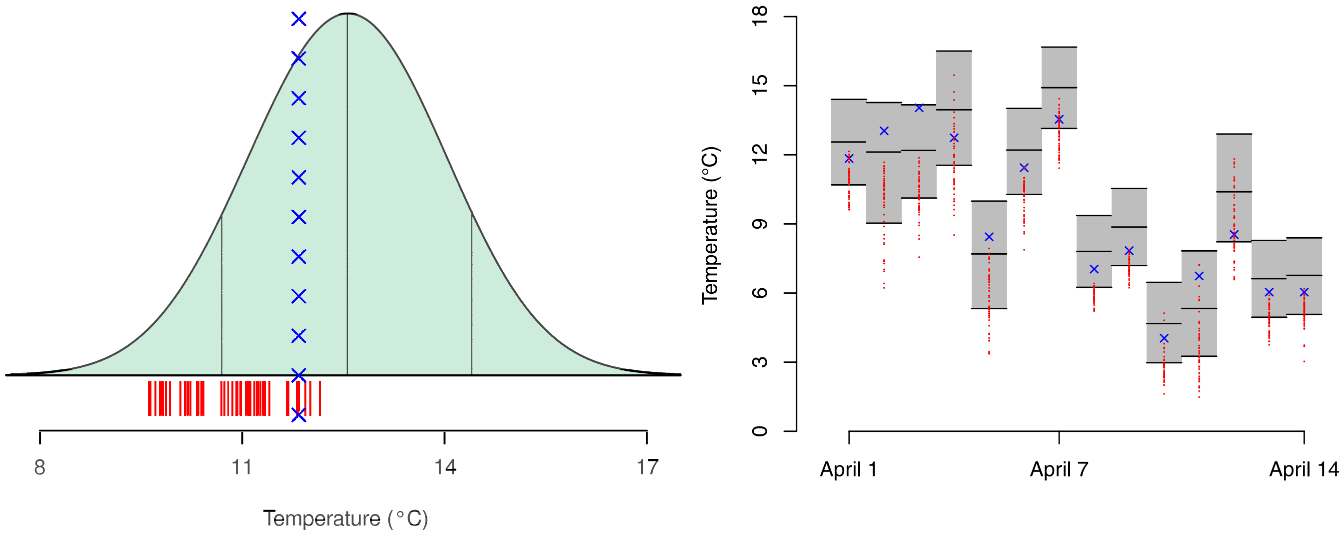


FIG. 3. 48*-hour ahead BMA postprocessed predictive distributions for temperature in Berlin based on the* 50*-member ECMWF ensemble. The ensemble forecast is shown in red, the realizing observation in blue. Left*: *predictive density valid* 2:00 *am on April* 1, 2011. *Right*: 10*th*, 50*th and* 90*th percentiles of the predictive distributions valid* 2:00 *am on April* 1–14, 2011.

(Wilks and Hamill, 2007; Hagedorn et al., 2012). Frequently, such methods apply to each weather variable at each location and each lead time individually and, therefore, they may fail to take cross-variable, spatial and temporal interactions properly into account. NWP models rely on discretizations of the equations that govern the physics of the atmosphere and, thus, multivariate dependence structures tend to be reasonably well represented in the raw ensemble system. However, these structures may fail to be retained if the univariate margins are postprocessed individually. In low-dimensional or highly structured settings, parametric approaches to the modeling of multivariate dependence structures in the forecast errors are feasible, such as in the recent work of Pinson (2012), Schuhen, Thorarinsdottir and Gneiting (2012) and Sloughter, Gneiting and Raftery (2013) on wind vectors, or in the approach of Gel, Raftery and Gneiting (2004) and Berrocal, Raftery

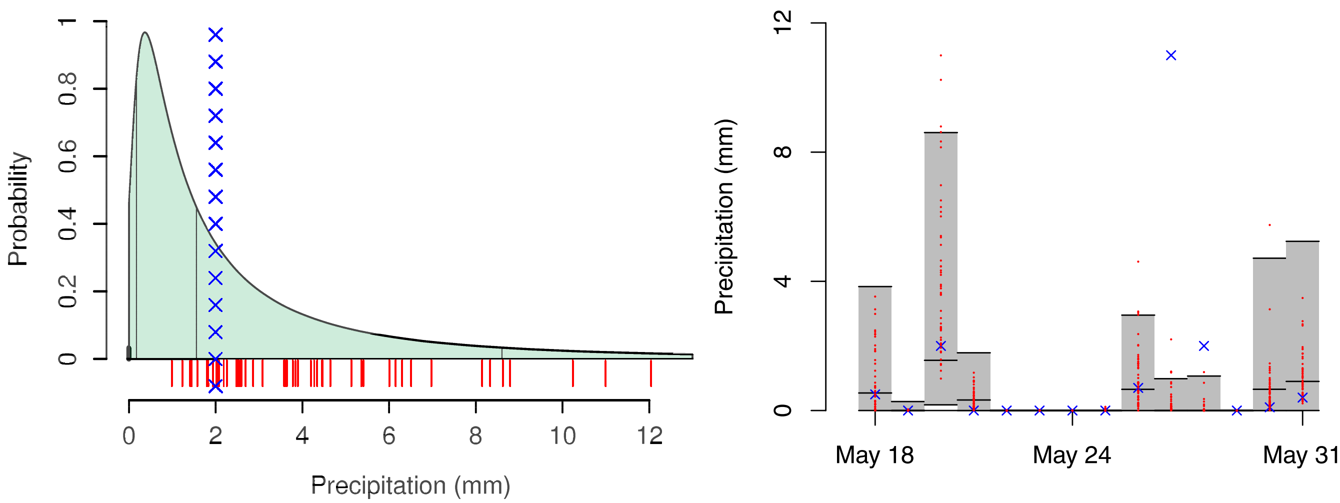


FIG. 4. 24*-hour ahead BMA postprocessed predictive distributions for six-hour precipitation accumulation in Frankfurt based on the* 50*-member ECMWF ensemble. The ensemble forecast is shown in red, the realizing observation in blue. Left*: *mixed discrete-continuous predictive distribution valid* 2:00 *am on May* 20, 2010, *comprising a point mass of* 0.033 *at zero, which is indicated by the thick black bar, and a density at positive accumulations, with mass* 0.967. *Right*: 10*th*, 50*th and* 90*th percentiles of the predictive distribution distributions valid* 2:00 *am on May* 18–31, 2010.

and Gneiting (2007) that relies on geostatistical models in spatial settings.

However, the statistical postprocessing of a full NWP ensemble forecast poses extremely high-dimensional problems. For instance, we might be interested in five weather variables at 500×500 grid boxes, ten vertical levels and 72 lead times, for a total of 900 million variables. While not all of them may need to be considered simultaneously, critical applications, such as air traffic control (Chaloulos and Lygeros, 2007), air quality (Delle Monache et al., 2006) and flood management (Cloke and Pappenberger, 2009; Schaake et al., 2010), depend on physically realistic probabilistic forecasts of spatio-temporal weather trajectories and therefore may entail much higher dimensions than can readily be incorporated into a parametric model.

To address this challenge, we propose and review a general multi-stage procedure called ensemble copula coupling (ECC), originally hinted at by Bremnes (2007) and Krzysztofowicz and Toth (2008), and recently investigated and developed by Schefzik (2011). The ECC approach allows for the multivariate rank dependence structure of the raw NWP ensemble to be preserved in the postprocessed ensemble, proceeding roughly as follows.

Univariate postprocessing. Apply statistical postprocessing techniques, such as ensemble BMA or NR, to obtain calibrated and sharp marginal predictive distributions for each weather variable, location and look-ahead time individually.

Quantization. Draw a discrete sample of the same size as the raw ensemble from each univariate, postprocessed predictive distribution.

Ensemble reordering. Arrange the sampled values in the rank order structure of the raw ensemble to obtain the ECC postprocessed ensemble.

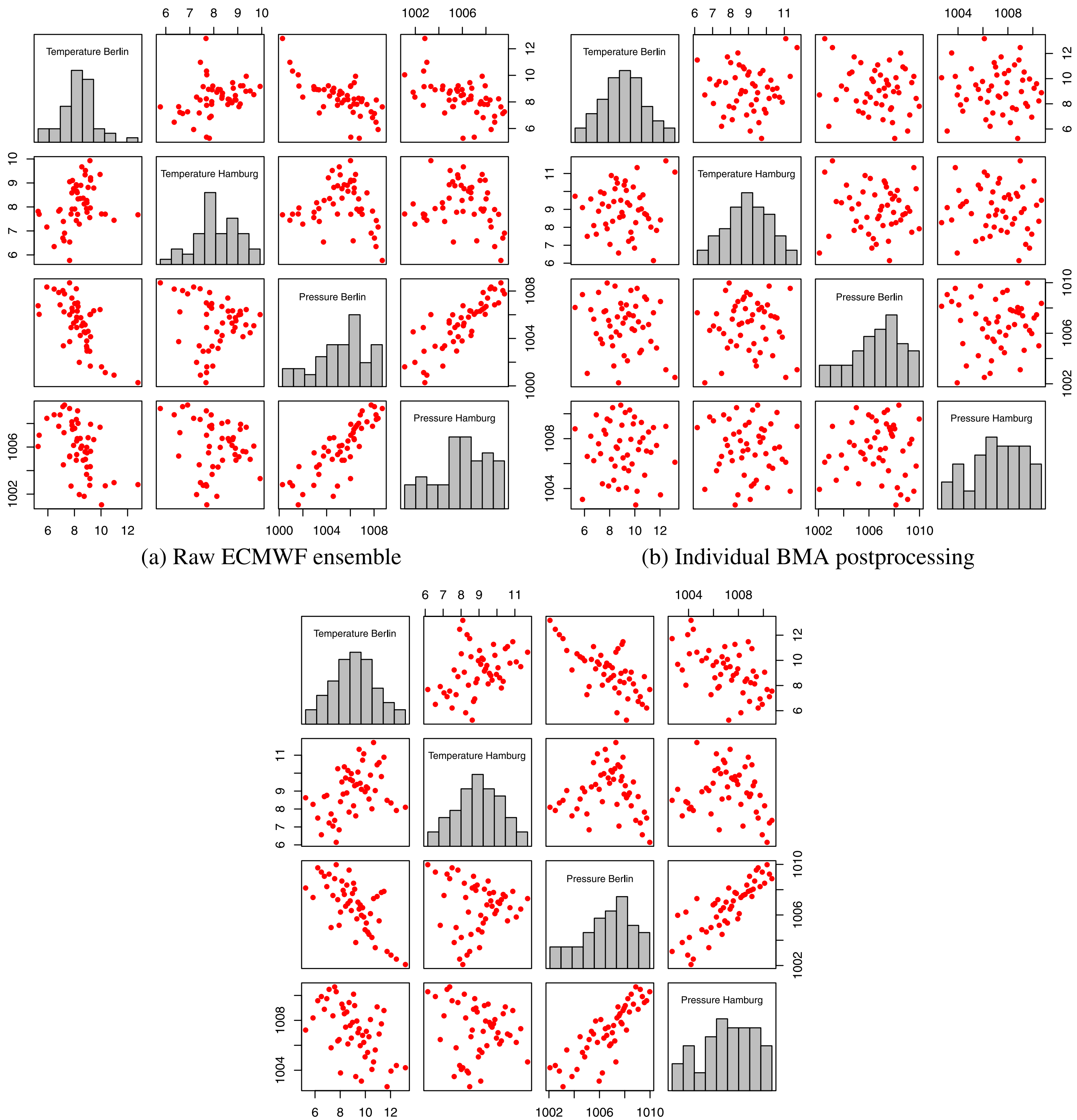
An illustration of the ECC approach is given in Figure 5, a dynamic version of which is available in the supplementary material (Schefzik, Thorarinsdottir and Gneiting, 2013). Here, the setting is four dimensional. We consider surface temperature and sea level pressure in Berlin and Hamburg, respectively. The scatterplot matrix in the top panel illustrates the 50-member ECMWF ensemble forecast at a 24 hours lead time. Clearly, there are dependencies between the margins; for example, there is a positive association between temperature in Berlin and temperature in Hamburg, and there are negative associations between temperature and pressure. The scatterplot matrix in the middle panel is constructed from samples of the individually BMA postprocessed predictive distributions. Here, the systematic errors in the margins have been corrected, at the cost of a loss of the error dependence structure. The bottom panel elucidates the effects of the ECC ensemble reordering; while the margins remain unchanged from the middle panel, the rank dependence structure of the raw ensemble is restored.

Owing to the intuitive appeal and striking simplicity, which incurs essentially no computational costs beyond the marginal postprocessing, approaches of ECC type are rapidly gaining prominence at weather centers worldwide, with variants recently having been implemented by Flowerdew (2012), Pinson (2012) and Roulin and Vannitsem (2012), among others. Our goal here is to interpret, fuse and consolidate these and other seemingly unrelated advances within the framework of ECC. As we will demonstrate, the common thread of the approaches lies in the adoption of the empirical copula of the raw ensemble, thereby restoring its rank dependence structure and justifying the term ensemble copula coupling.

The remainder of the paper is organized as follows. In Section 2 we review and discuss statistical postprocessing techniques for univariate NWP ensemble output. General copula approaches to the handling of multivariate output are discussed in Section 3, with subsequent focus on the ECC approach in Section 4, where we distinguish the ECC-Q, ECC-R and ECC-T variants, depending on the use of Quantiles, Random draws or Transformations at the quantization stage. Section 5 turns to a case study on probabilistic predictions of temperature, pressure, precipitation and wind over Germany, based on the ECMWF ensemble. The paper closes with Section 6, where we discuss benefits and limitations of the ECC approach and return to the general theme of uncertainty quantification for high-dimensional output from complex simulation models with intricate dependence structures.

2. UNIVARIATE POSTPROCESSING: BAYESIAN MODEL AVERAGING (BMA) AND NONHOMOGENEOUS REGRESSION (NR)

Following the pioneering work of Hamill and Colucci (1997), various types of statistical postprocessing techniques for the output of NWP ensemble forecasts have been developed, with Wilks and Hamill (2007), Bröcker and Smith (2008), Schmeits and Kok (2010) and Ruiz and Saulo (2012) providing critical reviews. As noted, postprocessing aims to correct for



(a) Raw ECMWF ensemble (b) Individual BMA postprocessing

(c) ECC postprocessed ensemble

FIG. 5. *24-hour ahead ensemble forecasts of temperature and pressure at Berlin and Hamburg*, *valid* 2:00 *am on May* 27, 2010. *The units used are degrees Celsius and hPa.*

biases and dispersion errors in the ensemble output, and state-of-the-art techniques can roughly be divided into mixture approaches, building on the ensemble Bayesian model averaging (BMA) approach of Raftery et al. (2005), and regression approaches, such as the nonhomogeneous regression (NR) method put forth by Gneiting et al. (2005).

Specifically, consider a univariate weather quantity of interest, y, and write $x_1, \ldots, x_M$ for the corresponding M ensemble member forecasts. As noted, the en-

TABLE 1
Ensemble BMA implementations for univariate weather quantities. In the case of precipitation amount, we refer to $y^{1/3} \in \mathbb{R}^+$, because the gamma kernels apply to cube root transformed precipitation accumulations. In the case of wind direction, $\mathbb{S}$ denotes the circle, z_m is a bias-corrected ensemble member value on the circle, and κ_m is a concentration parameter, for $m = 1, \ldots, M$

Weather quantity	Range	Kernel (f)	Mean	Variance
Temperature	$y \in \mathbb{R}$	Normal	$a_m + b_m x_m$	σ_m^2
Pressure	$y \in \mathbb{R}$	Normal	$a_m + b_m x_m$	σ_m^2
Precipitation amount	$y^{1/3} \in \mathbb{R}^+$	Gamma	$a_m + b_m x_m^{1/3}$	$c_m + d_m x_m$
Wind speed	$y \in \mathbb{R}^+$	Gamma	$a_m + b_m x_m$	$c_m + d_m x_m$
Wind direction	$y \in \mathbb{S}$	von Mises	z_m	κ_m^{-1}
Visibility	$y \in [0, 1]$	Beta	$a_m + b_m x_m^{1/2}$	$c_m + d_m x_m^{1/2}$

semble BMA approach uses mixture distributions of the general form

$$y|x_1, \ldots, x_M \sim \sum_{m=1}^{M} w_m f(y|x_m), \tag{2.1}$$

where the left-hand side refers to the conditional distribution of y given the ensemble member forecasts $x_1, \ldots, x_M$, and $f(y|x_m)$ is a parametric distribution that depends on x_m only.[1] The mixture weights $w_1, \ldots, w_m$ are nonnegative and sum to 1; they reflect the corresponding member's relative contributions to predictive skill over a training period. In contrast, the NR predictive distribution is a single parametric distribution of the general form

$$y|x_1, \ldots, x_M \sim g(y|x_1, \ldots, x_M), \tag{2.2}$$

where the right-hand side refers to a parametric family of probability distributions, with the parameters depending on all ensemble members simultaneously.

The particular choice of a parametric model for the BMA kernel f or the NR distribution g depends on the weather quantity at hand. Table 1 sketches ensemble BMA implementations for temperature and pressure (Raftery et al., 2005), where the kernel $f(y|x_m)$ is normal with mean $a_{0m} + a_{1m} x_m$ and variance σ_m^2, precipitation (Sloughter et al., 2007), wind speed (Sloughter, Gneiting and Raftery, 2010), wind direction (Bao et al., 2010) and visibility (Chmielecki and Raftery, 2011). Furthermore, ensemble BMA implementations are available for fog (Roquelaure and Bergot, 2008), visibility and ceiling (Chmielecki and Raftery, 2011). Frequently, the parameters in the specifications for the mean and the variance of the kernels are subject to constraints; for example, the variance parameters are often assumed to be constant across ensemble members. If the ensemble is generated in such a way that its members are statistically indistinguishable or exchangeable, as in the case of the ECMWF ensemble, the BMA weights as well as the BMA mean and variance parameters are assumed to be constant across ensemble members (Fraley, Raftery and Gneiting, 2010). Table 2 hints at NR implementations for temperature and pressure (Gneiting et al., 2005), where the postprocessed predictive distribution is normal with mean $a + b_1 x_1 + \cdots + b_M x_M$ and variance $c + dS^2$ where S^2 is the ensemble variance, for precipitation (Wilks, 2009; Scheuerer, 2013) and for wind speed (Thorarinsdottir and Gneiting, 2010; Thorarinsdottir and Johnson, 2012).

TABLE 2
NR implementations for univariate weather quantities. In the case of precipitation amount, we refer to the distinct approaches of Wilks (2009) and Scheuerer (2013)

Weather quantity	Range	Distribution (g)
Temperature	$y \in \mathbb{R}$	Normal
Pressure	$y \in \mathbb{R}$	Normal
Precipitation amount	$y \in \mathbb{R}^+$	Truncated logistic
	$y \in \mathbb{R}^+$	Generalized extreme value
Wind components	$y \in \mathbb{R}$	Normal
Wind speed	$y \in \mathbb{R}^+$	Truncated normal

[1] In the case of ensembles with nonexchangeable members the distribution f might depend on member specific statistical parameters. Furthermore, in some implementations f might depend on observed variables or on NWP model output for quantities other than y, such as in the approach of Glahn et al. (2009). Similar comments apply to the NR technique.

In the remainder of this section we provide a detailed description of the postprocessing methods for the weather variables temperature, pressure, precipitation and wind which are analyzed in our case study. Generally, the ensemble BMA method is more flexible, while the NR technique is more parsimonious. In terms of the predictive performance, the general experience is that the BMA and NR approaches yield comparable results. Software for estimation and prediction is available in the form of the `ensembleBMA` (Fraley et al., 2011) and `ensembleMOS` packages in `R`.[2]

2.1 Temperature and Pressure

For the weather variables temperature and pressure, Raftery et al. (2005) propose the ensemble BMA specification

$$y|x_1, \dots, x_M \sim \sum_{m=1}^{M} w_m \mathcal{N}(a_m + b_m x_m, \sigma_m^2), \tag{2.3}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The BMA weights $w_1, \dots, w_M$, the mean parameters $a_1, \dots, a_M$ and $b_1, \dots, b_M$, and the variance parameters $\sigma_1^2, \dots, \sigma_M^2$, which in the standard implementation are assumed to be constant across ensemble members, are estimated on training data. This type of mixture approach has been applied successfully at weather centers worldwide,[3] and we give an example in Figure 3.

Gneiting et al. (2005) propose an NR approach for temperature and pressure, in which the predictive distribution is normal,

$$\begin{aligned} &y|x_1, \dots, x_M \\ &\quad \sim \mathcal{N}(a + b_1 x_1 + \cdots + b_M x_M, c + dS^2), \end{aligned} \tag{2.4}$$

where $S^2 = \sum_{m=1}^{M} (x_m - \bar{x})^2 / M$ denotes the ensemble variance. If the ensemble members are exchangeable, it needs to be assumed that $b_1 = \cdots = b_M$. This approach has also been applied at weather centers internationally, as exemplified in the work of Hagedorn, Hamill and Whitaker (2008) and Kann et al. (2009).

[2]These packages are available for download at www.r-project.org.

[3]A real-time ensemble BMA implementation for predictions of temperature and precipitation over the Pacific Northwest region of the United States is available to the general public at www.probcast.com, based on the University of Washington mesoscale ensemble in the form described by Eckel and Mass (2005).

2.2 Precipitation

While of critical applied importance, probabilistic forecasts for quantitative precipitation pose technical challenges, in that the predictive distribution is mixed discrete-continuous, comprising both a point mass at zero and a density on the positive real axis, which might be considerably skewed.

Sloughter et al. (2007) propose an ensemble BMA model of the general form (2.1) for precipitation accumulation, where the kernel $f(y|x_m)$ is a Bernoulli–Gamma mixture. The Bernoulli component provides a point mass at zero via a logistic regression link, in that

$$\begin{aligned} \operatorname{logit} f[y = 0|x_m] &= \log \frac{f[y = 0|x_m]}{f[y > 0|x_m]} \\ &= \alpha_m + \beta_m x_m^{1/3} + \gamma_m \delta_m, \end{aligned} \tag{2.5}$$

where δ_m equals 1 if $x_m = 0$ and equals 0 otherwise. The continuous part of the kernel is a gamma distribution in terms of the cube root transformation, $y^{1/3}$, of the precipitation accumulation, so that

$$\begin{aligned} f(y^{1/3}|x_m) &= f[y = 0|x_m] \mathbb{1}_{\{y=0\}} \\ &\quad + f[y > 0|x_m] h(y^{1/3}|x_m) \mathbb{1}_{\{y>0\}}, \end{aligned} \tag{2.6}$$

where h denotes a gamma distribution with mean μ_m and variance σ_m^2, with

$$\mu_m = a_m + b_m x_m^{1/3} \quad \text{and} \quad \sigma_m^2 = c_m + d_m x_m, \tag{2.7}$$

and where $\mathbb{1}_A$ denotes the indicator function of the event A. Figure 4 shows an example of the resulting BMA postprocessed predictive distribution in terms of the nontransformed precipitation accumulation, y.

Turning to the NR approach, we follow Roulin and Vannitsem (2012) and interpret the logistic regression technique of Wilks (2009) in this setting. To put the method into context, forecasts for the probability of the precipitation amount exceeding a certain threshold have commonly been obtained using either quantile regression (Bremnes, 2004) or logistic regression (Wilks and Hamill, 2007; Hamill, Hagedorn and Whitaker, 2008). If a full predictive distribution is sought, such methods frequently fail, as they typically are inconsistent across thresholds, violating the monotonicity constraint for cumulative distribution functions. For quantile regression, Dette and Volgushev (2008) and Kneib (2013) describe possible solutions to this problem. In the case of the logistic regression approach, Wilks (2009) proposes an elegant remedy. In his method, the

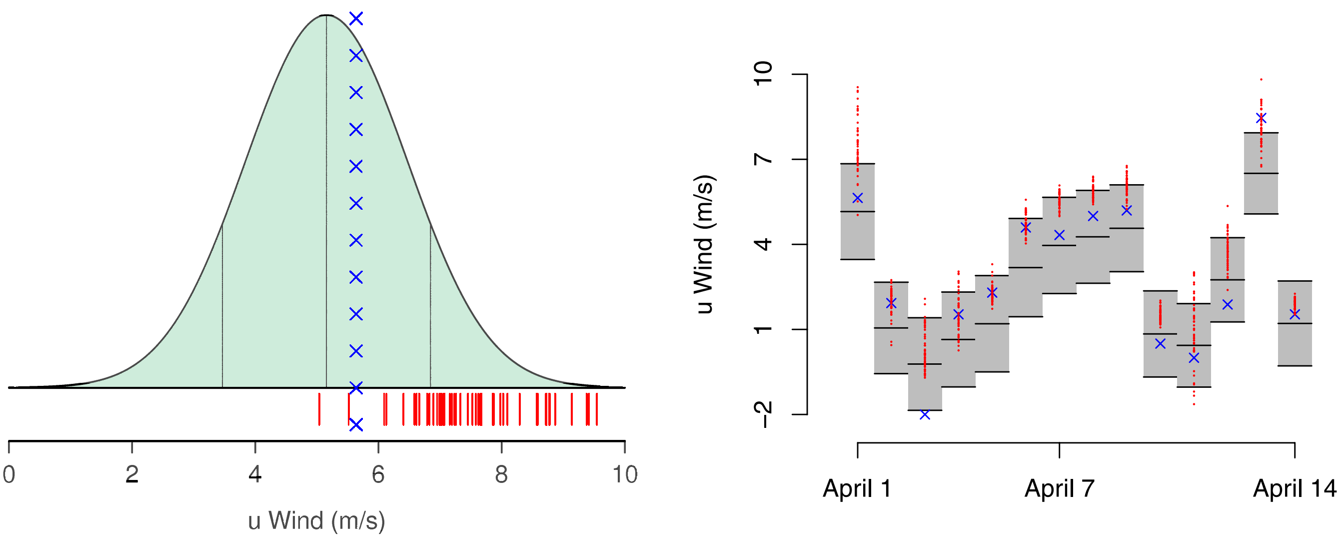


FIG. 6. *24-hour ahead NR postprocessed predictive distributions for the u wind component at Hamburg based on the* 50*-member ECMWF ensemble. The ensemble forecast is shown in red, the realizing observation in blue. Left*: *predictive density valid* 2:00 *am on April* 1, 2011. *Right*: 10*th*, 50*th and* 90*th percentiles of the predictive distributions valid* 2:00 *am on April* 1–14, 2011.

postprocessed predictive cumulative distribution function takes the form

$$G(y|x_1, \dots, x_M) = \frac{\exp(a + b_1 x_1 + \cdots + b_m x_M + h(y))}{1 + \exp(a + b_1 x_1 + \cdots + b_m x_M + h(y))}, \tag{2.8}$$

where h grows strictly monotonically and without bounds as a function of the precipitation accumulation $y \geq 0$. Linear choices for h result in mixtures of a point mass at zero and a truncated logistic distribution and, in light of the parametric family in (2.8), the technique can be interpreted as an NR approach. More general formulations that allow for interaction terms have recently been proposed by Ben Bouallègue (2013). As an alternative, Scheuerer (2013) introduces an NR approach in terms of generalized extreme value (GEV) distributions.

2.3 Wind

A wind vector can be represented by wind speed and wind direction or by its u (zonal or west–east) and v (meridional or north–south) velocity components. Wind speed is a nonnegative continuous variable. Sloughter, Gneiting and Raftery (2010) provide an ensemble BMA implementation, where the kernel is a gamma distribution with the mean and the variance being affine functions of the respective ensemble member forecast. Thorarinsdottir and Gneiting (2010) and Thorarinsdottir and Johnson (2012) develop an NR approach in which the predictive distribution is truncated normal. Wind direction is a circular quantity and Bao et al. (2010) propose an ensemble BMA specification where the kernel is a von Mises distribution.

When a wind vector is represented by its u and v components, the methods described in Section 2.1 for temperature and pressure become available, and examples of NR postprocessed predictive distributions of the form (2.4) for the u component are shown in Figure 6. In recent work, truly bivariate postprocessing techniques for wind vectors have become available, taking dependencies between the components into account (Pinson, 2012; Schuhen, Thorarinsdottir and Gneiting, 2012; Sloughter, Gneiting and Raftery, 2013). These methods are discussed in subsequent sections.

2.4 Estimation

Ensemble postprocessing techniques depend on the availability of training data for estimating the predictive model. Typically, optimum score approaches have been used for estimation (Gneiting et al., 2005), with the maximum likelihood technique being a special case thereof (Gneiting and Raftery, 2007), and Bayesian approaches offering alternatives (Di Narzo and Cocchi, 2010).

The training data are usually taken from a rolling training period consisting of the recent past, including the most recent available ensemble forecasts along with the corresponding realizing values. Common choices for the length of the training period range from 20 to 40 days. In schemes of this type, the training set is updated continually, thereby allowing the estimates to adapt to changes in the seasons and weather regimes. Clearly, there is a trade-off here, in that larger training periods may allow for better estimation in principle, thereby reducing estimation variances, but may introduce biases due to seasonal effects. More flexible, adaptive estimation approaches, such as recursive maximum likelihood techniques, have been proposed and studied by Pinson et al. (2009), Raftery, Kárný and Ettler (2010) and Pinson (2012).

In addition to deciding on the temporal extent of training sets, choices regarding their spatial composition are to be made. Local approaches use training data from the station location or grid box at hand only, resulting in distinct sets of coefficients that are tailored to the local terrain, while regional approaches composite training sets spatially, to estimate a single set of coefficients that is then used over an entire region (Thorarinsdottir and Gneiting, 2010). Recently, flexible spatially adaptive approaches have been developed that estimate coefficients at each station location individually, interpolating them to sites where no observational assets are available (Kleiber et al., 2011; Kleiber, Raftery and Gneiting, 2011).

Introduced by Hamill, Whitaker and Mullen (2006), reforecasts are retrospective weather forecasts with today's NWP models applied to past initialization and valid dates. As reforecasts are based on the model version that is currently run operationally, the availability of reforecast data sets results in massive enlargements of training sets for statistical postprocessing. The ensuing gains in the predictive performance can be substantial, as demonstrated by Hagedorn, Hamill and Whitaker (2008), Hamill, Hagedorn and Whitaker (2008) and Hagedorn et al. (2012), among others.

3. FROM UNIVARIATE TO MULTIVARIATE PREDICTIVE DISTRIBUTIONS: COPULA APPROACHES

The univariate postprocessing methods discussed thus far yield significant improvement in the predictive performance of raw NWP ensemble output. However, in many applications it is critical that multivariate dependencies in the forecast error, including the case of temporal, spatial and spatio-temporal weather trajectories, are accounted for. For example, winter road maintenance requires joint probabilistic forecasts of temperature and precipitation (Berrocal et al., 2010), air traffic control calls for probabilistic forecasts of wind fields (Chaloulos and Lygeros, 2007), the management of renewable energy resources hinges on spatio-temporal weather trajectories (Pinson, 2013), and NWP output is used to drive hydrologic models to address tasks such as flood warnings, the operation of waterways and releases from reservoirs, with Schaake et al. [(2010), pages 61–62] noting in this context that

> "relationships between physically dependent variables like, for example, precipitation and temperature should be respected."

If statistical postprocessing proceeds independently for each weather variable, location and look-ahead time, such relationships are ignored, and it is critical that they be restored.

Toward this end, we recall Sklar's theorem, which is of fundamental theoretical importance in dependence modeling, and we review Gaussian and other parametric copulas approaches to the statistical postprocessing of multivariate ensemble output. Then we turn to empirical copulas, which permit the adoption of a rank order structure from data records, as exemplified by the Schaake shuffle technique of Clark et al. (2004).

3.1 Handling Dependencies: Sklar's Theorem

Taking a technical perspective momentarily, suppose that we have a postprocessed predictive cumulative distribution function, F_l, for each univariate weather quantity Y_l, where $l = 1, \ldots, L$, with the multi-index $l = (i, j, k)$ referring to weather variable i, location j and look-ahead time k. What we seek is a physically realistic multivariate joint predictive cumulative distribution function F with margins $F_1, \ldots, F_L$.

Recall that a copula is a multivariate cumulative distribution function with standard uniform margins (Joe, 1997; Nelsen, 2006). Copulas have been employed successfully in a wealth of applications, such as in finance (McNeil, Frey and Embrechts, 2005), hydrology (Genest and Favre, 2007) and climatology (Schoelzel and Friederichs, 2008), to name but a few. Their relevance stems from the following celebrated theorem of Sklar (1959).

THEOREM 3.1 (Sklar). *For any multivariate cumulative distribution function F with margins $F_1, \ldots, F_L$ there exists a copula C such that*

$$F(y_1, \ldots, y_L) = C\big(F_1(y_1), \ldots, F_L(y_L)\big) \tag{3.1}$$

for $y_1, \ldots, y_L \in \mathbb{R}$. Furthermore, C is unique on the range of the margins.

In particular, Sklar's theorem demonstrates that univariate approaches to the statistical postprocessing of ensemble output can accommodate any type of joint dependence structure, provided that a suitable copula function is specified. As copula methods allow for the modeling of the marginal distributions and of the multivariate dependence structure, as embodied by the copula, to be decoupled, they are well suited for our problem.

3.2 Gaussian and Other Parametric Copula Approaches

If the dimension L of the output quantity is small, or if specific structure can be exploited, such as in spatial or temporal settings, parametric or semiparametric families of copulas can be employed.

The most common parametric approaches invoke a Gaussian copula framework, under which the multivariate cumulative distribution function F is of the form

$$\begin{aligned} &C(y_1, \dots, y_L | \Sigma) \\ &\quad = \Phi_L(\Phi^{-1}(F_1(y_1)), \dots, \Phi^{-1}(F_L(y_L)) | \Sigma), \end{aligned} \tag{3.2}$$

where $\Phi_L(\cdot|\Sigma)$ is the cumulative distribution function of an L-variate normal distribution with mean zero and correlation matrix Σ, and Φ^{-1} is the quantile function of the univariate standard normal distribution. The use of Gaussian copulas makes for a particularly tractable approach, as only the correlation matrix Σ needs to be modeled. In a recent paper, Möller, Lenkoski and Thorarinsdottir (2013) propose the use of Gaussian copulas to recover the cross-variable dependence structure for multi-variable forecasts at individual locations, where the ensemble BMA methodology is used to obtain the postprocessed marginal predictive distributions. The method is straightforward except that precipitation requires special treatment due to the mixed discrete-continuous nature of the variable. The recent work of Pinson (2012) and Schuhen, Thorarinsdottir and Gneiting (2012) on bivariate wind vectors invokes multivariate normal predictive distributions, corresponding to the special case in (3.2) in which the margins $F_1, \dots, F_L$ are normal.

The use of Gaussian copula methods has a long and well-established tradition in geostatistics, where the approach is referred to as anamorphosis; see Chilès and Delfiner (2012) and the references therein. In the spatial setting, the correlation matrix Σ in (3.2) is taken to be highly structured, satisfying assumptions such as spatial stationarity and/or isotropy, as exemplified by Gel, Raftery and Gneiting (2004) and Berrocal, Raftery and Gneiting (2007, 2008) in ensemble BMA approaches to temperature and precipitation field forecasting. Similarly, Gaussian copulas have been employed to capture dependencies over consecutive lead times in postprocessed predictive distributions (Pinson et al., 2009; Schoelzel and Hense, 2011). When the margins $F_1, \dots, F_L$ are normal, the underlying stochastic model is that of a Gaussian process or Gaussian random field, and choices in the parameterization of the correlation matrix Σ correspond to the selection of a parametric correlation model in spatial statistics (Stein, 1999; Cressie and Wikle, 2011).

While Gaussian copulas yield convenient, ubiquitous stochastic models, parametric or semiparametric alternatives are available, including but not limited to the use of elliptical copulas (Demarta and McNeil, 2005), Archimedian copulas (McNeil and Nešlehová, 2009), extremal copulas (Davison, Padoan and Ribatet, 2012) and pair copulas (Aas et al., 2009).

3.3 Empirical Copulas

In the common case in which the dimension L of the output quantity is huge and no specific structure can be exploited, parametric methods are bound to fail. We then need to resort to nonparametric approaches that depend on the use of empirical copulas. Here, let $\{(x_m^1, \dots, x_m^L) : m = 1, \dots, M\}$ denote a data set of size M with values in $\mathbb{R}^L$. Assuming for simplicity that there are no ties, let $\text{rk}(x_m^l)$ denote the rank of x_m^l within $x_1^l, \dots, x_M^l$. The corresponding empirical copula E_M is defined as

$$\begin{aligned} &E_M\left(\frac{i_1}{M}, \dots, \frac{i_L}{M}\right) \\ &\quad = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{\text{rk}(x_m^1) \le i_1, \dots, \text{rk}(x_m^L) \le i_L\} \end{aligned} \tag{3.3}$$

for integers $0 \le i_1, \dots, i_L \le M$; see Deheuvels (1979), who uses the term empirical dependence function, and Rüschendorf (2009) and the references therein.

Any empirical copula is an irreducible discrete copula in the sense described by Kolesárová et al. (2006), with Mayor, Suñer and Torrens (2007) providing a bivariate version of Sklar's theorem in this setting. As we will illustrate below, empirical copulas can be thought of as corresponding to Latin hypersquares. Asymptotic theory for the respective empirical processes has been developed by Rüschendorf (1976, 2009), Stute (1984), van der Vaart and Wellner (1996), Fermanian, Radulović and Wegkamp (2004) and Segers (2012), among other authors.

In the context of nonparametric approaches to the statistical postprocessing of multivariate NWP ensemble output, empirical copulas allow for the adoption of a multivariate rank order structure either from historical weather observations, as in the Schaake shuffle technique of Clark et al. (2004), or directly from the ensemble forecast, to be discussed in detail in Section 4.

3.4 The Schaake Shuffle

Clark et al. (2004) introduced the ingenious Schaake shuffle as a method for reconstructing physically realistic spatio-temporal structure in forecasted temperature and precipitation fields. Even though it has been presented as a reordering technique in the extant literature, an empirical copula interpretation of the Schaake shuffle is readily available.

Consider an output quantity taking values in $\mathbb{R}^L$ and suppose that we have univariate postprocessed predictive distributions $F_1, \ldots, F_L$ for the margins. Suppose, furthermore, that we have a set of M historical weather field observations for the $\mathbb{R}^L$-valued output quantity at hand. From the historical record, we can construct an empirical copula of the form (3.3), as illustrated in the right-hand panel of Figure 7, where we merely have $L = 2$ as corresponds to the components of a wind vector and $M = 20$.

To apply the Schaake shuffle, we take a discrete sample of size M from each of the univariate postprocessed predictive distributions $F_1, \ldots, F_L$, and then we reorder to match with the rank order structure in the historical record, which is also of size M. This procedure corresponds to the application of the empirical copula of the historical weather field record to the discrete samples from the univariate postprocessed predictive distribution, and in this sense it is natural to consider the Schaake shuffle as an empirical copula technique. The thus reordered forecast inherits the multivariate rank dependence structure and the pairwise Spearman rank correlation coefficients from the historical weather record at hand. A more technical discussion can be given in close analogy to what we describe in Section 4.2 within the related context of the ensemble copula coupling approach.

The Schaake shuffle has met great success in meteorological and hydrologic applications, where it recovers observed spatial and cross-variable dependence structures as well as temporal persistence (Clark et al., 2004; Schaake et al., 2007; Voisin et al., 2011). Nevertheless, there is a major limitation, in that the standard implementation fails to condition the multivariate dependence structure on current or predicted atmospheric conditions. Clark et al. [(2004), page 260] therefore describe a future extension of the Schaake shuffle, the idea of which is as follows:

> "to preferentially select dates from the historical record that resemble forecasted atmospheric conditions and use the spatial correlation structure from this subset of dates to reconstruct the spatial variability for a specific forecast."

In what follows we pursue a related empirical copula approach, in which the postprocessed forecast inherits the multivariate dependence structure from the raw NWP ensemble, rather than from a historical record of weather observations, thereby addressing the lack of atmospheric flow and time dependence in the standard Schaake shuffle.

4. ENSEMBLE COPULA COUPLING (ECC)

The ensemble copula coupling (ECC) approach draws on the rank order information available in the

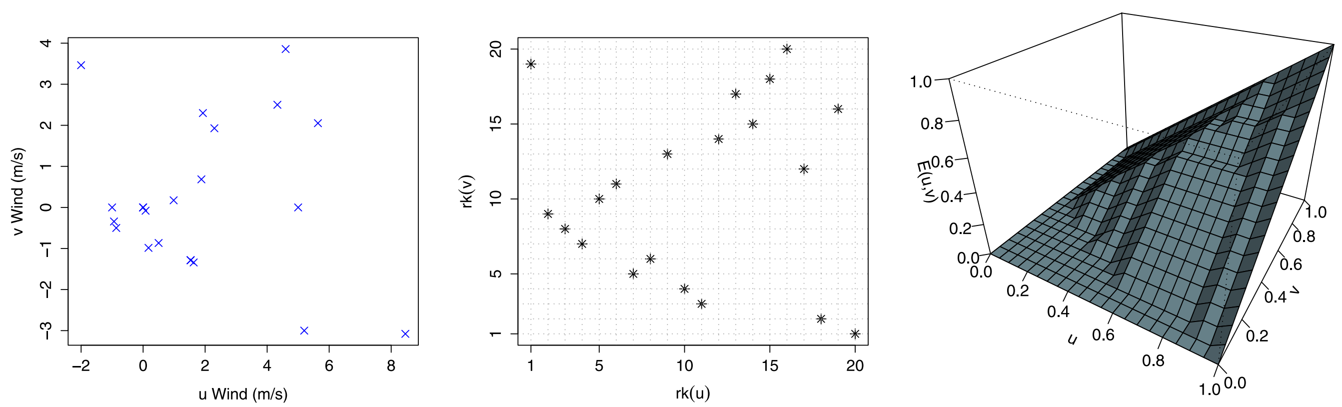


FIG. 7. *The bivariate empirical distribution of the observed u and v wind components at Hamburg at* 2:00 *am on April* 1–20, 2011. *Left*: *bivariate scatterplot*. *Middle*: *representation of the rank dependence structure by a Latin square*. *Right*: *empirical copula*.

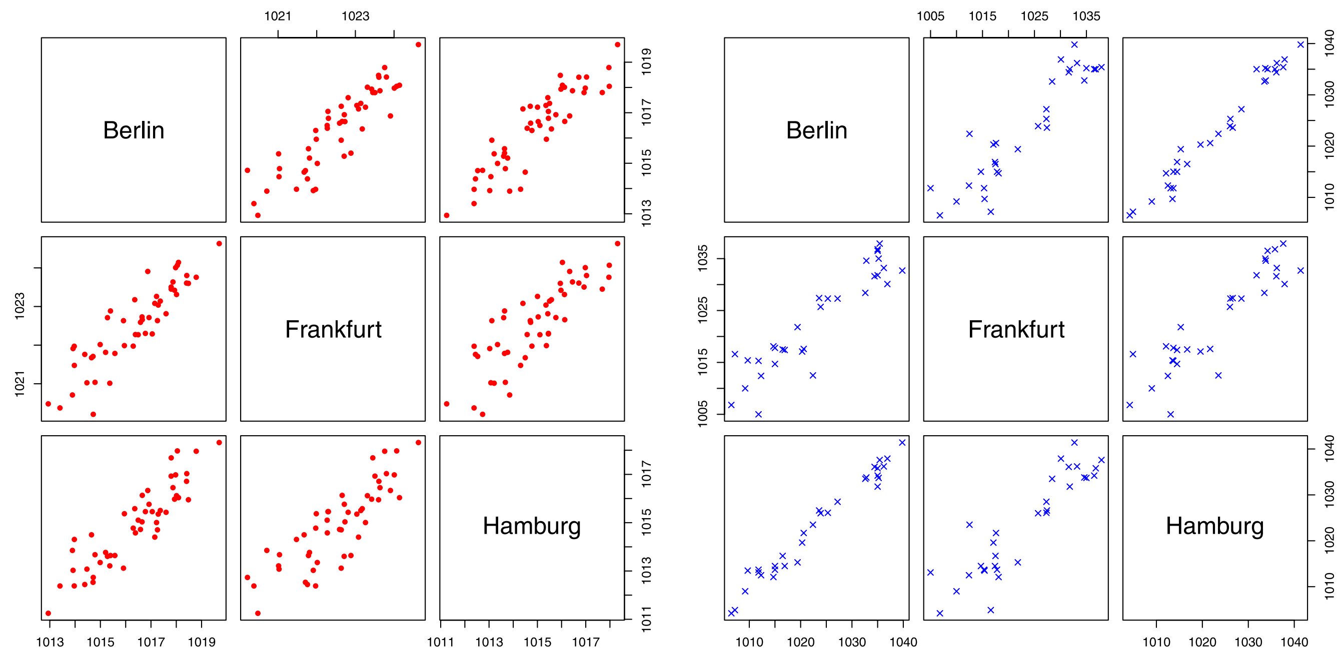


FIG. 8. *Scatterplot matrices for pressure at Berlin, Frankfurt and Hamburg. Left*: 48-*hour ahead ECMWF ensemble forecast valid* 2:00 *am on April* 1, 2011. *Right*: *empirical distribution of the pressure observations at the same hour over the period March* 1–31, 2011.

raw ensemble forecast, based on the implicit assumptions that its members are exchangeable and that the NWP ensemble is capable of representing observed cross-variable, spatial and temporal dependence structures. While the latter is to be expected, given that NWP models discretize the equations that govern the physics of the atmosphere, diagnostic checks are advisable, to assess empirically whether dependence structures in individual ensemble forecasts are compatible with observational records. We give a simple illustration in Figure 8, where the dependence structures within the ensemble forecast valid April 1, 2011 and those in the observational record over the preceding month resemble each other strongly.

4.1 The ECC Approach

The ECC approach is a general multi-stage procedure for the generation of a postprocessed ensemble of the same size, M, as the raw ensemble. We write $x_1^l, \ldots, x_M^l$ for the univariate margins of the raw ensemble, where the multi-index $l = (i, j, k)$ refers to weather variable $i \in \{1, \ldots, I\}$, location $j \in \{1, \ldots, J\}$ and lead time $k \in \{1, \ldots, K\}$, to comprise NWP output in $\mathbb{R}^L$, where the dimension is $L = I \times J \times K$. In order to generate an ECC postprocessed ensemble forecast, we proceed as follows.

Univariate postprocessing. For each margin l, obtain a postprocessed predictive distribution, F_l, by applying a univariate postprocessing technique, such as ensemble BMA or NR, to the raw ensemble output

$$x_1^l, \ldots, x_M^l. \tag{4.1}$$

Quantization. Represent each univariate predictive distribution F_l by a discrete sample of size M, say,

$$\tilde{x}_1^l, \ldots, \tilde{x}_M^l. \tag{4.2}$$

The discrete sample can be generated in various ways, to be discussed in detail in Section 4.3, where we distinguish the ECC-Q, ECC-R and ECC-T variants, depending on how the quantization is performed.[4]

Ensemble reordering. For each margin l, the order statistics[5] of the raw ensemble values,

$$x_{(1)}^l \le \cdots \le x_{(M)}^l$$

induce a permutation σ_l of the integers $\{1, \ldots, M\}$, defined by $\sigma_l(m) = \mathrm{rk}(x_m^l)$ for $m = 1, \ldots, M$. If there

[4]Note that the quantized values in (4.2) may be ordered, as in the case of the ECC-Q approach, or may not be ordered, as in the case of the ECC-R and ECC-T scheme, respectively.

[5]The kth order statistic of a sample is defined as its kth smallest value. For each margin l, we write $x_{(1)}^l \le \cdots \le x_{(M)}^l$ and $\tilde{x}_{(1)}^l \le \cdots \le \tilde{x}_{(M)}^l$ for the order statistics of the raw ensemble values in (4.1) and the quantized values in (4.2), respectively. The latter appear on the right-hand side of (4.3), where we define the ECC postprocessed ensemble.

are ties among the ensemble values, the corresponding ranks can be allocated at random.[6] The respective margin of the ECC postprocessed ensemble is then given by

$$\hat{x}_1^l = \tilde{x}^l_{(\sigma_l(1))}, \ldots, \hat{x}_M^l = \tilde{x}^l_{(\sigma_l(M))}. \tag{4.3}$$

Note that, while the permutation σ_l is determined by the order statistics of the raw ensemble, equation (4.3) applies this permutation to the postprocessed and quantized values.

The ECC approach is attractive computationally, in that the modeling of the multivariate dependence structure requires only the calculation of marginal ranks. In the recent literature, the approach has been introduced as a reordering technique, as described colorfully by Flowerdew (2012), page 15:

> "The key to preserving spatial, temporal and inter-variable structure is how this set of values is distributed between ensemble members. One can always construct ensemble members by sampling from the calibrated PDF, but this alone would produce spatially noisy fields lacking the correct correlations. Instead, the values are assigned to ensemble members in the same order as the values from the raw ensemble: the member with the locally highest rainfall remains locally highest, but with a calibrated rainfall magnitude."

That said, it is fruitful to interpret the ECC approach as a nonparametric copula technique, which permits us to fuse and consolidate seemingly unrelated, recent advances within a single, structured framework.

4.2 Empirical Copula Interpretation

Elaborating on our interpretation of the Schaake shuffle, we now demonstrate that the ECC approach can be considered an empirical copula technique. For convenience, we assume that there are no ties among the raw ensemble margins. We write $R_1, \ldots, R_L$ for the corresponding marginal empirical cumulative distribution functions, which take values in the set

$$I_M = \left\{0, \frac{1}{M}, \ldots, \frac{M-1}{M}, 1\right\}.$$

[6]While randomization is a natural approach in the case of ties, other allocation methods are feasible and do not pose technical problems. Regardless of the allocation, equation (3.3) continues to apply.

The multivariate empirical cumulative distribution function $R : \mathbb{R}^L \to I_M$ of the raw ensemble maps into I_M, too. According to the discrete version of Sklar's theorem described by Mayor, Suñer and Torrens (2007) in the bivariate case, there exists a uniquely determined empirical copula $E_M : I_M^L \to I_M$ such that

$$R(y_1, \ldots, y_L) = E_M\big(R_1(y_1), \ldots, R_L(y_L)\big) \tag{4.4}$$

for all $y_1, \ldots, y_L \in \mathbb{R}$, allowing for the same type of interpretation as illustrated in Figure 7 in the case of the Schaake shuffle.

Analogous considerations apply to the quantized independently postprocessed ensemble (4.2) and the ECC postprocessed ensemble (4.3). Using obvious notation, we write $\tilde{F}$ and $\hat{F}$ for the corresponding multivariate empirical cumulative distribution functions. Furthermore, we denote the marginal empirical cumulative distribution functions of the quantized independently postprocessed ensemble by $\tilde{F}_1, \ldots, \tilde{F}_L$, respectively, and we use the symbol $\tilde{E}_M$ to denote the corresponding copula. Then

$$\tilde{F}(y_1, \ldots, y_L) = \tilde{E}_M\big(\tilde{F}_1(y_1), \ldots, \tilde{F}_L(y_L)\big) \tag{4.5}$$

and

$$\hat{F}(y_1, \ldots, y_L) = E_M\big(\tilde{F}_1(y_1), \ldots, \tilde{F}_L(y_L)\big) \tag{4.6}$$

for all $y_1, \ldots, y_L \in \mathbb{R}$. As elucidated by equations (4.4), (4.5) and (4.6), the quantized independently postprocessed ensemble and the ECC postprocessed ensemble share the margins, whereas the raw ensemble and the ECC postprocessed ensemble share the copula, as illustrated in Figure 5. In particular, the ECC postprocessed ensemble honors and retains the flow-dependent multivariate rank dependence structure and bivariate Spearman rank correlation coefficients in the raw NWP ensemble output.

4.3 ECC-Q, ECC-R and ECC-T

We now discuss options for the generation of the discrete samples (4.2) at the quantization stage of the ECC approach. Perhaps the most natural way of obtaining a discrete sample of size M from the postprocessed predictive cumulative distribution function F_l is to take equidistant Quantiles of the form

$$\tilde{x}_1^l = F_l^{-1}\left(\frac{1}{M+1}\right), \ldots, \tilde{x}_M^l = F_l^{-1}\left(\frac{M}{M+1}\right), \tag{ECC-Q}$$

and we refer to this approach as ECC-Q.[7] Another option is to take a simple Random sample of the form

$$\text{(ECC-R)} \qquad \tilde{x}_1^l = F_l^{-1}(u_1), \ldots, \tilde{x}_M^l = F_l^{-1}(u_M),$$

where $u_1, \ldots, u_M$ are independent standard uniform random variates. We refer to this latter option as ECC-R.

Finally, we consider a quantile mapping or transformation approach that generalizes a recent proposal by Pinson (2012) in the case of wind vectors. In this technique, we adopt the ensemble smoothing approach of Wilks (2002) and fit a parametric, continuous cumulative distribution function S_l to the raw ensemble margin R_l. We then extract the quantiles from F_l that correspond to the percentiles of the raw ensemble values in S_l, in that

$$\tilde{x}_1^l = F_l^{-1}(S_l(x_1^l)), \ldots, \tilde{x}_M^l = F_l^{-1}(S_l(x_M^l)).$$

(ECC-T)

We refer to this Transformation approach for continuous variables as ECC-T. Frequently, as in the case of temperature, pressure and the u and v wind vector components, S_l can be taken to be normal, with mean equal to the ensemble mean and variance equal to the ensemble variance. In the special situation in which S_l and F_l belong to the same location-scale family, such that $S_l(x) = G((x - \mu)/\sigma)$ and $F_l(x) = G((x - \tilde{\mu})/\tilde{\sigma})$ for some continuous cumulative distribution function G, $\mu, \tilde{\mu} \in \mathbb{R}$ and $\sigma, \tilde{\sigma} > 0$, the transformation from x to

$$\text{(4.7)} \qquad \tilde{x} = F_l^{-1}(S_l(x)) = \tilde{\mu} + \frac{\tilde{\sigma}}{\sigma}(x - \mu)$$

becomes affine and, thus, the ECC-T postprocessed ensemble conserves the raw ensemble's bivariate Pearson product moment correlation coefficients, in addition to retaining its bivariate Spearman rank correlation coefficients.

[7] Bröcker (2012) provides theoretical arguments in support of the particular choice of the quantiles in (ECC-Q), which maintains the calibration of the univariate ensemble forecasts, well in line with the goal of maximizing the sharpness of the predictive distributions subject to calibration (Gneiting, Balabdaoui and Raftery, 2007). An alternative choice would be to set

$$\tilde{x}_1^l = F_l^{-1}\left(\frac{1/2}{M}\right), \tilde{x}_2^l = F_l^{-1}\left(\frac{3/2}{M}\right), \ldots, \tilde{x}_M^l = F_l^{-1}\left(\frac{M - 1/2}{M}\right),$$

which fails to maintain calibration in some respects, but is optimal in expectation if the predictive performance is measured by the continuous ranked probability score (Bröcker, 2012). Related optimality results can be found in the literature on the quantization of probability distributions as reviewed by Graf and Luschgy (2000).

The discussion in Bröcker (2012) provides theoretical support in favor of the ECC-Q approach, and so does our case study in Section 5.3, where we compare the predictive performance of the ECC-Q, ECC-R and ECC-T schemes. We therefore recommend the use of the natural ECC-Q approach.

4.4 Relationships to Extant Work

While the broad framework and the interpretation in terms of empirical copulas in our paper are original, the idea of the ECC approach is not new, with its recent appearances in the literature coming in various seemingly unrelated shades and flavors. In this context, the connections to the work of Pinson (2012) and Roulin and Vannitsem (2012) are of particular interest.

The method described in Section 2.c of Roulin and Vannitsem (2012) in the context of areal precipitation forecasts can be viewed as a variant of the ECC-Q scheme, as it extracts equally spaced quantiles from the postprocessed marginal predictive cumulative distribution functions, which are of logistic type, followed by a reordering with respect to the raw ensemble values, with adaptations to account for a point mass at zero.

Pinson (2012) proposes a transformation technique for the postprocessing of ensemble forecasts of wind vector components. In this method, each postprocessed margin is a translated and dilated version of the original margin, with the mapping being compatible with the ECC-T scheme in the special case in which both S_l and F_l are normal.

5. CASE STUDY

In this case study we exemplify the use of statistical postprocessing techniques, illustrate and assess the ECC approach, and compare the predictive performance of the ECC-Q, ECC-R and ECC-T schemes, respectively. All forecasts are based on the 50-member global NWP ensemble managed by the European Centre for Medium-Range Weather Forecasts (ECMWF), which operates at a horizontal resolution of approximately 32 km and lead times up to ten days ahead (Molteni et al., 1996; Leutbecher and Palmer, 2008). The differences between the ensemble members stem from random perturbations in initial conditions and stochastic physics parameterizations and, thus, the ensemble members are statistically indistinguishable and can be considered as exchangeable.

5.1 Setting

We restrict attention to the ECMWF ensemble run initialized at 00:00 Universal Time Coordinated (UTC) and consider forecasts for surface temperature, sea level pressure, precipitation and the u wind vector component at lead times of 24 and 48 hours, with emphasis on the international airports at Berlin–Tegel, Frankfurt am Main and Hamburg in Germany, where 00:00 UTC corresponds to 2:00 am local time in summer and 1:00 am local time in winter. The locations of the three airports are marked in the upper left panel in Figure 1. Our test period consists of the twelve month period ranging from May 1, 2010 through April 30, 2011. Forecasts and observations prior to May 1, 2010 are used as training data as needed.

To obtain postprocessed marginal predictive distributions for each weather variable, location and lead time individually, we apply the techniques described in Section 2. For temperature and pressure, we employ the ensemble BMA model (2.3) with a normal kernel, and for precipitation the Bernoulli–Gamma ensemble BMA model specified in (2.5), (2.6) and (2.7), respectively. For the wind vector components, we use the NR model (2.4). To fit the univariate predictive models, we use local data from a rolling training period consisting of the most recent available 30 days and employ the estimation techniques proposed by Raftery et al. (2005), Sloughter et al. (2007) and Gneiting et al. (2005). Then we apply the ECC-Q, ECC-R and ECC-T schemes as described in Section 4.

5.2 Evaluation Methods

Statistical postprocessing techniques aim at generating calibrated and sharp probabilistic forecasts from NWP ensemble output. As argued by Gneiting, Balabdaoui and Raftery (2007), the goal in probabilistic forecasting is to maximize the sharpness of the predictive distributions subject to calibration. Calibration is a multi-faceted, joint property of the forecasts and the observations; essentially, the forecasts are calibrated if the observations can be interpreted as random draws from the predictive distributions. Sharpness refers to the concentration of the predictive distributions, and thus is a property of the forecasts only.

In univariate settings, calibration is checked via the probability integral transform (PIT) or the verification rank. The PIT is simply the value that the predictive cumulative distribution function attains at the realizing observation (Dawid, 1984; Gneiting, Balabdaoui and Raftery, 2007), with suitable adaptations in the case of discrete distributions (Czado, Gneiting and Held, 2009). For an ensemble forecast, the verification rank is the rank of the realizing observation when pooled with the ensemble values (Hamill, 2001). When a predictive distribution is calibrated, the PIT or verification rank is uniformly distributed. Thus, calibration can be diagnosed by compositing over forecast cases, plotting a PIT or verification rank histogram, respectively, and checking for deviations from uniformity. Verification rank and PIT histograms are directly comparable, with a U-shape indicating underdispersion, an inverse U-shape indicating overdispersion, and skew pointing at biases in the predictive distributions.

Proper scoring rules provide decision theoretically coherent numerical measures of predictive performance that may assess calibration and sharpness simultaneously. Here we use the proper continuous ranked probability score (CRPS), defined by

$$\text{crps}(F, y) = \int_{-\infty}^{\infty} \big(F(z) - \mathbb{1}\{y \le z\}\big)^2 \, dz \tag{5.1}$$

$$= \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'|, \tag{5.2}$$

where F is a predictive cumulative distribution function with finite first moment, y is the verifying observation, and X and X' are independent random variables with distribution F (Gneiting and Raftery, 2007). If F corresponds to a point measure δ_x, the proper continuous ranked probability score reduces to the absolute error, $|x - y|$. If $F = F_{\text{ens}}$ is an ensemble forecast with members $x_1, \dots, x_M \in \mathbb{R}$, we interpret it as an empirical measure and compute the continuous ranked probability score as

$$\text{crps}(F_{\text{ens}}, y) = \frac{1}{M}\sum_{m=1}^{M}|x_m - y| - \frac{1}{2M^2}\sum_{n=1}^{M}\sum_{m=1}^{M}|x_n - x_m|. \tag{5.3}$$

We furthermore find the absolute error for the point forecast given by the median of the predictive distribution, which is the Bayes predictor under this loss function (Gneiting, 2011). Forecasting methods then are compared by averaging scores over the test set, with smaller values indicating better predictive performance.

To assess the calibration of ensemble forecasts of a multivariate quantity, we use the multivariate version of the rank histogram described by Gneiting et al. (2008). We also employ the proper energy score, which generalizes the continuous ranked probability score in the representation (5.2), and is defined as

$$\text{es}(F, y) = \mathbb{E}_F\|X - y\| - \tfrac{1}{2}\mathbb{E}_F\|X - X'\|, \tag{5.4}$$

where $\|\cdot\|$ denotes the Euclidean norm, F is a predictive distribution with finite first moments, X and X' are

TABLE 3

Mean continuous ranked probability score (CRPS) and mean absolute error (MAE) for univariate forecasts of temperature, pressure, precipitation and the u wind component at Berlin, Frankfurt and Hamburg, at lead times of 24 *and* 48 *hours, respectively, for a test period ranging from May* 1, 2010 *through April* 30, 2011

			CRPS			MAE		
			Berlin	**Frankfurt**	**Hamburg**	**Berlin**	**Frankfurt**	**Hamburg**
Temp.	24	ECMWF	1.21	1.23	1.01	1.50	1.53	1.26
(°C)		BMA	0.90	0.88	0.79	1.27	1.23	1.10
	48	ECMWF	1.25	1.26	1.06	1.62	1.62	1.39
		BMA	0.99	0.97	0.92	1.41	1.33	1.31
Pressure	24	ECMWF	0.54	0.55	0.51	0.75	0.75	0.71
(hPa)		BMA	0.43	0.43	0.39	0.62	0.61	0.54
	48	ECMWF	0.80	0.78	0.77	1.12	1.08	1.09
		BMA	0.77	0.74	0.73	1.08	1.03	1.03
Precip.	24	ECMWF	0.25	0.41	0.31	0.32	0.51	0.39
(mm)		BMA	0.23	0.40	0.37	0.30	0.49	0.44
	48	ECMWF	0.26	0.41	0.36	0.34	0.50	0.45
		BMA	0.26	0.43	0.39	0.32	0.52	0.48
u Wind	24	ECMWF	0.83	0.96	0.89	1.06	1.19	1.11
(m/s)		NR	0.70	0.60	0.68	0.97	0.81	0.96
	48	ECMWF	0.82	0.89	0.88	1.09	1.15	1.18
		NR	0.75	0.62	0.75	1.05	0.83	1.04

independent random vectors with distribution F, and y is the verifying observation (Gneiting and Raftery, 2007). For ensemble forecasts the natural analogue of the formula (5.3) applies. If the scales of the weather variables vary, the margins should be standardized before computing the joint energy score for these variables. This can be done using the marginal means and standard deviations of the observations in the test set.

The aforementioned techniques for the evaluation of probabilistic forecasts of multivariate quantities have been developed with low-dimensional quantities in mind (Gneiting et al., 2008), and we apply them in dimension $L \leq 3$ only. In higher dimension, these methods lose power, and there is a pronounced need for the development of theoretically principled evaluation techniques that are tailored to such settings (Pinson, 2013, Section 5.2).

5.3 Predictive Performance for Univariate Weather Quantities

Table 3 compares the predictive performance of the raw ECMWF ensemble and the postprocessed predictive distributions for temperature, pressure, precipitation and the u wind vector component at lead times of 24 and 48 hours at Berlin, Frankfurt and Hamburg, respectively. The BMA and NR postprocessing generally leads to a significant improvement in the predictive skill, as measured by the mean CRPS and the MAE, with exceptions in the case of precipitation.[8] Not unexpectedly, the performance generally is better at the shorter prediction horizon of 24 hours.

Figure 9 shows verification rank and PIT histograms for temperature, pressure, precipitation and u wind at a lead time of 48 hours at Frankfurt. The postprocessed forecasts show much better calibration, as evidenced by the nearly uniform PIT histograms, except perhaps in the case of precipitation, where a slight inverse U-shape of the PIT histogram may indicate overdispersion in the BMA postprocessed predictive distributions.

5.4 Predictive Performance for Multivariate Weather Quantities

We now give an illustration and initial evaluation of ECC postprocessed multivariate predictive distributions.

[8]The particularly good performance of the raw ensemble for precipitation accumulations at the stations considered and potential shortcomings in the details of the postprocessing technique (Scheuerer, 2013) may serve to explain these exceptions.

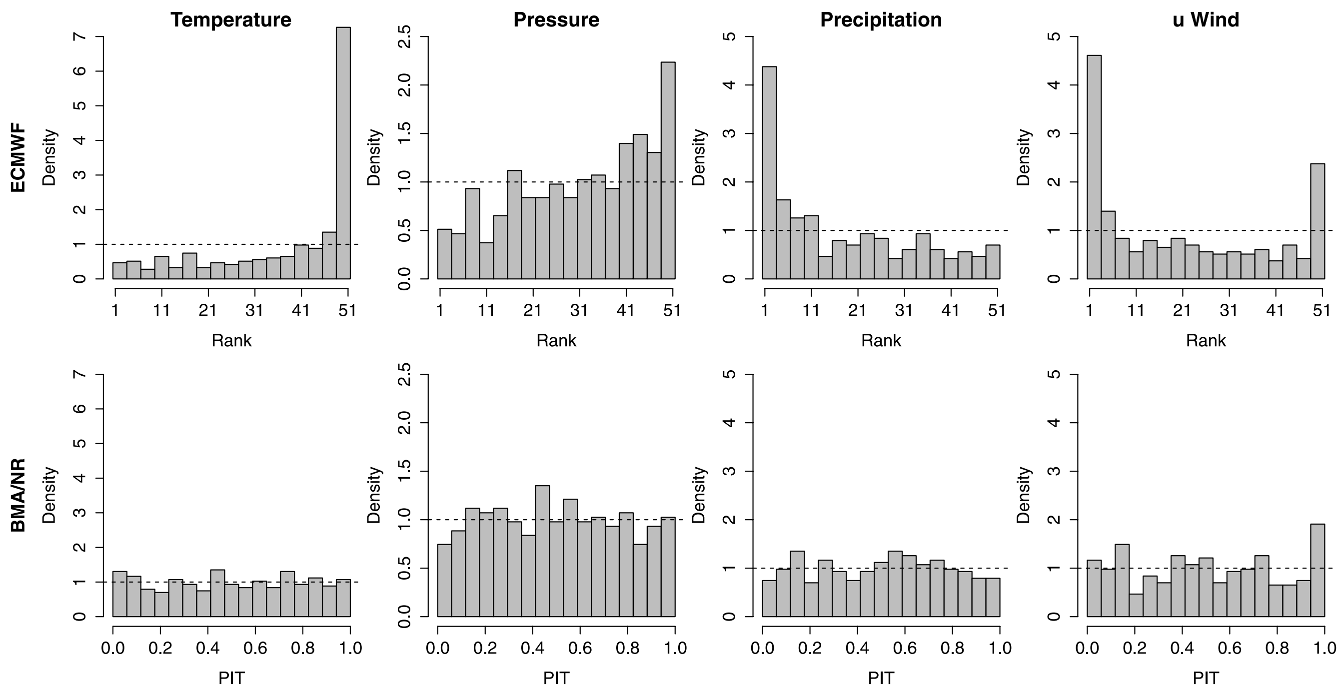


FIG. 9. *Calibration checks for* 48*-hour ahead forecasts of temperature*, *pressure*, *precipitation and u wind at Frankfurt*, *for a test period ranging from May* 1, 2010 *through April* 30, 2011. *Top*: *verification rank histograms for the ECMWF ensemble. Bottom*: *PIT histograms for BMA or NR postprocessed predictive distributions*.

Table 4 and Figure 10 concern temperature and pressure, with each of these variables being considered at Berlin, Frankfurt and Hamburg jointly. The distance from Frankfurt to either Berlin or Hamburg is on the order of 400 kilometers, and the distance between Berlin and Hamburg is approximately 250 kilometers. Wind and precipitation patterns vary at considerably smaller spatial scales and we thus do not expect ECC to make much of a difference here. In contrast, forecast errors for pressure can be expected to show pronounced long range dependencies, and perhaps to some lesser extent for temperature. The scores and multivariate rank histograms confirm the strongly positive effects of ECC in the case of pressure, where the ECC postprocessed trivariate predictive distributions are much better calibrated than either the raw ensemble or the independent BMA postprocessed predictive distributions. The ECC-Q quantization scheme outperforms the ECC-R and ECC-T approaches.

TABLE 4
Mean energy score for 48*-h ahead forecasts of temperature and pressure*, *each considered at Berlin*, *Frankfurt and Hamburg jointly*, *for a test period ranging from May* 1, 2010 *through April* 30, 2011. *The scores for the independent BMA and ECC-R techniques*, *which involve randomization*, *are averaged over* 100 *repetitions*

	Temperature (°C)	**Pressure (hPa)**
ECMWF	2.342	1.478
BMA	1.929	1.473
ECC-Q	1.927	1.428
ECC-R	1.945	1.454
ECC-T	1.934	1.442

While for temperature the BMA postprocessing improves strongly on the raw ensemble forecast, the effect of ECC is minor, if not negative, due to the correlations in the forecast errors being negligible at the distances considered here. That said, Figure 11 illustrates the strongly positive effects of ECC on temperature field forecasts, where dependencies at short and moderate distances are of critical importance. Here we consider $33 \times 37 = 1221$ NWP model grid boxes over Germany and adjacent areas, with the forecast made a day ahead for 2:00 am on April 25, 2011, for what promises to be a pleasant, unusually warm spring night.

The postprocessing uses a single BMA model of the form (2.3), which is trained on spatially pooled pairs of

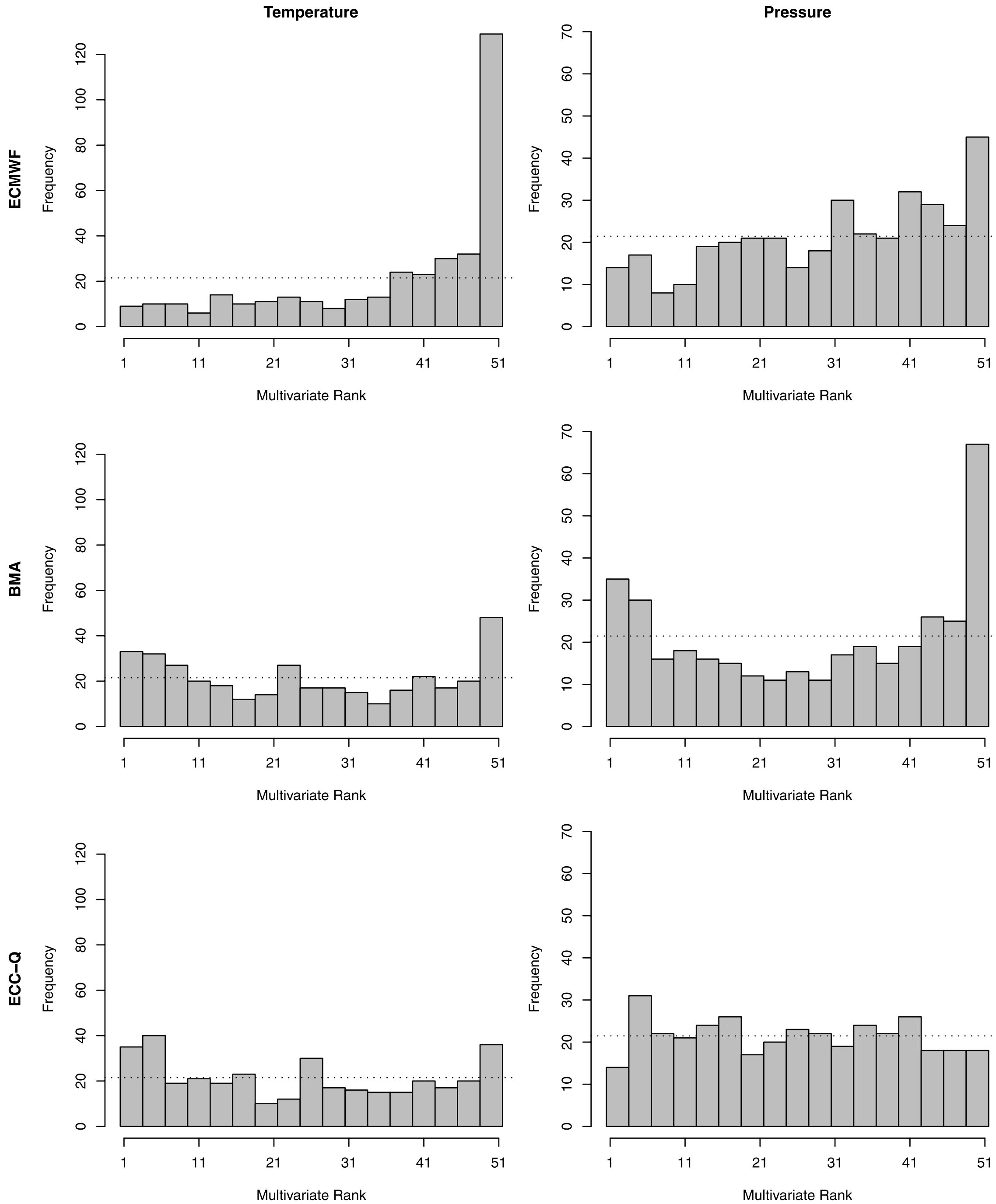


FIG. 10. *Multivariate rank histograms for* 48-*h ahead ensemble forecasts of temperature and pressure*, *each considered at Berlin*, *Frankfurt and Hamburg jointly*, *for a test period ranging from May* 1, 2010 *through April* 30, 2011.

ensemble forecasts and corresponding nowcasts from the previous 20 days. The nowcast[9] that serves as grid-based ground truth is the corresponding initialization

[9]Generally, the term nowcast is used for short-term weather forecasts, comprising prediction horizons from 0 to 6 hours ahead. Here we use it for the initialization of the ECMWFs control run—a distinguished NWP run outside the 50-member core ensemble considered here—that represents the best estimate of the state of the atmosphere at the initialization time, given recent and concurrent observational assets. In our specific usage, the term nowcast thus corresponds to a prediction horizon of 0 hours, and it provides a single-valued best estimate of the state of the atmosphere, rather than an ensemble.

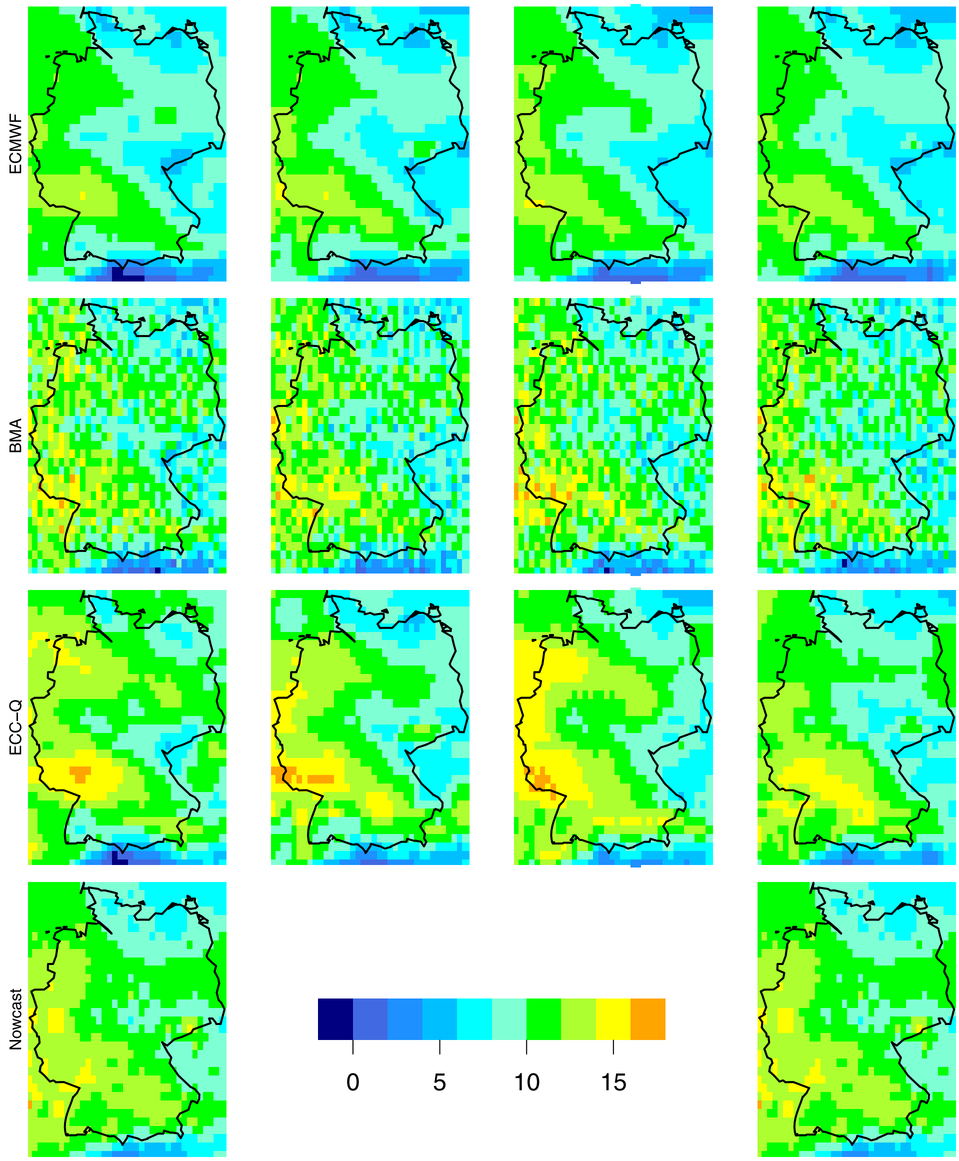


FIG. 11. 24-*hour ahead ensemble forecasts for temperature over Germany valid* 2:00 *am on April* 25, 2011, *in the unit of degrees Celsius*. *Top row*: *four randomly selected members of the raw ECMWF ensemble*. *Second row*: *independent BMA postprocessing—for each grid box*, *a random number from the corresponding BMA postprocessed predictive distribution is drawn*. *Third row*: *four members of the corresponding ECC ensemble*, *with rank order structures adopted from the respective raw ensemble members in the top row*. *Bottom row*: *single-valued nowcast as described in the text*, *shown both at left and at right*.

of the ECMWFs so-called control run (Molteni et al., 1996). The members of the unprocessed raw ECMWF ensemble appear to capture spatial structure fairly well, but they show an overall negative bias, especially in the mountainous Alps region in the south and in the central east of the country. While the BMA postprocessing addresses biases, and the use of a single BMA model avoids inconsistencies between the univariate postprocessed predictive distributions themselves, the independent samples result in noisy and incoherent spatial structure. The ECC postprocessed ensemble inherits the bias-corrected marginals from the independent BMA postprocessed forecast and simultaneously maintains the $L = 1221$ variate dependence structure in the raw ensemble.

While these examples concern the spatial case only, ECC is equally well suited to handling temporal and cross-variable dependencies, with Figure 5 illustrating the latter aspect. To generate physically realistic and consistent ensemble forecasts of temporal trajectories, constraints can be put on the BMA or NR parameters, so that they vary smoothly across lead times, which ensures the temporal consistency of the postprocessed marginal predictive distributions. Then, the ECC approach can be used to account for dependence structures across lead times. These settings are being investigated in ongoing work, and we expect to report quantitative results in due time.

6. DISCUSSION

The intensified attention to the quantification of uncertainty in the output of complex simulation models poses major challenges in a vast range of critical applications. In this paper, we have introduced the general uncertainty quantification framework of ensemble copula coupling (ECC), which we have illustrated on the key example of numerical weather prediction (NWP). The approach is conceptionally very simple and straightforward to implement in practice. Starting from raw ensemble output, ECC employs standard techniques to obtain postprocessed predictive distributions for each of the univariate margins individually. Then we quantize the postprocessed predictive distributions and adopt the rank dependence structure of the raw ensemble, as embodied by its empirical copula.

The defining feature of the ECC approach, namely, the adoption of the rank order structure of the raw ensemble, also sets its limitations. The number of members in the ECC postprocessed ensemble equals that of the raw ensemble, which typically is small, and ECC operates under a perfect model assumption with respect to the multivariate rank dependence structure. For state-of-the-art NWP models such an assumption seems defensible and reasonably adequate in practice, and it can be comfirmed by diagnostic checks, as we have illustrated in Figure 8, where the situation might be typical, but cannot be expected to be encountered each and every day. Generally, it seems realistic to assume that numerical models may show errors in dependence structures, which one may wish to diagnose and ameliorate to the extent possible. Future work in these directions is strongly encouraged.

Currently, approaches of the ECC type are being investigated and tested by weather centers internationally; see, for example, the recent work of Flowerdew (2012), Pinson (2012) and Roulin and Vannitsem (2012). We applaud these developments and call for case studies and quantitative comparisons to the Schaake shuffle (Clark et al., 2004), which also admits an empirical copula interpretation. In ECC, the multivariate dependence structure of the forecast errors derives from the ensemble forecast; in the Schaake shuffle, it derives from a record of historical weather observations. Judiciously designed combinations of the ECC and the Schaake shuffle approaches address the aforementioned problem of the statistical correction of systematic errors in dependence structures, and thus might lead to improved predictive performance.

If the model output under consideration is low-dimensional or strongly structured, parametric copula approaches become available, which may allow for the correction of any systematic errors in the ensemble's representation of conditional dependence structures. Here, the most prominent option lies in the use of Gaussian copulas, as in the general approach of Möller, Lenkoski and Thorarinsdottir (2013) and in the temporally or spatially structured settings of Gel, Raftery and Gneiting (2004), Berrocal, Raftery and Gneiting (2007, 2008) and Pinson et al. (2009). In such situations, it is to be expected that parametric techniques outperform the ECC approach and the Schaake shuffle, and comparative studies of the predictive abilities and relative merits of the various methods are strongly encouraged. Given its intuitive appeal and simplicity of implementation, the ECC approach offers a natural benchmark.

In Figure 11 we have given an example of how ECC can be used to restore spatial consistency in weather field forecasts directly on the model grid. The aforementioned parametric Gaussian approaches of Gel, Raftery and Gneiting (2004) and Berrocal, Raftery and Gneiting (2007) can achieve this, too, but require elaborate spatial statistical models to be fitted. In contrast, the computational and human resources necessitated by ECC are nearly negligible, and ECC can also handle temporal and cross-variable dependencies, for model output of nearly any dimensionality.

While we have focused on weather forecasting in this paper, the general framework of ECC as a multistage approach to the quantification of uncertainty in the output of complex simulation models with intricate multivariate dependence structures is likely to be useful in a vast range of applications. Essentially, ECC can be applied whenever an ensemble of simulation runs is available, the ensemble is capable of realistically representing multivariate dependence structures, and training data for the statistical correction of the univariate margins are at hand. In this general setting of uncertainty quantification, the goals articulated by

Gneiting, Balabdaoui and Raftery (2007) continue to provide guidance, in that we seek to gauge our incomplete knowledge of current, past or future quantities of interest by means of joint probability distributions, which ought to be as sharp as possible, subject to them being calibrated, in the broad sense of reality being statistically compatible with the postprocessed distributions.

ACKNOWLEDGMENTS

We are indebted to colleagues at Heidelberg University, the University of Washington, the German Weather Service (DWD), the European Centre for Medium-Range Weather Forecasts (ECMWF) and elsewhere, including but not limited to Konrad Bogner, Jonathan Flowerdew, Renate Hagedorn, Tom Hamill, Alex Lenkoski, Martin Leutbecher, Florian Pappenberger, Pierre Pinson, David Richardson and Johanna Ziegel, who have graciously shared their thoughts and expertise. In particular, Tom Hamill drew our attention to approaches of the ECC type during a stroll on the University of Washington campus in summer 2009, and Martin Leutbecher noted a data error in a poster version of our work presented at a conference in 2012. We gratefully acknowledge support by the Volkswagen Foundation and sfi^2, Statistics for Innovation in Oslo, and we thank the Editors and referees for their constructive feedback.

SUPPLEMENTARY MATERIAL

Dynamic version of Figure 5 (DOI: 10.1214/13-STS443SUPP; .pdf). In this version of Figure 5, the ensemble reordering step in the ECC approach is elucidated when switching back and forth between pages.

REFERENCES

AAS, K., CZADO, C., FRIGESSI, A. and BAKKEN, H. (2009). Pair-copula constructions of multiple dependence. *Insurance Math. Econom.* **44** 182–198. MR2517884

BAO, L., GNEITING, T., GRIMIT, E. P., GUTTORP, P. and RAFTERY, A. E. (2010). Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review* **138** 1811–1821.

BEN BOUALLÈGUE, Z. (2013). Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather and Forecasting* **28** 515–524.

BERROCAL, V. J., RAFTERY, A. E. and GNEITING, T. (2007). Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review* **135** 1386–1402.

BERROCAL, V. J., RAFTERY, A. E. and GNEITING, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Ann. Appl. Stat.* **2** 1170–1193. MR2655654

BERROCAL, V. J., RAFTERY, A. E., GNEITING, T. and STEED, R. C. (2010). Probabilistic weather forecasting for winter road maintenance. *J. Amer. Statist. Assoc.* **105** 522–537. MR2724842

BREMNES, J. B. (2004). Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Monthly Weather Review* **132** 338–347.

BREMNES, J. B. (2007). Improved calibration of precipitation forecasts using ensemble techniques. Part 2: Statistical calibration methods. Technical Report 04/2007, Norwegian Meteorological Institute.

BRÖCKER, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society* **138** 1611–1617.

BRÖCKER, J. and SMITH, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus Ser. A* **60** 663–678.

CHALOULOS, G. and LYGEROS, J. (2007). Effect of wind correlation on aircraft conflict probability. *Journal of Guidance, Control, and Dynamics* **30** 1742–1752.

CHILÈS, J.-P. and DELFINER, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed. Wiley, Hoboken, NJ. MR2850475

CHMIELECKI, R. M. and RAFTERY, A. E. (2011). Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Review* **139** 1626–1636.

CLARK, M., GANGOPADHYAY, S., HAY, L., RAJAGOPALAN, B. and WILBY, R. (2004). The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology* **5** 243–262.

CLOKE, H. L. and PAPPENBERGER, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology* **375** 613–626.

COOKE, R. (1906). Forecasts and verifications in Western Australia. *Monthly Weather Review* **34** 23–24.

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ. MR2848400

CZADO, C., GNEITING, T. and HELD, L. (2009). Predictive model assessment for count data. *Biometrics* **65** 1254–1261. MR2756513

DAVISON, A. C., PADOAN, S. A. and RIBATET, M. (2012). Statistical modeling of spatial extremes. *Statist. Sci.* **27** 161–186. MR2963980

DAWID, A. P. (1984). Statistical theory. The prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292. MR0763811

DEHEUVELS, P. (1979). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Acad. Roy. Belg. Bull. Cl. Sci.* (5) **65** 274–292. MR0573609

DELLE MONACHE, L., HACKER, J. P., ZHOU, Y., DENG, X. and STULL, R. B. (2006). Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research* **111** D24307.

DEMARTA, S. and MCNEIL, A. J. (2005). The *t* copula and related copulas. *International Statistical Review* **73** 111–129.

DETTE, H. and VOLGUSHEV, S. (2008). Non-crossing nonparametric estimates of quantile curves. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 609–627. MR2420417

DI NARZO, A. F. and COCCHI, D. (2010). A Bayesian hierarchical approach to ensemble weather forecasting. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 405–422. MR2756542

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

FERMANIAN, J.-D., RADULOVIĆ, D. and WEGKAMP, M. (2004). Weak convergence of empirical copula processes. *Bernoulli* **10** 847–860. MR2093613

FLOWERDEW, J. (2012). Calibration and combination of medium-range ensemble precipitation forecasts. Technical Report 567, United Kingdom Met Office.

FRALEY, C., RAFTERY, A. E. and GNEITING, T. (2010). Calibrating multi-model forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review* **138** 190–202.

FRALEY, C., RAFTERY, A. E., GNEITING, T., SLOUGHTER, J. M. and BERROCAL, V. J. (2011). Probabilistic weather forecasting in R. *R Journal* **3** 55–63.

GEL, Y., RAFTERY, A. E. and GNEITING, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method (with discussion and rejoinder). *J. Amer. Statist. Assoc.* **99** 575–590.

GENEST, C. and FAVRE, A. C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* **12** 347–368.

GLAHN, H. R., PEROUTKA, M., WEIDENFELD, J., WAGNER, J., ZYLSTRA, G. and SCHUKNECHT, B. (2009). MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review* **137** 246–268.

GNEITING, T. (2008). Editorial: Probabilistic forecasting. *J. Roy. Statist. Soc. Ser. A* **171** 319–321. MR2427336

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 243–268. MR2325275

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. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GNEITING, T., RAFTERY, A. E., WESTVELD, A. H. and GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review* **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 applications to ensemble predictions of surface winds (with discussion and rejoinder). *TEST* **17** 211–264.

GRAF, S. and LUSCHGY, H. (2000). *Foundations of Quantization for Probability Distributions*. *Lecture Notes in Math.* **1730**. Springer, Berlin. MR1764176

HAGEDORN, R., HAMILL, T. M. and WHITAKER, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Monthly Weather Review* **136** 2608–2619.

HAGEDORN, R., BUIZZA, R., HAMILL, T. M., LEUTBECHER, M. and PALMER, T. N. (2012). Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society* **138** 1814–1827.

HAMILL, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review* **129** 550–560.

HAMILL, T. M. and COLUCCI, S. J. (1997). Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review* **125** 1312–1327.

HAMILL, T. M., HAGEDORN, R. and WHITAKER, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly Weather Review* **136** 2620–2632.

HAMILL, T. M., WHITAKER, J. S. and MULLEN, S. L. (2006). Reforecasts: An important dataset for improving weather predictions. *Bulletin of the American Meteorological Society* **87** 33–46.

JOE, H. (1997). *Multivariate Models and Dependence Concepts*. *Monographs on Statistics and Applied Probability* **73**. Chapman & Hall, London. MR1462613

KALNAY, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge Univ. Press, Cambridge.

KANN, A., WITTMANN, C., WANG, Y. and MA, X. (2009). Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. *Monthly Weather Review* **137** 3373–3387.

KLEIBER, W., RAFTERY, A. E. and GNEITING, T. (2011). Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *J. Amer. Statist. Assoc.* **106** 1291–1303. MR2896836

KLEIBER, W., RAFTERY, A. E., BAARS, J., GNEITING, T., MASS, C. and GRIMIT, E. P. (2011). Locally calibrated probabilistic temperature foreasting using geostatistical model averaging and local Bayesian model averaging. *Monthly Weather Review* **139** 2630–2649.

KNEIB, T. (2013). Beyond mean regression (with discussion and rejoinder). *Statistical Modelling* **13** 275–385.

KOLESÁROVÁ, A., MESIAR, R., MORDELOVÁ, J. and SEMPI, C. (2006). Discrete copulas. *IEEE Transactions on Fuzzy Systems* **14** 698–705.

KRZYSZTOFOWICZ, R. and TOTH, Z. (2008). Bayesian processor of ensemble (BPE): Concept and implementation. Slides presented at the 4th NCEP/NWS Ensemble User Workshop, Laurel, MD.

LEUTBECHER, M. and PALMER, T. N. (2008). Ensemble forecasting. *J. Comput. Phys.* **227** 3515–3539. MR2400226

MAYOR, G., SUÑER, J. and TORRENS, J. (2007). Sklar's theorem in finite settings. *IEEE Transactions on Fuzzy Systems* **15** 410–416.

MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Univ. Press, Princeton, NJ. MR2175089

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

MÖLLER, A., LENKOSKI, A. and THORARINSDOTTIR, T. L. (2013). Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society* **139** 982–991.

MOLTENI, F., BUIZZA, R., PALMER, T. N. and PETROLIAGIS, T. (1996). The new ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society* **122** 73–119.

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

PALMER, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society* **128** 747–774.

PINSON, P. (2012). Adaptive calibration of (u, v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society* **138** 1273–1284.

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

PINSON, P., MADSEN, H., NIELSEN, H. A., PAPAEFTHYMIOU, G. and KLÖCKL, B. (2009). From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy* **12** 51–62.

RAFTERY, A. E., KÁRNÝ, M. and ETTLER, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics* **52** 52–66. MR2654989

RAFTERY, A. E., GNEITING, T., BALABDAOUI, F. and POLAKOWSKI, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133** 1155–1174.

ROQUELAURE, S. and BERGOT, T. (2008). A local ensemble prediction system for fog and low clouds: Construction, Bayesian model averaging calibration, and validation. *Journal of Applied Meteorology and Climatology* **47** 3072–3088.

ROULIN, E. and VANNITSEM, S. (2012). Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly Weather Review* **140** 874–888.

RUIZ, J. J. and SAULO, C. (2012). How sensitive are probabilistic precipitation forecasts to the choice of calibration algorithms and the ensemble generation method? Part I: Sensitivity to calibration methods. *Meteorological Applications* **19** 302–313.

RÜSCHENDORF, L. (1976). Asymptotic distributions of multivariate rank order statistics. *Ann. Statist.* **4** 912–923. MR0420794

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

SCHAAKE, J., DEMARGNE, J., HARTMAN, R., MULLUSKY, M., WELLES, E., WU, L., HERR, H., FAN, X. and SEO, D. J. (2007). Precipitation and temperature ensemble forecasts from single-valued forecasts. *Hydrology and Earth System Sciences Discussions* **4** 655–717.

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. *Atmospheric Science Letters* **11** 59–63.

SCHEFZIK, R. (2011). Ensemble copula coupling. Diploma thesis, Faculty of Mathematics and Informatics, Heidelberg Univ.

SCHEFZIK, R., THORARINSDOTTIR, T. L. and GNEITING, T. (2013). Supplement to "Uncertainty quantification in complex simulation models using ensemble copula coupling." DOI:10.1214/13-STS443SUPP.

SCHEUERER, M. (2013). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*. To appear. DOI:10.1002/qj.2183.

SCHMEITS, M. J. and KOK, K. J. (2010). A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Monthly Weather Review* **138** 4199–4211.

SCHOELZEL, C. and FRIEDERICHS, P. (2008). Multivariate non-normally distributed random variables in climate research—Introduction to the copula approach. *Nonlinear Processes in Geophysics* **15** 761–772.

SCHOELZEL, C. and HENSE, A. (2011). Probabilistic assessment of regional climate change in Southwest Germany by ensemble dressing. *Climate Dynamics* **36** 2003–2014.

SCHUHEN, N., THORARINSDOTTIR, T. L. and GNEITING, T. (2012). Ensemble model output statistics for wind vectors. *Monthly Weather Review* **140** 3204–3219.

SEGERS, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli* **18** 764–782. MR2948900

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8** 229–231. MR0125600

SLOUGHTER, M., GNEITING, T. and RAFTERY, A. E. (2010). Probabilistic wind spread forecasting using ensembles and Bayesian model averaging. *J. Amer. Statist. Assoc.* **105** 25–35. MR2757189

SLOUGHTER, J. M., GNEITING, T. and RAFTERY, A. E. (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review* **141** 2107–2119.

SLOUGHTER, J. M., RAFTERY, A. E., GNEITING, T. and FRALEY, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review* **135** 3209–3220.

STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York. MR1697409

STUTE, W. (1984). The oscillation behavior of empirical processes: The multivariate case. *Ann. Probab.* **12** 361–379. MR0735843

THORARINSDOTTIR, T. L. and GNEITING, T. (2010). Probabilistic forecasts of wind speed: Ensemble model ouput statistics by using heteroscedastic censored regression. *J. Roy. Statist. Soc. Ser. A* **173** 371–388. MR2751882

THORARINSDOTTIR, T. L. and JOHNSON, M. S. (2012). Probabilistic wind gust forecasting using non-homogeneous Gaussian regression. *Monthly Weather Review* **140** 889–897.

TREFETHEN, N. (2012). From the President: Discrete or continuous? *SIAM News* **45** 4.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

VOISIN, N., PAPPENBERGER, F., LETTENMAIER, D. P., BUIZZA, R. and SCHAAKE, J. C. (2011). Application of a medium-range global hydrologic probabilistic forecast scheme to the Ohio River basin. *Weather and Forecasting* **26** 425–446.

WILKS, D. S. (2002). Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society* **128** 2821–2836.

WILKS, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications* **16** 361–368.

WILKS, D. S. and HAMILL, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review* **135** 2379–2390.