Statistical Postprocessing and Multivariate Structuring of High-Resolution Ensemble Precipitation Forecasts

CHIAM VAN STRAATEN, KIRIEN WHAN, AND MAURICE SCHMEITS

R&D Weather and Climate Modeling, Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands

(Manuscript received 22 May 2018, in final form 19 September 2018)

ABSTRACT

A comparison of statistical postprocessing methods is performed for high-resolution precipitation forecasts. We keep hydrological end users in mind and thus require that the systematic errors of probabilistic forecasts are corrected and that they show a realistic high-dimensional spatial structure. The most skillful forecasts of 3-h accumulated precipitation in $3 \times 3 \text{ km}^2$ grid cells covering the land surface of the Netherlands were made with a nonparametric method [quantile regression forests (QRF)]. A parametric alternative [zero-adjusted gamma distribution (ZAGA)] corrected the precipitation forecasts of the short-range Grand Limited Area Model Ensemble Prediction System (GLAMEPS) up to +60 h less well, particularly at high quantiles, as verified against calibrated precipitation radar observations. For the subsequent multivariate restructuring, three empirical methods, namely, ensemble copula coupling (ECC), the Schaake shuffle (SSh), and a recent minimum-divergence sophistication of the Schaake shuffle (MDSSh), were tested and verified using both the multivariate variogram skill score (VSS) and the continuous ranked probability score (CRPS), the latter after aggregating the forecasts spatially. ECC and MDSSh were more skillful than SSh in terms of the CRPS and the VSS. ECC performed somewhat worse than MDSSh for summer afternoon and evening periods, probably due to the worse representation of deep convection in the hydrostatic GLAMEPS compared to reality. Overall, the high-resolution postprocessing comparison shows that skill for local precipitation amounts improves up to the 98th percentile in both the summer and winter season and that the high-dimensional joint distribution can successfully be restructured. Forecasting products like this enable multiple end users to derive their own desired aggregations.

1. Introduction

For a number of years, numerical weather prediction (NWP) forecasts have been increasingly produced in a probabilistic form instead of a deterministic one. This is needed because the expression of uncertainty about the future atmospheric state leads to enhanced decision making (e.g., Verkade and Werner 2011; Joslyn and LeClerc 2012; Ramos et al. 2013), made possible because increased computing power allowed the use of ensemble prediction systems (EPSs). Such an ensemble of deterministic NWP models is used to integrate trajectories of different initial conditions through the nonlinear dynamical system, that is, the atmosphere, taking its incomplete representation in numerical models also into account. The collection of weather states thus forms a distribution that expresses uncertainty of two inseparable sources: initial conditions and model errors (Leutbecher and Palmer 2008).

Ideally, the corresponding weather state that materializes lies within the forecast distribution, meaning that the distribution correctly covered the error between the forecast and observation. But this is often not the case, and ensemble forecasts exhibit biases and under- or overdispersion. Hence, *calibration* is needed: not a tuning of the NWP parameters, but a statistical correction of the forecast distribution such that it has the same statistical properties as the collection of events that materializes (e.g., Gneiting 2014). The correction method assumes that disparities between forecast distributions and observations are stationary. The systematic disparities differ per lead time and location and become more challenging to correct as the resolution of precipitation forecasts increases. Hence, the correction is uniquely built per lead time (Wilks and Hamill 2007) and possibly per location.

Over the past years a variety of statistical postprocessing methods have been applied to precipitation, but the variable poses a difficult challenge (e.g., Scheuerer and Hamill 2015a). The fact that precipitation is alternated by dry

Corresponding author: Dr. M. J. Schmeits, schmeits@knmi.nl

DOI: 10.1175/JHM-D-18-0105.1

© 2018 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

weather gives the distribution a discrete separation into a probability of being exactly zero and a skewed continuous distribution for the nonzero precipitation amounts. A simple bias correction will not do because additive terms affect the zero forecast, and multiplicative terms cannot alter the frequency of zero precipitation (Gneiting 2014). Also, the nonzero amounts exhibit more spread (i.e., uncertainty) with increased amounts of forecast precipitation (heteroskedasticity).

Several elaborate methods exist. Either they apply a direct transformation such as quantile mapping (Panofsky et al. 1968) or they construct a parametric two-stage model where the probability of precipitation is a logistic regression function of the predictors and a skewed distribution is used for the nonzero precipitation amounts (Sloughter et al. 2007; Berrocal et al. 2008; Schmeits and Kok 2010). The model characterizes the whole distribution at once or is used as kernel in an ensemble dressing method like Bayesian model averaging.

In parametric approaches to precipitation, the assumed underlying distributions can be ill fitted (Flowerdew 2014). Therefore, a nonparametric machine learning method is also applied in this study. The postprocessed distribution is derived from a collection of regression trees, called quantile regression forests (QRF; Meinshausen 2006; Taillardat et al. 2016, 2017; Whan and Schmeits 2018), which avoids the assumption of a certain parametric distribution. It is compared to an established parametric two-stage method, the zero-adjusted gamma distribution (ZAGA), which has a location parameter, a shape parameter, and a parameter for point mass at zero (Fig. 1; Rigby and Stasinopoulos 2010). Another parametric alternative is the (heteroskedastic) extended logistic regression (Wilks 2009; Messner et al. 2014), but it is not chosen due to its worse performance than ZAGA and QRF for extremes (Whan and Schmeits 2018).

As mentioned, the systematic disparities can differ per location. When a single statistical model is fitted to the varying predictor values of all locations (a spatial pool), it results in a single set of parameter values and multiple different local univariate distributions that do not guarantee local calibration. A first accommodation to this problem is to fit separate models, each with its own unique parameter values. Kleiber et al. (2011) apply this to precipitation and enforce spatial consistency through a geostatistical model for parameter values. A second accommodation is to transform the predictors in the statistical model into measures of local bias (Stauffer et al. 2017) and still fit a single model, which results in separate but locally calibrated distributions.

The fact that corrected univariate distributions are independent of each other is inconvenient because many users require a joint distribution through space and time

FIG. 1. Example of a zero-adjusted gamma distribution with a point mass at zero.

(Wilks 2015). Regarding precipitation, these can be forecasts of a compound quantity such as probability of precipitation somewhere in an area, or the generation of precipitation fields to force a hydrological forecast system (Verkade et al. 2013). They require multivariate descriptions that capture the spatial, temporal, or cross-variable dependence between univariate forecast distributions.

These descriptions are also known as *copulas* (Genest and Favre 2007) and can be applied in parametric form to precipitation forecasts (Berrocal et al. 2008; Möller et al. 2013), but because the additional parameters and assumptions become unfeasible in high-dimensional settings, empirical methods have also been developed. In those cases, the univariate distributions are sampled and reordered by applying the rank structure (a nonparametric, discrete copula) of a dependence template. In case of the Schaake shuffle (SSh; Clark et al. 2004), the template is a random selection of observations from the historical record, which implicitly assumes that the spatiotemporal rank correlation is independent of the forecast atmospheric state. In case of empirical copula coupling (ECC; Schefzik et al. 2013), the raw ensemble itself serves as template, which is clearly linked to the atmospheric state but can transfer spurious correlations of the NWP model fields into the postprocessed multivariate distribution (Flowerdew 2014). Both empirical methods thus risk the magnification of an incorrect dependence template when it is applied to the calibrated marginals. Recent research has therefore tried to sophisticate the methods (Ben Bouallègue et al. 2016; Schefzik 2016; Scheuerer et al. 2017; Bellier et al. 2017; Wu et al. 2018).

Previous multivariate comparisons have been made for various variables and often for settings with low dimensionality or resolution (Table 1). This study will compare the two fundamental empirical methods (ECC and SSh) and a sophisticated version of SSh, namely, the minimum-divergence Schaake shuffle (MDSSh; as in Scheuerer et al. 2017), but here the comparison will be in the context of high-resolution precipitation

TABLE 1. Studies on multivariate restructuring techniques (dimensionality is approximate).

Authors	Predictand	Resolution	Dimensionality	Structuring	(Empirical) technique
Clark et al. (2004)	Precipitation, temperature	Station, 24-h accumulation	756	Cross variable, space, time	SSh
Berrocal et al. (2008)	Precipitation	Station	197	Space	Parametric
Schefzik et al. (2013)	Temperature, pressure	Station	3	Space	ECC
Wilks (2015)	Temperature, dewpoint	Station	4	Cross variable, time	ECC, SSh
Shrestha et al. (2015)	Precipitation	Subbasin 80 km ² , 3-h accumulation	2960	Space, time	SSh
Feldmann et al. (2015)	Temperature	Station	514	Space	ECC and parametric
Hu et al. (2016)	Precipitation	Station, 24-h accumulation	4	Space	ECC
Ben Bouallègue et al. (2016)	Wind	Station	21	Time	Dual-ECC, ECC
Schefzik (2016)	Temperature	Station	2	Space	SSh, SimSSh, ECC
Schefzik (2017)	Temperature	Station	3	Space	ECC
Scheuerer et al. (2017)	Precipitation, temperature	Subbasin, 24-h accumulation	840	Space, time	SSh, MDSSh, ECC
Bellier et al. (2017)	Precipitation	Subbasin 290 km ² , 6-h accumulation	100	Space, time	SSh, AnSSh, AnSSh-ECC, MDSSh, ECC
Wu et al. (2018)	Precipitation	19 × 19 km ² grid, 6-h accumulations	2500	Space	SSh, SSh-ANA, ECC
This study	Precipitation	3 × 3 km ² grid, 3-h accumulation	4497	Space	SSh, MDSSh, ECC

forecasts that should benefit multiple users. We therefore compare two statistical postprocessing methods (ZAGA and QRF) to find the most skillful univariate forecasts, and compare three multivariate restructuring methods (ECC, SSh, and MDSSh) to find the most suited in the high-dimensional setting, that can transform the forecasts to precipitation fields from which desired compound quantities can be derived. Section 2 introduces the EPS and the high-resolution observations. Section 3 describes all applied methods and verification techniques. In section 4 the results are presented, and section 5 concludes and discusses the results.

2. Data

The NWP precipitation data are from the Grand Limited Area Model Ensemble Prediction System (GLAMEPS), which is essentially a multimodel dynamical downscaling of the ECMWF EPS for Europe. We have used the available unlagged forecasts of version 2, initiated at 0000 UTC each day, from October 2014 to December 2016. The ensemble is composed of the hydrostatic HIRLAM (Yang 2008) and Alaro (Gerard et al. 2009) models, each with two different parameterization schemes: HIRLAM with either the Kain–Fritsch or the Soft Transition Condensation (STRACO) convection scheme, and Alaro with either the ISBA or the Surface Externalisée (SURFEX)

surface scheme (Deckmyn 2014). This provides $4 \times 7 = 28$ different ensemble members with output intervals of 3 h up to a lead time of +60 h, on an original rotated-pole latitude–longitude grid of $0.075^\circ \times 0.075^\circ$.

The observational dataset spans from January 2008 to December 2016 and is composed of calibrated radar data covering the land surface of the Netherlands, with high temporal (1-h accumulation) and spatial resolution ($1 \times 1 \text{ km}^2$; Overeem et al. 2009). The measured reflectivity is often influenced by tall objects on land (land clutter). At some locations in the Netherlands this distorts the derived precipitation. By looking at the annual rainfall sums of 2012, 2014, and 2016, a total of 43 cells of $1 \times 1 \text{ km}^2$ were removed from the dataset. The clutter at these locations was caused by the ships in Rotterdam harbor; buildings in The Hague; and towers in Cabauw, Goes, and Hoogersmilde. For the occasional “beam occultation,” however, no suitable correction was available (Overeem et al. 2009). Also not corrected is a systematic error in the manual rain gauge data that is used to calibrate the radar data, as this was discovered only recently by KNMI.

The application and verification of the statistical postprocessing methods is only possible for the overlapping period of October 2014–December 2016 and is stratified into a warm half year (15 April–14 October) and a cold half year (15 October–14 April), which are respectively called “summer” and “winter” from now on. Both forecasts and radar data were transformed into

FIG. 2. The empirical cumulative distribution functions of the raw ensemble mean forecasts and the calibrated radar observations, pooled over all $3 \times 3 \text{ km}^2$ grid cells in the Netherlands and the time period from October 2014 to December 2016.

the high-resolution predictand: 3-h accumulated precipitation fields on a $3 \times 3 \text{ km}^2$ grid, by using the nearest neighbor GLAMEPS forecasts and spatially averaging the calibrated radar data. This transformation introduces a representativity error: the value in the $\pm 9 \times 9 \text{ km}^2$ GLAMEPS grid cell becomes the forecast for multiple $3 \times 3 \text{ km}^2$ grid cells, without displaying the larger variability due to lower degrees of aggregation. This causes (apart from the common underforecasting of extreme precipitation amounts by the EPS) fewer extremes to be present in the forecast than are observed (Fig. 2). The statistical postprocessing in this case thus bears similarity to statistical downscaling, which also requires the reconstruction of finescale variability (Maraun et al. 2010). Also visible in the figure is that the performance of the EPS changes with lead time: the common overforecasting of low amounts between 0 and 1 mm seems worse at +60 h than at +3 h.

In this study the raw precipitation forecast at time t is denoted by

$$x_t = \{(x_{1,t}^1, \dots, x_{1,t}^M), \dots, (x_{S,t}^1, \dots, x_{S,t}^M)\}, \quad (1)$$

where $M = 28$ is the number of members and S is the multivariate dimensionality of the simultaneously initialized members. It has a maximum of the number of

grid cells times the number of lead times, but because we correct and restructure per lead time (accounting only for spatial dependencies), S is limited to the 4497 grid cells s . Each time step t (unique combination of lead time and initialization time) of the raw forecast thus makes up an empirical univariate forecast distribution for predictand Y that is not yet calibrated:

$$F_{\text{raw},s,t}(y) = \Pr(y \leq Y) = \frac{1}{M} \sum_{m=1}^M \mathbb{1}(x_{s,t}^m \leq y), \quad (2)$$

where $\mathbb{1}$ is the indicator function and for which the assumption is made that the ensemble members are equally probable. Properties characterizing this distribution are derived for each location and time step (Table 2) and are used as potential predictors U in the statistical models.

3. Methods

a. Univariate postprocessing and verification methods

1) QUANTILE REGRESSION FORESTS

The first statistical postprocessing method, the non-parametric QRF, is basically a collection of regression trees (Meinshausen 2006; Taillardat et al. 2016). Such a

TABLE 2. Constructed predictors to characterize the precipitation distribution $F_{\text{raw},s,t}$ [(2)] of assumed equally probable ensemble members. The HirK and HirS subensembles are HIRLAM with either the Kain–Fritsch or STRACO convection schemes, respectively, and AloI and AloS are Alaro with either the ISBA or SURFEX surface schemes, respectively.

Name	Formulation
Mean	$\bar{x}_{s,t} = \frac{1}{M} \sum_{m=1}^M x_{s,t}^m$
Std dev	$\sigma_{s,t} = \sqrt{\frac{1}{M-1} \sum_{m=1}^M (x_{s,t}^m - \bar{x}_{s,t})^2}$
Skewness	$\frac{1}{M-1} \frac{1}{\sigma_{s,t}^3} \sum_{m=1}^M (x_{s,t}^m - \bar{x}_{s,t})^3$
Kurtosis	$-3 + \frac{1}{M-1} \frac{1}{\sigma_{s,t}^4} \sum_{m=1}^M (x_{s,t}^m - \bar{x}_{s,t})^4$
Q30, Q50, Q90	$F_{\text{raw},s,t}^{-1}(p)$, for $p = 0.3, 0.5, 0.9$
P03, P1, P3, P5	$\Pr(y > q) = 1 - F_{\text{raw},s,t}(q)$, for $q = 0.3, 1, 3, 5$ mm
Nrzero	Members with almost zero precipitation $\sum_{m=1}^M \mathbb{1}(x_{s,t}^m < 0.05 \text{ mm})$
subens _{mean}	Mean with $M = 7$ for the subensembles: HirK, HirS, AloI, and AloS
subens _{stddev}	Std dev of the subensembles
subens ₀₀₀	Control member of the subensembles

tree uses binary splits to divide the precipitation observations y_{obs} of the dependent training data into two homogeneous groups at each node. It does so by finding a suitable threshold in one of the predictor variables, and keeps on splitting until a stopping criterion is reached and all observations are collected in a set of terminal nodes (Breiman et al. 1984). A forecast in the independent set then consists of the analogous observations in the node to which the associated independent predictor values ($U = u_{s,t}$) lead. Because a small set of homogeneous values is too similar to be a well calibrated uncertainty distribution, QRF employs a large amount of differently fitted trees. One tree assigns equal weights to the analog observations with predictive power, and the forest averages those weights to w_i for each observation in the joint collection of n . Together, they form a smooth postprocessed empirical cumulative distribution function:

$$F_{\text{qrf}|U=u_{s,t}}(y) = \sum_{i=1}^n w_i(u_{s,t}) \mathbb{1}(y_{\text{obs},i} \leq y). \quad (3)$$

The chosen number of 500 trees differs in their random (bootstrapped) selection from the training set and in

their random selection of potential predictors. The stopping criterion was set to a minimum of 1200 observations in each terminal node, which led to the best splitting depth for this study's amount of training data. This amount depended on the choice to fit a single model per lead time and per season, as the systematic forecast error differs between (mostly convective) summer precipitation and (mostly nonconvective) winter precipitation. We opted for a spatial pool because of the relative homogeneity of the Netherlands, which provided us with 2×10^6 observations of which a spatial fraction of 0.04 was randomly selected because of computational limitations. The fitting was handled with a threefold cross validation, where each QRF model was trained with $2 \times 10^6 \times 2/3 \times 0.04 \approx 5.3 \times 10^4$ cases.

2) ZERO-ADJUSTED GAMMA DISTRIBUTION

The second statistical postprocessing method, ZAGA, was fitted to the exact same random selections of training data. Its two-stage parametric distribution has the following probability density function (pdf; Rigby and Stasinopoulos 2010):

$$f_{\text{zaga}|\mu,\sigma,\nu,U=u_{s,t}}(y) = \begin{cases} \nu, & \text{if } y = 0, \\ (1-\nu) \left[\frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y^{(1/\sigma^2)-1} e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)} \right], & \text{if } y > 0, \end{cases} \quad (4)$$

where its expected value is influenced by two parameters [$E(Y) = (1 - \nu)\mu$] and the variance by three [$\text{Var}(Y) = (1 - \nu)\mu^2(\nu + \sigma^2)$]. The previously mentioned heteroskedasticity in precipitation is captured in a power law similar to [Scheuerer and Hamill \(2015a\)](#), as $\log(\mu)$ and $\log(\sigma)$ are linearly related to predictors. The point mass at zero ν is fitted using a logit transformation. For a fixed two-predictor version “fixtwo,” this resulted in the following equations (see also the predictor list in [Table 2](#)):

$$\begin{aligned}\log(\mu) &= \alpha_1 \text{P03}_{s,t} + \alpha_2 \text{stdev}_{s,t}, \\ \log(\sigma) &= \beta_1 \text{nrzero}_{s,t} + \beta_2 \text{mean}_{s,t}, \quad \text{and} \\ \log\left(\frac{\nu}{1 - \nu}\right) &= \gamma_1 \text{nrzero}_{s,t} + \gamma_2 \text{P03}_{s,t}.\end{aligned}\quad (5)$$

The two chosen predictors for each parameter were partly based on a stepwise forward objective predictor selection that minimized the generalized Akaike information criterion on the dependent set, only partly because the objective method produced counterintuitive results. The ensemble mean, an important predictor in QRF, was, for instance, hardly selected by the objective model. This pushed us to create the fixed version above, in which (also counterintuitively) the ensemble mean was more suited for σ than for μ , and a fully objective version “seltwo.” The verification of both ZAGA models is presented in the results.

3) UNIVARIATE VERIFICATION METRICS

The verification of a probabilistic forecast concerns the joint distribution of forecasts and corresponding observations and is executed on the independent set. The question of whether the forecasts and observations are statistically consistent can be subdivided into aspects (e.g., [Toth et al. 2003](#); [Wilks 2011](#); [Thorarinsdottir and Schuhen 2017](#)). *Reliability* indicates whether the observed frequency of an event, conditional on a certain forecast probability, is equal to that probability. *Sharpness* concerns the concentration of forecast uncertainty in the predictive distributions and is a property of the forecast only. A climatological forecast is perfectly reliable as it gives a 1/100 chance to an event with return time 100. However, it lacks sharpness as the spread of the distribution remains as wide as the climatological distribution itself, and it lacks *resolution* as the averages of observed frequencies for the (in this case constant) forecast probabilities do not differ from each other. Forecasts without resolution do not resolve the occasions in which the event is more/less likely to occur.

Because classical assessments in the form of rank histograms or the probability integral transform are shown to be insufficient on their own ([Gneiting et al. 2007](#)), the

proper Brier score is used in this study to address both reliability and resolution ([Wilks 2011](#)). It assesses the mean square error (MSE) of probabilistic forecasts p_j (from 0 to 1) for binary events o_j (0 or 1) in the verification set of size n :

$$\text{BS} = \frac{1}{n} \sum_{j=1}^n (p_j - o_j)^2. \quad (6)$$

In the case of precipitation, such an event is often the exceedance of a certain threshold. The forecast exceedance probability $p_{j|U=u_{s,t}}$ for the predictor values at that location and time is related to the nonexceedance probability given by $F_{\text{zaga},s,t}$, a definite integral of (4), and directly by the smooth empirical cumulative distribution function (ecdf) in (3). This direct derivation prevents sampling effects on the scores of the postprocessed distributions ([Ferro et al. 2008](#); [Hu et al. 2016](#)). For the raw ensemble in (2), however, the score is affected by the rugged 28-member ecdf that leads to less precise derivation of the exceedance probabilities ([Hemri et al. 2014](#)). The Brier score is transformed to a Brier skill score (BSS) relative to the (resolutionless) local sample climatology (seasonally stratified for the period from October 2014 to December 2016):

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{\text{clim}}}. \quad (7)$$

The assessment of reliability, resolution, and sharpness is complemented with reliability diagrams (e.g., [Wilks 2011](#)) for the same selected thresholds. Because of computational reasons, the verification was only performed for a random selection of 20% of the cells in the Netherlands. This still provided enough forecast–observation pairs (1×10^5) for each season–lead time combination.

b. Multivariate structuring and verification methods

1) EMPIRICAL STRUCTURING

The univariate distributions $F_{\text{qrf},s,t}$ or $F_{\text{zaga},s,t}$ signify forecast uncertainty and can be understood in two ways: either as a set of precipitation values corresponding to slightly different situations of which we cannot say which one will materialize, or as a distribution of forecast error around an expected value. When, for example, fields for hydrological models are generated, these two interpretations illustrate that there is no way to do so without a multivariate structure: at no moment in time will the same quantile materialize in all of the univariate distributions in the field ([Fig. 3](#), middle row), as this would imply the same forecast error everywhere.

FIG. 3. An example of the observed precipitation field on the right, with the left and center panels showing (top) two raw ensemble members, (middle) two quantile samples, and (bottom) two random samples from postprocessed distributions for 1500 UTC 2 Sep 2015. Presentation after [Feldmann et al. \(2015\)](#).

For calibrated marginals it makes more sense that the realization is above the expected value at some neighboring places and below at other neighboring places. And indeed, the observed materialization (Fig. 3, right) has a distinct structure and is certainly not lacking structure like the nonphysical independent random sampling (Fig. 3, bottom row).

The restoring of structure by empirical methods is always by means of the rank dependence ([Genest and Favre 2007](#); [Scheffzik et al. 2013](#)). The methods differ only in where the template for rank dependence is taken from. The general steps are sample, rank, and reorder.

First, as many values are sampled from the calibrated marginal distributions as there were ensemble members $M = 28$. This sampling from the postprocessed distribution can take multiple forms. Previous studies found that equidistant quantile sampling leads to better results

than random sampling ([Scheffzik et al. 2013](#); [Wilks 2015](#)), but the equidistant sampling permits multiple plotting position estimators α ([Wilks 2011](#)):

$$\hat{y}_{s,t}^m = F_{s,t}^{-1}\left(\frac{m - \alpha}{M + 1 - 2\alpha}\right), \quad \text{for } m = 1, \dots, M. \quad (8)$$

When the Weibull plotting position estimator is used ($\alpha = 0$), the tails can be underrepresented as mentioned in [Hu et al. \(2016\)](#). But this is less so for the Hazen estimator ($\alpha = 0.5$) employed in this study. This sampling leads to M predicted precipitation values $\hat{y}$ that are already ordered from the smallest to the largest quantile:

$$\hat{y} = \{(\hat{y}_1^1, \dots, \hat{y}_1^M), \dots, (\hat{y}_S^1, \dots, \hat{y}_S^M)\}. \quad (9)$$

Then, templates are sought that have the same dimensionality S as the multivariate forecast and also

contain M members/scenarios. Because this study only restructures the univariate forecasts in space,¹ the template is formed by 28 fields with $S = 4497$:

$$z = \{(z_1^1, \dots, z_1^M), \dots, (z_S^1, \dots, z_S^M)\}. \quad (10)$$

At each location s the 28 values in the template are ranked with respect to the other values in the marginal template distribution:

$$r_s^m = \text{rank of } z_s^m \text{ in } z_s^1, \dots, z_s^M. \quad (11)$$

When executed for all S locations in one materialization m , this gives a field of ranks in the template marginals and makes us select the r_s^m th value in each of the S forecast marginals (that were already ordered due to quantile sampling). The copied multivariate correlation structure provides a plausible materialized field:

$$\tilde{y}^m = \hat{y}_1^{r_1^m}, \dots, \hat{y}_S^{r_S^m}. \quad (12)$$

Overall, the M scenarios do not change precipitation values but select fields from the sampled univariate postprocessed distributions with a consistent rank correlation structure. Together they form the reordered multivariate forecast:

$$\tilde{y} = \{(\hat{y}_1^{r_1^1}, \dots, \hat{y}_1^{r_1^M}), \dots, (\hat{y}_S^{r_S^1}, \dots, \hat{y}_S^{r_S^M})\}. \quad (13)$$

2) ECC, SSH, AND MDSSH

The empirical structuring by rank dependence is similar for ECC, SSH, and MDSSH, while only the dependence templates differ. In this study, ECC is straightforwardly implemented by using the raw GLAMEPS forecasts [(1)] as the template [(10)]. The template of SSH is randomly sampled from the complete historical record (2008–16) and narrowed down to observed fields at the same time of the day and within a ± 12 -day window, similar to the original implementation by Clark et al. (2004). MDSSH also samples from the complete historical record but compares the marginal distributions of the selected fields to the postprocessed univariates. In this way the template will contain similar marginals, from which it makes more sense to copy the rank structure. The window is set to ± 22 days, and fields are omitted stepwise until 28 remain. For the computational details the reader is referred to Scheuerer et al. (2017).

3) MULTIVARIATE VERIFICATION METRICS

The simplest way to verify a multivariate probabilistic forecast is to aggregate it to a univariate compound quantity that is sensitive to the dependence structure (Livezey 2003; Scheuerer et al. 2017) and evaluate that with the continuous ranked probability score (CRPS):

$$\text{CRPS}(F, y) = \mathbb{E}_F |X - y| - \frac{1}{2} \mathbb{E}_p |X - X'|, \quad (14)$$

where F is the forecast distribution, y is the verifying observation, and X and X' are independent random variables from the distribution F . For example, hydrological end users, such as water boards in the Netherlands, could aggregate the reordered precipitation fields to the average accumulation in their catchment. In this study, we have used averaged precipitation in each of the 12 provinces of the Netherlands. Other examples of aggregation to a univariate quantity are differences between two locations (Feldmann et al. 2015) or a compound quantity like a heatwave index (Wilks 2015).

In addition to the province-aggregated CRPS, we directly score the multivariate distributions. There are several multivariate scores (Gneiting and Raftery 2007; Thorarinsdottir et al. 2013) of which only one can be applied in this setting. The Dawid and Sebastiani score (Dawid and Sebastiani 1999) is based on the covariance matrix that is too hard to estimate in this high dimensional setting, when it does not directly follow from the univariate correction (Wilks 2017). The energy score is relatively insensitive to differences in the multivariate correlation structures (Scheffzik 2016), which is not desired because the univariate marginals are the same for all compared restructuring methods and the only difference is in the structure. As an alternative, Scheuerer and Hamill (2015b) have proposed the variogram score (VS), which is proper and specifically compares all point-by-point differences in the observation vector $y_{\text{obs},t}$ with the expected differences in the M forecast fields:

$$\text{VS}(F_t, y_{\text{obs},t}) = \sum_{i,j=1}^S w_{ij} (|y_{\text{obs},i} - y_{\text{obs},j}|^p - E|\hat{y}_i - \hat{y}_j|^p)^2. \quad (15)$$

They found that the implementation of order $p = 0.5$ had favorable sampling properties and is therefore also implemented here, and just like Scheffzik (2016), the inverse of the Euclidean spatial distance is used as weights:

$$w_{ij} = \frac{1}{\text{dist}(i,j)} \cdot \frac{1}{\sum_{i,j=1}^S \frac{1}{\text{dist}(i,j)}}. \quad (16)$$

¹ The subscript t is dropped for convenience, though the procedure is executed for each unique combination of initialization time and lead time.

For comparison, the VS is transformed to a variogram skill score (VSS) with ECC as reference [similar to (7)]. It thereby scores only the realism of the spatial structure, because, as mentioned, this study disregards temporal restructuring. Contrary to the univariate verification, we use the complete field and not just a sample of 20%.

4. Results

a. Univariate postprocessing

The thresholds for which the BSS was computed correspond to the climatological quantiles in Table 3. An accumulation of more than 7 mm is extreme as it occurs in only 0.6% of the cases, due to the relatively short accumulation periods (3 h). The spatial mean BSS up to this threshold is plotted against lead time in Fig. 4 and is computed for the full period from October 2014 to December 2016 using threefold cross validation.

When we regard skill of the raw and postprocessed forecasts relative to the dashed climatological reference, we see four things: 1) skill generally decreases with lead time, 2) skill decreases with threshold, 3) skill in summer is generally lower than in winter, and 4) the skill in summer exhibits a diurnal cycle. First, the decreasing skill with lead time is explained by error growth in the nonlinear system. With increasing lead time the ensemble members start to diverge and eventually approach the zero skill climatological distribution. An additional effect is that systematic disparities between the raw forecast distributions and observations disappear, such that statistical postprocessing cannot add any skill. Second, as uncertainty increases with the amount of expected precipitation, it is understandable that the infrequent exceedance of high thresholds is less predictable than the more frequent exceedance of lower thresholds. In Fig. 4 the skill of all models drops with increasing threshold and the lines become more spiked, indicating a greater uncertainty in BSS computation. In winter the first occurrence of negative skill is for multiple lead times at the 7-mm threshold. Third, for equivalent situations in summer the skill has already disappeared. Instead of the large-scale wintertime precipitation patterns, the summer regime is more dominated by deep convection due to daytime heating, which is harder to predict and easily displaced at this high spatial resolution. Finally, the presence of this mechanism appears also as diurnal skill cycle: only after 1500 UTC [i.e., 1700 central European summer time (CEST)], when convective showers had a chance to develop, skill starts to drop.

When we regard the skill added by postprocessing for moderate events, there is always one method that adds

TABLE 3. Climatological quantiles, associated with the 3-h accumulation thresholds used for verification, computed by pooling observations in summer and winter.

Threshold (mm)	Quantile
0.3	Q.84
1	Q.90
3	Q.96
5	Q.98
7	Q.994

skill to the raw forecast at each lead time. For thresholds above the 5-mm threshold (Q.98), the postprocessed forecasts are comparable to or worse than the raw forecasts. At small thresholds (0.3 and 1 mm), QRF and ZAGA seem to behave similarly, meaning that even though ZAGA uses only two predictors, far fewer than the number of predictors used in QRF, it apparently captures the systematic relations and improves forecasts over the full 60-h lead time range. The ZAGA methods perform worse than QRF for a number of lead times above the threshold of 3 mm. Especially in the 7-mm panel for summer, we see that QRF remains close to the raw ensemble and zero skill climatology, while the skill of the parametric ZAGA methods becomes negative. QRF is a nonparametric method that cannot predict outside its sample training data range, while for ZAGA the fit for the tails can be poor because the model fit is mostly determined by the bulk of the distribution (Whan and Schmeits 2018).

It is interesting to note that the objective predictor selection does not always perform best, as demonstrated by the 3-, 5-, and 7-mm thresholds in summer and the 7-mm threshold in winter, where ZAGafixtwo performs better. Apparently the ability of the objective method to optimize itself for each training sample can lead to overfitting when verified on the independent set. Therefore, in the remainder of this section, only verification results are shown for ZAGafixtwo and QRF.

When the BSS for +9 h is plotted spatially on the employed 20% of the grid cells (Fig. 5), we see in the top row (0.3 mm) that both QRF and ZAGafixtwo raise skill uniformly over the country. However, for the more extreme threshold of 7 mm, when postprocessing methods on average do not add skill, the performance varies spatially. ZAGafixtwo seems especially to decrease skill at several locations and increase it at a few others. This can be the result of sampling variability but might also be an effect of the spatially pooled fit, indicating that even in an apparently homogeneous country like the Netherlands the postprocessing might benefit from corrections of local errors instead of pooled disparities. However, for correct local fits a larger

FIG. 4. Spatial mean of the BSS relative to local sample climatology (dashed line), presented for the raw ensemble and the models fitted separately for each season and lead time.

temporal range of training data should be available, especially for extreme thresholds.

Reliability diagrams for +12 h show that all methods improve forecasts at the threshold of 0.3 mm (Fig. 6, top row). The overforecast low precipitation accumulations (see also Fig. 2) are corrected and reliability improves (i.e., the postprocessed forecasts move closer to the 1:1 line compared to the raw forecasts). However, at the 3-mm threshold in summer (Fig. 6, left-middle panel) differences emerge between QRF and ZAGAfxtwo for forecast exceedance probabilities $p_i > 0.4$. QRF has less resolution and does not issue high probability forecasts, while ZAGAfxtwo does issue them but is not more skillful than QRF and the raw forecasts (Fig. 4). In winter QRF issues higher probabilities and appears to be reliable. These patterns at 3 mm are amplified in the

bottom row which concerns the 5 mm threshold (Q.98), with little predictability in summer and a large sampling variability due to fewer occurrences.

Although QRF, as a nonparametric method, does not forecast the full range of p_i , it appears to be a better univariate postprocessing method than ZAGA. We evaluate which of its predictors are most important: whose splits, if randomly permuted, deteriorate the forecast the most in terms of mean square error. Because the values of this importance are dependent on the actual precipitation amounts (Taillardat et al. 2016), and thus differ for the two seasons, we ranked the values to study the predictors' relative importance. For both seasons (Fig. 7) it appears that the ensemble mean is very important (which we expect because this is a smoothed summary of the ensemble that captures the mean atmospheric patterns), while the

FIG. 5. Spatial pattern of the BSS for the (left) raw, (center) QRF, and (right) ZAGA (with two fixed predictors) relative to local sample climatology, presented for the +9-h lead time and the following seasons and thresholds: (top) 0.3 mm in summer, (middle) 3 mm in winter, and (bottom) 7 mm in winter.

higher-order moments (skewness and kurtosis) matter less. The Q.90 appears informative, as it often represents the nonzero precipitation amounts. Alaro predictors are more important in winter than in summer. The reverse is true for the HIRLAM predictors.

b. Multivariate structuring

We use only QRF to compare the multivariate restructuring methods as it was the most skillful method

in the univariate setting. We sample 28 members from the postprocessed distributions and restructure them using the different templates created by each method (ECC, SSh, MDSSh).

First, we verify the reshuffled 28-member distributions using the multivariate VSS relative to ECC [(15)]. According to this metric, MDSSh scores higher than SSh for almost all lead times in both seasons (Fig. 8). Relative to ECC, we see that MDSSh has positive scores in

FIG. 6. Reliability diagrams for the raw and postprocessed forecasts (raw: blue, QRF: red, and ZAGAfixtwo: green) for a lead time of +12 h, and three precipitation thresholds, pooled for all locations. Plots are made for (left) summer and (right) winter with the (top) 0.3-mm, (middle) 3-mm, and (bottom) 5-mm thresholds. Insets show the relative frequency as a function of the forecast probability bins.

about 50% of the lead times in summer. The somewhat better performance of MDSSh over ECC is clearest for the late afternoon and early evening in summer, when convective storms often occur. In these cases, the coarse structure of the ensemble, with $9 \times 9 \text{ km}^2$ cells (as nearest neighbor downscaling does not add detail) and parameterized deep convection, clearly gives way to a finely structured template of selected radar observations at $3 \times 3 \text{ km}^2$. An example is given by the 3-h precipitation forecasts at 2100 UTC 29 June 2016 (Fig. 9),

where spatial rank correlation was used as a metric to select the members from ECC and MDSSh that most resemble the observed precipitation field. In this case MDSSh shows a more realistic precipitation pattern. The choice of preferred restructuring method thus depends on the season, as MDSSh is somewhat more skillful during the convective season. If, however, the goal is to postprocess for winter, one might prefer ECC, which is more skillful than MDSSh, except for the first forecast day, and is almost always more skillful than SSH.

FIG. 7. Ranked importance of predictors in the QRF method. The largest rank means that a random permutation of its values increases the MSE in the forest more than permutations of the other predictors. The box-and-whisker plots are drawn for 1 season $\times$ 20 lead times $\times$ 3 cross validation folds = 60 fitted forests.

It thus appears that the problem with ECC (incorrect dependencies from the raw ensemble are introduced and are magnified when postprocessing enlarges the spread in the forecast distribution; Flowerdew 2014) is in this case less influential than that of SSh (a dissimilar template of randomly selected dates introduces correlations that are not representative for the atmospheric state). Because MDSSh accommodates this SSh problem, it can compete with ECC. This does not change qualitatively when an unweighted version of the variogram score is used.

We have also computed the CRPS for the restructured 28-member distributions after aggregating into provinces. ECC and MDSSh are consistently the most skillful reshuffling methods with lower CRPS values than SSh in all seasons and at all lead times (Fig. 10). The differences, however, are generally small. The diurnal cycle in forecast skill, already seen in the BSS (Fig. 4), is also evident in the CRPS, particularly in summer. Contrary to the VSS, however, we do not see a somewhat better summertime performance of MDSSh over ECC, probably as the aggregation to provinces eliminates the difference between fine- and coarse-scale restructuring for the convective regime.

5. Conclusions and discussion

a. Conclusions

To our knowledge, this study was the first to compare existing multivariate restructuring methods for

postprocessed ensemble forecasts at a high-resolution grid. Three-hourly accumulated precipitation at $3 \times 3 \text{ km}^2$ is a difficult predictand, especially because displacement errors are common and skillful forecasts are hard to make. In winter, the raw ensemble has negative skill for multiple lead times at the 7-mm threshold. In summer, precipitation is generally even less predictable due to deep convection. The clearest improvement by statistical postprocessing is for the low thresholds, where the overforecasting of low precipitation amounts was substantially corrected over the whole 60-h forecast range. Added skill by postprocessing is distinguishable up to 5 mm (Q.98) in winter and summer. Overall, QRF appeared to be a more skillful statistical postprocessing method than ZAGA. ZAGA deviated most at the extreme thresholds and had most difficulty with local calibration. Also, its implementation with a purely objective predictor selection was not necessarily the best.

The high-resolution predictand also meant that the multivariate restructuring took place in a high-dimensional setting. Here, the variogram skill scores showed that ECC is a better basic empirical method than SSh and that only the sophisticated search for similar templates (MDSSh) makes the latter a competitive method. A univariate scoring with CRPS, after aggregation of the restructured forecasts into spatial regions (i.e., the 12 provinces of the Netherlands), also indicated that MDSSh is better than SSh, while the CRPS values of ECC and MDSSh are similar.

FIG. 8. VSS of the spatial fields of SSh (red) and MDSSh (blue) relative to ECC for (top) summer and (bottom) winter.

b. Discussion

1) UNIVARIATE

Postprocessing clearly adds skill for moderate and high quantiles, where the ZAGA models perform as well as the nonparametric QRF. But for the extreme quantiles (>99th percentile) ZAGA is less skillful than QRF and can even deteriorate the raw ensemble predictions. In such cases the predictor values lead to predictive distributions that do not tend to climatology. Apparently the parametric skewed continuous nonzero distributions that were earlier regarded as a competitive option (e.g., Kleiber et al. 2011) are not so good for extremes, as was also concluded by Whan and Schmeits (2018).

Besides, the high resolution more easily permits displacement errors. Even though an ensemble of situations is used, NWP models might generally simulate a thunderstorm too early (regarded from one location) or spatially displaced (regarded from one time step). Perhaps the observations of one location should therefore not be related to the predictor values of that location

only (van der Plas et al. 2017). Options to include information from neighboring cells are a spatial weighting of all predictors or an extra predictor that expresses flow-dependent uncertainty due to displacement (Scheuerer and Hamill 2015a). The same study also showed that the desired behavior of tending to climatology for long lead times can be introduced by letting predictors modify the climatological values of shape parameters.

In the current implementation with a spatial pool, the models were fitted to the average systematic disparities. The BSS maps showed that local calibration is in that case not guaranteed. On the other hand, the alternative of fitting separate models to each location might suffer from insufficient training data, especially for extremes. Stauffer et al. (2017) implemented a solution for temperature forecasts by transforming predictor values into anomalies relative to their average in the training period. Combined with a local observed climatology, this can handle spatial differences in forecast bias, while still doing a single spatially pooled fit.

FIG. 9. (bottom right) Observed precipitation at 2100 UTC 29 Jun 2016, plotted together with the most similar members from the respective ensemble forecasts, at a lead time of 21 h. (bottom left) Before correction, (top left) after correction as restructured with ECC, and (top right) after correction as restructured with MDSSh. The similarity criterion was spatial rank correlation.

This option and the local fitting techniques addressed in, for instance, Kleiber et al. (2011) and Lerch and Baran (2017) can be explored in future research, in addition to the displacement predictor formulation mentioned above. This might improve the competitiveness of parametric postprocessing relative to QRF. On the other hand, local fitting is also not a guarantee for success. The decision trees of QRF in this study contained the possibility to spatially separate the postprocessing correction by splitting on latitude and longitude thresholds. That these predictors did not seem relevant (Fig. 7) indicates that the forecast bias does not exhibit a systematic spatial pattern in the Netherlands.

2) MULTIVARIATE

The explored empirical structuring is currently the only way to deal with high-dimensional distributions of precipitation that are non-Gaussian and intermittent in

nature (Gneiting 2014; Scheuerer et al. 2017). The conducted comparison between the two basic empirical methods favors ECC over SSh, both in terms of the VSS and the CRPS of aggregated forecasts. But in other studies the flaws of both methods might balance out differently.

The standard SSh may, for instance, not preserve the spatial or temporal gradients. The randomly selected dates are not conditional on the flow situation and are therefore not a collection of slightly different realizations under similar meteorological circumstances. Because the template fields are dissimilar, the ranks of a field tend to uniformity and reordering can result in unrealistic fields of the same quantile (Scheffzik 2016) (Fig. 3, middle row). On the other hand, ECC can introduce and magnify incorrect NWP dependencies when postprocessing increases the spread in the forecast distribution (Flowerdew 2014). Wilks (2015), for instance,

FIG. 10. Box plot of CRPS of the 28-member precipitation forecast distributions, after spatially averaging each member in each of the 12 provinces, for each forecast lead time. The 28 members, sampled from the forecast distributions, are reshuffled with ECC (red), SSh (green), and MDSSh (blue) and aggregated in each province in (a) summer and (b) winter. In the box plot, the horizontal line shows the median, and the box indicates the 25th–75th percentiles of the CRPS distribution of 12 values.

found that ECC leads to worse structures than SSh. Attempts are made to handle the flaw of NWP error autocorrelation in a method called dual-ECC (Ben Bouallègue et al. 2016), but the benefit of dependence on atmospheric state still greatly depends on the quality of the template. Remember, for instance, that the GLAMEPS values of $9 \times 9 \text{ km}^2$ cells are mapped to $3 \times 3 \text{ km}^2$ with the nearest neighbor method, and that the lack of high-resolution structure makes it less suitable for reordering. We saw this in the diminishing performance during summertime convection. Scheuerer et al. (2017) and Bellier et al. (2017) note similar ECC problems when the forecast resolution is too coarse for finescale structures.

Why then does ECC stand on equal footing with MDSSh? Why does the matching of high-resolution templates to the forecast marginal distributions not completely nullify incorrect dependence structures and their magnification? A first possibility is that precipitation

values of zero have too much weight in the matching of distributions and that the template therefore ends up with more zeros than the forecast. The method then distributes the tied ranks randomly and this results in fields with scatter (a random rank of the template, usually associated with zero, is suddenly associated with a nonzero value). This problem is also encountered in the unconditional template of standard SSh where the amount of zeros is often equal to the climatological probability of it being dry (Wu et al. 2018). A second possibility is that the statistical matching in MDSSh to the 4497 marginal distributions does not necessarily lead to the selection of meteorological analogs (Bellier et al. 2017). Correct analogs would be even more difficult to find when lead time is included, and the matching is extended for full space–time trajectories.

This study deliberately restructured only in space because high-dimensionality itself was our main goal (Table 1). For operational applications, however, it is

clear that the lost structure by postprocessing, except by the member-by-member methods (Van Schaeybroeck and Vannitsem 2015; Schefzik 2017), should be restored to a full spatiotemporal consistency (Gneiting 2014). ECC already automatically restructures across lead times, but the applied versions of Schaake shuffle in this study need the extension as in Scheuerer et al. (2017) and Bellier et al. (2017).

In future studies, the range of multivariate verification methods could also be extended. It is common that different scores rank the restructuring methods differently, so insight might be gained when, for instance, multivariate rank histograms are also used, whose “band depth” and “average rank” implementations have different properties (Thorarinsdottir et al. 2016; Wilks 2017). Beware, however, that they are sensitive to miscalibration in the univariate distributions, which is difficult to avoid in high-resolution settings.

The reason to continue with the approach of this study, that is, the postprocessing at high resolution and the subsequent joint restructuring, is that calibrated forecasts become more universally applicable. Compound quantities like the average in a province, or other aggregations in time and space, can be easily derived. A future study could investigate whether this approach is competitive compared to directly postprocessing to the aggregated resolution in time and space.

Acknowledgments. We thank Michael Scheuerer (NOAA) for providing software for the MDSSH method and for a useful discussion of preliminary results. John Bjørnar Bremnes and Thomas Nipen from Met Norway are thanked for many useful discussions about calibrating GLAMEPS during multiple working weeks. Kees Kok and two reviewers are thanked for their suggestions and comments that helped improve this paper.

REFERENCES

- Bellier, J., G. Bontron, and I. Zin, 2017: Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting. *Water Resour. Res.*, **53**, 10 085–10 107, <https://doi.org/10.1002/2017WR021245>.
- Ben Bouallègue, Z., T. Heppelmann, S. E. Theis, and P. Pinson, 2016: Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Mon. Wea. Rev.*, **144**, 4737–4750, <https://doi.org/10.1175/MWR-D-15-0403.1>.
- Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2008: Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Ann. Appl. Stat.*, **2**, 1170–1193, <https://doi.org/10.1214/08-AOAS203>.
- Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen, 1984: *Classification and Regression Trees*. CRC Press, 368 pp.
- Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.*, **5**, 243–262, [https://doi.org/10.1175/1525-7541\(2004\)005<0243:TSSAMF>2.0.CO;2](https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2).
- Dawid, A. P., and P. Sebastiani, 1999: Coherent dispersion criteria for optimal experimental design. *Ann. Stat.*, **27**, 65–81, <https://doi.org/10.1214/aos/1018031101>.
- Deckmyn, A., 2014: Introducing GLAMEPSv2. *ALADIN Forecasters Meeting*, Ankara, Turkey, ALADIN Consortium, 17 pp., http://www.umn-cnrn.fr/aladin/meshtml/FM2014/presentation/AladinFm_AD_be.pdf.
- Feldmann, K., M. Scheuerer, and T. L. Thorarinsdottir, 2015: Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.*, **143**, 955–971, <https://doi.org/10.1175/MWR-D-14-00210.1>.
- Ferro, C. A., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteor. Appl.*, **15**, 19–24, <https://doi.org/10.1002/met.45>.
- Flowerdew, J., 2014: Calibrating ensemble reliability whilst preserving spatial structure. *Tellus*, **66A**, 22662, <https://doi.org/10.3402/tellusa.v66.22662>.
- Genest, C., and A.-C. Favre, 2007: Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.*, **12**, 347–368, [https://doi.org/10.1061/\(ASCE\)1084-0699\(2007\)12:4\(347\)](https://doi.org/10.1061/(ASCE)1084-0699(2007)12:4(347)).
- Gerard, L., J.-M. Piriou, R. Brožková, J.-F. Geleyn, and D. Banciu, 2009: Cloud and precipitation parameterization in a meso-gamma-scale operational weather prediction model. *Mon. Wea. Rev.*, **137**, 3960–3977, <https://doi.org/10.1175/2009MWR2750.1>.
- Gneiting, T., 2014: Calibration of medium-range weather forecasts. ECMWF Tech. Memo. 719, 30 pp., <https://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf>.
- , and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- , F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>.
- Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden, 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205, <https://doi.org/10.1002/2014GL062472>.
- Hu, Y., M. J. Schmeits, S. J. van Andel, J. S. Verkade, M. Xu, D. P. Solomatine, and Z. Liang, 2016: A stratified sampling approach for improved sampling from a calibrated ensemble forecast distribution. *J. Hydrometeorol.*, **17**, 2405–2417, <https://doi.org/10.1175/JHM-D-15-0205.1>.
- Joslyn, S. L., and J. E. LeClerc, 2012: Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *J. Exp. Psychol. Appl.*, **18**, 126, <https://doi.org/10.1037/a0025185>.
- Kleiber, W., A. E. Raftery, and T. Gneiting, 2011: Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *J. Amer. Stat. Assoc.*, **106**, 1291–1303, <https://doi.org/10.1198/jasa.2011.ap10433>.
- Lerch, S., and S. Baran, 2017: Similarity-based semilocal estimation of post-processing models. *J. Roy. Stat. Soc.*, **66C**, 29–51, <https://doi.org/10.1111/rssc.12153>.

- Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, <https://doi.org/10.1016/j.jcp.2007.02.014>.
- Livezey, R., 2003: Categorical events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Joliffe and D. B. Stephenson, Eds., Wiley, 77–96.
- Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.*, **48**, RG3003, <https://doi.org/10.1029/2009RG000314>.
- Meinshausen, N., 2006: Quantile regression forests. *J. Mach. Learn. Res.*, **7**, 983–999.
- Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks, 2014: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, <https://doi.org/10.1175/MWR-D-13-00271.1>.
- Möller, A., A. Lenkoski, and T. L. Thorarinsdottir, 2013: Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quart. J. Roy. Meteor. Soc.*, **139**, 982–991, <https://doi.org/10.1002/qj.2009>.
- Overeem, A., I. Holleman, and A. Buishand, 2009: Derivation of a 10-year radar-based climatology of rainfall. *J. Appl. Meteor. Climatol.*, **48**, 1448–1463, <https://doi.org/10.1175/2009JAMC1954.1>.
- Panofsky, H. A., G. W. Brier, and W. H. Best, 1968: *Some Applications of Statistics to Meteorology*. The Pennsylvania State University Press, 224 pp.
- Ramos, M., S. Van Andel, and F. Pappenberger, 2013: Do probabilistic forecasts lead to better decisions? *Hydrol. Earth Syst. Sci.*, **17**, 2219, <https://doi.org/10.5194/hess-17-2219-2013>.
- Rigby, R., and D. Stasinopoulos, 2010: A flexible regression approach using GAMLSS in R. University of Lancaster, 282 pp.
- Schefzik, R., 2016: A similarity-based implementation of the Schaake shuffle. *Mon. Wea. Rev.*, **144**, 1909–1921, <https://doi.org/10.1175/MWR-D-15-0227.1>.
- , 2017: Ensemble calibration with preserved correlations: unifying and comparing ensemble copula coupling and member-by-member postprocessing. *Quart. J. Roy. Meteor. Soc.*, **143**, 999–1008, <https://doi.org/10.1002/qj.2984>.
- , T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, <https://doi.org/10.1214/13-STS443>.
- Scheuerer, M., and T. M. Hamill, 2015a: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.
- , and —, 2015b: Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Mon. Wea. Rev.*, **143**, 1321–1334, <https://doi.org/10.1175/MWR-D-14-00269.1>.
- , —, B. Whitin, M. He, and A. Henkel, 2017: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.*, **53**, 3029–3046, <https://doi.org/10.1002/2016WR020133>.
- Schmeits, M. J., and K. J. Kok, 2010: A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Mon. Wea. Rev.*, **138**, 4199–4211, <https://doi.org/10.1175/2010MWR3285.1>.
- Shrestha, D. L., D. E. Robertson, J. C. Bennett, and Q. Wang, 2015: Improving precipitation forecasts by generating ensembles through postprocessing. *Mon. Wea. Rev.*, **143**, 3642–3663, <https://doi.org/10.1175/MWR-D-14-00329.1>.
- Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, <https://doi.org/10.1175/MWR3441.1>.
- Stauffer, R., N. Umlauf, J. W. Messner, G. J. Mayr, and A. Zeileis, 2017: Ensemble postprocessing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. *Mon. Wea. Rev.*, **145**, 955–969, <https://doi.org/10.1175/MWR-D-16-0260.1>.
- Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, <https://doi.org/10.1175/MWR-D-15-0260.1>.
- , A. Fougères, P. Naveau, and O. Mestre, 2017: Forest-based methods and ensemble model output statistics for rainfall ensemble forecasting. ArXiv, <https://arxiv.org/abs/1711.10937>.
- Thorarinsdottir, T. L., and N. Schuhen, 2017: Verification: Assessment of calibration and accuracy. Norsk Regnesentral Note SAMBA/17/2017, 38 pp., <https://www.nr.no/en/nrpublication?query=/file/1498830259/VerificationChapter-Thorarinsdottir.pdf>.
- , T. Gneiting, and N. Gissibl, 2013: Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncertainty Quantif.*, **1**, 522–534.
- , M. Scheuerer, and C. Heinz, 2016: Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Stat.*, **25**, 105–122, <https://doi.org/10.1080/10618600.2014.977447>.
- Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Joliffe and D. B. Stephenson, Eds., Wiley, 137–163.
- van der Plas, E., M. Schmeits, N. Hooijman, and K. Kok, 2017: A comparative verification of high resolution precipitation forecasts using model output statistics. *Mon. Wea. Rev.*, **145**, 4037–4054, <https://doi.org/10.1175/MWR-D-16-0256.1>.
- Van Schaeybroeck, B., and S. Vannitsem, 2015: Ensemble postprocessing using member-by-member approaches: Theoretical aspects. *Quart. J. Roy. Meteor. Soc.*, **141**, 807–818, <https://doi.org/10.1002/qj.2397>.
- Verkade, J., and M. Werner, 2011: Estimating the benefits of single value and probability forecasting for flood warning. *Hydrol. Earth Syst. Sci.*, **15**, 3751–3765, <https://doi.org/10.5194/hess-15-3751-2011>.
- , J. Brown, P. Reggiani, and A. Weerts, 2013: Postprocessing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *J. Hydrol.*, **501**, 73–91, <https://doi.org/10.1016/j.jhydrol.2013.07.039>.
- Whan, K., and M. J. Schmeits, 2018: Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Mon. Wea. Rev.*, **146**, 3651–3673, <https://doi.org/10.1175/MWR-D-17-0290.1>.
- Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, <https://doi.org/10.1002/met.134>.
- , 2011: *Statistical Methods in the Atmospheric Sciences*, 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

- , 2015: Multivariate ensemble model output statistics using empirical copulas. *Quart. J. Roy. Meteor. Soc.*, **141**, 945–952, <https://doi.org/10.1002/qj.2414>.
- , 2017: On assessing calibration of multivariate ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **143**, 164–172, <https://doi.org/10.1002/qj.2906>.
- , and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using gfs reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, <https://doi.org/10.1175/MWR3402.1>.
- Wu, L., Y. Zhang, T. Adams, H. Lee, Y. Liu, and J. Schaake, 2018: Comparative evaluation of three Schaake shuffle schemes in postprocessing GEFS precipitation ensemble forecasts. *J. Hydrometeor.*, **19**, 575–598, <https://doi.org/10.1175/JHM-D-17-0054.1>.
- Yang, X., 2008: Status of the reference system. *HIRLAM Newsletter*, No. 54, HIRLAM Programme, De Bilt, Netherlands, 202–203, http://hirlam.org/index.php/component/docman/doc_view/152-hirlam-newsletter-no-54-paper-27-yang?Itemid=70.