Spatial Postprocessing of Ensemble Forecasts for Temperature Using Nonhomogeneous Gaussian Regression

KIRA FELDMANN

Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

MICHAEL SCHEUERER

National Oceanic and Atmospheric Administration, Boulder, Colorado

THORDIS L. THORARINSDOTTIR

Norwegian Computing Center, Oslo, Norway

ABSTRACT

Statistical postprocessing techniques are commonly used to improve the skill of ensembles from numerical weather forecasts. This paper considers spatial extensions of the well-established nonhomogeneous Gaussian regression (NGR) postprocessing technique for surface temperature and a recent modification thereof in which the local climatology is included in the regression model to permit locally adaptive postprocessing. In a comparative study employing 21-h forecasts from the Consortium for Small Scale Modelling ensemble predictive system over Germany (COSMO-DE), two approaches for modeling spatial forecast error correlations are considered: a parametric Gaussian random field model and the ensemble copula coupling (ECC) approach, which utilizes the spatial rank correlation structure of the raw ensemble. Additionally, the NGR methods are compared to both univariate and spatial versions of the ensemble Bayesian model averaging (BMA) postprocessing technique.

1. Introduction

The first ensemble prediction systems were developed in the early 1990s to account for various sources of uncertainty in numerical weather prediction (NWP) model outputs (Lewis 2005). Such systems have now become state-of-the-art in meteorological forecasting (Leutbecher and Palmer 2008). Additionally, the ensemble forecasts are commonly postprocessed using statistical techniques to improve calibration and correct for potential biases, and a diverse range of postprocessing techniques has been proposed (e.g., Gneiting et al. 2005; Raftery et al. 2005; Wilks and Hamill 2007; Bröcker and Smith 2008). While these methods have been shown to greatly improve the predictive performance, many are only applicable to univariate weather

quantities and neglect forecast error dependencies over time or between different observational sites. However, correct multivariate dependence structure is often important in applications, especially when considering composite quantities such as minima, maxima, or an aggregated total. These quantities are crucial, for example, for highway maintenance operations or flood management, where subsequent risk calculations based on the forecast require a calibrated probabilistic forecast for both the original weather variable and the composite quantity.

In this paper, we focus on spatial extensions of the nonhomogeneous Gaussian regression (NGR) or ensemble model output statistics method for surface temperature, originally proposed by Gneiting et al. (2005). NGR is a parsimonious postprocessing technique that, for temperature, returns a Gaussian predictive distribution where the mean value is an affine function of the ensemble member forecasts while the variance is an affine function of the ensemble variance. The parameters of the model are estimated based on recent forecast

Corresponding author address: Kira Feldmann, Heidelberg Institute for Theoretical Studies, Schlosswolfsbrunnenweg 35, 69118 Heidelberg, Germany.
E-mail: kira.feldmann@h-its.org

errors jointly over a region or separately at each observation location (Gneiting et al. 2005; Thorarinsdottir and Gneiting 2010; Hagedorn et al. 2008; Kann et al. 2009). Recently, Scheuerer and König (2014) proposed a modification of the NGR methods of Gneiting et al. (2005) in which the postprocessing at individual locations varies in space by parameterizing the predictive mean and variance in terms of the local forecast anomalies rather than the forecasts themselves.

To obtain a multivariate predictive distribution based on a deterministic temperature forecast, Gel et al. (2004) propose the geostatistical output perturbation (GOP) method to generate spatially consistent forecasts fields in which the forecast error field is described through a Gaussian random field model. Berrocal et al. (2007) combine GOP with the univariate postprocessing method ensemble Bayesian model averaging (BMA) of Raftery et al. (2005). Ensemble BMA for temperature dresses each bias-corrected ensemble member with a Gaussian kernel and returns a predictive distribution given by a weighted average of the individual kernels. By merging ensemble BMA and GOP, calibrated probabilistic forecasts of entire weather fields are produced. We propose a similar conceptualization, combining the NGR methods of Gneiting et al. (2005) and Scheuerer and König (2014) with a Gaussian random field error model in an approach we refer to as spatial NGR.

As an alternative multivariate method, we consider the nonparametric ensemble copula coupling (ECC) approach of Schefzik et al. (2013). ECC returns a postprocessed ensemble of the same size as the original raw ensemble. The prediction values at each location are samples from the univariate postprocessed predictive distribution at that location. Multivariate forecast fields are subsequently generated using the rank correlation structure of the raw ensemble. ECC thus assumes that the ensemble prediction system correctly describes the spatial dependence structure of the weather quantity. The method applies equally to any multivariate setting and comes at virtually no additional computational cost once the univariate postprocessed predictive distributions are available.

Figure 1 illustrates temperature field forecasts obtained from the raw ensemble, the standard univariate NGR method, NGR combined with ECC, and spatial NGR. The raw ensemble is depicted in the first row. The NWP model output has a physically consistent spatial structure, but as we shall see later, it is strongly underdispersive and does not adequately represent the true forecast uncertainty. The samples in rows 2–4 all share the same NGR marginal predictive distributions that have larger uncertainty bounds than the raw ensemble.

In the second row, the realizations have been sampled independently for each grid point (i.e., no spatial dependence structure is present). This results in unrealistic temperature fields and, when considering compound quantities, forecasts that are statistically inappropriate. The combination of NGR and ECC in the third row gives forecast fields with similar spatial structures as the raw ensemble even though there is larger spread both within each field and between the realized fields. As a consequence of spatial correlations being modeled through a discrete copula, the resulting temperature fields feature some sharp transitions at locations where the ranks of the raw ensemble change. The bottom row depicts temperature field simulations obtained with spatial NGR. Here, the spatial dependence between forecast errors at different locations is modeled by a statistical correlation model and the physical consistency is implicitly learned from the data.

In a comparative study, we apply the various extensions of NGR as well as ensemble BMA and its spatial extension to 21-h forecasts of surface temperature over Germany issued by the German Weather Service through their Consortium for Small Scale Modelling (COSME-DE) ensemble prediction system. The remainder of the paper is organized as follows. The forecast and observation data are described in the next section 2. The univariate NGR postprocessing methods are introduced in section 3, while the multivariate methods are described in section 4. Forecast evaluation methods are discussed in section 5. In section 6 we report the results of the case study and a discussion is provided in section 7. Finally, computational details regarding the calculation of the evaluation methods are given in the appendix.

2. Forecast and observation data

The COSMO-DE forecast dataset consists of a 20-member ensemble. Forecasts are made for lead times from 0 to 21 h on a 2.8-km grid covering Germany with a new model run being started every 3 h. The ensemble is based on a convection-permitting configuration of the NWP model COSMO (Steppeler et al. 2003; Baldauf et al. 2011). It has a 5×4 factorial design with five different perturbations in the model physics and four different initial and boundary conditions provided by global forecasting models (Gebhardt et al. 2011; Peralta and Buchhold 2011). The ensemble members are thus not exchangeable. The preoperational phase of the COSMO-DE ensemble prediction system started on 9 December 2010 and the operational phase was launched on 22 May 2012.

We employ 21-h forecasts from the preoperational phase initialized at 0000 UTC; our entire dataset consists
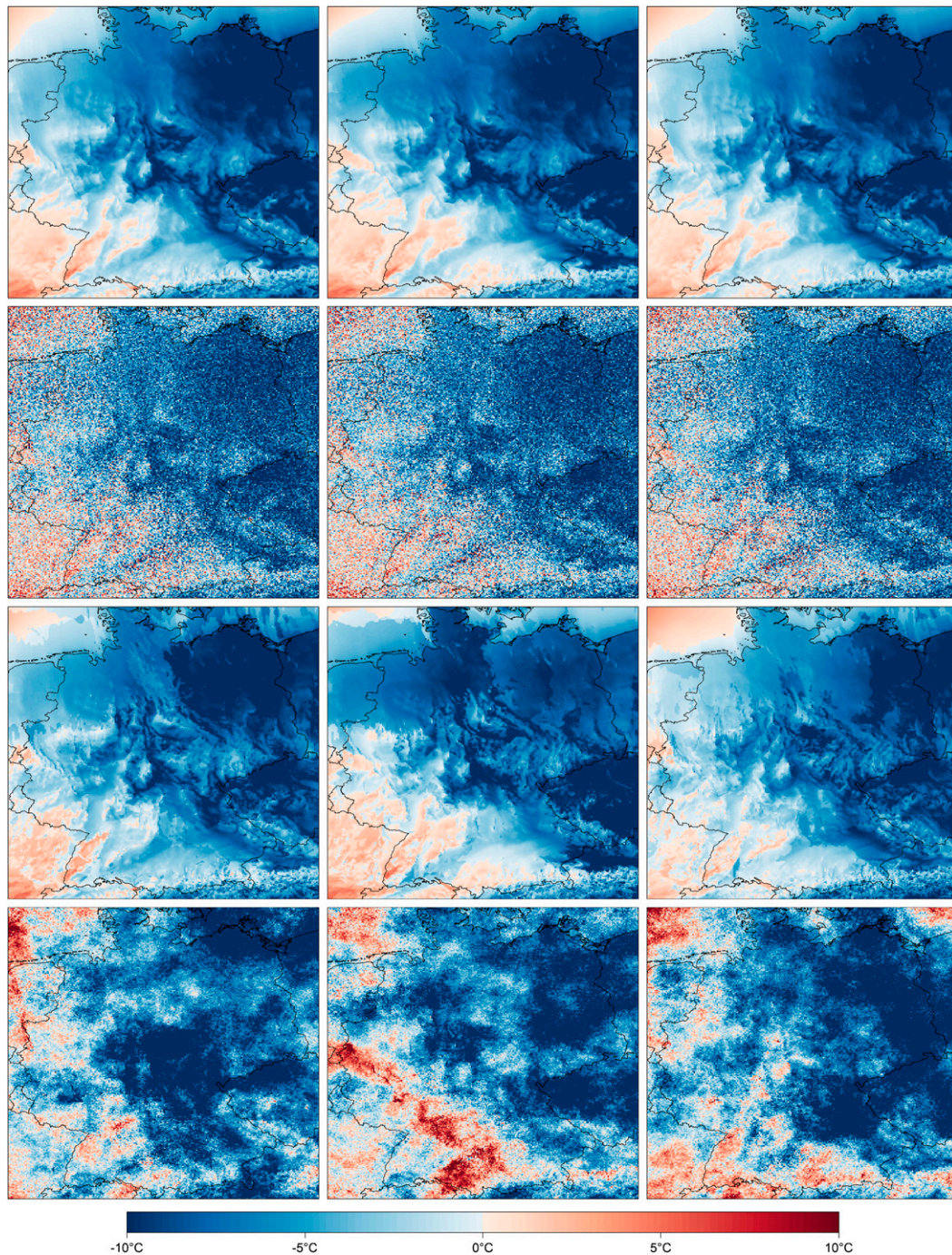
FIG. 1. The 21-h temperature forecasts over Germany at 2100 UTC 5 Jan 2011. (from top to bottom) Three members from the COSMO-DE ensemble prediction system, the postprocessed nonhomogeneous Gaussian regression (NGR) forecast, the NGR combined with ensemble copula coupling, and examples of the multivariate spatial NGR$_+$ forecast, respectively.

of forecasts from 10 December 2010 to 30 November 2011. As we use a rolling training period of 25 days to fit the parameters of the statistical postprocessing methods, the evaluation period runs from 5 January 2011 to 30 November 2011. If at least one ensemble member forecast is missing at all observation locations on a specific day, we omit this day from the dataset. This way, 10 days are eliminated with 346 days remaining. The
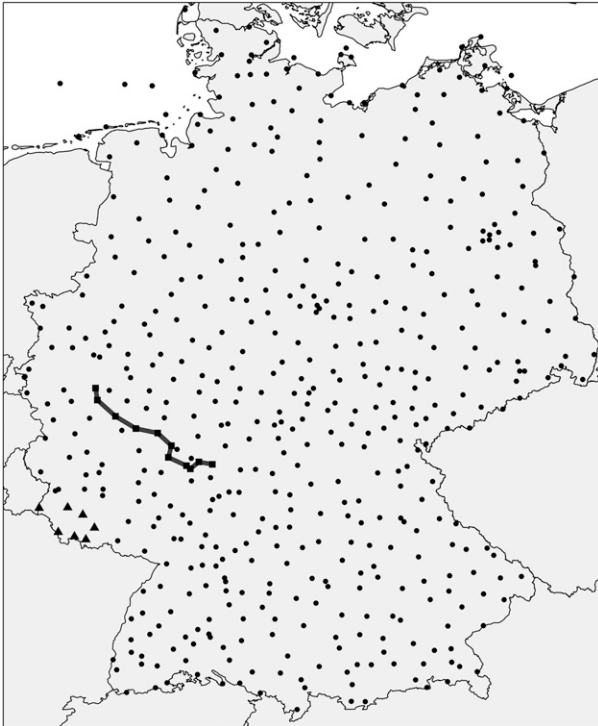
FIG. 2. Map of Germany showing the location of a total of 514 SYNOP observation stations. The gray line illustrates a section of the highway A3 with the surrounding observation stations indicated by solid squares. Seven observation stations in the state of Saarland are represented by solid triangles. Other stations are indicated by solid circles.

temperature observations we employ stem from 514 synoptic observation (SYNOP) stations over Germany. Their locations are shown in Fig. 2. The forecasts are interpolated from the forecast grid to the station locations using bilinear interpolation. Many of the stations have some missing data. In total, we evaluate forecasts for 117 879 verifying observations over 346 days. The COSMO model uses a rotated spherical coordinate system in order to project the geographical coordinates to the plane with distortions as small as possible (Doms and Schättler 2002, their section 3.3), with 421×461 equidistant grid points in the meridional and zonal direction. We adopt this coordinate system to calculate horizontal distances within the framework of our spatial correlation models.

3. Univariate postprocessing

a. NGR for temperature

The NGR method of Gneiting et al. (2005) generalizes the common model output statistics (MOS) postprocessing technique (see e.g., Wilks 2011). The distribution of the future state y_s of the temperature at location s is

modeled as a Gaussian distribution with parameters depending on the M ensemble forecasts $f_{1s}, \ldots, f_{Ms}$:

$$y_s | f_{1s}, \ldots, f_{Ms} \sim \mathcal{N}(a + b_1 f_{1s} + \cdots + b_M f_{Ms}, c + d S_s^2), \tag{1}$$

where S_s^2 is the ensemble variance, $a, b_1, \ldots, b_M \in \mathbb{R}$ are regression coefficients, and $c, d \in \mathbb{R}_+$ are nonnegative coefficients. NGR is thus a linear model with the ensemble forecasts as predictors and a nonhomogeneous error term that is modeled as an affine function of the ensemble variance S_s^2. This modeling setup counteracts a possible over- or underdispersion of the ensemble while exploiting a positive spread-error correlation. The normal predictive distribution presents a reasonable model for variables like temperature or surface pressure. While the Gaussian assumption is not appropriate for all weather variables, the basic idea of NGR can also be used with other types of predictive distributions (Thorarinsdottir and Gneiting 2010; Thorarinsdottir and Johnson 2012; Lerch and Thorarinsdottir 2013; Scheuerer 2014).

In the formulation in (1), the regression coefficients $b_1, \ldots, b_M$ can take any value in $\mathbb{R}$. However, as negative values are difficult to interpret, Gneiting et al. (2005) suggest an alternative formulation restricting the coefficients $b_1, \ldots, b_M$ to be nonnegative by iteratively removing those ensemble members f_m from the linear model for which the coefficients b_m are negative. We follow Thorarinsdottir and Gneiting (2010) and obtain nonnegative coefficients by setting $b_1 = \beta_1^2, \ldots, b_M = \beta_M^2$ with $\beta_1, \ldots, \beta_M \in \mathbb{R}$. The (normalized) coefficients can then be interpreted as weights and reflect the relative performance of the ensemble members during the training period. In the following, we refer to this approach as NGR$_+$.

b. Locally adaptive NGR

The NGR postprocessing in (1) makes the same adjustments of ensemble mean and variance at all locations. However, it has been argued that systematic model biases may vary in space due to incomplete resolution of the orography or different land-use characteristics. Similarly, the prediction uncertainty may differ between locations in a way that is not represented by the ensemble spread (Kleiber et al. 2011; Scheuerer and Büermann 2014; Scheuerer and König 2014).

As an alternative to the NGR model in (1), we consider the locally adaptive NGR method of Scheuerer and König (2014). Here, the local adaptation is obtained by incorporating information about the short-term local climatology in the covariates of the predictive distribution, thereby using forecast anomalies rather than the original forecasts as covariates. Specifically, let $\bar{y}_s^T$ denote

the average observed local temperature at location s over all days t in the training period $\mathcal{T}$ and, correspondingly, denote by $\overline{f}_{ms}^{\mathcal{T}}$ the average temperature forecast by the mth ensemble member. The predictive distribution then equals

$$
\begin{aligned}
y_s \,|\, f_{1s}, \ldots, f_{Ms}, \mathcal{D}_s^{\mathcal{T}} \sim \mathcal{N}(\overline{y}_s^{\mathcal{T}} + b_1(f_{1st} - \overline{f}_{1s}^{\mathcal{T}}) + \cdots \\
+ \, b_M(f_{Ms} - \overline{f}_{Ms}^{\mathcal{T}}), c\xi_s^2 + dS_s^2), \quad (2)
\end{aligned}
$$

where $\mathcal{D}_s^{\mathcal{T}}$ denotes the forecast and observation data at location s during the training period and ξ_s^2 is a predictor variable for the location specific uncertainty. It is given by the mean squared residuals:

$$
\xi_s^2 = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left[y_{st} - \overline{y}_s^{\mathcal{T}} - \sum_{m=1}^{M} \hat{b}_m(f_{mst} - \overline{f}_{ms}^{\mathcal{T}}) \right]^2,
$$

where $|\mathcal{T}|$ denotes the number of days in the training period, and $\hat{b}_1, \ldots, \hat{b}_M$ are estimates of the regression coefficients obtained in a first step by a variant of weighted least squares (see section 3d). The parameterization of the model in (2) is a compromise between flexibility and parsimony. On the one hand, it retains the assumption of universal slope parameters $b_1, \ldots, b_M$ and variance parameters c and d in order to keep the number of location specific parameters low and to avoid overfitting. On the other hand, it utilizes the parameters $\overline{y}_s^{\mathcal{T}}$ and ξ_s^2, which depend on the local climatology, permit location specific bias correction and uncertainty quantification, and thus improve local calibration. At locations without an observation station, those two local parameters cannot be calculated, but predictive means and variances can be obtained by spatial interpolation as described in Scheuerer and König (2014). We refer to this inclusion of local climatological information as NGR$_c$.

c. Ensemble BMA

The univariate ensemble BMA method of Raftery et al. (2005) is a kernel dressing approach where, for temperature, each bias-corrected ensemble member is dressed with a Gaussian kernel with a fixed variance. That is,

$$
y_s \,|\, f_{ms} \sim \mathcal{N}(a_m + b_m f_{ms}, \sigma^2),
$$

for $m = 1, \ldots, M$, and $s \in \mathcal{S}$. The predictive density is then given by a weighted average

$$
\sum_{m=1}^{M} \omega_m \varphi(a_m + b_m f_{ms}, \sigma^2), \quad (3)
$$

where φ denotes the Gaussian density. The weights $\omega_1, \ldots, \omega_M$ are nonnegative with $\sum_{m=1}^{M} \omega_m = 1$, and reflect the skill of each ensemble member in the training period $\mathcal{T}$.

d. Parameter estimation in the univariate setting

We focus on the NGR$_+$ formulation of the NGR method in (1) and the NGR$_c$ extension in (2). The parameter estimation for both methods proceeds in a similar manner. It is assumed that the forecast error statistics change only slowly over time and a rolling training window $\mathcal{T}$ of the $|\mathcal{T}|$ most recent dates with forecasts and observations available is used to estimate the model parameter. The model fitting is carried out with all available data from the set of observation sites $\mathcal{S}$ within the training window $\mathcal{T}$. We denote this set by $\mathcal{S}$ in the equations below even though observations might not always be available at all observation sites. Note that with NGR$_+$ we can easily generate postprocessed forecasts at locations outside $\mathcal{S}$ since the parameters of the postprocessing techniques are not site specific. When NGR$_c$ is used, this can be achieved through an additional spatial interpolation step.

We follow Gneiting et al. (2005) and estimate the NGR$_+$ parameters by minimizing the continuous ranked probability score (CRPS) (e.g., Gneiting and Raftery 2007) over the training set. That is, we chose them as a solution to

$$
\min_{a,b_1,\ldots,b_M,c,d} \frac{1}{|\mathcal{T}||\mathcal{S}|} \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{S}} \mathrm{CRPS}(\Phi_{st}, y_{st}), \quad (4)
$$

where Φ_{st} is the Gaussian distribution function in (1) on training day t at site s, y_{st} is the corresponding verifying observation, and

$$
\mathrm{CRPS}(\Phi_{st}, y_{st}) = \int_{-\infty}^{\infty} [\Phi_{st}(x) - \mathbb{1}_{[y_{st}, \infty)}(x)]^2 \, dx, \quad (5)
$$

with $\mathbb{1}_{[y_{st}, \infty)}(x)$ equal to 1 if $x \in [y_{st}, \infty)$ and 0 otherwise. For Gaussian distributions, the integral in (5) can be expressed in a closed form that minimizes the computational costs (Gneiting et al. 2005). [Software for the estimation and prediction is available through the ensemble MOS package in R (R Core Team 2013), which can be downloaded at www.r-project.org.]

The parameters of the NGR$_c$ method are estimated in two steps. In a first step, the regression parameters $b_1, \ldots, b_M$ in (2) are estimated by weighted least squares using a penalized version of the loss function to prevent overfitting [see Scheuerer and König (2014) for details]. The estimated parameters $\hat{b}_1, \ldots, \hat{b}_1$ are then kept fixed, and in a second step the variance parameters c and d are estimated via CRPS minimization as in (4) above.

We estimate the ensemble BMA parameters using the R package ensemble BMA employing the same training period $\mathcal{T}$ as for the univariate NGR methods.

4. Multivariate methods

a. Ensemble copula coupling

The ECC method of Schefzik et al. (2013) employs the rank order structure of the raw ensemble to obtain a postprocessed ensemble of forecasts fields with the same multivariate correlation structure as the raw ensemble, while retaining the univariate NGR marginals. It is thus a semiparametric copula approach with continuous marginals and the nonparametric empirical copula. For each $s \in \mathcal{S}$, we create a sample of size M from the predictive distribution $\hat{\Phi}_s$ given by (1) or (2) of the following form:

$$\hat{f}_{1s} = \hat{\Phi}_s^{-1}\left(\frac{1}{M+1}\right), \ldots, \hat{f}_{Ms} = \hat{\Phi}_s^{-1}\left(\frac{M}{M+1}\right). \quad (6)$$

That is, it holds $\hat{f}_{1s} \leq \cdots \leq \hat{f}_{Ms}$. Let ρ_s denote a permutation of the integers $\{1, \ldots, M\}$ defined by $\rho_s(m) = \text{rank}(f_{ms})$ for $m = 1, \ldots, M$ with ties resolved at random. Then it follows that the sample $\{\hat{f}_{\rho_s(1)s}, \ldots, \hat{f}_{\rho_s(M)s}\}$ has the same rank order structure as the raw ensemble $\{f_{1s}, \ldots, f_{Ms}\}$. The ECC ensemble of postprocessed forecast fields is thus given by

$$\{\hat{f}_{\rho_s(m)s}\}_{s \in \mathcal{S}}, \quad m = 1, \ldots, M. \quad (7)$$

Schefzik et al. (2013) discuss and compare several alternative methods for generating a sample from each univariate predictive distribution. We use the quantiles in (6) as it follows from results in Bröcker (2012) that this sample maintains the calibration of the univariate forecasts.

b. Spatial NGR

The GOP (Gel et al. 2004) approach was originally introduced as an inexpensive substitute of a dynamical ensemble based on a single numerical weather prediction. It dresses the deterministic forecast with a simulated forecast error field according to a spatial random process, thus perturbing the outputs of the NWP models rather than their inputs. We propose a spatial NGR method that adopts the ideas from GOP and combines them with the univariate NGR methods described in sections 3a and 3b. The result is a multivariate predictive distribution that generates spatially coherent forecast fields, while retaining the univariate NGR marginals. The spatial NGR method can thus also be seen as a fully parametric Gaussian copula approach (Möller et al. 2013; Schefzik et al. 2013).

Denote by $\mathbf{Y} = \{y_s : s \in \mathcal{S}\}$ the vector whose components represent the temperature at each location in $\mathcal{S}$, and by $\mathbf{F}_m = \{f_{ms} : s \in \mathcal{S}\}$ the corresponding weather field forecast by the mth ensemble member. The vector $\boldsymbol{\mu}$ of predictive means obtained by marginal NGR postprocessing is given by

$$\boldsymbol{\mu} = \begin{cases} a\mathbf{1} + b_1\mathbf{F}_1 + \cdots + b_M\mathbf{F}_M & (\text{NGR}_+), \\ \overline{\mathbf{Y}}^{\mathcal{T}} + b_1(\mathbf{F}_1 - \overline{\mathbf{F}}_1^{\mathcal{T}}) + \cdots + b_M(\mathbf{F}_M - \overline{\mathbf{F}}_M^{\mathcal{T}}) & (\text{NGR}_c), \end{cases} \quad (8)$$

where $\mathbf{1}$ is a vector of length $|\mathcal{S}|$ with all entries equal to 1, $\overline{\mathbf{Y}}^{\mathcal{T}}$ is the average observed temperature vector over the training period $\mathcal{T}$, and $\overline{\mathbf{F}}_m^{\mathcal{T}}$ denotes the vector consisting of the average temperature forecast by the mth ensemble member over $\mathcal{T}$, see also (2). Similarly, denote by $\mathbf{D}$ the diagonal matrix of the univariate predictive standard deviations σ_s with $\sigma_s^2 = c + dS_s^2$ for NGR_+ and $\sigma_s^2 = c\xi_s^2 + dS_s^2$ for NGR_c.

The spatial NGR multivariate predictive distribution corresponds to the sum of the bias-corrected forecast mean vector given by (8), a scaled version of a zero-mean random vector $\mathbf{E}_1$ with correlated components, and a scaled version of an additional zero-mean random vector $\mathbf{E}_2$ with uncorrelated components representing small-scale variations that cannot be resolved with the available data. That is,

$$\mathbf{Y} \mid \mathbf{F}_1, \ldots, \mathbf{F}_M, \mathcal{D}_{\mathcal{S}}^{\mathcal{T}} = \boldsymbol{\mu} + \mathbf{D}\tilde{\mathbf{E}}, \quad (9)$$

where

$$\tilde{\mathbf{E}} = \sqrt{1-\theta}\,\mathbf{E}_1 + \sqrt{\theta}\,\mathbf{E}_2, \quad \theta \in [0, 1].$$

If all components of $\mathbf{E}_1$ and $\mathbf{E}_2$ have unit variance, the multiplication with $\mathbf{D}$ scales the components of $\tilde{\mathbf{E}}$ such that their variances match those predicted by the univariate NGR postprocessing methods. In particular, the resulting multivariate model features spatially varying predictive variances.

For the spatial correlations we follow Gel et al. (2004) and assume a stationary and isotropic correlation function $C_{\theta,r}$ of the exponential type. That is, we assume that the correlation between two components of $\tilde{\mathbf{E}}$ corresponding to locations s_i and s_j depends only on their Euclidean distance $\|s_i - s_j\|$ and is given by

$$C_{\theta,r}(s_i, s_j) = (1-\theta)\exp\left(-\frac{\|s_i - s_j\|}{r}\right) + \theta\delta_{ij}, \quad (10)$$

where δ_{ij} denotes to the Kronecker delta function, which is equal to 1 if $i = j$ and 0 otherwise. The parameter $\theta \in [0, 1]$ has already been introduced above and controls the relative contribution of the spatially correlated random vector $\mathbf{E}_1$ and the spatially uncorrelated random vector $\mathbf{E}_2$ to the overall variance. The range parameter $r > 0$ determines the rate at which the spatial correlations of $\mathbf{E}_1$ decay with distance. Once these parameters have been estimated (see below), the correlation matrix $\mathbf{P}$ of the random vector $\tilde{\mathbf{E}}$ can be defined via $\mathbf{P}_{ij} = C_{\theta,r}(s_i, s_j)$, and the resulting spatial NGR multivariate predictive distribution at locations within $\mathcal{S}$ is given by

$$\mathbf{Y} \mid \mathbf{F}_1, \ldots, \mathbf{F}_M, \mathcal{D}_\mathcal{S} \sim \mathcal{N}_{|\mathcal{S}|}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{with} \quad \boldsymbol{\Sigma} = \mathbf{DPD},$$
(11)

where $\mathcal{N}_{|\mathcal{S}|}$ denotes a multivariate normal distribution of dimension $|\mathcal{S}|$. Note that this definition can be extended to locations outside the set $\mathcal{S}$. As pointed out above, both $\boldsymbol{\mu}$ and $\mathbf{D}$ can be defined for any set of locations where ensemble forecasts are available: for NGR$_+$ by plugging the estimated, location-unspecific model parameters into (1), for NGR$_c$ via spatial interpolation. Now, since (10) presents a well-defined correlation function over the entire Euclidean plane, $C_{\theta,r}$ can be evaluated for arbitrary pairs of locations, and, hence, $\mathbf{P}$ can be defined for any set of locations.

c. Spatial BMA

The spatial BMA approach by Berrocal et al. (2007) combines ensemble BMA with the GOP method of Gel et al. (2004) in a similar way as the spatial NGR methods described in the previous section except that M error field models are constructed, one for each ensemble member. It, thus, also differs from spatial NGR in the manner in which realizations of the multivariate predictive distribution are simulated. For simulating a temperature forecast field under spatial BMA, we first randomly choose a member of the dynamical ensemble according to the ensemble BMA weights in (3), and then dress the corresponding bias-corrected forecast field with an error field that has a stationary covariance structure specific to this member. As the forecast field is chosen at random and the covariance function is member specific, the final covariance structure becomes nonstationary. This comes at the expense of having to estimate M different covariance functions. The spatial NGR approach, on the contrary, is based on a single correlation function, which results in a rather simple spatial dependence structure. Here, the corresponding realizations become nonstationary through the scaling of the stationary error field $\tilde{\mathbf{E}}$ according to (9).

d. Estimating the correlation parameters

To estimate the correlation parameters θ and r in (10), we consider the standardized forecast errors $\tilde{e}_{st} := (y_{st} - \mu_{st})/\sigma_{st}$ at all locations $s \in \mathcal{S}$ and on all training days $t \in \mathcal{T}$, and study their half-squared differences $1/2(\tilde{e}_{s_i t} - \tilde{e}_{s_j t})^2$, $s_i, s_j \in \mathcal{S}$. The stationary and isotropic correlation model in (10) implies that $C_{\theta,r}(s_i, s_j)$ is a function of the distance $\|s_i - s_j\|$ only, and, hence, we can write the variogram (e.g., Chilès and Delfiner 2012) of $\tilde{\mathbf{E}}$:

$$\gamma_{\theta,r}(\|s_i - s_j\|) := 1 - C_{\theta,r}(s_i, s_j) = \frac{1}{2}\text{Var}(\tilde{\mathbf{E}}_{s_i} - \tilde{\mathbf{E}}_{s_j}),$$

as a function of a single, scalar argument:

$$\gamma_{\theta,r}(h) = (1 - \theta)\left[1 - \exp\left(-\frac{h}{r}\right)\right] + \theta 1_{(0,\infty)}(h). \quad (12)$$

This allows us to aggregate the half squared differences not only over all training days $t \in \mathcal{T}$, but also over all pairs of locations with similar distance. Specifically, we fix a certain cutoff distance $h_{\max}$ and partition the left-open interval $(0, h_{\max}]$ into a family of left-open, disjoint subintervals $B_1, \ldots, B_L$ ("bins") with midpoints $h_1 < h_2 < \cdots < h_L$. If we denote by $\mathcal{I}_l$ the set of all pairs (i, j) such that $\|s_i - s_j\| \in B_l$ and $n_l = |\mathcal{I}_l|$ its size, then

$$\hat{\gamma}_l := \frac{1}{2n_l|\mathcal{T}|} \sum_{(i,j) \in \mathcal{I}_l} \sum_{t \in \mathcal{T}} (\tilde{e}_{s_i t} - \tilde{e}_{s_j t})^2$$

is an empirical approximation of $\gamma_{\theta,r}(h_l)$. For the calculation of $\hat{\gamma}_1, \ldots, \hat{\gamma}_L$, we employ the R package RandomFields by Schlather (2011), choosing the number and size of the bins such that the sets $\mathcal{I}_1, \ldots, \mathcal{I}_L$ have approximately the same size. Every day, we allocate about 800 pairs of locations in 300 bins each. Now, the parameters θ and r can be estimated by fitting a theoretical variogram of the form in (12) to the pairs $(h_1, \hat{\gamma}_1), \ldots, (h_L, \hat{\gamma}_L)$ that define the empirical variogram. We follow Berrocal et al. (2007) by using weighted least squares fitting (Cressie 1985), minimizing the function:

$$S(\theta, r) = \sum_{l=1}^{L} n_l \left[\frac{\hat{\gamma}_l - \gamma_{\theta,r}(h_l)}{\gamma_{\theta,r}(h_l)}\right]^2.$$

To solve this minimization problem numerically we use the optimization algorithm by Byrd et al. (1995) as implemented in the R function optim (R Core Team 2013). The range parameter r is constrained to be positive and not larger than the maximum distance over the entire domain, which equals 890 km. Averaged estimates over

the entire forecasting period obtained in earlier experiments are used as starting values for the optimization.

Alternatively, the random field parameters could be estimated via maximum likelihood. Under the assumption of independent forecast errors between different days, the log-likelihood that comes with the predictive distribution model in (11) is given by

$$
\ell(\theta, r) = - \sum_{t \in \mathcal{T}} \left\{ \frac{|\mathcal{S}|}{2} \log(2\pi) + \frac{1}{2} \log \det(\mathbf{D}_t \mathbf{P} \mathbf{D}_t) \right.
$$
$$
\left. + \frac{1}{2} (\mathbf{Y}_t - \boldsymbol{\mu}_t)^t (\mathbf{D}_t \mathbf{P} \mathbf{D}_t)^{-1} (\mathbf{Y}_t - \boldsymbol{\mu}_t) \right\},
$$

where $\mathbf{Y}_t$ and $\boldsymbol{\mu}_t$ are the vectors of observations and predictive means, and $\mathbf{D}_t$ is the diagonal matrix of predictive standard deviations on training day t. The correlation matrix $\mathbf{P}$ of the standardized forecast errors depends on the two unknown parameters θ and r, and maximizing $\ell(\theta, r)$ yields, under ideal conditions, statistically more efficient estimates than the variogram-based approach presented above. The latter is, however, more robust to outliers and computationally less expensive, and, therefore, we prefer it over maximum likelihood estimation in line with Berrocal et al. (2007). Indeed, results obtained with maximum likelihood estimation (not shown here) slightly reduced the predictive performance of the spatial NGR forecasting methods.

5. Forecast evaluation methods

Statistical postprocessing aims at correcting systematic biases and/or misrepresentation of the forecast uncertainty in the raw ensemble and, in our case, returns full probabilistic distributions. To evaluate the predictive performance of the methods under consideration, we follow Gneiting et al. (2007) who state that the goal of probabilistic forecasting is to maximize the sharpness of the predictive distribution subject to calibration.

a. Assessing calibration

Calibration refers to the statistical compatibility between forecasts and observations; the forecast is calibrated if the observation cannot be distinguished from a random draw from the predictive distribution. For continuous univariate distributions, calibration can be assessed empirically by plotting the histogram of the probability integral transform (PIT)—the value of the predictive cumulative distribution function in the observed value (Dawid 1984; Gneiting et al. 2007)—over all forecast cases. A forecasting method that is calibrated on average will return a uniform histogram,

a $\cap$-shape indicates overdispersion and a $\cup$-shape indicates underdispersion, while a systematic bias results in a triangular shape histogram. The discrete equivalent of the PIT histogram, which applies to ensemble forecasts, is the verification rank histogram (Anderson 1996; Hamill and Colucci 1997). It shows the distribution of the ranks of the observations within the corresponding ensembles and has the same interpretation as the PIT histogram. To facilitate direct comparison of the various methods, we only employ the rank histogram. That is, for the continuous predictive distributions, we create a 20-member ensemble given by 20 random samples from the distribution.

For multivariate settings, we employ the band depth rank histogram proposed by Thorarinsdottir et al. (2014). This approach ranks the observation within the sample of M ensemble forecasts by assessing the centrality of the observation within the sample. Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{M+1}\} = \{\mathbf{Y}, \mathbf{F}_1, \ldots, \mathbf{F}_M\}$ denote the set consisting of the observation vector $\mathbf{Y}$ and forecast vectors $\mathbf{F}_1, \ldots, \mathbf{F}_M$. Here, the dimension $|\mathcal{S}|$ of these vectors corresponds to the number of locations considered simultaneously. To calculate the band depth rank of the observation $\mathbf{Y}$ in $\mathcal{X}$, we first apply the prerank function:

$$
r(\mathbf{x}) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \sum_{1 \le i_1 < i_2 \le M+1} \mathbb{1}_{[\min\{x_{i_1 s}, x_{i_2 s}\}, \max\{x_{i_1 s}, x_{i_2 s}\}]}(x_s)
$$
$$
= \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} [M + 1 - \operatorname{rank}_{\mathcal{X}}(x_s)][\operatorname{rank}_{\mathcal{X}}(x_s) - 1] + M,
$$

to all vectors $\mathbf{x} \in \mathcal{X}$, where $\operatorname{rank}_{\mathcal{X}}(x_s)$ is the rank of the sth component of the vector $\mathbf{x}$ within the set $\mathcal{X}$. The band depth rank of the observation $\mathbf{Y} = \mathbf{x}_1$ is then given by the rank of $r(\mathbf{x}_1)$ in $\{r(\mathbf{x}_1), \ldots, r(\mathbf{x}_{M+1})\}$ with ties resolved at random. Calibrated forecasts should result in a uniform histogram. However, the interpretation of miscalibration in the band depth rank histogram is somewhat different than that of the classic univariate rank histogram. A skew histogram with too many high ranks is an indication of an overdispersive ensemble, while too many low ranks can result from either an underdispersive or biased ensemble. Furthermore, too high correlations in the ensemble produce a $\cap$-shaped histogram, while a $\cup$-shaped histogram is an indication of a lack of correlation in the ensemble.

Alternatively, we also investigate the fit of the correlation structure by investigating the calibration of predicted temperature differences between close-by stations that we define to be all stations within a 50-km neighborhood of the station under consideration. The form of the predictive distribution of the temperature

differences under the various models is given in the [appendix](). If the strength of spatial correlations implied by the respective postprocessing approach is adequate, the predictive distributions of temperature differences are calibrated and the corresponding PIT values are uniformly distributed on [0, 1]. Underestimating the correlation strength would entail ∩-shaped PIT histograms (i.e., PIT values would tend to accumulate around 0.5). Conversely, overestimating the correlation strength would yield PIT values closer to 0 or 1. A station-specific PIT histogram may thus be summarized by the mean absolute deviations (MADs) of the PIT values from 0.5 over all verification days and all temperature differences between this station and stations within the 50-km neighborhood. A flat histogram translates into an MAD of 0.25, smaller values go along with ∩-shaped histograms, and larger values go along with ∪-shaped histograms.

The information provided by a rank histogram may also be summarized numerically by the reliability index (RI), which is defined as

$$\text{RI} = \sum_{i=1}^{I} \left| \zeta_i - \frac{1}{I} \right|,$$

where I is the number of (equally sized) bins in the histogram and ζ_i is the observed relative frequency in bin $i = 1, \ldots, I$. The reliability index, thus, measures the departure of the rank histogram from uniformity ([Delle Monache et al. 2006]()).

b. Scoring rules

While rank histograms are a useful calibration diagnostic tool, they do not yield information on the sharpness of the predictive distributions. The latter can be evaluated by studying the average width of prediction intervals, which should be as small as possible, provided that the empirical coverage is close to the nominal coverage. As a quantitative measure for predictive performance that takes both calibration and sharpness into account, we employ several proper scoring rules ([Gneiting and Raftery 2007]()). The different scores assess different aspects of the forecasts. However, they are all negatively oriented in that a smaller score indicates a better forecast. For events with a binary outcome (e.g., "the temperature y does not exceed a threshold x"), we use the Brier score (BS; [Brier 1950]()):

$$\text{BS}_x(G, y) = [G(x) - \mathbb{1}_{[y,\infty)}(x)]^2,$$

where $G(x)$ is the predicted probability for $y \leq x$.

The continuous ranked probability score (CRPS) in [(5)]() is the integral of the Brier scores over all thresholds

$x \in \mathbb{R}$ and, thus, an overall performance measure. For multivariate distributions, the energy score (ES) provides a similar measure of predictive skill ([Gneiting and Raftery 2007]()). It is given by

$$\text{ES}(G, \mathbf{Y}) = \mathbb{E}\|\mathbf{X} - \mathbf{Y}\| - \frac{1}{2}\mathbb{E}\|\mathbf{X} - \mathbf{X}'\|, \quad (13)$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\mathbf{X}$ and $\mathbf{X}'$ are independent random vectors that are distributed according to G.

It has been noted ([Pinson and Tastu 2013]()) that the sensitivity of the energy score to misrepresentation of the correlation structure is rather limited. As an additional score, we therefore consider the Dawid–Sebastiani (DS) score that depends on the predictive mean vector $\boldsymbol{\mu}_G$ and the predictive covariance matrix $\boldsymbol{\Sigma}_G$ of the multivariate predictive distribution G via

$$\text{DS}(G, \mathbf{Y}) = -\log \det\boldsymbol{\Sigma}_G - (\mathbf{Y} - \boldsymbol{\mu}_G)^t \boldsymbol{\Sigma}_G^{-1}(\mathbf{Y} - \boldsymbol{\mu}_G)$$
$$(14)$$

([Dawid and Sebastiani 1999](); [Gneiting and Raftery 2007]()).

Finally, we provide some error measures of the deterministic forecasts that are obtained as functionals (e.g., mean or median) of the predictive distributions. For univariate probabilistic forecasts, the mean absolute error (MAE) and the root-mean-squared error (RMSE) assess the average proximity of the observation to the center of the predictive distribution. The absolute error is calculated as the absolute difference between the observation and the median of the predictive distribution, while the squared error is calculated using the mean of the predictive distribution ([Gneiting 2011]()). The Euclidean error (EE) is the natural generalization of the absolute error to higher dimensions. It is given by the Euclidean distance between the observation and the median of the predictive distribution.

Approaches to calculate the various scores under our prediction models are discussed in the [appendix]().

c. Confidence intervals for score differences

To assess the statistical significance of the score difference between two different approaches, we provide 95% confidence intervals for some of the more interesting pairings. Following [Efron and Tibshirani (1993)]() and [Hamill (1999)]() we generate 10 000-member bootstrap sample of the daily score differences (univariate scores are averaged over all locations), take the average over each sample, and report the 2.5% and 97.5% quantiles of the 10 000 average score differences obtained in this way. The implicit assumption that score

TABLE 1. Mean CRPS, MAE, and RMSE for 21-h temperature forecasts aggregated over all 514 stations and 346 days in the test set. Also reported here are the average width (PI-W) and coverage (PI-C) of 90.5% prediction intervals aggregated over the entire test set and the mean (RI-mean), minimum (RI-min), and maximum (RI-max) station-specific reliability indices.

	CRPS (°C)	MAE (°C)	RMSE (°C)	PI-W (°C)	PI-C (%)	RI-mean	RI-min	RI-max
Raw ensemble	1.56	1.77	2.27	1.50	26.0	1.29	0.72	1.63
BMA	1.04	1.46	1.86	5.91	88.9	0.35	0.14	0.97
NGR$_+$	1.04	1.46	1.87	5.76	88.0	0.35	0.12	0.93
NGR$_c$	0.96	1.35	1.73	5.17	86.6	0.23	0.13	0.84

differences are approximately independent from one day to the next seems justified in our setting where a forecast lead time of less than 1 day is considered, which implies that the underlying NWP model is re-initialized between two consecutive forecast days. We consider the score difference between two methods significant if zero is outside the 95% confidence interval.

6. Results

In this section we present the results of applying the univariate NGR$_+$ and NGR$_c$ postprocessing methods as well as their spatial extensions to forecasts from the COSME-DE ensemble prediction system, described in section 2. Additionally, we provide a comparison to the univariate ensemble BMA method of Raftery et al. (2005) and the multivariate spatial BMA approach, proposed by Berrocal et al. (2007).

a. Univariate predictive performance

Measures of univariate predictive performance of the raw COSMO-DE ensemble and the postprocessed forecasts under NGR$_+$, NGR$_c$, and BMA are given in Table 1. A simple approach to assess calibration and sharpness of univariate probabilistic forecasts is to calculate the nominal coverage and width of prediction intervals. If the ensemble members and the observation are exchangeable, the probability that the observation lies within the ensemble range is $19/21 \times 100\% \approx 90.5\%$, and so we take this as the nominal level of the considered prediction intervals. While the raw ensemble returns very sharp forecasts, it is severely underdispersive as can be seen by the insufficient coverage. This is also reflected in the numerical scores that are significantly better for all three postprocessing methods. NGR$_+$ and ensemble BMA return essentially identical scores, improving upon the ensemble by 34% in terms of the CRPS and by approximately 18% in terms of both MAE and RMSE. Ensemble BMA returns minimally wider prediction intervals than NGR$_+$, but yields an empirical coverage that is closest to the nominal 90.5%. The locally adaptive postprocessing of NGR$_c$ is slightly underdispersive, but yields the best overall scores and approximately 10% narrower prediction intervals than

NGR$_+$ on average. The station-specific reliability indices indicate that the postprocessing improves the calibration consistently across the country with the postprocessing methods always yielding lower indices than the raw ensemble. The 95% confidence intervals (not shown here) for the differences in CRPS, MAE, and RMSE between the different methods show that the performance of NGR$_+$ and BMA is not statistically significant; all other score differences observed in Table 1 are statistically significant.

b. Spatial calibration

In Fig. 3 we assess the calibration of the joint forecast fields at all 514 observation stations in Germany using multivariate band depth rank histograms. Without additional spatial modeling (i.e., assuming independent forecast errors at the different stations) the multivariate calibration of BMA, NGR$_+$, and NGR$_c$ is rather poor, despite their good marginal calibration. The three spatial forecasts that are based on parametric modeling of the error field (spatial BMA, spatial NGR$_+$, and spatial NGR$_c$) significantly improve upon the calibration of the univariate methods, in particular spatial NGR$_c$. However, the strength of the correlations seems somewhat too low as the observed field is too often either the most central or the most outlying field resulting in a ∪-shaped histogram [see also section 4 of Thorarinsdottir et al. (2014)]. In contrast, the combination of ECC and NGR produces forecast fields where the strength of the correlations appears slightly too high. This result is consistent with the spatial correlation patterns portrayed in Fig. 1 where the raw ensemble—and thus also the ECC fields—appears to have distinctly less spatial variability than the estimated Gaussian error fields.

In our spatial NGR$_+$/NGR$_c$ model we made the simplifying assumption of a stationary and isotropic correlation function. To check whether this assumption is appropriate or whether correlation strengths vary strongly over the domain considered here, we study the PITs of predicted temperature differences between close-by stations. We focus on the NGR$_c$ model and its spatial extensions where we can assume that the univariate predictive distributions have no local biases and reflect the local prediction uncertainty reasonably well
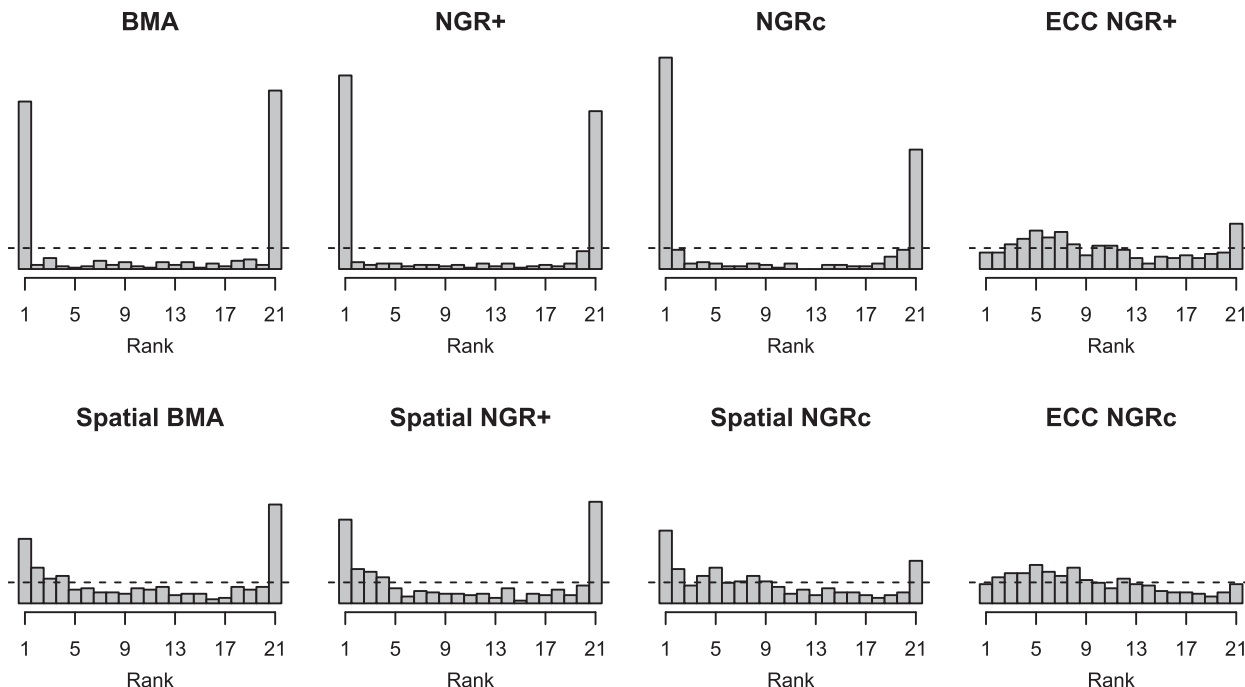
FIG. 3. Multivariate band depth rank histograms to assess the calibration of joint forecast fields at 514 stations in Germany aggregated over the 346 days in the test set.

(Scheuerer and König 2014). Departures from uniformity can then be attributed to misspecifications of the correlation strength. Figure 4 depicts, for each station, the mean absolute deviations of the PIT values from 0.5

over all verification days and all temperature differences between this station and stations within a 50-km neighborhood. As expected, in the absence of a spatial model the magnitude of temperature differences is overestimated.
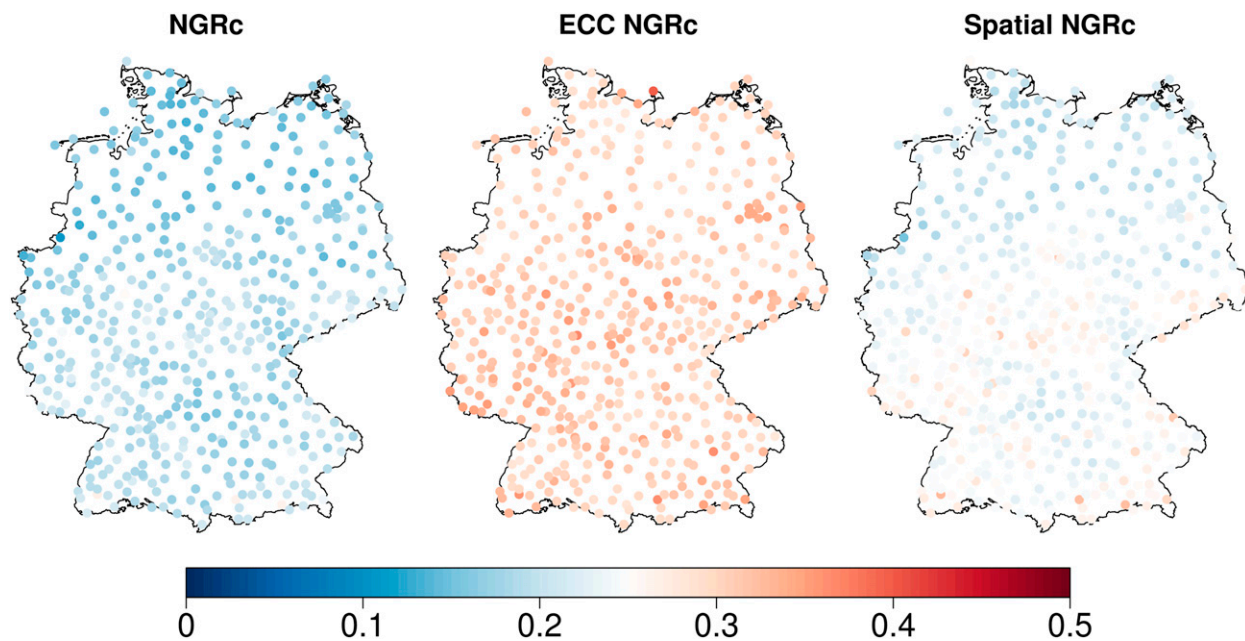

FIG. 4. Mean absolute deviations (from 0.5) of the PIT values of temperature differences between each station and all stations within a radius of 50 km. The depicted values are averaged over the respective neighborhoods and the 346 days in the test set.
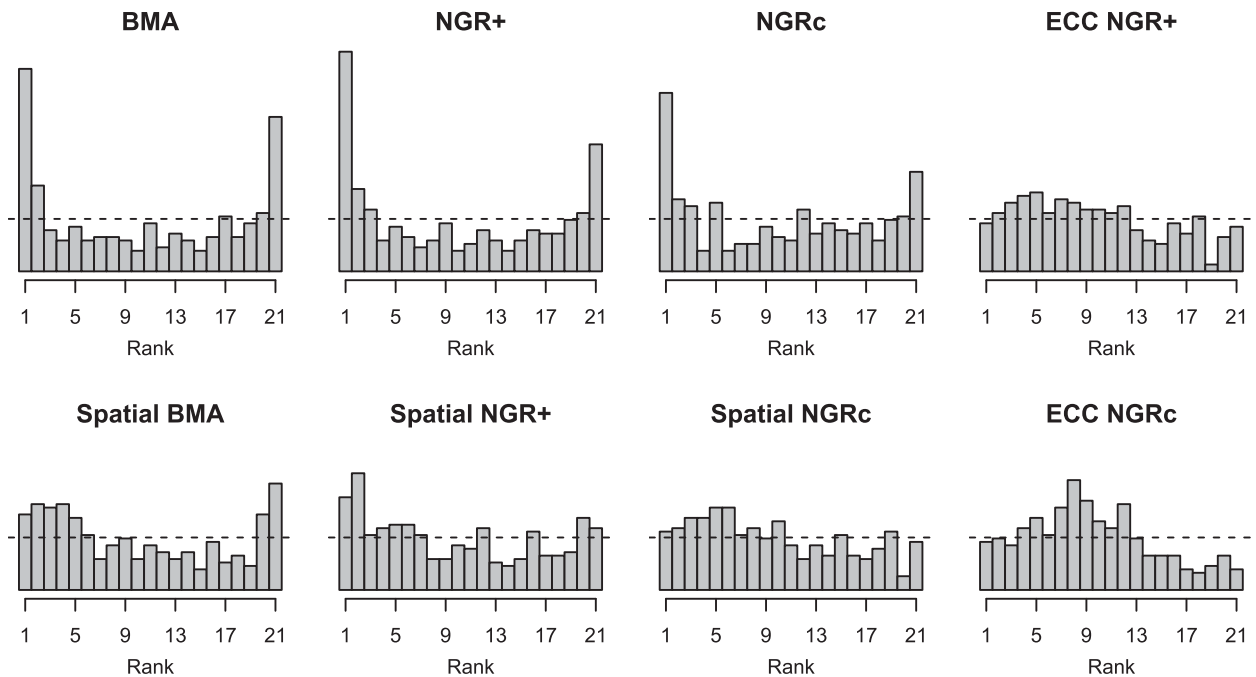
FIG. 5. Multivariate band depth rank histograms to assess the calibration of joint temperature forecasts at the seven observation stations in the state of Saarland in Germany aggregated over the 346 days in the test set.

When ECC is used to restore the rank correlations of the raw ensemble, it is underestimated (i.e., spatial correlations are too strong), which is in line with our conclusions from Fig. 3. On average, the mean absolute deviations from 0.5 of the PIT values corresponding to spatial NGR_c are closest to the value 0.25, which corresponds to perfect calibration. However, the adequate correlation strength varies across the domain. The assumption of stationarity and isotropy of our statistical correlation model entails too weak correlations over the north German plain and too strong correlations near the Alpine foothills and in the vicinity of the various low mountain ranges. That is, (10) presents a good first approximation, but a more sophisticated, nonstationary correlation model may yield further improvement.

c. Case study I: Predictive performance in Saarland

For a more quantitative assessment of multivariate predictive performance, we focus on two smaller subsets of the 514 stations. This is necessary because in our own experience, the lack of sensitivity of the energy score in (13) to misspecifications of the spatial correlation structure (Pinson and Tastu 2013) becomes even worse as the dimension of locations considered simultaneously increases.

First, we consider the joint predictive distribution at the seven stations in the state of Saarland (see Fig. 2). The corresponding multivariate band depth rank histograms in Fig. 5 confirm the conclusions from the preceding subsection in that spatial modeling significantly improves the joint calibration of the standard (nonspatial) postprocessing methods. The histograms for spatial BMA and spatial NGR_+ are still somewhat ∪-shaped, while the one for ECC NGR_c is ∩-shaped. Those for spatial NGR_c and ECC NGR_+ are slightly ∩-shaped, but closest to uniformity, which suggests that the corresponding predictions have the best calibration overall. In all histograms the lower ranks are somewhat more populated than the higher ranks, which is in line with our conclusion from Table 1 that the postprocessed forecasts tend to be slightly underdispersive.

Table 2 shows the multivariate scores over this region, and while these results are subject to some sampling

TABLE 2. Average ES, EE, and DS of joint temperature forecasts at seven observation stations in the state of Saarland in Germany over all 346 days in the test set.

	ES (°C)	EE (°C)	DS
Raw ensemble	5.27	5.86	21 210.9
BMA	3.59	4.90	20.6
NGR_+	3.59	4.91	20.7
NGR_c	3.28	4.55	17.4
Spatial BMA	3.56	4.90	16.8
Spatial NGR_+	3.57	4.91	16.9
Spatial NGR_c	3.25	4.55	14.2
ECC NGR_+	3.69	4.92	1112.9
ECC NGR_c	3.36	4.55	1665.1

TABLE 3. The 95% bootstrap confidence intervals for differences in the average ES and EE between selected postprocessing methods for seven observation stations in the state of Saarland in Germany over all 346 days in the test set.

	ΔES (°C)	ΔEE (°C)
BMA vs spatial BMA	[0.016, 0.039]	[−0.003, 0.004]
NGR$_+$ vs spatial NGR$_+$	[0.018, 0.041]	[−0.004, 0.002]
NGR$_c$ vs spatial NGR$_c$	[0.018, 0.043]	[−0.003, 0.004]
NGR$_+$ vs ECC-NGR$_+$	[−0.120, −0.077]	[−0.031, 0.003]
NGR$_c$ vs ECC-NGR$_c$	[−0.099, −0.066]	[−0.021, 0.013]
Spatial BMA vs spatial NGR$_+$	[−0.016, 0.006]	[−0.024, 0.006]
Spatial BMA vs spatial NGR$_c$	[0.190, 0.432]	[0.188, 0.510]
Spatial NGR$_+$ vs spatial NGR$_c$	[0.193, 0.439]	[0.198, 0.519]

variability, they show again a clear tendency of the spatial models yielding better multivariate performance than their univariate counterparts with spatial NGR$_c$ being especially competitive. Somewhat surprisingly, the energy scores of ECC NGR$_c$ and ECC NGR$_+$ are larger than those of NGR$_c$ and NGR$_+$. A look at Fig. 4 suggests that in the particular region considered here the overestimation of spatial dependence by the ECC technique might be more serious than its underestimation by completely ignoring spatial correlations. While the latter seems to have a strong impact on the band depth rank histograms (see Fig. 5), the energy score seems to be more sensitive to overestimation of spatial dependence, putting the ECC-based spatial models to the rear places in the performance ranking. The confidence intervals in Table 3 show that the energy score differences observed in Table 2 are statistically significant with the exception of the difference between spatial BMA and spatial NGR$_+$. With respect to the Euclidean error of the predictive medians, on the contrary, there are no significant differences between spatial and nonspatial methods. Here, it is mainly the local adaptivity of the NGR$_c$ approach that yields a significant improvement over NGR$_+$ and BMA. Finally, the Dawid–Sebastiani scores confirm the ranking between the spatial and nonspatial variants of BMA, NGR$_+$, and NGR$_c$. They do not permit a reasonable comparison with the ECC ensembles and the raw ensemble, though. The latter consist of only 20 members—as opposed to a very large sample that can be generated by all other methods—which does not warrant a stable estimation of the empirical covariance matrix. This can be disastrous when calculating the Dawid–Sebastiani score in (14) and it shows that it can be problematic in certain contexts that ECC NGR$_+$ and ECC NGR$_c$ inherit the sometimes close to singular correlation matrices from the raw COSMO-DE ensemble forecasts.

d. Case study II: Minimum temperature along the highway A3

As a second example in which the multivariate aspect of the predictive distributions becomes noticeable, we consider the task of predicting the minimum temperature along a section of the highway A3, which connects the two cities Frankfurt am Main and Cologne, Germany. For consistency with the forecasts at the individual stations and with other composite quantities, we do not set up a separate postprocessing model for minimum temperature, but derive it by taking the minimum over 11 stations along this section of the A3.

Since the minimum of several random variables depends not only on their means and variances, but also on their correlations, we expect that only the spatial postprocessing methods can provide calibrated probabilistic forecasts. Indeed, the histograms in Fig. 6 show that without spatial modeling the minimum temperature is systematically underestimated. This is a consequence of the fact that the minimum over independent random variables is on average much smaller than the minimum over positively correlated random variables. This systematic underestimation is largely avoided by spatial BMA, spatial NGR$_+$, and spatial NGR$_c$ while the ECC techniques here yield the histograms closest to uniformity. This clear advantage of postprocessing methods that account for spatial correlations is further confirmed by the CRPS and MAE scores in Table 4, and the corresponding confidence intervals in Table 5.

As an application and example of the relevance of spatial modeling in practice, consider the decision problem of dispatching or not dispatching salt spreaders when the temperatures along the considered section of the A3 are predicted to fall below 0°C. The event "temperature falls below 0°C at at least one location along the A3" is equivalent to "minimum temperature along the A3 falls below 0°C," and good decisions are, therefore, taken if this event is predicted accurately. The last column of Table 4 shows the corresponding average Brier scores (BS) over the verification days in the winter months of January, February, and November, and illustrates once again that appropriate consideration of spatial dependence is required to take full advantage of statistical postprocessing.

7. Discussion

In this paper we have proposed a postprocessing method for temperature that uses the information of a dynamical ensemble as inputs and generates a calibrated statistical ensemble as an output. By following this approach, it not only yields calibrated marginal
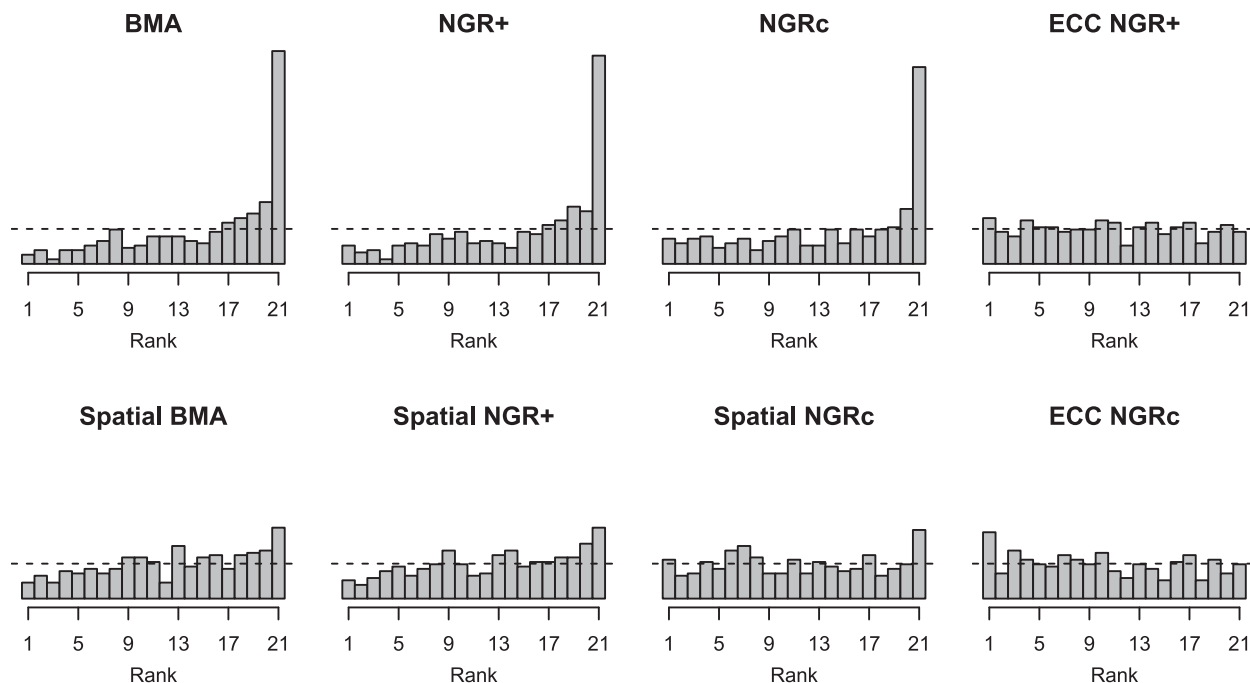
FIG. 6. Verification rank histograms for forecasts of the minimum temperature over 11 stations along a section of the highway A3, where the different cases correspond to the days of the verification period.

predictive distributions but entire temperature forecast fields, thus aiming for multivariate calibration. The importance of this property is underlined by the results presented in section 6 where forecasts of spatially aggregated quantities are studied and spatial correlations have to be considered. Our spatial NGR$_+$ approach performs similar to the spatial BMA approach of Berrocal et al. (2007). However, it is conceptually simpler and computationally more efficient; the estimation of the spatial correlation structure of spatial BMA is M times more expensive than that of spatial NGR$_+$, where M is the size of the original ensemble. This makes it an attractive alternative, especially since further extensions—such as the spatial NGR$_c$ method presented here—are also easier to implement.

In our case study using the ensemble forecasts of the COSMO-DE-EPS, the performance of the parametric spatial methods was overall slightly better than the results obtained by modeling spatial dependence via ECC. However, this result may not hold in all cases. When the (spatial) correlation structure of the ensemble represents the true multivariate uncertainty well, methods that use or retain the rank correlations (Roulin and Vannitsem 2012; Schefzik et al. 2013; Van Schaeybroeck and Vannitsem 2014) have the potential advantage that they can feature flow-dependent dependence structures while the statistical models presented here rely on the assumption that correlations are constant over a certain

period of time. A statistical approach, on the other hand, has the advantage that it determines the correlation structure based on both forecasts and observations, and thus does not inherit (or even amplify) spurious and wrong correlations that may be present in the ensemble.

The exponential correlation function used by Gel et al. (2004), Berrocal et al. (2007), and in the present paper is, of course, a somewhat simplistic model. While replacing it by a function from the more general Matérn class, which nests the exponential model as a special case, did not improve the performance of our method, Fig. 4 suggests that a nonstationary correlation function

TABLE 4. CRPS and MAE for minimum temperature forecasts over 11 stations along the highway A3 averaged over all verification days. The last column gives the BS for the event that the temperature drops below freezing (0°C) at at least one of these stations, averaged over the subset of verification days in January, February, and November 2011.

	CRPS (°C)	MAE (°C)	BS$_0$
Raw ensemble	1.74	1.92	0.120
BMA	1.08	1.41	0.121
NGR$_+$	1.05	1.37	0.114
NGR$_c$	0.96	1.27	0.103
Spatial BMA	0.86	1.20	0.079
Spatial NGR$_+$	0.86	1.20	0.082
Spatial NGR$_c$	0.82	1.16	0.079
ECC NGR$_+$	0.84	1.18	0.085
ECC NGR$_c$	0.84	1.17	0.083

TABLE 5. The 95% bootstrap confidence intervals for differences in the CRPS, MAE, and BS between selected postprocessing methods for minimum temperature forecasts over 11 stations along the highway A3 averaged over all verification days.

	ΔCRPS (°C)	ΔMAE (°C)	ΔBS$_0$
BMA vs spatial BMA	[0.170, 0.260]	[0.151, 0.264]	[0.013, 0.073]
NGR$_+$ vs spatial NGR$_+$	[0.150, 0.238]	[0.119, 0.225]	[0.007, 0.059]
NGR$_c$ vs spatial NGR$_c$	[0.091, 0.185]	[0.049, 0.171]	[−0.003, 0.052]
NGR$_+$ vs ECC-NGR$_+$	[0.134, 0.288]	[0.094, 0.294]	[−0.009, 0.068]
NGR$_c$ vs ECC-NGR$_c$	[0.049, 0.190]	[0.001, 0.187]	[−0.013, 0.057]
Spatial BMA vs spatial NGR$_+$	[−0.008, 0.009]	[−0.010, 0.016]	[−0.008, 0.002]
Spatial BMA vs spatial NGR$_c$	[0.009, 0.074]	[−0.006, 0.103]	[−0.012, 0.012]
Spatial NGR$_+$ vs spatial NGR$_c$	[0.010, 0.072]	[−0.006, 0.095]	[−0.009, 0.013]
Spatial NGR$_+$ vs ECC-NGR$_+$	[−0.022, 0.054]	[−0.039, 0.081]	[−0.020, 0.013]
Spatial NGR$_c$ vs ECC-NGR$_c$	[−0.047, 0.011]	[−0.064, 0.029]	[−0.016, 0.009]

might yield a better approximation of the true spatial dependence structure. There are a number of non-parametric modeling approaches that can potentially deal with these kinds of effects (Anderes 2011; Lindgren et al. 2011; Jun et al. 2011; Kleiber et al. 2013). However, this is rather challenging and left for future research.

A further extension of the approach presented here concerns correlations between different lead times. Instead of modeling spatial correlations only once, one would need to set up a model that captures correlations in both space and time. Similarly, some applications require appropriate correlations between different weather variables. This presents yet another multivariate aspect that has been addressed by Möller et al. (2013). Taking all three aspects—space, time, and different variables—into account would be the ultimate goal in multivariate modeling. At the same time, this further increases the level of complexity so that in this very general setting the ECC approach might be preferred just for the sake of simplicity.

APPENDIX

Predictive Distributions for Temperature Differences

Under the multivariate predictive distribution model in (11), the predictive distribution of the temperature difference between location s_i and s_j is given by

$$(\Delta y)_{s_i s_j} \mid f_{1s_i}, f_{1s_j}, \ldots, f_{Ms_i}, f_{Ms_j}, \mathcal{D}_{\mathcal{S}}^{\mathcal{T}}$$
$$\sim \mathcal{N}(\mu_{s_i} - \mu_{s_j}, \sigma_{s_i}^2 - 2\rho_{s_i s_j}\sigma_{s_i}\sigma_{s_j} + \sigma_{s_j}^2), \quad \text{(A1)}$$

where μ_{s_i}, μ_{s_j} and $\sigma_{s_i}^2$, $\sigma_{s_j}^2$ are the predictive means and variances at s_i and s_j and $\rho_{s_i s_j} = C_{\theta,r}(s_i, s_j)$ is the correlation between the forecast errors at those two locations. For each station, we calculate the observed temperature differences between this station and all stations within a radius of 50 km, and calculate the PIT values of the predictive distributions given by (A1). In the absence of a spatial model, we take $\rho_{s_i s_j} = 0$ for all $s_i, s_j \in \mathcal{S}$. For a combination of ECC and NGR where the multivariate distribution is represented by an ensemble, we approximate (A1) by the empirical cumulative distribution function (CDF) of the temperature differences predicted by the individual ensemble members.

Assuming no local biases and marginal calibration, the calibration of the temperature difference forecasts mainly depends on the correct specification of $\rho_{s_i s_j}$. It will be underdispersive if $\rho_{s_i s_j}$ is overestimated and overdispersive if $\rho_{s_i s_j}$ is underestimated. If the strength of spatial correlations implied by the respective postprocessing approach is adequate, the predictive distributions of temperature differences are calibrated and the corresponding PIT values are uniformly distributed on [0, 1].

Calculation of scoring rules

When the integral defining the continuous ranked probability score (CRPS) in (5) is not available in a closed form, the equivalent formulation

$$\text{CRPS}(G, y) = \mathbb{E}|X - y| - \frac{1}{2}\mathbb{E}|X - X'| \quad \text{(A2)}$$

may be employed instead (Gneiting and Raftery 2007). Here, $\mathbb{E}$ denotes expectation, $|\cdot|$ stands for the absolute value, and X and X' are independent copies of a random variable with cumulative distribution function G. To estimate the expression in (A2), we generate two independent samples $\mathbf{x} = \{x_j\}_{j=1}^J$ and $\mathbf{x}' = \{x_j'\}_{j=1}^J$ from the predictive distribution and calculate

$$\widehat{\mathrm{CRPS}}(G, y) = \frac{1}{J} \sum_{j=1}^{J} |x_j - y| - \frac{1}{2J} \sum_{j=1}^{J} |x_j - x_j'|,$$

where we typically set $J = 5000$. The energy score (ES) in (13) may be approximated in a similar manner.

For multivariate Gaussian distributions such as the spatial NGR predictive distributions, the Dawid–Sebastiani score in (14) is equal to the negative log-likelihood and may be calculated directly. For spatial BMA, the mean and covariance matrix may be calculated as follows. Let $\mathbf{Y}$ be a random vector with a distribution that is given by a mixture of M Gaussian distributions each with mean $\boldsymbol{\mu}_m$, covariance $\boldsymbol{\Sigma}_m$, and weight ω_m for $m = 1, \ldots, M$. Then it holds that

$$\mathbb{E}(Y_i) = \sum_{m=1}^{M} \omega_m \mu_{mi}$$

and

$$\mathbb{E}(Y_i Y_j) = \sum_{m=1}^{M} \omega_m [(\boldsymbol{\Sigma}_m)_{ij} + \mu_{mi} \mu_{mj}].$$

The former formula can now be used to calculate the mean $\boldsymbol{\mu}_G$, while the covariance matrix may be calculated by noting that $(\boldsymbol{\Sigma}_G)_{ij} = \mathbb{E}(Y_i Y_j) - \mathbb{E}(Y_i)\mathbb{E}(Y_j)$. When $\boldsymbol{\Sigma}_G$ must be estimated nonparametrically from a sample, such as for ECC, the calculations may be numerically unstable. In this case, we add 0.000 01 to all elements on the diagonal in order to improve the numerical stability (Rasmussen and Williams 2006).

The Euclidean error (EE) requires the median of a multivariate predictive distribution. It is estimated using the functionality of the R package ICSNP (Nordhausen et al. 2014).

REFERENCES

Anderes, E. B., and M. L. Stein, 2011: Local likelihood estimation for nonstationary random fields. *J. Multivar. Anal.,* **102,** 506–520, doi:10.1016/j.jmva.2010.10.010.

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

Baldauf, M., A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer, and T. Reinhardt, 2011: Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Mon. Wea. Rev.,* **139,** 3887–3905, doi:10.1175/MWR-D-10-05013.1.

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

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Bröcker, J., 2012: Evaluating raw ensembles with the continous ranked probability score. *Quart. J. Roy. Meteor. Soc.,* **138,** 1611–1617, doi:10.1002/qj.1891.

——, and L. A. Smith, 2008: From ensemble forecasts to predictive distribution functions. *Tellus,* **60A,** 663–678, doi:10.1111/j.1600-0870.2008.00333.x.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu, 1995: A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.,* **16,** 1190–1208, doi:10.1137/0916069.

Chilès, J.-P., and P. Delfiner, 2012: *Geostatistics: Modeling Spatial Uncertainty.* 2nd ed. John Wiley & Sons, 734 pp.

Cressie, N. A. C., 1985: Fitting variogram models by weighted least squares. *Math. Geol.,* **17,** 563–586, doi:10.1007/BF01032109.

Dawid, A. P., 1984: Statistical theory: The prequential approach (with discussion and rejoinder). *J. Roy. Stat. Soc. A,* **147,** 278–292, doi:10.2307/2981683.

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

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2006: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.,* **111,** D24307, doi:10.1029/2005JD006917.

Doms, G., and U. Schättler, 2002: A description of the non-hydrostatic regional model LM: Dynamics and numerics. Tech. Rep., Deutscher Wetterdienst, 134 pp.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap.* Chapman & Hall/CRC, 456 pp.

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

Gel, Y., A. E. Raftery, and T. Gneiting, 2004: Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method (with discussion and rejoinder). *J. Amer. Stat. Assoc.,* **99,** 575–590, doi:10.1198/016214504000000872.

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

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

——, ——, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118, doi:10.1175/MWR2904.1.

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

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.,* **136,** 2608–2619, doi:10.1175/2007MWR2410.1.

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

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

Jun, M., I. Szunyogh, M. G. Genton, F. Zhang, and C. H. Bishop, 2011: A statistical investigation of the sensitivity of ensemble-based Kalman filters to covariance filtering. *Mon. Wea. Rev.,* **139,** 3036–3051, doi:10.1175/2011MWR3577.1.

Kann, A., C. Wittmann, Y. Wang, and X. Ma, 2009: Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. *Mon. Wea. Rev.,* **137,** 3373–3387, doi:10.1175/2009MWR2793.1.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass, and E. P. Grimit, 2011: Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Wea. Rev.,* **139,** 2630–2649, doi:10.1175/2010MWR3511.1.

——, R. Katz, and B. Rajagopalan, 2013: Daily minimum and maximum temperature simulation over complex terrain. *Ann. Appl. Stat.,* **7,** 588–612, doi:10.1214/12-AOAS602.

Lerch, S., and T. L. Thorarinsdottir, 2013: Comparison of nonhomogeneous regression models for probabilistic wind speed forecasting. *Tellus,* **65A,** 21206, doi:10.3402/tellusa.v65i0.21206.

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

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

Lindgren, F., H. Rue, and J. Lindström, 2011: An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach (with discussion). *J. Roy. Stat. Soc.,* **73B,** 423–498, doi:10.1111/j.1467-9868.2011.00777.x.

Möller, A., A. Lenkoski, and T. L. Thorarinsdottir, 2013: Multivariate probabilistic forecasting using Bayesian model averaging and copulas. *Quart. J. Roy. Meteor. Soc.,* **139,** 982–991, doi:10.1002/qj.2009.

Nordhausen, K., S. Sirkia, H. Oja, and D. E. Tyler, 2014: ICSNP: Tools for multivariate nonparametrics, version 1.0-9. R package. [Available online at http://CRAN.R-project.org/web/packages/ICSNP.]

Peralta, C., and M. Buchhold, 2011: Initial condition perturbations for the COSMO-DE-EPS. *COSMO Newsl.,* **11,** 115–123.

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

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174, doi:10.1175/MWR2906.1.

Rasmussen, C. E., and C. K. I. Williams, 2006: *Gaussian Processes for Machine Learning*. The MIT Press, 266 pp.

R Core Team, 2013: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing,Vienna, Austria. [Available online at http://www.R-project.org/.]

Roulin, E., and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Wea. Rev.,* **140,** 874–888, doi:10.1175/MWR-D-11-00062.1.

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

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.,* **140,** 1086–1096, doi:10.1002/qj.2183.

——, and L. Büermann, 2014: Spatially adaptive post-processing of ensemble forecasts for temperature. *J. Roy. Stat. Soc.,* **63C,** 405–422, doi:10.1111/rssc.12040.

——, and G. König, 2014: Gridded, locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.,* **140,** 2582–2590, doi:10.1002/qj.2323.

Schlather, M., 2011: RandomFields: Simulation and analysis of random fields, version 3.0.44. R package. [Available online at http://CRAN.R-project.org/package=RandomFields.]

Steppeler, J., G. Doms, U. Schättler, H. W. Bitzer, A. Gassmann, U. Damrath, and G. Gregoric, 2003: Meso-gamma scale forecasts using the nonhydrostatic model LM. *Meteor. Atmos. Phys.,* **82,** 75–96, doi:10.1007/s00703-001-0592-9.

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

——, and M. S. Johnson, 2012: Probabilistic wind gust forecasting using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.,* **140,** 889–897, doi:10.1175/MWR-D-11-00075.1.

——, M. Scheuerer, and C. Heinz, 2014: Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Stat.,* doi:10.1080/10618600.2014.977447, in press.

Van Schaeybroeck, B., and S. Vannitsem, 2014: Ensemble postprocessing using member-by-member approaches: Theoretical aspects. *Quart. J. Roy. Meteor. Soc.,* doi:10.1002/qj.2397, in press.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier Academic Press, 704 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390, doi:10.1175/MWR3402.1.