Multivariate postprocessing techniques for probabilistic hydrological forecasting

S. Hemri1, D. Lisniak2, and B. Klein2

1Heidelberg Institute for Theoretical Studies, Heidelberg, Germany, 2German Federal Institute of Hydrology, Koblenz, Germany

Abstract Hydrologic ensemble forecasts driven by atmospheric ensemble prediction systems need statistical postprocessing in order to account for systematic errors in terms of both location and spread. Runoff is an inherently multivariate process with typical events lasting from hours in case of floods to weeks or even months in case of droughts. This calls for multivariate postprocessing techniques that yield well-calibrated forecasts in univariate terms and ensure a realistic temporal dependence structure at the same time. To this end, the univariate ensemble model output statistics (EMOS) postprocessing method is combined with two different copula approaches that ensure multivariate calibration throughout the entire forecast horizon. The domain of this study covers three subcatchments of the river Rhine that represent different sizes and hydrological regimes: the Upper Rhine up to the gauge Maxau, the river Moselle up to the gauge Trier, and the river Lahn up to the gauge Kalkofen. In this study, the two approaches to model the temporal dependence structure are ensemble copula coupling (ECC), which preserves the dependence structure of the raw ensemble, and a Gaussian copula approach (GCA), which estimates the temporal correlations from training observations. The results indicate that both methods are suitable for modeling the temporal dependencies of probabilistic hydrologic forecasts.

1. Introduction

Reliable hydrologic forecasts are crucial for a wide range of activities like, for instance, the operation of hydropower plants, shipping, flood prevention, and leisure activities. Information about the predictive uncertainty of the predictand, i.e., runoff or water level, is required for rational decision making. The predictive uncertainty is defined as the uncertainty of a future realization of a predictand, the quantity of interest, conditional on all available information and knowledge [Krzysztofowicz, 1999; Todini, 2008]. The available knowledge about the future realization in hydrologic forecasting is generally embedded in one or more hydrological model forecasts. One of the main sources of uncertainty is the meteorological uncertainty of the short to medium-range development of weather patterns. In order to take account of this uncertainty in hydrologic forecasting, atmospheric ensemble forecasts are used as forcing to the hydrologic forecasting systems. Consequently, hydrologic forecast ensembles are generated. Atmospheric ensemble forecasts for surface variables, such as surface temperature and precipitation, usually lack reliability. In many cases, they are underdispersive and biased [Bougeault et al., 2010; Park et al., 2008]. Hence, predictive uncertainty is not represented adequately by the meteorological input ensemble. Additionally, hydrological uncertainties are typically neglected in rainfall-runoff modeling. Thus, statistical postprocessing is needed in order to estimate predictive uncertainty and, in particular, to obtain reliable forecasts. In other words, the main goal of postprocessing is to achieve well-calibrated and yet sharp probabilistic predictions [Raftery et al., 2005; Gneiting et al., 2007]. In case of well-calibrated forecasts, the theoretical levels of prediction intervals are equal to the relative frequency of the observations to lie within the corresponding forecast intervals. Sharpness relates only to the forecasts and denotes how “narrow” prediction intervals are at a given nominal level.

In order to put this paper in a broader context, a selection of studies that focus on statistical postprocessing of hydrologic ensemble forecasts is presented first. Earlier studies on statistical postprocessing proposed Bayesian models to quantify the uncertainties of hydrological forecasts. Krzysztofowicz [1999, 2002] introduced the Bayesian forecasting system (BFS) to produce probabilistic forecasts from deterministic hydrological forecasts. The hydrological uncertainty processor (HUP) that aggregates the hydrological model uncertainties is a component of the BFS [Krzysztofowicz and Kelly, 2000]. Reggiani et al. [2009] extended the HUP for postprocessing of ensemble forecasts for the river Rhine on the Dutch-German border. Madadgar
et al. [2012] postprocessed ensemble forecasts by applying copula techniques that fit a bivariate distribution to forecasts and observations. Bayesian model averaging (BMA) [Raftery et al., 2005] has been used for the probabilistic combination of (ensemble) runoff forecasts in many cases. For instance, Ajami et al. [2007] or Duan et al. [2007] showed that the combination of hydrologic forecasts using BMA led to both, quantitative statements on prediction uncertainty and improvements in terms of deterministic verification measures. Fraley et al. [2010] introduced an adapted BMA version that is able to take account of ensemble forecasts with exchangeable members as typically encountered with meteorological ensemble forecasts. Recent developments allow to use flexible predictive distributions [Parish et al., 2012; Rings et al., 2012] and to postprocess forecasts over an entire range of lead times simultaneously [Hemri et al., 2013; Engeland and Steinsland, 2014]. Other alternatives for statistical postprocessing are the model conditional processor [Todini, 2008; Coccia and Todini, 2011], which has recently been extended to handle ensembles [Todini et al., 2015], and quantile regression [Weerts et al., 2011]. A nonparametric approach for the postprocessing of hydrological ensemble forecasts which is similar to indicator cokriging was proposed by Brown and Seo [2010]. This list gives an overview over the different postprocessing methods used in hydrology, but is by no means exhaustive.

The first but minor goal of this study is to achieve well calibrated and yet sharp marginal predictive densities. The term marginal refers to the univariate predictive distribution for a particular lead time. To this end, we adapt the ensemble model output statistics (EMOS) [Gneiting et al., 2005] postprocessing method, which is frequently used for meteorological variables, so that it becomes suitable for probabilistic river discharge forecasts. More specifically, the truncated EMOS method [Thorarinsdottir and Gneiting, 2010] is modified [see also Hemri et al., 2014] and then applied to ensemble runoff forecasts for three subcatchments of river Rhine.

According to Pinson and Girard [2012] knowing not only the marginal predictive distributions for each individual lead time, but also the dependence structure among different lead times, is crucial to making optimal decisions based on probabilistic forecasts. This applies in particular to runoff which is highly autocorrelated. After univariate postprocessing using EMOS, the information about the temporal (spatiotemporal in the case of several gauges in a river basin) dependence structure of the raw ensemble is lost. If one is interested in forecast runoff trajectories, then a sound representation of the dependence structure has to be added. Forecast runoff trajectories are, for instance, required for the optimization of reservoir operation or, as in the case presented here, used as input to a hydrodynamic model to forecast water levels. Hence, the second and main goal of this study is to obtain forecasts that are not only marginally well calibrated, but from which it is possible to obtain also runoff scenarios over the entire forecast horizon. If the observed trajectory and the scenarios are likely to follow the same multivariate distribution, the forecast model is said to exhibit good multivariate calibration. This study compares two different types of multivariate postprocessing methods: ensemble copula coupling (ECC) [Schefzik et al., 2013] and the Gaussian copula approach (GCA) [Pinson and Girard, 2012]. The nonparametric ECC approach is similar to the Schaake Shuffle [Clark et al., 2004] in that postprocessed forecast trajectories are reordered using exogenous information. In case of the Schaake Shuffle, this information stems from past observations, in case of ECC from the raw ensemble. As ECC accounts for both temporal and spatial dependencies, it is suitable for parallel postprocessing of forecasts for different subcatchments. This is required, for instance, if the postprocessed forecast trajectories are used as inputs to a hydrodynamic model to calculate water level forecasts further downstream. GCA is a parametric approach that estimates the correlation structure from training observations. GCA is expected to outperform ECC in cases, where a large number of forecast scenarios is required, or where it is doubtful whether the raw ensemble captures the correct correlation structure. The GCA variant described in this paper accounts only for temporal dependencies though it may easily be extended such that it is able to model spatiotemporal dependencies.

In this study, EMOS, ECC, and GCA are verified based on runoff forecasts from the operational forecasting system of the German Federal Institute of Hydrology (BfG) for river Rhine [Meißner and Rademacher, 2010]. Three different subcatchments of river Rhine with different characteristics are considered: river Upper Rhine up to gauge Maxau, river Moselle up to gauge Trier, and river Lahn up to gauge Kalkofen.

In section 2, the study areas and the observed runoff data are presented. The different types of forecasts used in this paper as well as the methods used for model fitting and verification are introduced in section 3. The results in section 4 are followed by a discussion in section 5. Note that all analyses have been performed using the statistical software R [R Development Core Team, 2014].
2. Study Areas and Runoff Data

The catchments in this study are selected such that different runoff regimes and catchment sizes are covered. Figure 1 shows the locations of the considered subcatchments within the Rhine river basin. The runoff of the Upper Rhine at the gauge Maxau (referenced as Upper Rhine catchment hereinafter) is dominated by the alpine part of the catchment. This explains its pronounced, single peak mountain snow (glacial-nival) regime with maximum in summer and minima in late autumn and winter. The catchments of the rivers Moselle and Lahn have a rainfall dominated runoff (pluvial) regime with maximum in winter and minimum in late summer. Catchment area as well as mean and maximum runoff are listed in the upper part of Table 1. Catchment area decreases in the following order: Upper Rhine > Moselle > Lahn. Water level measurements from 1 November 1998 to 10 January 2013, which are converted into runoff by means of rating curves, serve as observations. The forecast data are discussed in section 3.

3. Methods

The three probabilistic forecasts compared in this study—raw ensemble, climatological, and EMOS forecasts—are presented in the following. The raw ensemble and the climatological forecasts serve as benchmark models for verification of the EMOS forecasts.

3.1. Raw Ensemble Forecasts

At the BfG, the conceptual, semidistributed rainfall-runoff model HBV-96 [Bergström, 1995; Lindström et al., 1997] is used for operational runoff forecasting. The Rhine river basin is divided into 134 subbasins which

---

**Table 1.** Features of the Considered Catchments (Top)\(^a\) and the Meteorological Input Models (Bottom)\(^b\)

<table>
<thead>
<tr>
<th>Gauge</th>
<th>Catchment</th>
<th>Area (km(^2))</th>
<th>MQ (m(^3)/s)</th>
<th>HQ (m(^3)/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxau (MAXA)</td>
<td>Upper Rhine</td>
<td>50,196</td>
<td>1,247</td>
<td>4,293</td>
</tr>
<tr>
<td>Trier (TRIE)</td>
<td>Moselle</td>
<td>23,857</td>
<td>322</td>
<td>2,880</td>
</tr>
<tr>
<td>Kalkofen (KALK)</td>
<td>Lahn</td>
<td>5,304</td>
<td>48</td>
<td>598</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Number of Models</th>
<th>Lead Times</th>
<th>Spatial Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSMO-LEPS</td>
<td>16</td>
<td>1–132 h</td>
<td>10 km</td>
</tr>
<tr>
<td>DWD-GME</td>
<td>1 (deterministic)</td>
<td>1–174 h</td>
<td>20 km</td>
</tr>
<tr>
<td>DWD-MER</td>
<td>1 (deterministic)</td>
<td>1–78 h (174 h)</td>
<td>1–7 km (20 km)</td>
</tr>
<tr>
<td>ECMWF-IFS</td>
<td>1 (deterministic)</td>
<td>1–240 h</td>
<td>16 km</td>
</tr>
</tbody>
</table>

\(^a\)MQ (mean discharge) and HQ (maximum discharge) are calculated over the period from 1 November 1998 to 31 October 2011.

\(^b\)DWD-MER stands for a model run based on COSMO-EU forcing up to lead time 78 h and based on DWD-GME thereafter (corresponding forecast horizon and resolution are reported within parentheses). Note also that the forecast horizon of the hydrologic forecasts based on COSMO-LEPS is only 114 h, though the meteorological model forecasts up to 132 h.
are further subdivided into hydrological response units (HRUs) according to land use and elevation classes. The hydrologic processes for runoff formation are calculated on those HRUs [Meißner and Rademacher, 2010]. The model calculates runoff with a temporal resolution of 1 h using temperature and precipitation fields that have been interpolated over the subbasins as meteorological input. Runoff forecasts are obtained by running the hydrological model with meteorological inputs from several different numerical weather prediction systems (NWPs). Forecasts of NWP models can be either deterministic or probabilistic. In the first case, uncertainty is neglected and a single forecast trajectory is provided. In contrast, ensemble forecasts try to represent uncertainty by several model runs with different initial conditions, boundary conditions and physical parameter values. The term ensemble forecast comprises the collection of these distinct model runs. In current operational use, the ensemble forecasts from the hydrological model HBV-96 are used as boundary conditions and lateral inflows to a hydrodynamic model to calculate water level forecasts for stations along the river Rhine.

All hydrological forecast data used here are generated by hindcasting the hydrological model with archived operational meteorological forecasts. As summarized in the lower part of Table 1, the meteorological forcing models vary in both the number of ensemble members and the forecast time horizon. The hydrologic model is run with these models as forcing leading to a hydrologic 19 member multimodel ensemble, hereafter referred to as the raw ensemble. It is composed of the 16 COSMO-LEPS members [Montani et al., 2011], and the three deterministic models DWD-GME [Majewski et al., 2002, 2012], DWDMER, and ECMWF-IFS, i.e., the deterministic high-resolution run of the ECMWF ensemble [Molteni et al., 1996]. The hydrologic model provides hourly forecasts up to 240 h (ECMWF-IFS), 174 h (DWD-GME and DWDMER), and 114 h (COSMO-LEPS), respectively. DWDMER uses meteorological forcing from the COSMO-EU model [Steppeler et al., 2002; Schulz and Schättler, 2011] up to lead time 78 h, and data from the DWD-GME model from lead time 79 h. Hence, two members of the raw ensemble rely on the same meteorological inputs from lead time 79 h onward. The hydrologic forecasts are initialized on a daily basis from 1 November 2008 to 25 January 2011 at 6 h GMT. The initial conditions of the hydrologic model are generated by a continuous simulation up to the forecast issue date using observed meteorological input. Finally, the raw runoff ensemble forecasts are statistically corrected based on the observations available up to the forecast date using an autoregressive model [Boersen and Weerts, 2005]. For the following analyses, only lead times up to 114 h are considered because of dropping out ensemble members. At lead time 114 h, the forecast horizon of the first model, COSMO-LEPS, is reached. Hence, it drops out of the raw ensemble. After dropping out of DWD-GME and DWDMER at lead time 174 h, only ECMWF-IFS is remaining up to lead time 240. This problem of dropping out ensemble members would require a more detailed analysis [see also Hemri et al., 2013].

3.2. Climatological Forecasts
The climatological forecasts used in this study account for seasonal runoff variation. Seasonality is included by calculating daily climatological forecasts. Based on an hourly time series of observed runoff from 1 November 1998 to 10 January 2013, the climatological forecasts are obtained by calculating the empirical distribution of the observations that lie within $\pm x$ days of the calendar date of the verification day, but not in the same year. The dependence of runoff on the time of day is neglected. The drawback of mixing different times of day is more than compensated for by the increase in the sample size from which the climatology is constructed. Having explored intervals with $x \in \{15, 30, 45\}$ days, we select $x = 45$ for river Upper Rhine, $x = 30$ for river Moselle, and $x = 15$ for river Lahn as these values lead to the best climatological forecasts in terms of the continuous ranked probability score (CRPS, see equation (4)). The smaller the catchment, the narrower is the time frame to be used for calculating the climatological forecasts. Note that the climatological forecasts are probabilistic which results directly from constructing those using empirical distributions of historical observations.

3.3. EMOS Postprocessing
3.3.1. Preliminaries
Before presenting the truncated EMOS approach, the Box-Cox transformation for data normalization and the selection of training periods are discussed. Runoff data are undoubtedly non-Gaussian. In order to be able to resort to postprocessing methods relying on Gaussian distributions, both the observations and the raw ensemble predictions have to be transformed such that they are approximately normal. Like Duan et al.
we use the Box-Cox transformation [Box and Cox, 1964] for that purpose. Details on how the Box-Cox transformation has been implemented for this study can be found in Appendix A1. The effect of the Box-Cox transformation becomes clear from comparing the EMOS predictive densities of the same forecast on the transformed and on the original space as shown in Figures 2a and 2b.

Now, the selection of training and verification periods is outlined. The verification set comprises forecast/observation pairs from 1 November 2008 to 31 October 2011. As the forecasts cover lead times from 1 to 114 h, the forecast initialization dates range from 1 November 2008 to 25 October 2011. That is, the verification set consists of 1085 initialization days and 114 lead times. As the behavior of the hydrological system varies over the year, e.g., snow accumulation in winter, snow melt in spring, low flow situations in summer, the parameters of the EMOS model have to be estimated for each meteorological season separately. That is, for the verification of a forecast issued on a particular date, the forecast/observation pairs issued on days that are in the same season but not in the same year are used as training data. For instance, if the forecast to be verified is issued in March 2009, the training period comprises the forecasts issued in spring 2010 and

Figure 2. Example forecasts for river Moselle at Trier for a high flow event issued on 5 January 2011 at 6 h GMT. Univariate raw ensemble and EMOS probability density forecasts with a lead time of 48 h are shown in subfigures (a) on the Box-Cox transformed space and (b) on the original space. The horizontal line of distinct dots represents the raw ensemble members, and the vertical purple line shows the observed value. (c–f) The multivariate forecasts covering lead times 1–114 h. (c) The trajectories of the raw ensemble, (d) the quantiles of the EMOS forecast, and (e and f) the trajectories of the EMOS forecasts with correlation structure by ECC-T or GCA-exp, respectively.

[2007], we use the Box-Cox transformation [Box and Cox, 1964] for that purpose. Details on how the Box-Cox transformation has been implemented for this study can be found in Appendix A1. The effect of the Box-Cox transformation becomes clear from comparing the EMOS predictive densities of the same forecast on the transformed and on the original space as shown in Figures 2a and 2b.
are each a group of one member. The index model, that is COSMO-LEPS constitutes a group of 16 members, while DWD-GME, DWD-MER, and ECMWF-IFS are transformed using the Box-Cox transformation.

Non-normally distributed forecasts are right truncated at two times the maximum of the observations from 1 November 1998 to 31 October 2008. This truncated EMOS variant is closely related to the one proposed by Thorarinsdottir and Gneiting [2010] with the difference that the left-truncation at zero is replaced by a right truncation. Note that truncated EMOS also accounts for heteroscedasticity, i.e., heterogeneity in variances. Figures 2a and 2b illustrate the conversion of raw ensemble members into a predictive density function. The parameters of the EMOS model are estimated by minimization of the CRPS over the training period. The CRPS is a widely used verification measure that considers both calibration and sharpness. A more detailed summary of the truncated EMOS method, which follows mainly Hemri et al. [2014], is given in the following.

Before summarizing the details of the truncated EMOS method, the concept of exchangeable ensemble members is introduced here. Meteorological ensemble systems, such as COSMOS-LEPS, give an estimate of the forecast uncertainty by providing a finite sample of forecast scenarios. Each scenario is represented by an ensemble member. For instance, all COSMO-LEPS ensemble members are equally likely and exchangeable in that they lack individually distinguishable physical features. Statistical postprocessing methods have to take account of exchangeable ensemble members [Fraley et al., 2010; Gneiting et al., 2005]. Therefore, EMOS model parameters are constrained to be equal within each exchangeable group. In this study, COSMO-LEPS provides the only exchangeable group. This can also be seen from Figures 2a and 2b. In the following, the index \( i = 1, \ldots, I \) differentiates between groups of the raw ensemble according to the meteorological input model, that is COSMO-LEPS constitutes a group of 16 members, while DWD-GME, DWD-MER, and ECMWF-IFS are each a group of one member. The index \( j = 1, \ldots, m_i \) differentiates between the individual members of group \( i \) with size \( m_i \). For COSMO-LEPS, \( m_i = 16 \), whereas \( m_i = 1 \) for the other models. With this, \( r_{ij} \) denotes runoff predicted by the \( j \)th member of the \( i \)th model. And \( r_i \) is then the mean runoff of group \( i \). In case of deterministic models, \( m_i = 1 \) and hence \( r_{ij} = r_{ij} \). As stated above, the runoff forecasts and observations are transformed using the Box-Cox transformation \( f_{ij} = h(r_{ij}) \) with \( f_i = m_i^{-1} \sum_{j=1}^{m_i} f_{ij} \). The upper limit \( b \) of the predictive truncated normal distribution is set to two times the Box-Cox transform of the maximum observed value. For the three catchments considered, there is no need for a lower limit. In case of the river Moselle, the estimated Box-Cox parameter \( \lambda \) is negative, which means that \( -\infty \) on the Box-Cox transformed space maps to zero on the original space. \( \lambda \) is positive for the rivers Upper Rhine and Lahn, but the predictive probabilities for negative runoff are negligible. They are numerically zero for the vast majority of verification days and lead times. The highest probabilities attained are 3.7E-4 and 7.7E-87 for the rivers Upper Rhine and Lahn, respectively. With \( N^b(\mu, \sigma^2) \) denoting a right truncated normal distribution with support \( [-\infty, b] \), the truncated EMOS predictive density of the variable \( X \), here Box-Cox transformed runoff, can be written as:

\[
p(x|f_{i1}, \ldots, f_{im}, f_{2i}, \ldots, f_{im}) = N^b(\mu, \sigma^2), \tag{1}
\]

where \( \mu = a_0 + a_1 f_{i1} + a_2 f_{i2} + \cdots + a_{m_i} f_{im} \) and \( \sigma^2 = c_1 + c_2 S^2 \) depends on the ensemble variance

\[
S^2 = \frac{1}{\sum_{i=1}^{m} m_i} \sum_{j=1}^{m} \left( f_{ij} - \bar{f}_i \right)^2, \tag{2}
\]

with ensemble mean

\[
\bar{f}_i = \frac{1}{m_i} \sum_{j=1}^{m} f_{ij}.
\]

**Table 2. Examples of Pairs of Verification and Training Periods**

<table>
<thead>
<tr>
<th>Verification Period</th>
<th>Training Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAM 2009</td>
<td>MAM 2010, MAM 2011</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

*SON denotes September, October, November; SO September, October; DJF December, January, February; MAM March, April, May.*
Assign unique ordered indices 1, . . . , M to the Box-Cox transformed raw ensemble members \( f^i_m \), so that the ensemble can be rewritten as \( (f^i_1, . . . , f^i_M) \). Even though the actual order of the indices does not matter, the index assigned to a particular raw ensemble member has to remain constant over all lead times.

2. Obtain for each lead time the reordered EMOS forecasts:

\[
\hat{y}^i_m = \hat{F}^{-1}_{m,j}(\hat{S}_j(f^i_m)), \quad m=1, . . . , M, \quad i=1, . . . , L,
\]

where \( \hat{S}_j \) is the fitted CDF of a suitable parametric distribution to the Box-Cox transformed raw ensemble and \( \hat{F}^{-1}_{m,j} \) denotes the inverse of the marginal EMOS CDF. Here a right truncated normal distribution with mean \( \mu_j \), variance \( \sigma_j^2 \), and upper threshold \( b \) (cf. equation (1)) is fitted using maximum-likelihood
estimation. In order to avoid unrealistically extreme quantiles in cases of very low raw ensemble spread, the variance of $S_l$ is set to

$$\max \left\{ \sigma_l^2, \left[ h((1+d)\tau^l) - h((1-d)\tau^l) \right]^2 \right\},$$

where $\tau^l$ is the mean of the raw ensemble at lead time $l$, $h$ denotes the Box-Cox transformation, and $d$ is a tuning parameter. This heuristic approach ensures that the minimal variance is linked to the mean of the raw ensemble and applicable on the Box-Cox transformed space. After having compared different example forecast trajectories and verification scores, we have set $d = 0.0005$ for this study.

3.4.2. Gaussian Copula Approach
As already mentioned, GCA is a parametric approach for modeling the correlation structure among different lead times. One first estimates a parametric correlation function from training data and then the respective multivariate normal distribution of dimension equal to the number of lead times. By sampling several times from this distribution and then evaluating the CDF of the univariate standard normal distribution at the sampled values, trajectories of probabilities are obtained. Extracting the corresponding quantiles from the EMOS distributions results in the GCA EMOS trajectories. As shown in Figures 2c and 2f, the rank-order structure of the GCA EMOS trajectories is independent from the rank-order structure of the raw ensemble. Technically, GCA can be summarized as follows:

1. Calculate the empirical correlogram among lead times 1 to $L$ from the observations in the training period.
2. This correlogram is then interpreted as an empirical variogram of a process with variance 1. Hence, fit a theoretical variogram to the empirical correlogram. The exponential, the Matérn, and the generalized Cauchy correlation models (see Schlather [1999] for a comprehensive review on correlation functions) are candidates for the data at hand. A figure of the empirical correlograms and the corresponding fitted correlation functions is available as supplemental material to this paper. The correlation parameters are estimated using the R package geoR [Diggle and Ribeiro, 2007; Ribeiro and Diggle, 2001].
3. Sample $K$ realizations, $(x_{1l}^1, x_{2l}^1, \ldots, x_{Kl}^1)$ with $k = 1, \ldots, K$, from a standard $L$-variate Gaussian distribution $\mathcal{N}(\mu = 0, \Sigma)$ with diagonal elements $\Sigma_{l(l)} = 1$ and correlation structure from 2.
4. Using the inverse, $\tilde{F}_l^{-1}$, of the marginal EMOS CDF for each individual lead time, multivariate scenarios with marginal distributions inherited from the univariate fits are obtained:

$$\tilde{y}_l^k = \tilde{F}_l^{-1}(\Phi(x_{kl}^k)).$$

In principle, GCA allows to sample infinitely many forecast runoff trajectories. For this paper, the number of GCA trajectories is set to be equal to the size of the raw ensemble, i.e., $K = M$. Preliminary tests have shown that GCA does not depend much on the parameterization of the covariance function. For the rest of this paper, we consider only exponential GCA, which relies on an exponential covariance function, and hence is the simplest GCA model. Henceforth, the term GCA refers to exponential GCA.

3.5. Example Forecast
In order to illustrate EMOS, ECC, and GCA the hydrographs of an example prediction are discussed now. To this end, the forecasts issued on 5 January 2011 for river Moselle have been selected, which cover a high flow event at a forecast lead time of about three days. Though the raw ensemble is able to predict the magnitude of the event, all members underestimate runoff during the rising limb of the hydrograph as shown in Figure 2c. The EMOS probability forecasts shown in Figure 2d clearly improve the prediction compared to the raw ensemble. ECC-T yields quite realistic forecast trajectories with the same rank-order structure as the raw ensemble. As demonstrated by the runoff trajectories in Figures 2e and 2f, GCA is more flexible than ECC-T. On the one hand, the 19 randomly selected quantiles cover the observed trajectory better than the raw ensemble or ECC-T. On the other hand, the forecast trajectories are a bit too wiggly. Additionally, there is a remarkably high outlier trajectory. Similar plots for additional issue dates for all three considered catchments in low and high flow conditions are available as supporting information to this paper.

3.6. Verification Techniques
For the assessment of univariate skill, we use the continuous ranked probability skill score (CRPSS, see equation (A2) in Appendix A2), which is the skill score associated to the CRPS. CRPSS is positively oriented. The
best attainable value is 1, while a negative score indicates superiority of the reference forecast compared to the forecast at hand. Univariate calibration is assessed via the probability integral transform (PIT) [Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007]. The PIT value \( z \) for an individual verification day and lead time is defined as the value of the predictive cumulative density function (CDF) evaluated at the observation. According to Rosenblatt [1952] well-calibrated forecasts imply that \( z \sim U(0, 1) \). In the present context, this means that the observations should look like random samples from the predictive distribution. When translating into bins and calculating the relative frequencies over the entire verification period, the PIT can be visualized by a histogram. A flat histogram indicates well-calibrated forecasts, whereas underdispersion is indicated by a \( U \)-shape and overdispersion by an \( \cap \)-shape, respectively. Sharpness can be assessed by verifying the widths of prediction intervals at a given nominal level. Here we focus on 90% prediction intervals.

Multivariate calibration, i.e., the correct representation of the dependence structure among different lead times by the forecasts, can be assessed visually and by numerical scores. For visual verification in high-dimensional settings, Thorarinsdottir et al. (2014) proposed the average rank and the band depth rank histogram. These rank histogram methods are further developments of the concept of the multivariate rank histogram [Gneiting et al., 2008]. As both the average and the band depth rank histogram lead to similar results, but the latter is more difficult to interpret, we limit ourselves to the former. A detailed description of the average rank histogram is given in Appendix A2.

A variety of scores are used in order to obtain numerical verification measures of the multivariate forecasts. They all focus on different aspects. The energy score (ES) introduced by Gneiting and Raftery (2007) can be used as an overall measure of quality for multidimensional probabilistic forecasts. The ES discriminates well between forecasts with different mean vectors, and shows satisfying discrimination ability with regard to variance specification. But its ability to detect errors in the correlation structure is quite poor [Pinson and Girard, 2012; Pinson and Tastu, 2013; Scheuerer and Hamill, 2015]. For this study, the ES is calculated from the forecast vector of the lead times 24, 48, 72, and 96 h in order to avoid issues of high dimensionality. Scheuerer and Hamill (2015) developed the \( p \)-variogram score as a complement to the ES. Its main advantage is the much better discrimination ability between correct and misspecified correlation structures. In a nutshell, the \( p \)-variogram score relates to the dissimilarity between variograms of order \( p \) of observations and forecasts. For this study, the variograms are temporal, i.e., they are measuring the amount of correlation between different lead times. Additional insights into the properties of the multivariate forecasts are gained by assessing the CRPS values of one-dimensional functionals, namely the sum, the minimum, and the maximum runoff value over the range of lead times. For instance, the CRPS of the sum functional assesses, how well the multivariate forecasts predict the total observed runoff over the entire range of lead times. Formulas of the ES and the \( p \)-variogram score are given in Appendix A2.

4. Results

In the following subsections, truncated EMOS, ECC, and GCA are verified using statistical verification methods. The considered runoff forecasts and observations are standardized by catchment size, i.e., the corresponding unit is \([\text{m}^3\text{km}^{-2}\text{s}^{-1}]\).

4.1. Univariate Verification

Based on the verification methods presented above, the forecasts are now verified over the entire verification period and lead times 1–114 h. As a sound assessment of multivariate forecast properties relies on marginally well-calibrated forecasts, we start with univariate verification of the predictive distributions for each individual lead time. Figure 3a shows the CRPSS values for the raw ensemble and EMOS forecasts with the daily climatological forecasts as reference. Skill in terms of CRPSS is much improved by postprocessing in case of all three catchments. The gain in skill by the raw ensemble forecasts over the climatological forecasts decreases with decreasing catchment size and increasing lead time. However, the EMOS forecasts for the river Lahn exhibit equal performance in terms of CRPSS as the EMOS forecasts for the substantially larger catchment of river Moselle.

After having discussed general prediction skill in terms of CRPSS, let us now have a closer look at calibration and sharpness. For all three catchments, the raw ensemble forecasts are highly underdispersed as depicted...
by the 3-D PIT histograms in Figure 3b. With increasing lead time underdispersion slightly decreases. Note also the time-of-day-dependent oscillation of raw ensemble calibration of the forecasts for river Rhine at Maxau. This oscillation most likely arises from the intraday operation of the Swiss lakes, which are not included in the hydrological model. EMOS postprocessing flattens the PIT histograms regardless of catchment size and lead time. Generally, EMOS leads to well-calibrated forecasts as shown in Figure 3c. However, the [0.9,1] quantile interval is still overrepresented. Differences in calibration of the postprocessed forecasts between the different catchments can hardly be detected.

Calibration is only meaningful together with sharpness. For the assessment of forecast sharpness, the empirical distribution of the widths of the 90% prediction intervals is constructed from the entire verification period. Following Gneiting et al. (2007), important quantiles of that distribution are plotted in a Box-Plot like manner. Figure 4 reveals that sharpness is clearly deteriorated by EMOS in case of river Upper Rhine.

Figure 3. (a) CRPSS against lead time of the raw ensemble and the EMOS forecasts. The daily climatological forecasts serve as reference model. (b) Three-dimensional PIT histograms of the raw ensemble forecasts for the rivers Upper Rhine, Moselle, and Lahn. (c) Three-dimensional PIT histograms of the truncated EMOS forecasts for the corresponding catchments.
For the rivers Moselle and Lahn, EMOS deteriorates sharpness at the short lead time of 24 h, whereas the effect of EMOS on sharpness for higher lead times is less pronounced. However, EMOS turns the very poor coverage of the raw ensemble forecasts into almost perfect coverage.

4.2. Multivariate Verification

Even though the GCA forecasts look a bit less realistic, they perform slightly better than ECC-T in terms of multivariate statistical verification. But note that the differences in verification results between EMOS with either GCA or ECC-T are minor, compared to the differences to the raw ensemble. The average rank histograms shown in Figure 5 indicate that ECC-T lacks in multivariate calibration. The \( L \)-shaped histograms for all three catchments indicate either a too low correlation among lead times or forecast trajectories that are marginally underdispersive, in that the predictive densities for the individual lead times are too narrow. The correlations of GCA are too strong as can be seen from the rather \( \cap \)-shaped histograms. In order to highlight the effects of a misspecified correlation structure, a forecast ensemble consisting of 19 runoff trajectories drawn from the marginal EMOS distributions for the individual lead times with zero correlation between lead times (INDEP) is evaluated as well in the following. According to Table 3, INDEP performs quite well in terms of ES, but very poor in terms of the CRPS of the sum, minimum and maximum functionals, and the \( p \)-variogram score. ECC and GCA perform better than INDEP in any combination of verification score and catchment, but the differences are very low in case of the ES. Furthermore, the \( p \)-variogram score indicates that ECC-T outperforms GCA in terms of correlation structure in case of river Upper Rhine, while GCA outperforms ECC-T for the rivers Moselle and Lahn. The CRPS values of the minimum functional favor GCA over ECC-T for all catchments. The sum functional tends to favor GCA as well. However, ECC-T outperforms GCA in terms of the CRPS of the maximum functional in case of the large catchments Upper Rhine and Moselle.

5. Discussion and Conclusions

The results confirm that univariate postprocessing using EMOS improves skill of the probabilistic runoff forecasts over the entire range of lead times. In particular, univariate calibration is greatly improved. However, the main focus of this study was on multivariate calibration. Our results demonstrate that temporal
Figure 5. Average rank histograms comparing raw ensemble, ECC-T, and GCA with exponential covariance structure forecasts for the rivers Upper Rhine, Moselle, and Lahn.

<table>
<thead>
<tr>
<th>Catchment</th>
<th>Method</th>
<th>ES</th>
<th>CRPS_SUM</th>
<th>CRPS_MIN</th>
<th>CRPS_MAX</th>
<th>p-vario_0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper Rhine</td>
<td>Raw ensemble</td>
<td>3.75</td>
<td>156</td>
<td>1.02</td>
<td>2.01</td>
<td>36.0</td>
</tr>
<tr>
<td></td>
<td>INDEP</td>
<td>3.23</td>
<td>155</td>
<td>3.08</td>
<td>3.70</td>
<td>144</td>
</tr>
<tr>
<td></td>
<td>ECC-T</td>
<td>3.18</td>
<td>126</td>
<td>0.94</td>
<td>1.49</td>
<td>28.9</td>
</tr>
<tr>
<td></td>
<td>GCA-exp</td>
<td><strong>3.13</strong></td>
<td><strong>124</strong></td>
<td><strong>0.86</strong></td>
<td><strong>1.52</strong></td>
<td><strong>29.1</strong></td>
</tr>
<tr>
<td>Moselle</td>
<td>Raw ensemble</td>
<td>3.60</td>
<td>147</td>
<td>0.94</td>
<td>2.24</td>
<td>44.1</td>
</tr>
<tr>
<td></td>
<td>INDEP</td>
<td>3.11</td>
<td>144</td>
<td>1.48</td>
<td>4.44</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>ECC-T</td>
<td>3.07</td>
<td>119</td>
<td>0.69</td>
<td>1.97</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>GCA-exp</td>
<td><strong>3.07</strong></td>
<td><strong>119</strong></td>
<td><strong>0.62</strong></td>
<td><strong>2.02</strong></td>
<td><strong>40.2</strong></td>
</tr>
<tr>
<td>Lahn</td>
<td>Raw ensemble</td>
<td>3.57</td>
<td>146</td>
<td>0.98</td>
<td>2.19</td>
<td>49.7</td>
</tr>
<tr>
<td></td>
<td>INDEP</td>
<td>2.78</td>
<td>131</td>
<td>1.23</td>
<td>4.00</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>ECC-T</td>
<td>2.77</td>
<td>109</td>
<td>0.70</td>
<td>1.57</td>
<td>37.5</td>
</tr>
<tr>
<td></td>
<td>GCA-exp</td>
<td><strong>2.71</strong></td>
<td><strong>107</strong></td>
<td><strong>0.63</strong></td>
<td><strong>1.53</strong></td>
<td><strong>36.3</strong></td>
</tr>
</tbody>
</table>

*ES, CRPS of the sum, minimum, and maximum functionals as well as the 0.5 variogram scores comparing raw ensemble and EMOS forecasts with independent, ECC-T, and GCA-exp correlation structure for the rivers Upper Rhine, Moselle, and Lahn. The best values for each catchment and verification score are in bold.
dependence structures can mostly be represented adequately by either ECC-T or GCA. On average, GCA performs slightly better than ECC-T in terms of statistical verification measures. However, this is expected, because GCA retains the univariate predictive distributions, while the ECC-T trajectories depend on the raw ensemble. For instance, the ECC-T spread is zero if the ensemble spread is zero even in cases where the variance of the marginal predictive distribution for the particular lead time is large. Nevertheless, in combination with the potential to model spatiotemporal dependencies between subcatchments and lead times, EMOS with ECC-T is a suitable approach for postprocessing of subcatchment ensemble forecasts. The postprocessed subcatchment trajectories can then be used as boundary conditions and lateral inflows for a hydrodynamic model. In the present case, this would lead to well-specified forecast scenarios of runoff, and hence also water levels, in the river Rhine. Such multivariate, probabilistic forecasts may, for instance, be useful for shipping companies. Furthermore, the results suggest that the relative performance of ECC-T compared to GCA deteriorates with decreasing catchment size. This is in line with the results by Pappenberger et al. [2010] who showed that the performance of ensemble river discharge forecasts based on similar settings of coupled atmospheric and hydrologic ensemble models decreases with decreasing catchment size. Hence, it is reasonable to assume that quality of the correlation structure of the raw ensemble is highest for the large catchment of the Upper Rhine, moderate for the medium-sized catchment of the river Moselle, and lowest for the small catchment of the river Lahn. Further analyses are needed in order to confirm these results.

The EMOS models have been optimized on the Box-Cox transformed space. This approach has been chosen in order to be able to use Gaussian distributions. Nevertheless, one has to keep in mind that the predictive distributions have to be back-transformed. Hence, distances between equidistant quantiles on the Box-Cox transformed space are transformed to quantiles with increasing distances with increasing runoff volume on the original space. This in turn influences CRPS optimization, i.e., on the Box-Cox transformed space, the lower parts of the predictive distributions have more influence on CRPS calculation than on the original space. Considering the—though slight—miscalibration of the EMOS models in the [0.9,1.0] decile, an optimization procedure that gives more weight to the higher quantiles may be desirable. A first test has shown that refining the parameter estimates on the original space may slightly increase verification scores. However, numerical CRPS optimization drastically increases computational cost. Another way to approach this problem would be to apply EMOS methods that are based on positively skewed non-Gaussian distributions. Promising approaches might rely, for instance, on generalized extreme value distributions [Scheuerer, 2014; Lerch and Thorarinsdottir, 2013].

In summary, this study confirms that EMOS along with the multivariate extensions, ECC-T and GCA, provides reasonably sharp probabilistic runoff forecasts that are well calibrated in terms of univariate calibration, and from which realistic runoff scenarios over the entire range of lead times can be extracted in a straightforward manner.

Appendix A: Technical Details

A1. Box-Cox Transformation

The Box-Cox transformation is given by:

\[ h(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(x) & \text{if } \lambda = 0, \end{cases} \]

where \( x \) is on the original space and \( \lambda \) is the Box-Cox coefficient. For this study, the same estimated parameter \( \hat{\lambda} \) is used for both observations and forecasts. For each catchment considered, the estimation has been performed using the complete time series of observations and corresponding simulations from 1 November 1998 to 31 October 2008, which corresponds to the period prior to the verification period. That is, for each catchment, the estimate \( \hat{\lambda} \) is constant throughout the entire study. The actual parameter estimation has been performed by minimizing the Kolmogorov-Smirnov test statistics of a normal distribution with appropriate mean and variance and the empirical distribution of the differences between the transformed observations and the corresponding hydrological model simulations using observed meteorological input by applying the R function ks.test. The estimates \( \hat{\lambda} \) are 0.61, −0.04, and 0.03 for the rivers Upper Rhine, Moselle,
A2. Verification Techniques

A2.1. CRPSS

The CRPS has already been introduced in equation (4). Its associated skill score is given by:

$$\text{CRPSS} = \frac{\text{CRPS}_{\text{ref}} - \text{CRPS}_{\text{forc}}}{\text{CRPS}_{\text{ref}}}$$  \hspace{1cm} (A2)

where the climatological forecasts serve as reference. Note that CRPS$_{\text{ref}}$ and CRPS$_{\text{forc}}$ are averages over the entire verification period.

A2.2. Average Rank Histogram

The average rank histogram can be obtained as follows:

1. Obtain $M$ randomly sampled forecast trajectories (here over lead times 1–114 h, i.e., dimension $L = 114$) from the multivariate predictive distribution, where $M$ corresponds to the size of the raw ensemble.
2. Add the observed trajectory to the set of sampled forecast trajectories, leading to the set $S = \{x_1, \ldots, x_M, x_{M+1}\}$ of trajectories of dimension $L$ with $x_m = (x_{m1}, \ldots, x_{ml})$ for $m = 1, \ldots, M+1$. The observed trajectory $y$ is now denoted by $x_{M+1}$.
3. Calculate preranks using the average prerank function

$$\rho_S(x_l) = \frac{1}{M} \sum_{i=1}^{M} \text{rank}_S(x_i),$$

where $\text{rank}_S(x_i)$ denotes the rank of member $x$ at lead time $l$.
4. Obtain the rank of $x_{M+1}$ by first calculating $\rho_S(x_{M+1})$ and then determining its rank in $\{\rho_S(x_1), \ldots, \rho_S(x_M), \rho_S(x_{M+1})\}$ with ties resolved at random.

Calculating the above ranks for each day in the verification period allows to plot PIT-like histograms. Though they look like univariate PIT histograms, their interpretation is somewhat different. Assuming the forecasts to be marginally, i.e., for each individual lead time, well calibrated, U-shaped average rank histograms indicate too low correlations among lead times, whereas ∩-shaped histograms indicate too high correlations. But note that these histograms are highly sensitive to marginal miscalibration. Refer to Thorarinsdottir et al. [2014] for further details.

A3. ES and $p$-Variogram Score

The ES is given by:

$$\text{es}(F, y) = \mathbb{E}_p|X - y| - \frac{1}{2} \mathbb{E}_p||X - X'||,$$  \hspace{1cm} (A4)

where $|| : ||$ denotes the Euclidian norm and $\mathbb{E}$ denotes expectation. Here $X$ and $X'$ are independent random vectors following the predictive distribution $F$ and the observation vector is denoted by $y$. The ES is negatively oriented, proper, and a generalization of the CRPS.

The $p$-variogram score of order $p$ is defined by:

$$V_p(F, y) = \sum_{i,j = 1}^{L} w_{ij} (y_i - y_j)^p - \mathbb{E}_p|X_i - X_j|^p,$$  \hspace{1cm} (A5)

where $F$ denotes the $L$-variate predictive distribution, $y$ is the observation vector of length $L$, $X_i$ and $X_j$ are the $i$th and $j$th component of a random vector $X \sim F$, and $w_{ij} \geq 0$ are weights. As proposed in Scheuerer and Hamill [2015] pairs of far distant lead times are down-weighted in order to increase the signal to noise ratio. This is done by setting $w_{ij}$ to be proportional to the inverse distance between $i$ and $j$. Additionally, they have demonstrated by simulation experiments that setting $p = 0.5$ leads to the best discrimination ability of the $p$-variogram score. For further details on the $p$-variogram score, we refer to Scheuerer and Hamill [2015].
Acknowledgments

The authors are indebted to the German Federal Institute of Hydrology (BfG) for funding, support, and providing their data. Data used in this paper can be obtained on request from the BfG for research purposes. Please contact Bastian Klein: klein@bafg.de. Furthermore, the authors are grateful to T. Gneiting of the Heidelberg Institute for Theoretical Studies (HITS) and the Institute for Stochastics at Karlsruhe Institute of Technology for valuable help and comments. Furthermore, they like to thank D. Meißner of the BfG, M. Scheurer of the Physical Sciences Division at NOAA, as well as A. Jordan and R. Scheffiz (HITS) for helpful discussions. Additionally, they are grateful to the two anonymous reviewers for their helpful comments. Finally, they thank T. Thorarinsson of the Norwegian Computing Center for sharing her EMOS scripts.

References


Diggle, P. J., and P. J. Ribeiro Jr. (2008), Model Based Geostatistics, Springer, N. Y.


