

Water Resources Research

RESEARCH ARTICLE

10.1029/2018WR024028

Key Points:

- Doubly truncated Bayesian model averaging (BMA) is introduced as a postprocessing method
- The proposed BMA approach, which is tailored to bounded variables, outperforms an ensemble model output statistics reference method
- The benefit from BMA is considerable for rolling training periods but shrinks drastically for analog-based training periods

Supporting Information:

- Supporting Information S1

Correspondence to:

S. Hemri,
stephan.hemri@meteoswiss.ch

Citation:

Baran, S., Hemri, S., & El Ayari, M. (2019). Statistical postprocessing of water level forecasts using Bayesian model averaging with doubly truncated normal components. *Water Resources Research*, 55, 3997–4013. <https://doi.org/10.1029/2018WR024028>




Received 3 SEP 2018

Accepted 3 APR 2019

Accepted article online 15 APR 2019

Published online 9 MAY 2019

Statistical Postprocessing of Water Level Forecasts Using Bayesian Model Averaging With Doubly Truncated Normal Components

Sándor Baran¹ , Stephan Hemri² , and Mehrez El Ayari¹ 

¹Faculty of Informatics, University of Debrecen, Debrecen, Hungary, ²Federal Office of Meteorology and Climatology MeteoSwiss, Zürich, Switzerland

Abstract Accurate and reliable probabilistic forecasts of hydrological quantities like runoff or water level are beneficial to various areas of society. Probabilistic state-of-the-art hydrological ensemble prediction models are usually driven with meteorological ensemble forecasts. Hence, biases and dispersion errors of the meteorological forecasts cascade down to the hydrological predictions and add to the errors of the hydrological models. The systematic parts of these errors can be reduced by applying statistical postprocessing. For a sound estimation of predictive uncertainty and an optimal correction of systematic errors, statistical postprocessing methods should be tailored to the particular forecast variable at hand. Former studies have shown that it can make sense to treat hydrological quantities as bounded variables. In this paper, a doubly truncated Bayesian model averaging (BMA) method, which allows for flexible postprocessing of possibly multimodel ensemble forecasts of water level, is introduced. A case study based on water levels for a gauge of river Rhine reveals a good predictive skill of doubly truncated BMA compared both to the raw ensemble and the reference ensemble model output statistics approach. Using rolling training periods, BMA considerably outperforms ensemble model output statistics. However, this gap shrinks drastically when using analog-based training periods.

1. Introduction

Hydrological forecasts are important for a heterogeneous group of users such as the operators of hydrological power plants, flood prevention authorities, or shipping companies. For rational decision making based on cost-benefit analyses, an estimate of the predictive uncertainty (Krzysztofowicz, 1999; Todini, 2008) needs to be provided with any forecast. The state-of-the-art approach of using a set of parallel runs of a hydrological model driven by meteorological ensemble forecasts provided by numerical weather prediction models (Cloke & Pappenberger, 2009) gives a first estimate of the meteorological input uncertainty. However, numerical weather prediction ensembles are usually biased and underdispersed (Bougeault et al., 2010; Buizza et al., 2005; Park et al., 2008). Moreover, additional sources of uncertainty like hydrological model formulation, boundary, and initial condition uncertainty as well as measurement uncertainties are typically neglected. Hence, statistical postprocessing is important in order to reduce systematic errors and to obtain an appropriate estimate of the predictive uncertainty (Buizza, 2018).

In the last decade, various methods of statistical calibration of ensemble forecasts for different weather variables have been developed (see, e.g., Ruiz & Saulo, 2012; Schmeits & Kok, 2010; Williams et al., 2014; Wilks, 2018), and parametric methods such as ensemble model output statistics (EMOS; Gneiting et al., 2005) or Bayesian model averaging (BMA; Raftery et al., 2005) provide full predictive distributions. The EMOS predictive distribution is given by a single parametric probability law with parameters depending on the ensemble, whereas the BMA predictive probability density function (PDF) is a weighted mixture of PDFs corresponding to the individual ensemble members. EMOS and BMA models for various weather quantities differ in the applied parametric distribution family. Once the predictive distribution is given, its functionals (e.g., median or mean) can be considered as classical point forecasts.

Besides the successful application, for example, to ensemble forecasts for temperature (Gneiting et al., 2005), wind speed (Baran & Lerch, 2015; Lerch & Thorarinsdottir, 2013; Thorarinsdottir & Gneiting, 2010), or precipitation (Baran & Nemoda, 2016; Scheuerer, 2014; Scheuerer & Hamill, 2015), EMOS-based statistical postprocessing turned out to improve the predictive performance of hydrological ensemble forecasts for

different gauges along river Rhine (Hemri et al., 2015; Hemri & Klein, 2017). EMOS is a quite parsimonious postprocessing method that basically links a parametric forecast distribution to ensemble statistics like the ensemble mean and the ensemble variance. Therefore, its performance is limited by (i) how well the true process can be represented by a parametric distribution family and (ii) to what extent the complete information from the ensemble can be summarized in a limited set of ensemble statistics. For instance, a typical EMOS approach based on a Gaussian or a Gamma distribution family is not able to model bimodal forecast distributions. However, BMA, which has also been applied to hydrological ensemble forecasts (see, e.g., Duan et al., 2007; Hemri et al., 2013), is more flexible in that it converts a (multimodel) raw ensemble to a mixture distribution which allows multimodal shapes. Accordingly, we hypothesize that BMA may be able to outperform EMOS.

As the use of the BMA approach is very convenient in a Gaussian framework, both Duan et al. (2007) and Hemri et al. (2013) perform a Box-Cox transformation prior to applying BMA in order to achieve approximate normality despite the positive skewness of water levels. Additionally, it is important to ensure that the resulting water level quantiles of the predictive distribution are within realistic physical bounds. At the upper bound of the distribution water levels should be lower than a water level threshold resulting from an extreme flood with a small exceedance probability; at the lower bound water levels should be higher than a water level threshold resulting from an extreme long-lasting low water period with a small nonexceedance probability. In order to ensure realistic values while still being able to benefit from the mathematical simplicity of Gaussian models, we use a lower and upper truncated normal distribution. Generally, the data of the water level gauges are defined to avoid water levels below zero. Accordingly, we are using an ad hoc lower truncation limit of half of the lowest value ever recorded. Though this ad hoc limit turned out to be appropriate for the gauge considered, in general, we recommend to base the derivation of the lower truncation limit on a physical property like the cease-to-flow stage. Even though river physics do not provide any hard upper limit for water level, as a result of the Box-Cox transformation, we need to apply also an upper truncation limit. Otherwise, due to the skewness of water level, back transformation from the Box-Cox transformed to the original space may lead to predicted water levels that are way above the range of the rating curve. For this study we have set the upper truncation limit to two times the maximum gauge level ever recorded. However, in general, we recommend to base the upper limit based on assessing the range of validity of the rating curve. Above this range, the hydrometric gauge may be bypassed, and hence, the rating curve becomes obsolete.

To our best knowledge, up to now, there is no study that has applied a doubly truncated normal BMA approach. In this study, the work by Baran (2014), which introduces a one-sided truncated normal BMA method, is extended to a two-sided truncated normal BMA approach. Its performance and its suitability for hydrological ensemble forecasts is assessed through the example of multimodel ensemble forecasts of water level at gauge Kaub at river Rhine.

Doubly truncated BMA is introduced in section 2 on calibration methods and forecast evaluation, followed by a brief description of the data in section 3. The results are presented in section 4, and conclusions are drawn in section 5.

2. Calibration Methods and Forecast Evaluation

2.1. Bayesian Model Averaging

As mentioned in section 1, for the BMA postprocessing approach, the predictive distribution of a future weather quantity is a weighted mixture of probability laws corresponding to the individual ensemble members. In general, if f_1, \dots, f_K denote the ensemble forecast of a given weather or hydrological quantity X for a given location, time, and lead time, the BMA predictive PDF (Raftery et al., 2005) of X equals

$$p(x|f_1, \dots, f_K; \theta_1, \dots, \theta_K) := \sum_{k=1}^K \omega_k g(x|f_k, \theta_k), \quad (1)$$

where $g(x|f_k, \theta_k)$ is the component PDF from a parametric family corresponding to the k th ensemble member f_k with parameter (vector) θ_k to be estimated and ω_k is the corresponding weight linked to the relative performance of this particular member during the training period. Note that the weights should form a probability distribution, that is, $\omega_k \geq 0, k = 1, \dots, K$ and $\sum_{k=1}^K \omega_k = 1$.

Recently, most operational ensemble predictions systems incorporate ensembles where at least some members can be considered as statistically indistinguishable and in this way exchangeable. This is the case with

the 51 member operational European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble (Leutbecher & Palmer, 2008; Molteni et al., 1996), where the 50 members generated using perturbed initial conditions form an exchangeable group. Further, in this context one can also consider multimodel ensemble predictions systems such as the GLAMEPS ensemble (Iversen et al., 2011) or the THORPEX Interactive Grand Global Ensemble (Swinbank et al., 2016). Obviously, using ensemble weather forecasts consisting of groups of exchangeable members as inputs of a hydrological model results in hydrological ensemble forecasts with exchangeable members, which is the case with the water level data at hand described in section 3. To account for the existence of groups with exchangeable members, Fraley et al. (2010) suggest to use the same weights and parameters within a given group. Thus, if we have M ensemble members divided into K groups, where the k th group contains $M_k \geq 1$ exchangeable ensemble members ($\sum_{k=1}^K M_k = M$) and $f_{k,\ell}$ denotes the ℓ th member of the k th group, model (1) is replaced by

$$p(x|f_{1,1}, \dots, f_{1,M_1}, \dots, f_{K,1}, \dots, f_{K,M_K}; \theta_1, \dots, \theta_K) := \sum_{k=1}^K \sum_{\ell=1}^{M_k} \omega_k g(x|f_{k,\ell}, \theta_k). \quad (2)$$

For the sake of simplicity, in the remaining part of this section we provide results and formulae only for model (1) as their extension to model (2) is rather straightforward.

2.2. Truncated Normal BMA Model

For weather variables such as temperature or pressure, BMA models with Gaussian components provide a reasonable fit (Fraley et al., 2010; Raftery et al., 2005), whereas wind speed calls for nonnegative and skewed distributions such as gamma (Sloughter et al., 2010) or truncated normal with truncation from below at zero (Baran, 2014). However, water levels are typically non-Gaussian (see, e.g., Duan et al., 2007); moreover, as stated above, they are bounded both from below and from above. These constraints should also be taken into account during model formulation. A general procedure is to normalize the forecasts and observations using, for instance, Box-Cox transformation

$$h_\lambda(x) := \begin{cases} (x^\lambda - 1) / \lambda, & \lambda \neq 0, \\ \log(x), & \lambda = 0, \end{cases} \quad (3)$$

with some coefficient λ , perform postprocessing, and then backtransform the results using the inverse Box-Cox transformation (Duan et al., 2007; Hemri et al., 2014, 2015). Following the ideas of Hemri and Klein (2017), for modeling Box-Cox transformed water levels we use a doubly truncated normal distribution $\mathcal{N}_a^b(\mu, \sigma^2)$, with PDF

$$g_{a,b}(x|\mu, \sigma) := \frac{\frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}, \quad x \in [a, b], \quad (4)$$

and $g_{a,b}(x|\mu, \sigma) := 0$ otherwise, where a and b are the lower and upper bounds and φ and Φ denote the PDF and the cumulative distribution function (CDF) of the standard normal distribution, respectively. The mean and variance of $\mathcal{N}_a^b(\mu, \sigma^2)$ are

$$\begin{aligned} \kappa &= \mu + \sigma \frac{\varphi\left(\frac{a-\mu}{\sigma}\right) - \varphi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \quad \text{and} \\ \phi^2 &= \sigma^2 \left(1 + \frac{\frac{a-\mu}{\sigma} \varphi\left(\frac{a-\mu}{\sigma}\right) - \frac{b-\mu}{\sigma} \varphi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} - \left(\frac{\varphi\left(\frac{a-\mu}{\sigma}\right) - \varphi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \right)^2 \right), \end{aligned} \quad (5)$$

respectively. The proposed BMA predictive PDF is

$$p(x|f_1, \dots, f_K; \alpha_1, \dots, \alpha_K; \beta_1, \dots, \beta_K; \sigma) = \sum_{k=1}^K \omega_k g_{a,b}(x|\alpha_k + \beta_k f_k, \sigma), \quad (6)$$

where we assume that the location of the k th mixture component is an affine function of the corresponding ensemble member f_k and scale parameters are assumed to be equal for all component PDFs. The latter

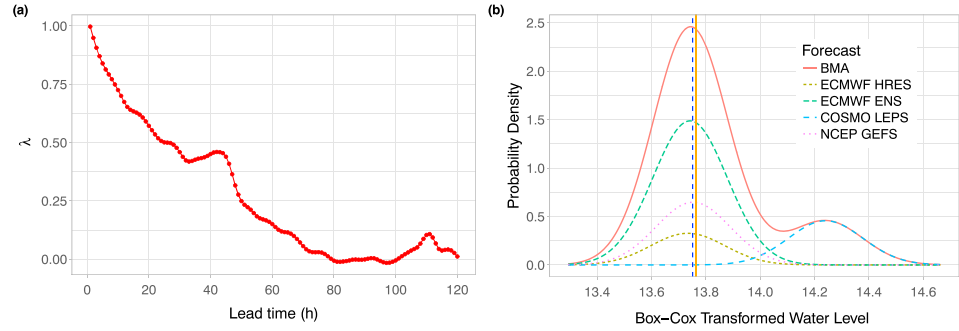


Figure 1. Box-Cox transformation parameter λ as function of the lead time (a); BMA predictive distribution and its components of Box-Cox transformed water levels for 30 July 2008 corresponding to 48-hr lead time (b). Vertical solid line: transformed verifying observation; dashedline: BMA median. BMA = Bayesian model averaging; ECMWF HRES = ECMWF high-resolution; ECMWF ENS = ECMWF ensemble; COSMO LEPS = consortium for small-scale modeling limited-area ensemble prediction system; NCEP GEFS = National Center for Environmental Prediction global ensemble forecast system.

assumption is for the sake of simplicity and is common in BMA modeling (see, e.g., Raftery et al., 2005), whereas the form of the location parameter is in line with the truncated normal BMA model of Baran (2014). An example of a BMA predictive distribution (6) of Box-Cox transformed water levels for 30 July 2008 is given in Figure 1b, where the overall predictive PDF and the component PDFs are plotted, together with the BMA median and the transformed validating observation.

2.3. Parameter Estimation

Location parameters α_k , β_k ; weights ω_k , $k = 1, \dots, K$; and scale parameter σ can be estimated from training data, which consists, for instance, of ensemble members and validating observations from the preceding n days. In section 2.6 several more sophisticated approaches to training data selection are provided. In the BMA approach, estimates of location parameters are typically obtained by regressing the validating observations on the ensemble members, whereas weights and scale parameter(s) are obtained via maximum likelihood (ML) estimation (see, e.g., Raftery et al., 2005; Slougher et al., 2007, 2010), where the log likelihood function of the training data is maximized using the EM algorithm for mixture distributions (Dempster et al., 1977; McLachlan & Krishnan, 1997). However, the regression approach assumes the location parameters to be simple functions of the mean, which is obviously not the case for the truncated normal distribution; see (5). Hence, we propose a pure ML method estimating all model parameters by maximizing the likelihood function, which ideas have already been considered, for example, by Slougher et al. (2010) or Baran (2014).

In what follows, for a given location $s \in S$ and time $t \in \mathcal{T}$ let $f_{k,s,t}$ denote the k th ensemble member and denote by $x_{s,t}$ the corresponding validating observation. Here S denotes the set of locations sharing the same BMA model parameters and \mathcal{T} is the set of training dates. In the case study of section 4 S consists of a single location; however, for more complex ensemble domains different choices of training data are possible; for more details, see, for example, Lerch and Baran (2017). Further, as in the case study of section 4 the different lead times are treated separately; reference to the lead time of the forecast is omitted. By assuming the conditional independence of forecast errors with respect to the ensemble members in space and time, the log likelihood function for model (6) corresponding to all forecast cases (s, t) in the training set equals

$$\ell(\omega_1, \dots, \omega_K, \alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_K, \sigma) = \sum_{s,t} \log \left[\sum_{k=1}^K \omega_k g_{a,b}(x_{s,t} | \alpha_k + \beta_k f_{k,s,t}, \sigma) \right]. \quad (7)$$

To obtain the ML estimates, we apply EM algorithm for truncated Gaussian mixtures proposed by Lee and Scott (2012) with a mean correction. In line with the classical EM algorithm for mixtures (McLachlan & Krishnan, 1997), we first introduce latent binary indicator variables $z_{k,s,t}$ identifying the mixture component where the observation $x_{s,t}$ comes from, that is, $z_{k,s,t}$ is 1 or 0 accordingly as whether $x_{s,t}$ follows or not the k th component distribution. Using these indicator variables, one can provide the complete data log likelihood

corresponding to (7) in the form

$$\begin{aligned} \ell_C(\omega_1, \dots, \omega_K, \alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_K, \sigma) \\ = \sum_{s,t} \sum_{k=1}^K z_{k,s,t} [\log(\omega_k) + \log(g_{a,b}(x_{s,t} | \mu_{k,s,t}, \sigma))], \end{aligned} \quad (8)$$

with $\mu_{k,s,t} := \alpha_k + \beta_k f_{k,s,t}$. After specifying the initial values of the parameters, the EM algorithm alternates between an expectation (E) and a maximization (M) step until convergence. As first guesses $a_k^{(0)}$ and $b_k^{(0)}$, $k = 1, \dots, K$, for the location parameters, we suggest to use the coefficients of the linear regression of $x_{s,t}$ on $f_{k,s,t}$, so $\mu_{k,s,t}^{(0)} = \alpha_k^{(0)} + \beta_k^{(0)} f_{k,s,t}$. Initial scale $\sigma^{(0)}$ can be the standard deviation of the observations in the training data set or the average residual standard deviation from the above regression, whereas the initial weights might be chosen uniformly, that is, $\omega_k^{(0)} = 1/K$, $k = 1, \dots, K$. Then in the E step the latent variables are estimated using the conditional expectation of the complete log likelihood on the observed data, while in the M step the parameter estimates are updated by maximizing ℓ_C given the actual values of the latent variables.

For the doubly truncated normal model specified by (4) and (6), the E step of the $(j + 1)$ st iteration is

$$z_{k,s,t}^{(j+1)} := \frac{\omega_k^{(j)} g_{a,b}(x_{s,t} | \mu_{k,s,t}^{(j)}, \sigma^{(j)})}{\sum_{i=1}^K \omega_i^{(j)} g_{a,b}(x_{s,t} | \mu_{i,s,t}^{(j)}, \sigma^{(j)})}. \quad (9)$$

Once the estimates of the indicator variables (which are not necessary 0 or 1 any more) are given, the first part of the M step updating the weights is obviously

$$\omega_k^{(j+1)} := \frac{1}{N} \sum_{s,t} z_{k,s,t}^{(j+1)}, \quad (10)$$

where N is the total number of forecast cases in the training set.

Further, nonlinear equations $\frac{\partial \ell_C}{\partial \alpha_k} = 0$ and $\frac{\partial \ell_C}{\partial \beta_k} = 0$, $k = 1, \dots, K$, lead us to update formulae

$$\begin{aligned} \alpha_k^{(j+1)} &:= \left[\sum_{s,t} z_{k,s,t}^{(j+1)} \right]^{-1} \sum_{s,t} z_{k,s,t}^{(j+1)} \left\{ \left(x_{k,s,t} - \beta_k^{(j)} f_{k,s,t} \right) + \sigma^{(j)} \frac{\varphi\left(\frac{b - \mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \varphi\left(\frac{a - \mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)}{\Phi\left(\frac{b - \mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \Phi\left(\frac{a - \mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)} \right\}, \\ \beta_k^{(j+1)} &:= \left[\sum_{s,t} z_{k,s,t}^{(j+1)} f_{k,s,t} \right]^{-1} \sum_{s,t} z_{k,s,t}^{(j+1)} f_{k,s,t} \left\{ \left(x_{k,s,t} - \alpha_k^{(j)} \right) + \sigma^{(j)} \frac{\varphi\left(\frac{b - \mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \varphi\left(\frac{a - \mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)}{\Phi\left(\frac{b - \mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \Phi\left(\frac{a - \mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)} \right\}, \end{aligned} \quad (11)$$

respectively. However, using then simply $\mu_{k,s,t}^{(j+1)} := \alpha_k^{(j+1)} + \beta_k^{(j+1)} f_{k,s,t}$ as the update of the location results in an unstable parameter estimation process due to numerical issues. Hence, similar to Baran (2014), where the same problem occurred in case studies with wind speed data, we introduce a mean correction of form

$$\mu_{k,s,t}^{(j+1)} := \mu_{k,s,t}^{(0)} - \sigma^{(j)} \frac{\varphi\left(\frac{a - \alpha^{(j+1)} - \beta^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\right) - \varphi\left(\frac{b - \alpha^{(j+1)} - \beta^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\right)}{\Phi\left(\frac{b - \alpha^{(j+1)} - \beta^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\right) - \Phi\left(\frac{a - \alpha^{(j+1)} - \beta^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\right)}, \quad (12)$$

which reflects the difference between the location and mean of a truncated normal distributions; see (5). Finally, from $\frac{\partial \ell_C}{\partial \sigma} = 0$ we obtain the last update formula

$$\begin{aligned} \sigma^{2(j+1)} &:= \frac{1}{N} \sum_{s,t} \sum_{k=1}^K z_{k,s,t}^{(j+1)} \left\{ \left(x_{s,t} - \mu_{k,s,t}^{(j+1)} \right)^2 \right. \\ &\quad \left. + \sigma^{(j)} \frac{\left(b - \mu_{k,s,t}^{(j+1)} \right) \varphi\left(\frac{b - \mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\right) - \left(a - \mu_{k,s,t}^{(j+1)} \right) \varphi\left(\frac{a - \mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\right)}{\Phi\left(\frac{b - \mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\right) - \Phi\left(\frac{a - \mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\right)} \right\}. \end{aligned} \quad (13)$$

Note that without truncation ($-a = b = \infty$) the terms of (11) and (13) depending on $\sigma^{(j)}$ disappear, so location (mean) and scale (standard deviation) are updated separately, no mean correction is required, and we get back the classical EM algorithm for normal mixtures.

As a more simple alternative approach, one can omit the update step (11) for α_k and β_k , simplify the mean correction step (12) to

$$\mu_{k,s,t}^{(j+1)} := \mu_{k,s,t}^{(0)} - \sigma^{(j)} \frac{\varphi\left(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \varphi\left(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)}{\Phi\left(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right) - \Phi\left(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\right)}, \quad (14)$$

and only after the EM algorithm stops, estimate location parameters α_k and β_k from a linear regression of the final value of $\mu_{k,s,t}$ on $f_{k,s,t}$.

Finally, one can also try the classical naive approach, where location parameters α_k and β_k are not updated at all, that is, $\mu_{k,s,t}^{(j+1)} \equiv \alpha_k^{(0)} + \beta_k^{(0)} f_{k,s,t}$. In the case study of section 4.1, the latter two approaches do not show significantly different forecast skills in terms of the verification scores defined in section 2.4, so only the results for the naive and pure ML approaches are reported. The two simple approaches provide very similar location and scale parameters; the corresponding predictive distributions mainly differ in weights. In contrast, the pure ML method results in completely different location parameters. In terms of computation costs the ranking is naive, naive with mean correction and pure ML. A more detailed analysis of computation times of these approaches can be found in Baran (2014).

2.4. Verification Scores

In probabilistic forecasting, the principal aim is to access the maximal sharpness of the predictive distribution subject to calibration (Gneiting et al., 2007), where the latter means a statistical consistency between the predictive distributions and the validating observations, whereas the former refers to the concentration of the predictive distribution. One of the simplest tools for getting a first impression about the calibration of forecast distributions is the probability integral transform (PIT) histogram. By definition, the PIT is the value of predictive CDF at the validating observation (Raftery et al., 2005), which in case of proper calibration should follow a uniform distribution on the $[0, 1]$ interval. In this way the PIT histogram is the continuous counterpart of the verification rank histogram for the raw ensemble, which is defined as histogram of ranks of validating observations with respect to the corresponding ensemble forecasts (see, e.g., Wilks, 2011, section 7.7.2). Again, for a properly calibrated ensemble the ranks should be uniformly distributed.

Predictive performance can be quantified with the help of scoring rules, which are loss functions assigning numerical values to pairs of forecasts and observations. In hydrology and atmospheric sciences, one of the most popular scoring rules is the continuous ranked probability score (CRPS; Gneiting & Raftery, 2007; Wilks, 2011), as it assesses calibration and sharpness simultaneously. For a (predictive) CDF $F(y)$ and real value (observation) x the CRPS is defined as

$$\begin{aligned} \text{CRPS}(F, x) &:= \int_{-\infty}^{\infty} (F(y) - \mathbb{1}_{\{y \geq x\}})^2 dy = \int_{-\infty}^x F^2(y) dy + \int_x^{\infty} (1 - F(y))^2 dy \\ &= \mathbb{E}|X - x| - \frac{1}{2} \mathbb{E}|X - X'|, \end{aligned} \quad (15)$$

where $\mathbb{1}_H$ denotes the indicator of a set H , whereas X and X' are independent random variables with CDF F and finite first moment. CRPS is a negatively oriented proper scoring rule (Gneiting & Raftery, 2007), that is, the smaller the better, and the right-hand side of (15) shows that it can be expressed in the same unit as the observation. For truncated normal distribution the CRPS has a simple closed form (see, e.g., the R package `scoringRules`; Jordan & Krüger), whereas for the truncated normal mixture (6), similar to the mixture model of Baran and Lerch (2016), the second integral expression in the definition (15) should be evaluated numerically. Moreover, in our case study each calibration approach provides a predictive CDF F for the Box-Cox transformed water level $X \in [a, b]$. Thus, the CRPS corresponding to the predictive CDF $G(y) := F(h_\lambda(y))$ of the original water level $Y = h_\lambda^{-1}(X) \in [h_\lambda^{-1}(a), h_\lambda^{-1}(b)]$ and a real value y equals

$$\text{CRPS}(G, y) = \int_{h_\lambda^{-1}(a)}^y F^2(h_\lambda(u)) du + \int_y^{h_\lambda^{-1}(b)} (1 - F(h_\lambda(u)))^2 du, \quad (16)$$

which integral should again be approximated numerically. Further, in order to get the CRPS of the raw ensemble, the empirical CDF is considered as predictive distribution.

In case studies, competing forecast methods can be compared by the mean CRPS value

$$\overline{\text{CRPS}} := \frac{1}{N} \sum_{n=1}^N \text{CRPS}(F_i, x_i)$$

over all pairs $(F_i, x_i), i = 1, \dots, N$, of forecasts and observations in the verification data. Further, the improvement in CRPS with respect to a reference method can be quantified with the help of the continuous ranked probability skill score (CRPSS; Gneiting & Raftery, 2007; Murphy, 1973), defined as

$$\text{CRPSS} := 1 - \frac{\overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{ref}}},$$

where $\overline{\text{CRPS}}_{\text{ref}}$ denotes the mean CRPS value corresponding to the reference approach. In contrast to the CRPS, the CRPSS is positively oriented, that is, the larger the better.

Calibration and sharpness of a predictive distribution can also be investigated using the coverage and average width of the $(1 - \alpha)100\%$, $\alpha \in (0, 1)$, central prediction interval, respectively (Gneiting et al., 2007). As the coverage we consider the proportion of validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive CDF, and level α should be chosen to match the nominal coverage of the raw ensemble, that is, $(K - 1)/(K + 1)100\%$, where K is the ensemble size (see, e.g., Baran & Lerch, 2015; Raftery et al., 2005). As the coverage of a calibrated predictive distribution should be around $(1 - \alpha)100\%$, such a choice of α allows direct comparison with the raw ensemble.

Further, as point forecasts we consider the medians of the predictive distributions and the raw ensemble, which are evaluated with the help of the mean absolute error (MAE).

Finally, as suggested by Gneiting and Ranjan (2011), statistical significance of the differences between the verification scores is assessed by utilizing the Diebold-Mariano (DM; Diebold & Mariano, 1995) test, which allows accounting for the temporal dependencies in the forecast errors. In the case study of section 4 we report the p values of the test. The p values less than 0.05 indicate significant difference in scores at a 5% level. The detailed description of the DM test can be found, for example, in Baran and Lerch (2016).

2.5. Truncated Normal EMOS Model

As a reference postprocessing method for calibration of Box-Cox transformed ensemble forecasts for water levels, we consider the truncated normal EMOS model of Hemri and Klein (2017). In this approach the predictive distribution is a single doubly truncated normal distribution $\mathcal{N}_a^b(\mu, \sigma^2)$ defined by (4), and the ensemble members are just linked to the location μ and scale σ via equations

$$\mu = a_0 + a_1 f_1 + \dots + a_K f_K \quad \text{and} \quad \sigma^2 = b_0 + b_1 S^2, \quad (17)$$

where S^2 denotes the variance of the transformed ensemble. In case of existence of groups of exchangeable ensemble members the equation for the location in (17) is replaced by

$$\mu = a_0 + a_1 \bar{f}_1 + \dots + a_K \bar{f}_K, \quad (18)$$

where \bar{f}_k denotes the mean value of the k th group. According to the optimum score estimation principle of Gneiting and Raftery (2007), location parameters $a_0, a_1, \dots, a_K \in \mathbb{R}$ and scale parameters $b_0, b_1 \geq 0$ are estimated from the training data by optimizing a proper verification score, which is usually the CRPS defined by (15).

2.6. Analog-Based Approaches to Choosing Training Data

Following Hemri and Klein (2017), we rely on the series distance method (SD; Ehret & Zehe, 2011; Seibert et al., 2016), the hydrograph matching algorithm (HMA; Ewen, 2011), and dynamic time warping (DTW; Sakoe & Chiba, 1978) to select analog-based training periods. The goal of SD and HMA is to mimic the human hydrologist in quantifying the difference between two hydrological time series. SD matches peaks, troughs, and the falling and rising limbs in between. The amount of corrections needed to match the two time series with regard to both timing and amplitude is used as a measure of similarity. HMA is similar; it connects each element of the first time series to the other time series allowing moderate time shifts. The total

length of all connections constitutes another measure of similarity. DTW provides a measure of similarity by finding the minimal amplitude error between the two time series that can be obtained by temporal stretching and compression only.

While these methods are typically used to compare simulated with observed hydrological time series, we use them here to find training data with a forecast time series that is similar to the one at the verification date of interest. Accordingly, for each verification date and for each of SD, HMA, and DTW, we select the 100 training dates with the highest similarity in the ECMWF ENS mean trajectories. Obviously, training dates with forecast trajectories that overlap with one of the verification dates are excluded from the training set. Refer to Hemri and Klein (2017) for details on these approaches.

3. Data

BMA and EMOS calibration approaches are tested on ensemble forecasts for water level (cm) at gauge Kaub of river Rhine (546 km) and the corresponding validating observations. Predictions for an 8-year period between 1 January 2008 and 31 December 2015 are investigated with lead times from 1 to 120 hr with a time step of 1 hr. The minimum and maximum recorded water levels at this particular gauge are 35 and 825 cm, respectively. Our 79-member multimodel water level ensemble is obtained by plugging ensemble forecasts for the relevant weather variables produced by different ensemble prediction systems into the hydrological model HBV-96 (Lindström et al., 1997), which is run at the German Federal Institute of Hydrology (BfG) for operational runoff forecasting. We consider the ECMWF high-resolution (HRES) forecast, the 51-member ECMWF forecast (ENS) (Leutbecher & Palmer, 2008; Molteni et al., 1996), the 16-member COSMO LEPS forecast of the limited-area ensemble prediction system of the consortium for small-scale modeling (Montani et al., 2011), and the 11-member NCEP GEFS forecast of the reforecast version 2 of the global ensemble forecast system of the National Center for Environmental Prediction (Hamill et al., 2013). The runoff forecasts are then converted into water level forecasts for the navigation-relevant gauges, including gauge Kaub, using a hydrodynamic model. All ensemble forecasts are initialized at 6 UTC. We remark that the data set at hand is part of the data studied in Hemri and Klein (2017), where we refer to for further details.

4. Results

As mentioned in sections 2.2 and 2.5, BMA and EMOS postprocessing is applied for modeling Box-Cox transformed water levels. As in Hemri and Klein (2017), each lead time has an individual Box-Cox parameter λ (see Figure 1a) maximizing the in-sample skill of seasonally fitted EMOS models in terms of the CRPS relative to the raw ensemble, where data from the same season of other years are used for training. These estimates are then averaged over the training periods in order to obtain one estimate per lead time. Obviously, for a given lead time the same coefficient is applied both for the forecasts and observations.

Similar to Hemri and Klein (2017), we assume that water levels are in the interval spanned by half of the minimum and double of the maximum recorded water level, that is, they are between 17.5 and 1,650 cm, so the Box-Cox transforms of these values serve as lower and upper bounds for the truncated normal distribution used both in BMA and EMOS modeling.

The generation of the hydrological ensemble forecast described in section 3 induces a natural grouping of the ensemble members. One contains just the forecast based on the ECMWF HRES, the other 51-member group corresponds to the ECMWF ensemble (ENS), whereas forecasts based on COSMO LEPS and NCEP GEFS ensemble weather forecasts form two other groups of sizes 16 and 11, respectively. Hence, Box-Cox transformed water level forecasts are calibrated using the truncated normal BMA model for exchangeable ensemble members specified by (2) and (4) and truncated normal EMOS given by (17) and (18) with $K = 4$ and $M_1 = 1$, $M_2 = 51$, $M_3 = 16$, $M_4 = 11$. This means that the BMA model has 12 free parameters to be estimated, whereas the corresponding EMOS model has 7. To ensure a reasonably stable parameter estimation, BMA and EMOS models are trained using forecasts and observations of 100 days. We consider both rolling training periods (RTP) and analog-based approaches.

While BMA and EMOS models are fit to Box-Cox transformed values, to ensure comparability, we provide verification scores for the original forecasts and observations. This means that for quantile-based scores (MAE, coverage, and average width), before evaluating the score, the inverse Box-Cox transformation is

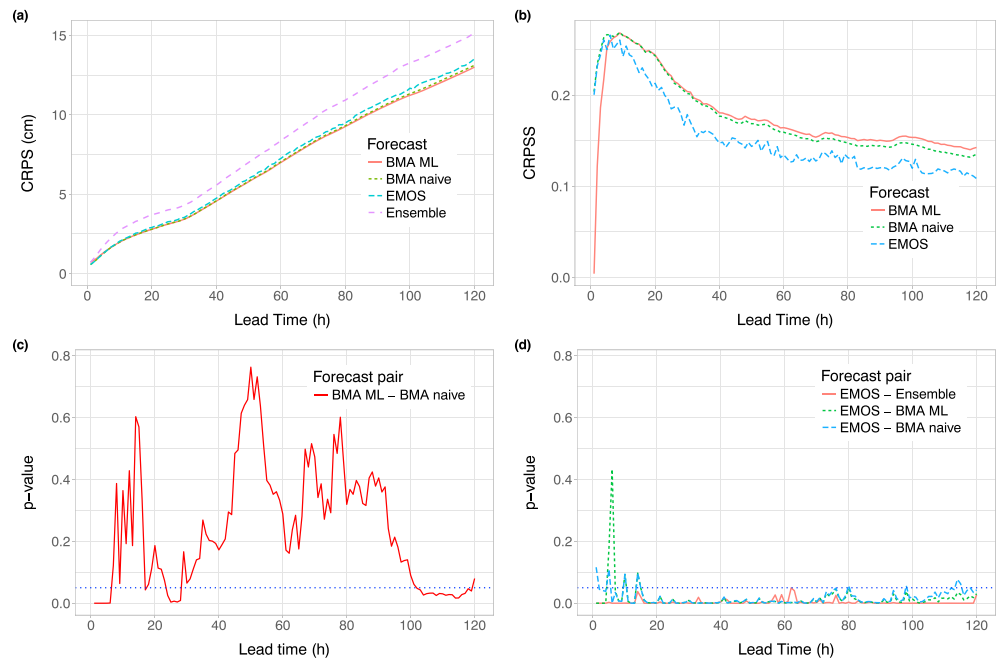


Figure 2. Mean CRPS values (a) and CRPSS with respect to the raw ensemble (b). *p* values of Diebold-Mariano tests for equality of mean CRPS of the two BMA approaches (c) and of all models compared to EMOS (d). Horizontal dotted lines of (c) and (d) indicate a 5% level of significance. BMA = Bayesian model averaging; ML = maximum likelihood; EMOS = ensemble model output statistics; CRPS = continuous ranked probability score; CRPSS = continuous ranked probability skill score.

applied to the appropriate quantiles of the predictive distribution, whereas CRPS is calculated with the help of (16).

4.1. BMA Versus EMOS Using A RTP

To investigate the forecast skill of the truncated normal BMA model introduced in section 2.2, first, we use the standard approach in BMA and EMOS postprocessing (Gneiting et al., 2005; Raftery et al., 2005) and estimate model parameters using a RTP of length 100 days. Thus, BMA and EMOS models are verified on the period 10 April 2008 to 31 December 2015 (2822 calendar days). Further, we consider 1 day ahead calibration for all lead times. This means that for modeling water level, for example, for 1 January 2015, we use forecasts and observations for the preceding 100 days ending at 31 December 2014. For 24-hr lead time the last forecasts are initialized at 30 December 2014, whereas 120-hr lead time at 26 December 2014.

In Figure 2a the mean CRPS values of the different postprocessing approaches and the raw ensemble are plotted as functions of the lead time. Note that compared to the raw ensemble, all calibration approaches reduce the mean CRPS and the gap increases together with the lead time. The differences between the forecast skills are more pronounced in Figure 2b showing the CRPSS values with respect to the raw ensemble forecast. Note that all three presented methods have their maximal skill score at hour 9. This reflects that the relative gap in CRPS between raw and postprocessed forecasts is increasing up to hour 9 and decreasing again thereafter. However, it does not imply that the absolute forecast skill increases with lead time between hours 1 and 9. For shorter lead times this increase is very fast and naive BMA shows the best predictive performance, whereas for longer lead times the pure ML BMA starts dominating. Obviously, longer lead times are also associated with larger forecast uncertainty which should be taken into account when one compares predictive performance. According to the results of DM tests for equal predictive performance, naive BMA significantly outperforms the raw ensemble for all lead times and the same holds for the pure ML BMA except hour 1. In general, in terms of the mean CRPS the two BMA approaches differ significantly mainly for very short and long lead times, as can be observed on the graph of *p* values displayed in Figure 2c. EMOS also significantly outperforms the raw ensemble for all lead times and except for the first couple of hours underperforms the BMA approaches, as depicted in Figure 2d.

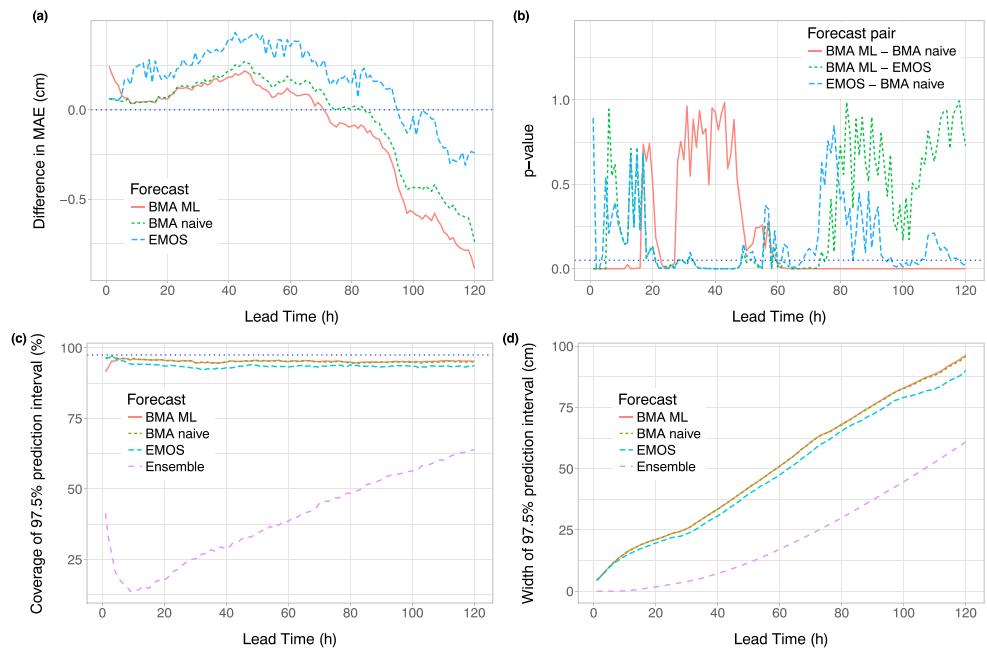


Figure 3. Top row: Difference in MAE values from the raw ensemble (a) and p values of Diebold-Mariano tests for equality of MAE of the various postprocessing approaches (b). Horizontal dotted lines indicate the reference raw ensemble (a) and a 5% level of significance (b). Bottom row: Coverage (c) and average width (d) of nominal 97.5% central prediction intervals. In panel (c) the ideal coverage is indicated by the horizontal dotted line. MAE = mean absolute error; BMA = Bayesian model averaging; ML = maximum likelihood; EMOS = ensemble model output statistics.

There is much less variety in the performance of BMA and EMOS calibrated medians in terms of the MAE. According to Figure 3a showing the difference in MAE with respect to the raw ensemble, the pure ML BMA has the best forecast skill; however, even this approach underperforms the raw ensemble until hour 70. Note that DM tests for equality of MAE values indicate that all differences plotted in Figure 3a are significant (DM test results are not reported), which is definitely not the case if we compare the performance of the three postprocessing methods; see the p values of Figure 3b.

The positive effect of postprocessing on calibration can be clearly observed in Figure 3c showing the coverages of nominal 97.5% central prediction intervals as functions of the lead time. All postprocessing approaches for all lead times result in almost perfect coverage, whereas the coverage of the raw ensemble is much lower and strongly depends on the lead time. The coverage values of the two BMA approaches are almost identical, and after hour 4 they are closer to the nominal value than those of the EMOS. Finally, as depicted in Figure 3d, the raw ensemble produces the sharpest forecasts for all lead times, however, at the cost of being uncalibrated. This is fully in line with the verification rank histograms of the raw ensemble and PIT histograms of postprocessed forecasts for lead times 24, 72, and 120 hr plotted in Figure 4a. All verification rank histograms are strongly U shaped (and the same holds for other lead times, not reported), indicating that the raw ensemble is strongly underdispersive and requires postprocessing. BMA and EMOS approaches significantly improve the statistical calibration of the forecast and result in more uniform PIT histograms, although for hour 120 naive BMA and EMOS still show a slight underdispersion. Figure 4b displays the values of the test statistic of the Kolmogorov-Smirnov test for uniformity of PIT values for different postprocessing approaches (the smaller the test statistic, the better the fit). Although the uniformity of the PIT values of pure ML BMA, naive BMA, and EMOS can be accepted at a 5% level of significance for only 9 (5, 6, 7, 14, 17, 72, 75, 77, and 79 hr), 6 (4, 5, 6, 7, 14, and 17 hr), and 4 (5, 6, 7, and 9 hr) different lead times (test statistic values under the dotted line), respectively, Figure 4b nicely illustrates the ranking of different approaches in terms of goodness of fit of PIT.

4.2. Analog-Based Versus RTP

According to the results of Hemri and Klein (2017), the analog-based choice of training data significantly improves the predictive performance of EMOS compared with the use of RTPs. Here we investigate whether

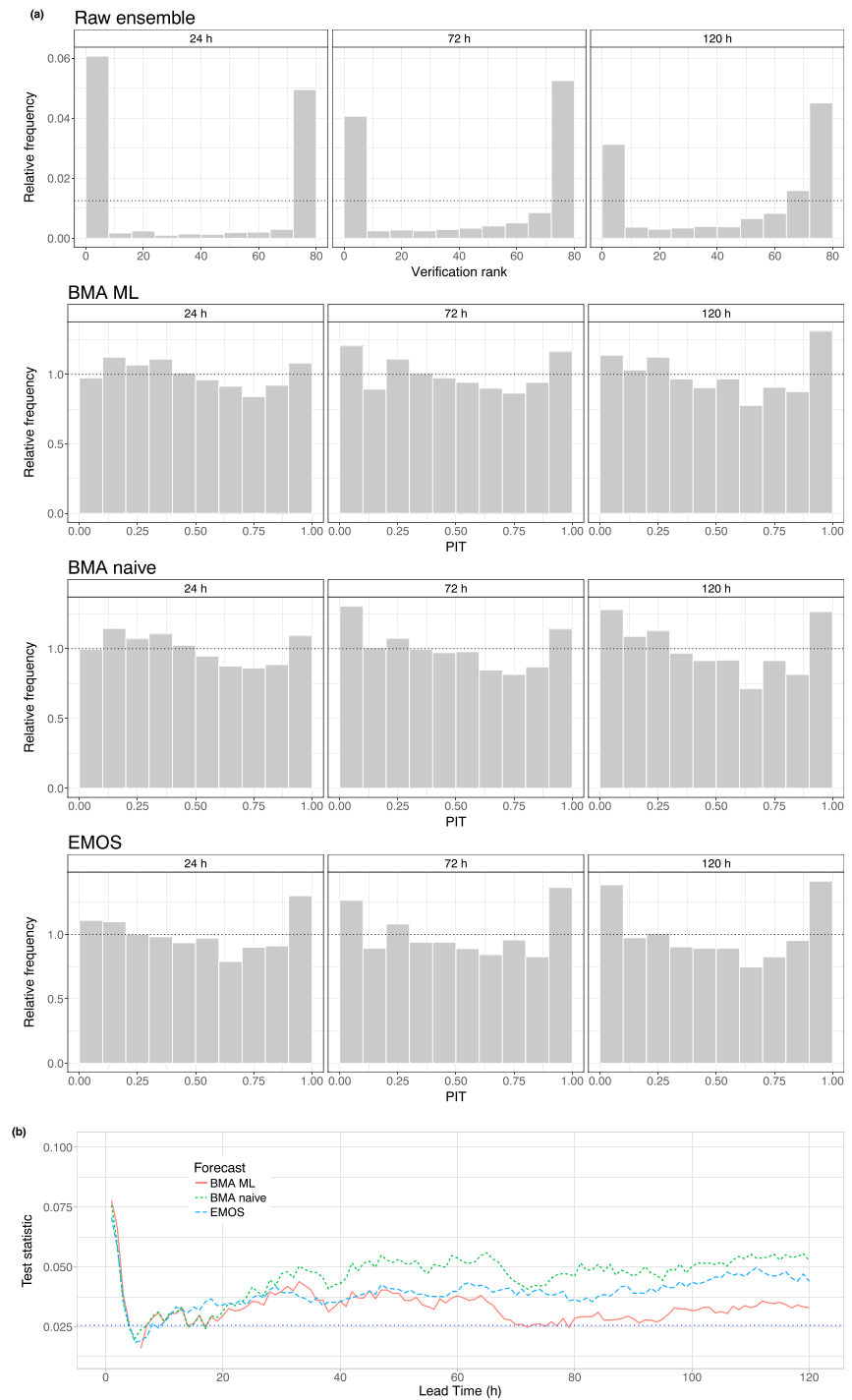


Figure 4. Verification rank histogram of the raw ensemble and probability integral transform histograms of the BMA and EMOS postprocessed forecasts for lead times 24, 72, and 120 hr (a). Values of the test statistic of Kolmogorov-Smirnov tests for uniformity of probability integral transform values (b). Smaller values indicate better fit; dotted horizontal line corresponds to 5% level of significance. BMA = Bayesian model averaging; EMOS = ensemble model output statistics.

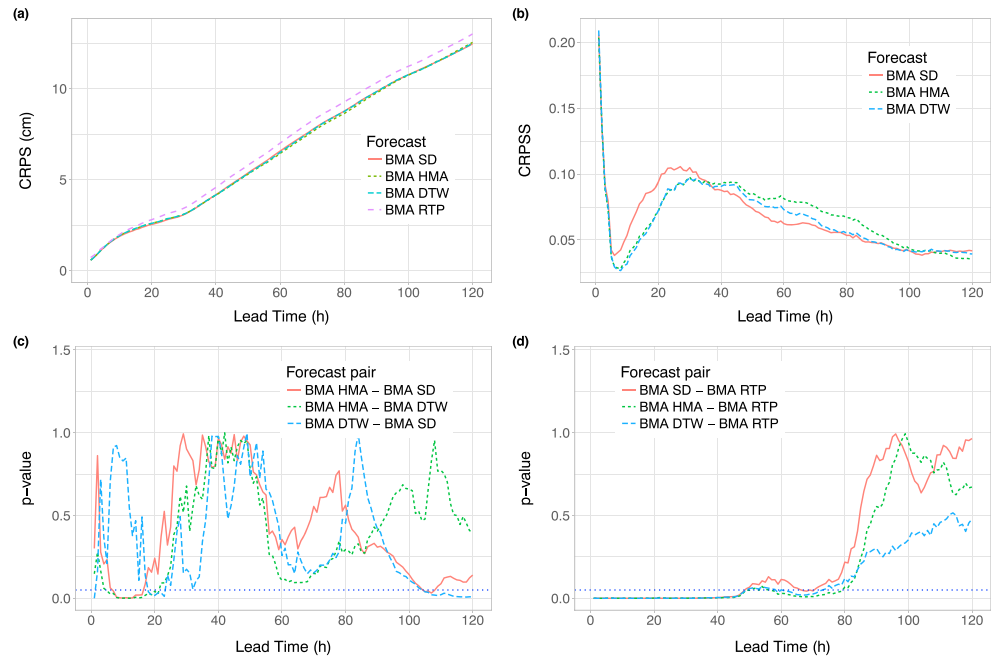


Figure 5. Mean CRPS values of BMA forecasts (a) and CRPSS with respect to the BMA with rolling training period (BMA RTP) (b). *p* values of Diebold-Mariano tests for equality of mean CRPS of the analog-based BMA approaches (c) and analog-based models compared to BMA RTP (d). Horizontal dotted lines of (c) and (d) indicate a 5% level of significance. CRPS = continuous ranked probability score; BMA = Bayesian model averaging; SD = series distance method; HMA = hydrograph matching algorithm; DTW = dynamic time warping.

this holds for the doubly truncated normal BMA approach, too. We compare the forecast skill of the pure ML BMA with training data chosen according to the SD, HMA, and DTW approaches described in section 2.6 (BMA SD, BMA HMA, and BMA DTW) and using a BMA RTP. To ensure comparability, all models are verified on data of the same 2,822 calendar days between 10 April 2008 and 31 December 2015 as in section 4.1.

Figure 5a shows the mean CRPS values of the four investigated BMA approaches as functions of the lead time, whereas in Figure 5b the corresponding skill scores with respect to the BMA RTP are plotted. According to the results of DM tests for equal predictive performance depicted in Figures 5c and 5d, there is no significant difference between the analog-based methods in terms of the mean CRPS, however, until hour 50 all of them significantly outperform the BMA RTP.

The use of analogs in model training results in more gain in terms of the MAE, as can clearly be observed in Figure 6a. MAE values of all three analog-based models are significantly lower than those of the BMA RTP for all lead times (DM test results are not reported). For lead times until 35 hr the BMA SD, for lead times between 40 and 104 hr the BMA HMA, whereas for longer lead times again the BMA SD yields the smallest MAE values. However, after 80 hr the differences in MAE between the analog-based methods are not significant; see Figure 6b.

In terms of coverage there is no big difference between the analog-based models and they are able to outperform BMA RTP only for very short lead times (Figure 6c). Further, according to Figure 6d, BMA SD, BMA HMA, and BMA DTW models result in slightly sharper central prediction intervals than BMA RTP.

Finally, the PIT histograms of the analog-based models plotted in Figure 7a are rather similar to the corresponding histograms in the second row of Figure 4a. These histograms, at least for lead times 24, 72, and 120 hr, indicate similar calibration of all BMA approaches. One can get more insight into the behavior of PIT values by comparing Figure 7b, showing the values of the test statistic of the Kolmogorov-Smirnov test for uniformity for the analog-based BMA approaches, with the corresponding line (BMA ML) of Figure 4b. The uniformity of PIT of BMA SD, BMA HMA, and BMA DTW can be accepted at a 5% level of significance for 25, 21, and 24 lead times, respectively, whereas for the BMA RTP (BMA ML in Figure 4b) there are just 9 such lead times.

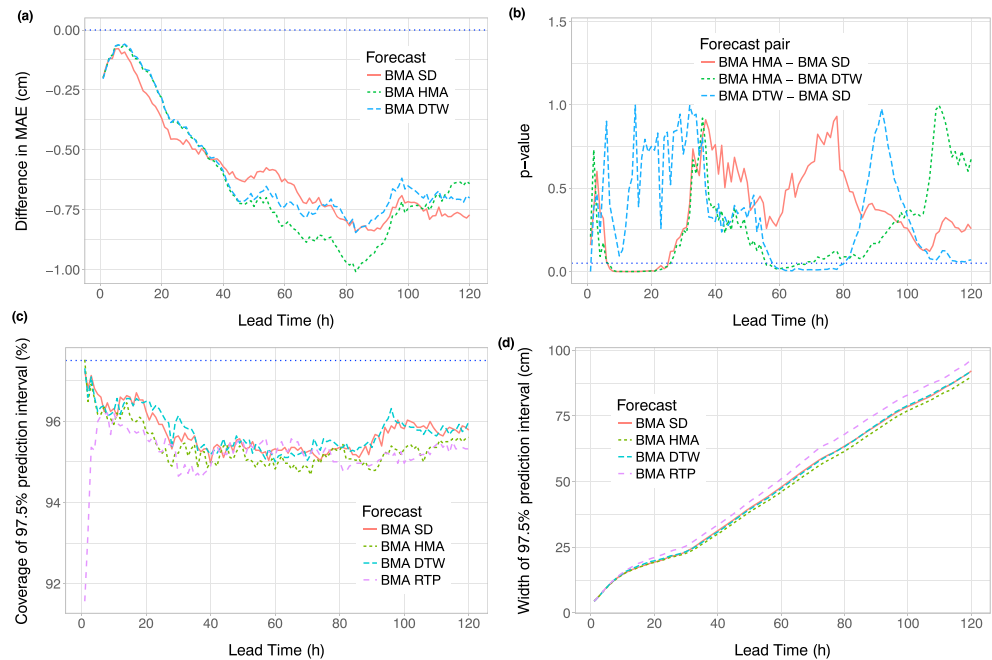


Figure 6. Top row: Difference in MAE values from the BMA RTP (a) and p values of Diebold-Mariano tests for equality of the various analog-based BMA approaches (b). Horizontal dotted lines indicate the reference BMA RTP (a) and a 5% level of significance (b). Bottom row: Coverage (c) and average width (d) of nominal 97.5% central prediction intervals. In panel (c) the ideal coverage is indicated by the horizontal dotted line. MAE = mean absolute error; BMA = Bayesian model averaging; SD = series distance method; HMA = hydrograph matching algorithm; DTW = dynamic time warping; RTP = rolling training period.

As reported above, the analog-based BMA approaches significantly outperform BMA RTP at shorter lead times up to about 40 hr. Up to about 80 hr this outperformance is still borderline significant, while this is not the case for longer lead times. This indicates that the gain by analog selection of training periods is restricted to shorter lead times. This is not surprising insofar as that a higher predictability at shorter lead times, which is usually the case compared to longer lead times, implies also that similarity in forecasts is more likely to be connected to similarity in the observations. Any analog-based approach assumes that such a connection can be established.

4.3. Analog-Based BMA Versus Analog-Based EMOS

According to the results of section 4.1, when the estimation of the postprocessing model coefficients is based on a simple RTP, pure ML BMA significantly outperforms EMOS in terms of the mean CRPS for almost all lead times (see Figures 2b and 2d) and in terms of MAE for very short (1–5 hr) and medium (21–75 hr) lead times (see Figures 3a and 3b).

The use of analog-based training data drastically changes the situation. In Figure 8a the CRPS values of the analog-based BMA models with respect to the corresponding analog-based EMOS approaches are plotted as functions of the lead time. Note that the skill scores vary in a very short interval between -0.0117 and 0.0084 . BMA SD, BMA HMA, and BMA DTW outperform EMOS SD, EMOS HMA, and EMOS DTW for only 60, 82, and 80 different lead times, respectively; however, for lead times 27, 28, 31, 43–46, and 71–117 hr none of the differences are significant (see Figure 8b). In this way, for SD, HMA, and DTW approaches there are 46, 47, and 41 different lead times, respectively, when BMA is significantly better in terms of mean CRPS and 11, 8, and 14 when EMOS performs better.

Further, Figure 8c shows that BMA SD, BMA HMA, and BMA DTW result in lower MAE values than their EMOS counterparts in 77, 82, and 77 cases, respectively. However, as one can observe in Figure 8d where the results of the corresponding DM tests are given, most of the differences are nonsignificant at a 5% level.

The above results indicate that analog-based BMA approaches are still slightly outperforming the corresponding analog-based EMOS models; however, the differences are far less pronounced than in the case of RTPs. Obviously, EMOS, which is based on a rather simple parametric predictive distribution, benefits a lot

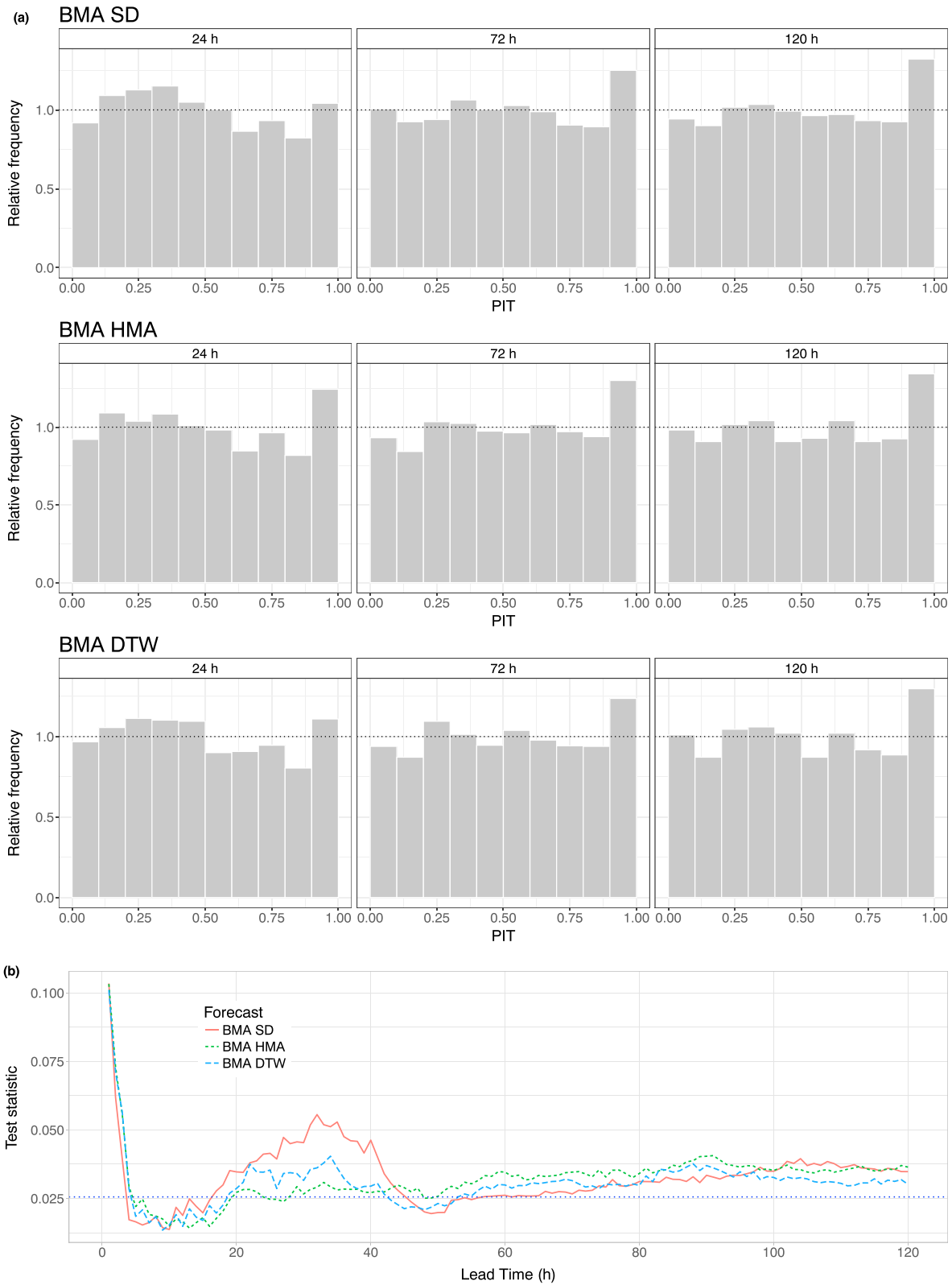


Figure 7. Probability integral transform histograms of the analog-based BMA postprocessed forecasts for lead times 24, 72, and 120 hr (a). Values of the test statistic of Kolmogorov-Smirnov tests for uniformity of probability integral transform values (b). Smaller values indicate better fit; dotted horizontal line corresponds to 5% level of significance. BMA = Bayesian model averaging; SD = series distance method; HMA = hydrograph matching algorithm; DTW = dynamic time warping.

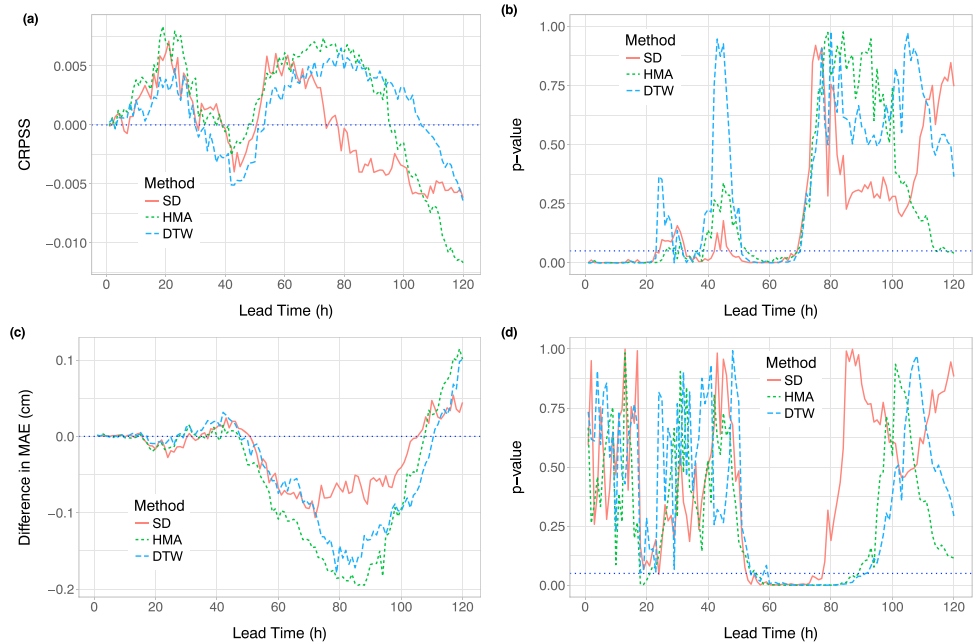


Figure 8. Top row: CRPSS values of the analog-based Bayesian model averaging (BMA) models with respect to the corresponding analog-based ensemble model output statistics (EMOS) approaches (a). p values of Diebold-Mariano tests for equality of mean continuous ranked probability score of BMA and EMOS postprocessed forecasts (b). Horizontal dotted line of panel (b) indicates a 5% level of significance. Bottom row: Difference in MAE values of various analog-based BMA models from the corresponding EMOS approaches (c) and p values of Diebold-Mariano tests for equality of MAE (d). Horizontal dotted lines indicate the reference EMOS model (c) and a 5% level of significance (d). CRPSS = continuous ranked probability skill score; SD = series distance method; HMA = hydrograph matching algorithm; DTW = dynamic time warping; MAE = mean absolute error.

from analog-based selection of training periods. This is most probably due to the fact that analog-based training periods allow for a stronger dependence between raw ensemble means and observations. This improves sharpness, at the cost of slightly deteriorating calibration (Hemri & Klein, 2017). In contrast, BMA leads to a quite flexible predictive distribution, which can have several modes in our case study. When using RTPs, this flexibility leads to a strong improvement in forecast skill compared to EMOS. Analog-based selection of training periods is also able to sharpen the BMA predictive distribution, but its effect on forecast skill is lower than in the case of EMOS. This may indicate that by using analog-based training periods, the postprocessed forecast skill is close to the theoretically achievable skill based on the forecast models at hand, and accordingly, the gain by using the more sophisticated BMA approach is rather marginal. Though out of scope of this paper, future work on postprocessing of water levels may benefit from identifying hydrological and statistical features like catchment size, climatic region, size of the data set available for training, or shape of the climatological distribution of water levels that may be able to predict which of the different postprocessing approaches is likely to perform best.

5. Conclusions

We introduce a new BMA model for calibrating Box-Cox transformed hydrological ensemble forecasts for water level, providing a predictive distribution which is a weighted mixture of doubly truncated normal distributions. The model with three different parameter estimation approaches is tested on the 79-member ensemble forecast of BfG for water level at gauge Kaub of river Rhine for 120 different lead times. For verification we use the CRPS of the probabilistic forecast distributions and the MAE of the corresponding median forecasts. Further, we analyze coverage and the average width of nominal central prediction intervals, which serves as a measure of sharpness. Furthermore, the forecast skill of the BMA model is compared to that of the recently introduced EMOS model of Hemri and Klein (2017) and the raw ensemble.

Based on the results of the presented case study, one can conclude that compared with the raw ensemble, postprocessing always improves the calibration of probabilistic and accuracy of point forecasts. With rolling

window training periods the BMA model outperforms the reference EMOS approach considerably. However, when using the analog-based selection of training periods from Hemri and Klein (2017) the gap in forecast skill decreases drastically leaving only a small advantage of BMA compared to EMOS. This indicates that using a more sophisticated postprocessing approach or the use of a smarter selection of training periods is fairly redundant. Accordingly, we recommend to use EMOS with analog-based training periods if a sufficiently long set of hydrological hindcasts is available and BMA otherwise.

Further, following the ideas of Hemri et al. (2015) and Bellier et al. (2018), one can combine the BMA calibrated forecasts corresponding to different locations and lead times either into temporally or both spatially and temporally coherent multivariate predictions. This can be done with the help of modern techniques such as the ensemble copula coupling (Scheffzik et al., 2013) or the Gaussian copula approach (Pinson & Girard, 2012). However, these studies are beyond the scope of the present paper.

Acknowledgments

The data on which this paper is based can be downloaded for noncommercial research purposes from HydroShare: Hemri (2019), <http://www.hydroshare.org/resource/0afabbae43d44472804b6c03c643dc10>. The authors are grateful to the German Federal Office of Hydrology (BfG), and in particular Bastian Klein, for providing the data and valuable inputs. Example R scripts for applying doubly truncated BMA are available as supporting information. The authors also thank Tilmann Gneiting and Sebastian Lerch for valuable comments and suggestions. Essential part of this work was made during the visit of Sándor Baran at the Heidelberg Institute of Theoretical Studies in the framework of the visiting scientist program. Sándor Baran was also supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and the National Research, Development and Innovation Office under grant NN125679. He is grateful to Sebastian Lerch for his useful suggestions, remarks, and help with the EMOS code. Sándor Baran and Mehrez El Ayari were supported by the EFOP-3.6.3-VEKOP-16-2017-00002 project. The project was cofinanced by the Hungarian Government and the European Social Fund. Sándor Baran and Stephan Hemri also acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) grant MO 3394/1-1 "Statistical postprocessing of ensemble forecasts for various weather quantities". Last but not least the authors are very grateful to the Associate Editor and the three Reviewers for their valuable comments.

References

- Baran, S. (2014). Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics & Data Analysis*, *75*, 227–238.
- Baran, S., & Lerch, S. (2015). Log-normal distribution based EMOS models for probabilistic wind speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, *141*, 2289–2299.
- Baran, S., & Lerch, S. (2016). Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, *27*, 116–130.
- Baran, S., & Nemoda, D. (2016). Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, *27*, 280–292.
- Bellier, J., Zin, I., & Bontron, G. (2018). Generating coherent ensemble forecasts after hydrological postprocessing: adaptations of ECC-based methods. *Water Resources Research*, *54*, 5741–5762. <https://doi.org/10.1029/2018WR022601>
- Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., et al. (2010). The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, *91*, 1059–1072.
- Buizza, R. (2018). Ensemble forecasting and the need for calibration. In S. Vannitsem, D. S. Wilks, & J. W. Messner (Eds.), *Statistical postprocessing of ensemble forecasts* (pp. 15–48). Amsterdam: Elsevier.
- Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M., & Zhu, Y. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, *133*, 1076–1097.
- Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*, 613–626.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–39.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.
- Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*, 1371–1386.
- Ehret, U., & Zehe, E. (2011). Series distance—An intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. *Hydrology and Earth System Sciences*, *15*, 877–896.
- Ewen, J. (2011). Hydrograph matching method for measuring model performance. *Journal of Hydrology*, *408*, 178–187.
- Fraley, C., Raftery, A. E., & Gneiting, T. (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, *138*, 190–202.
- Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, *69*, 243–268.
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, *102*, 359–378.
- Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*, 1098–1118.
- Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, *29*, 411–422.
- Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., et al. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, *94*, 1553–1565.
- Hemri, S. (2019). Gauge level forecasts and observations for gauge Kaub (river Rhine) used in Baran et al. (2019). HydroShare. <http://www.hydroshare.org/resource/0afabbae43d44472804b6c03c643dc10>
- Hemri, S., Fundel, M., & Zappa, M. (2013). Simultaneous calibration of ensemble river flow predictions over an entire range of lead times. *Water Resources Research*, *49*, 6744–6755. <https://doi.org/10.1002/wrcr.20542>
- Hemri, S., & Klein, B. (2017). Analog based post-processing of navigation-related hydrological ensemble forecasts. *Water Resources Research*, *53*, 9059–9077. <https://doi.org/10.1002/2017WR020684>
- Hemri, S., Lisniak, D., & Klein, B. (2014). Ermittlung probabilistischer abflussvorhersagen unter berücksichtigung zensierter daten. *HyWa*, *58*, 84–94.
- Hemri, S., Lisniak, D., & Klein, B. (2015). Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, *51*, 7436–7451. <https://doi.org/10.1002/2014WR016473>
- Iversen, T., Deckmin, A., Santos, C., Sattler, K., Bremnes, J. B., Feddersen, H., & Frogner, I.-L. (2011). Evaluation of 'GLAMEPS'—A proposed multimodel EPS for short range forecasting. *Tellus A*, *63*, 513–530.
- Jordan, A., Krüger, F., & Lerch, S. (2017). Evaluating probabilistic forecasts with the R package scoringRules. arxiv 1709.04743. <https://arxiv.org/abs/1709.04743>, [Accessed on 17 January 2019].
- Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, *35*, 2739–2750.
- Lee, G., & Scott, C. (2012). EM algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, *56*, 2816–2829.

- Lerch, S., & Baran, S. (2017). Similarity-based semi-local estimation of EMOS models. *Journal of the Royal Statistical Society, Series C*, *66*, 29–51.
- Lerch, S., & Thorarinsdottir, T. L. (2013). Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, *65*, 21206.
- Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*, 3515–3539.
- Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*, 272–288.
- McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.
- Molteni, F., Buizza, R., & Palmer, T. N. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, *122*, 73–119.
- Montani, A., Cesari, D., Marsigli, C., & Paccagnella, T. (2011). Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: Main achievements and open challenges. *Tellus A*, *63*, 605–624.
- Murphy, A. H. (1973). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, *12*, 215–223.
- Park, Y.-Y., Buizza, R., & Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, *134*, 2029–2050.
- Pinson, P., & Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, *96*, 12–20.
- Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*, 1155–1174.
- Ruiz, J. J., & Saulo, C. (2012). How sensitive are probabilistic precipitation forecasts to the choice of calibration algorithms and the ensemble generation method? Part I: Sensitivity to calibration methods. *Meteorological Applications*, *19*, 302–313.
- Sakoe, H., & Chiba, S. (1978). Dynamic-programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*, 43–49.
- Scheffzik, R., Thorarinsdottir, T. L., & Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, *28*, 616–640.
- Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1086–1096.
- Scheuerer, M., & Hamill, T. M. (2015). Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, *143*, 4578–4596.
- Schmeits, M. J., & Kok, K. J. (2010). A comparison between raw ensemble output, (modified) Bayesian model averaging and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Monthly Weather Review*, *138*, 4199–4211.
- Seibert, S. P., Ehret, U., & Zehe, E. (2016). Disentangling timing and amplitude errors in streamflow simulations. *Hydrology and Earth System Sciences*, *20*, 3745–3763.
- Sloughter, J. M., Gneiting, T., & Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, *105*, 25–37.
- Sloughter, J. M., Raftery, A. E., Gneiting, T., & Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, *135*, 3209–3220.
- Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., et al. (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, *97*, 49–67.
- Thorarinsdottir, T. L., & Gneiting, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society. Series A*, *173*, 371–388.
- Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, *6*, 123–137.
- Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (3rd ed.). Amsterdam: Elsevier.
- Wilks, D. S. (2018). Univariate ensemble forecasting. In S. Vannitsem, D. S. Wilks, & J. W. Messner (Eds.), *Statistical postprocessing of ensemble forecasts* (pp. 49–89): Elsevier.
- Williams, R. M., Ferro, C. A. T., & Kwasniok, F. (2014). A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1112–1120.