Simultaneous calibration of ensemble river flow predictions over an entire range of lead times

S. Hemri,^{1,2} F. Fundel,^{1,3} and M. Zappa¹

Received 13 June 2013; revised 1 September 2013; accepted 18 September 2013; published 17 October 2013.

[1] Probabilistic estimates of future water levels and river discharge are usually simulated with hydrologic models using ensemble weather forecasts as main inputs. As hydrologic models are imperfect and the meteorological ensembles tend to be biased and underdispersed, the ensemble forecasts for river runoff typically are biased and underdispersed, too. Thus, in order to achieve both reliable and sharp predictions statistical postprocessing is required. In this work Bayesian model averaging (BMA) is applied to statistically postprocess ensemble runoff raw forecasts for a catchment in Switzerland, at lead times ranging from 1 to 240 h. The raw forecasts have been obtained using deterministic and ensemble forcing meteorological models with different forecast lead time ranges. First, BMA is applied based on mixtures of univariate normal distributions, subject to the assumption of independence between distinct lead times. Then, the independence assumption is relaxed in order to estimate multivariate runoff forecasts over the entire range of lead times simultaneously, based on a BMA version that uses multivariate normal distributions. Since river runoff is a highly skewed variable, Box-Cox transformations are applied in order to achieve approximate normality. Both univariate and multivariate BMA approaches are able to generate well calibrated probabilistic forecasts that are considerably sharper than climatological forecasts. Additionally, multivariate BMA provides a promising approach for incorporating temporal dependencies into the postprocessed forecasts. Its major advantage against univariate BMA is an increase in reliability when the forecast system is changing due to model availability.

Citation: Hemri, S., F. Fundel, and M. Zappa (2013), Simultaneous calibration of ensemble river flow predictions over an entire range of lead times, *Water Resour. Res.*, 49, 6744–6755, doi:10.1002/wrcr.20542.

1. Introduction

[2] Bayesian model averaging (BMA) has been introduced by *Raftery et al.* [2005] for postprocessing of meteorological forecasts. It generates a single forecast PDF for a quantity like temperature or rainfall by combining several forecasts from different models. The goal of statistical postprocessing is to obtain calibrated and sharp predictive continuous PDF's of the quantity to be forecast given uncalibrated raw predictive ensembles [*Raftery et al.*, 2005; *Gneiting et al.*, 2007].

[3] In hydrology probabilistic runoff forecasts are typically generated by running a hydrologic model several

times using atmospheric variables from different meteorological models as forcing. In recent years BMA has been used increasingly for the combination of multiple rainfall-runoff models [*Ajami et al.*, 2007; *Duan et al.*, 2007; *Vrugt and Robinson*, 2007; *Diks and Vrugt*, 2010; *Parrish et al.*, 2012]. Besides the work of *Bogner et al.* [2013] there has so far been no similar BMA study accounting for the correlation structure between different lead times in the field of hydrologic forecasting. Note that *Yuan and Wood* [2012] successfully applied a different Bayesian postprocessing approach to probabilistic hydrologic forecasts. On the basis of *Berrocal et al.* [2007], combining BMA with the geo-statistical output perturbation (GOP) method [*Gel et al.*, 2004] in order to achieve spatially correlated BMA forecasts, we will use this method to account for the correlations among lead times. That is, space is replaced by lead times. For simplicity this BMA method is called multivariate BMA from now on.

[4] While multivariate BMA is supposed to improve probabilistic forecasts in terms of multivariate verification measures like the energy score *Gneiting et al.* [2008], it seems also to slightly improve the marginal predictive distributions at single lead times. *Pinson et al.* [2008] and *Pinson and Girard* [2012] introduced a similar method used for postprocessing of probabilistic forecasts of wind power. Like multivariate BMA, it resorts to the multivariate

Additional supporting information may be found in the online version of this article.

¹Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland.

²Now at Institute for Applied Mathematics, Heidelberg University, Heidelberg, Germany.

³Deutscher Wetterdienst, Offenbach, Germany.

Corresponding author: S. Hemri, Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 294, DE-69120 Heidelberg, Germany. (stephanhemri@gmx.ch)

©2013. American Geophysical Union. All Rights Reserved.
0043-1397/13/10.1002/wrcr.20542

Gaussian distribution and estimates covariance matrices from the series of prediction errors.

[5] For this study, the meteorological forcing consists of a mixture of different models. On the one hand, both global and regional models are used. On the other hand, the models can be divided into deterministic forecasts and ensemble prediction systems (EPS). Each member of an EPS stands for a model run with different initial states, boundary conditions or model formulations [Palmer, 2000]. Ideally, these model runs are exchangeable, i.e., they are supposed to be statistically identical. One of the main goals of this study is to combine forecasts stemming from such a bunch of different models into a sound probabilistic density forecast.

[6] In the following, we apply the methods introduced above to river discharge forecasts for Thur River in Eastern Switzerland. Hourly runoff forecasts forced by a mixture of deterministic models and EPS's that cover different ranges of lead times are combined to one single multivariate PDF using multivariate BMA. Thus, multivariate BMA features modifications that allow on the one hand to account for temporally correlated prediction errors and on the other hand to produce time consistent forecasts though the different forcings cover unequal ranges of lead times. Note that any analysis has been performed using the statistical software package R [R Development Core Team, 2011].

[7] Section 2 presents the data set used in this study along with an example of a particular forecast. BMA, and especially multivariate BMA, are explained in detail in section 3. The forecasts obtained by univariate and multivariate BMA over a period of two and a half years are verified in section 4. Lastly, we discuss the results shortly in section 5 and make some concluding remarks in section 6.

2. Data Set With Example Forecast

[8] In this study, we consider Thur River, which drains a medium-sized subcatchment of the Swiss Rhine. It covers an area of 1696 km². Ranging from 356 at the gauging station in Andelfingen to 2503 at the highest mountain peak its mean altitude is 769 m a.s.l. Due to its cool climate, runoff generation from autumn to spring are influenced by snow dynamics. As a result of snow melting monthly discharges are usually maximal in April, though precipitation amounts are highest in summer [Fundel and Zappa, 2011; Fundel et al., 2013]. Any observed runoff values have been provided by the Swiss Federal Office for the Environment (FOEN).

[9] The discharge forecasts consist of a hydrometeorological ensemble prediction system (HEPS). HEPS are obtained by running a hydrologic model several times with different meteorological forcing [Addor et al., 2011; Zappa et al., 2011]. The runoff predictions, which serve as raw forecasts for BMA in this study, have been generated by the hydrologic model "Precipitation-Runoff-Evapotranspiration HRU Model" (PREVAH) [Viviroli et al., 2007a, 2007b, 2009]. Reforecasts, i.e., historical forecasts with a consistent model version, from several models of different size have been available as meteorological forcing. Table 1 gives an overview over the 69 model members that provide forecasts on an hourly basis for different ranges of lead times. For more details on the particular models we refer to Molteni et al.

Table 1. Features of the Meteorological Input Models

Name	Number of Models	Lead times (h)	Spatial Resolution ~ (km)
COSMO-2	1	1–24	2.2
COSMO-7	1	1–72	6.6
COSMO-LEPS	16	1–120	10
ECMWF	51	1–240	50

[1996] for the ECMWF-EPS, to Marsigli et al. [2004, 2005]; Tibaldi et al. [2006]; Montani et al. [2011] for COSMO-LEPS and to [Consortium for Small-scale Modeling, 2011] for COSMO-2 and COSMO-7. PREVAH is run on a daily basis, initialized at 00:00 UTC, in order to forecast hourly discharges. Each run corresponds to an ensemble member of an EPS or to a deterministic model. The period from 17 July 2007 to 17 November 2009 serves as verification period. For any verification day x , a calibration period of 30 days running from day $x - 40$ to day $x - 11$ has been used for model fitting. This guarantees no overlapping of calibration and verification period for lead times up to 240 h.

[10] Now an example of a BMA forecast for lead times 1–240 h is given here. Figure 1 shows forecast trajectories of the raw ensemble, and both the univariate and the multivariate BMA models. The predictions have been initialized at 00:00 UTC on 18 April 2008. Note that COSMO-2, COSMO-7, and COSMO-LEPS drop out at lead times 24 h, 72 h, and 120 h, respectively. In case of the BMA forecasts the trajectories are obtained by random sampling of 69 different scenarios from the BMA distributions. The number of 69 samples corresponds to the number of raw ensemble members available at the first 24 lead times. For lead times up to 120 h the forecast trajectories of the raw ensemble, the univariate and the multivariate BMA forecast are plausible, if we verify them with the observations. Dropping out of COSMO-LEPS strongly affects the raw ensemble and the univariate BMA forecast. As they now rely only on the ECMWF members, which forecast low runoff, they completely miss the observed runoff pattern. However, the multivariate approach still relies on the COSMO-LEPS members through the covariances. For additional illustration predicted quantiles of raw ensemble and both BMA forecasts initialized at fifteen different days, of which five represent high flow events, five low flow events, and five are chosen randomly, are available as supporting information for this report.

3. Methods

3.1. Standard, Univariate BMA

[11] Raftery et al. [2005] introduced a Bayesian model averaging (BMA) approach that allows combining dynamic models. In its simplest form BMA combines forecasts $r_1, \dots, r_K$ from K different models to one mixture forecast probability density function (PDF) given by

$$p(y|r_1, \dots, r_K) = \sum_{k=1}^K w_k g_k(y|r_k), \quad (1)$$

where w_k is the weight of model k and g_k is the conditional PDF of y , e.g., the forecast runoff, given that model k is

Figure 1. Forecast trajectories of the raw ensemble, univariate and multivariate BMA forecasts initialized at 00:00 UTC on 18 April 2008 against lead time. BMA trajectories correspond to 69 random samples from the forecast distributions.

best. Note that this form accounts neither for bias correction nor for transformation of r_k . Additionally, lead times are implicitly assumed to be completely independent.

[12] As parts of the raw forecasts stem from EPS's BMA needs to deal with exchangeable ensemble members. *Fraley et al.* [2010] provide a BMA variant which guarantees by model design that parameters of exchangeable ensemble members are forced to be equal. Hence, equation (1) has to be rewritten as

$$p(y|r_1, \dots, r_K) = \sum_{i=1}^I \sum_{j=1}^{m_i} w_i g_i(y|r_{i,j}), \quad (2)$$

where the index j denotes a specific member of ensemble model i .

[13] Stream flow data are undoubtedly non-Gaussian. In order to be able to resort to BMA methods relying on Gaus-

sian distributions, both the observations and the raw ensemble predictions have to be transformed such that they are approximately normal. *Duan et al.* [2007] used the Box-Cox transformation [*Box and Cox*, 1964] to make runoff data approximately normal. The Box-Cox transformation is given by:

$$h(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(x) & \text{if } \lambda = 0, \end{cases} \quad (3)$$

where x are the original data and λ is the Box-Cox coefficient. For this study the parameter $\hat{\lambda}$ has been estimated by maximizing the likelihood of the Box-Cox transformed climatological observations using a normal distribution centered at the mean of the transformed observations using the

Figure 2. Quantile-quantile plot against the theoretical quantiles of a normal distribution with appropriate mean and variance verifying normality of the Box-Cox transformed climatology of Thur River. The Box-Cox parameter estimate is $\hat{\lambda} = 0.16$.

R-function `boxcox` of the package `Mass`. The term climatology denotes here the collection of the observed runoff values from January 1974 to Mai 2007. This period does not overlap with the verification and the associated training periods mentioned in section 2. Hence, the estimate of $\hat{\lambda}$ is independent from any model parameter values that were used to simulate discharge dynamics. Figure 2 shows a normal QQ-plot for the Box-Cox transformed climatology of Thur River. In the following we will use the Box-Cox parameter estimate $\hat{\lambda} = -0.16$ for both, observations and forecasts.

[14] The Box-Cox transformed raw forecasts $h(r_{i,j})$ still need to be bias corrected. This may be done by linear regression of the Box-Cox transformed observations on the Box-Cox transformed forecasts of the training period as in *Raftery et al.* [2005]. The complete model in the lead time independent case with Box-Cox transformed and bias corrected data is then:

$$p(\varphi | \{r_{i,j}\}_{i=1, \dots, I, j=1, \dots, m_i}) = \sum_{i=1}^I \sum_{j=1}^{m_i} w_i \mathcal{N}(\varphi | a_i + b_i h(r_{i,j}), \sigma_i^2), \quad (4)$$

where φ denotes the Box-Cox transformed observations. a_i and b_i are regression parameters. In this study b_i is set to 1 as we consider only simple additive bias correction. Tests with both regression methods, standard, and robust, performed worse. The weights w_i and the variances σ_i^2 are estimated by maximum likelihood using the Expectation Maximization (EM) algorithm [*Dempster et al.*, 1977; *McLachlan and Krishnan*, 1997; *Duan et al.*, 2007; *Fraley et al.*, 2010].

3.2. Multivariate BMA

[15] In traditional BMA, different lead times are considered as completely independent from each other. Hence,

ensemble models not covering the same range of lead times do not cause any problems. If a model drops out at a particular lead time, one just estimates the BMA model without using it anymore. Now the independence assumption is relaxed by replacing univariate BMA by a multivariate BMA method that is based on multivariate normal distributions. *Berrocal et al.* [2007] combined BMA with the geostatistical output perturbation (GOP) method [*Gel et al.*, 2004] in order to generate calibrated probabilistic spatial forecasts of whole weather fields. In the present multivariate approach space is replaced by lead time. That is, the goal is to estimate a BMA model that is able to generate calibrated probabilistic forecasts over an entire range of lead times simultaneously.

3.2.1. BMA Fit

[16] In a first step a BMA model is fitted as described in section 3.1. But now, instead of estimating the model parameters for each lead time separately, lead times are combined into blocks of the same model configuration. Within a block the number of available ensembles is constant. Additionally, the variance is assumed to be constant within a block, such that all lead times can be treated equally. Only $G < L$ BMA models are now estimated per verification day, where G is the number of blocks and L denotes the number of lead times considered. All equations remain the same, but Tr , i.e., the set of dates and lead times in the training period, has now another meaning. It is now the set of all forecasting lead times in a particular block and for all dates in the training period.

3.2.2. GOP Fit

[17] In order to connect the GOP method with the BMA fits, the joint transformed forecast needs to be written as

$$\varphi | \mathbf{r}_k \sim \mathcal{N}(a_k + b_k h(\mathbf{r}_k), \Sigma_k^0), \quad (5)$$

where $\varphi | \mathbf{r}_k$ is the conditional distribution of Box-Cox transformed runoffs for a block of lead times given the vector of forecasts $\mathbf{r}_k$ from model member k for the same block of lead times. a_k and b_k are bias correction coefficients, $h(\mathbf{r}_k)$ are the Box-Cox transformed raw forecasts from model k . Σ_k^0 is up to a constant given by Σ_k , the covariance matrix defining the error structure of the forecasts. In the following this covariance matrix needs to be estimated from the training data. As mentioned above, this can be done in the same way as in *Gel et al.* [2004] and *Berrocal et al.* [2007]. Their GOP model is stationary, isotropic and based on the assumption of an exponential variogram. Note that the more general Matérn variogram may be an alternative. Σ_k can thus be estimated by:

$$\sigma_{i,j} = \rho_k^2 \delta_{i,j} + \tau_k^2 e^{-\frac{d_{i,j}}{\tau_k}}, \quad (6)$$

where $\sigma_{i,j}$ is the (i, j) th element of Σ_k , $\delta_{i,j}$ is 1 if $i = j$ and 0 otherwise, and $d_{i,j}$ is the lag between lead time i and lead time j . The coefficients ρ_k^2 , τ_k^2 , and τ_k can be interpreted as the coefficients of a geostatistical model [*Chilès and Delfiner*, 1999; *Cressie*, 1993]. Hence, the following exponential variogram can be fitted:

$$\gamma_k(d) = \rho_k^2 + \tau_k^2 \left(1 - e^{-\frac{d}{\tau_k}}\right), \quad (7)$$

where d is the lag between two lead times.

[18] The procedure for estimation of those parameters is explained in detail in *Berrocal et al.* [2007]. As we use a slightly adapted version, accounting for exchangeable ensemble members, a short sketch of the procedure is given here:

[19] 1. Obtain the bias corrected forecasts, where $\mathbf{f}_{i,j,t}$ denotes the vector of bias corrected forecasts from ensemble i and member j on training day t of the training period Tr . The elements of $\mathbf{f}_{i,j,t}$ stand for the different lead times. $\mathbf{f}_{i,j,t}$ can easily be partitioned according to the block of lead times. The bias correction parameters depend now also on the blocks g , and hence, $a_{i,g}$ and $b_{i,g}$ are calculated for each ensemble using all lead times within block g . In this study, we have set $b_{i,g} = 1$.

[20] 2. Then the empirical variogram needs to be estimated. To this end, calculate first the matrix of forecast errors $\mathbf{err}^{i,t}$ for each model i and each training day t , where the rows of $\mathbf{err}^{i,t}$ indicate lead times and the columns indicate ensemble members. In case of a deterministic model, $\mathbf{err}^{i,t}$ is a vector. From the forecast errors the empirical variogram $\hat{\gamma}_i(d)$ for ensemble i can be calculated using:

$$\frac{1}{2T(L-d)m_i} \sum_{t \in Tr} \sum_{l=1}^{L-d} \sum_{j=1}^{m_i} \left(\mathbf{err}_{l,j}^{i,t} - \mathbf{err}_{l+d,j}^{i,t} \right)^2, \quad (8)$$

for d in $1, \dots, L-1$, and where L is the largest lead time of a given model, m_i is the number of members of model i and T is the number of days in the training period Tr . Then the optimal exponential variogram parameters, now denoted by ρ_i^2 , τ_i^2 , and r_i , are estimated. This is done numerically by minimizing the following expression:

$$U(\rho_i^2, \tau_i^2, r_i) = \sum_{d=1}^{L-1} (L-d) \left\{ \frac{\hat{\gamma}_i(d) - \left(\rho_i^2 + \tau_i^2 \left(1 - e^{-\frac{d}{r_i}} \right) \right)}{\rho_i^2 + \tau_i^2 \left(1 - e^{-\frac{d}{r_i}} \right)} \right\}^2. \quad (9)$$

[21] 3. Then Σ_i can easily be obtained using equation (6).

[22] 4. Σ_i does not account for the fact that the BMA variances σ_i^2 are conditional on the assumption that model i is the best. Hence, corrected covariance matrices $\Sigma_i^0 = \alpha_i \Sigma_i$ have to be calculated, where $\alpha_i = \frac{\sigma_i^2}{\rho_i^2 + \tau_i^2}$ is the deflation factor.

[23] As an illustration let us take a closer look at the predictions for Thur River initialized at 00:00 UTC on 18 April 2008. As explained above BMA models over entire blocks of lead times, i.e., 1–24, 25–72, 73–120, and 121–240 h in case of the Thur data set, are estimated first. In a second step, the correlation structure of the errors among forecasts for different lead times within each model ensemble needs to be estimated. This is done by fitting an exponential variogram to the empirical semivariates of the bias-corrected raw forecast errors, as shown in Figure 3. Obviously, the exponential variograms fit quite well to the empirical variograms for that particular training period. Though not shown here, variogram fits for other training periods are comparable.

[24] In a next step the multivariate forecast density distributions have to be constructed by simulation. The challenge here is to combine the separate block BMA estimates across block boundaries while accounting for dependencies among different lead times. This problem is solved by

Figure 3. Empirical variograms and corresponding exponential fits for COSMO-2, COSMO-7, COSMO-LEPS, and ECMWF forecast errors calculated with data from 9 March to 7 April 2008.

using the property of the multivariate normal distribution that subvectors of the multivariate normal given the remaining elements are again multivariate normal with specified mean and covariance matrix. This property is explained in detail in, e.g., Krzanowski [2000]. Now a description of how the property is used in order to combine the BMA fits is given here. The set $\mathcal{F}_t = \{\mathbf{f}_{1,1}, \dots, \mathbf{f}_{1,m_1}, \dots, \mathbf{f}_{L,1}, \dots, \mathbf{f}_{L,m_L}\}$ denotes the set of all vectors of bias corrected forecasts for verification day t . The random vector of forecast Box-Cox transformed runoffs over the entire range of lead times, $\boldsymbol{\varphi}_{1:L}$, follows a L -variate normal distribution. The vector $\boldsymbol{\varphi}_{1:L}$ can now be partitioned as follows:

$$\boldsymbol{\varphi}_{1:L} = \begin{pmatrix} \boldsymbol{\varphi}_{1:l_1} \\ \boldsymbol{\varphi}_{l_1+1:l_2} \\ \vdots \\ \boldsymbol{\varphi}_{l_{G-1}+1:L} \end{pmatrix},$$

where l_g denotes the largest lead time of block g , and hence $L = l_G$. The distribution of $\boldsymbol{\varphi}_{1:L}$ can now be written as:

$$\begin{aligned} P[\boldsymbol{\varphi}_{1:L} | \mathcal{F}_t] &= \overbrace{P[\boldsymbol{\varphi}_{1:l_1} | \mathcal{F}_t]}^{\text{block1}} \cdot \overbrace{P[\boldsymbol{\varphi}_{l_1+1:l_2} | \boldsymbol{\varphi}_{1:l_1}, \mathcal{F}_t]}^{\text{block2}} \cdots \\ &\quad \times \overbrace{P[\boldsymbol{\varphi}_{l_{G-1}+1:L} | \boldsymbol{\varphi}_{1:l_{G-1}}, \mathcal{F}_t]}^{\text{blockG}}. \end{aligned}$$

[25] In addition the bias corrected forecasts and the covariance matrices need to be partitioned accordingly. The forecast vector $\mathbf{f}_{i,j}$ of an ensemble covering lead times up to block g is partitioned by:

$$\mathbf{f}_{1:l_g} = \begin{pmatrix} \mathbf{f}_{1:l_1} \\ \mathbf{f}_{l_1+1:l_2} \\ \vdots \\ \mathbf{f}_{l_{g-1}+1:l_g} \end{pmatrix},$$

where $\mathbf{f}_{i,j} = \mathbf{f}$ for better readability. The covariance matrix can be partitioned as follows:

$$\boldsymbol{\Sigma}_i^0 = \begin{pmatrix} \boldsymbol{\Sigma}_{i,11}^0 & \boldsymbol{\Sigma}_{i,12}^0 & \cdots & \boldsymbol{\Sigma}_{i,1g}^0 \\ \boldsymbol{\Sigma}_{i,21}^0 & \boldsymbol{\Sigma}_{i,22}^0 & \cdots & \boldsymbol{\Sigma}_{i,2g}^0 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{i,g1}^0 & \boldsymbol{\Sigma}_{i,g2}^0 & \cdots & \boldsymbol{\Sigma}_{i,gg}^0 \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{i,11}^0$ denotes the covariances among lead times from 1 to l_1 , $\boldsymbol{\Sigma}_{i,22}^0$ among lead times $l_1 + 1$ to l_2 , and $\boldsymbol{\Sigma}_{i,gg}^0$ among lead times $l_{g-1} + 1$ to l_g . Ensembles covering only block 1 do not need any partitioning.

[26] Now the resulting forecast distribution for any verification day, t (index t omitted in the following), can be simulated by iterating the following steps:

[27] 1. Select a member (i, j) by sampling with weights w_i from the BMA fit for block 1.

[28] 2. Obtain a simulated realization $\tilde{\boldsymbol{\varphi}}_{1:l_1}$ from the distribution $\mathcal{N}_{l_1}(\mathbf{f}_{i,j,1:l_1}, \boldsymbol{\Sigma}_{i,11}^0)$

[29] 3. Select a member (i, j) by sampling with weights w_i from the BMA fit for block 2. Note that the number of available ensemble models usually decreases from block to

block and that the weights w_i are now different from those in step 1.

[30] 4. Obtain $\tilde{\boldsymbol{\varphi}}_{l_1+1:l_2}$ by simulating from:

$$\mathcal{N}_{l_2-l_1} \left(\mathbf{f}_{i,j,l_1+1:l_2} + \boldsymbol{\Sigma}_{i,21}^0 \boldsymbol{\Sigma}_{i,11}^{0^{-1}} (\tilde{\boldsymbol{\varphi}}_{1:l_1} - \mathbf{f}_{i,j,1:l_1}), \boldsymbol{\Sigma}_{i,22}^0 - \boldsymbol{\Sigma}_{i,21}^0 \boldsymbol{\Sigma}_{i,11}^{0^{-1}} \boldsymbol{\Sigma}_{i,12}^0 \right). \quad (10)$$

[31] 5. For blocks $g \in \{3, \dots, G\}$ do the following:

[32] 1. Select a member (i, j) by sampling with weights w_i from the BMA fit for block g .

[33] 2. Obtain $\tilde{\boldsymbol{\varphi}}_{l_{g-1}+1:l_g}$ by simulating from:

$$\mathcal{N}_{l_g-l_{g-1}} \left(\mathbf{f}_{i,j,l_{g-1}+1:l_g} + \mathbf{A}, \mathbf{B} \right), \quad (11)$$

where

$$\begin{aligned} \mathbf{A} &= \left(\boldsymbol{\Sigma}_{i,g1}^0 \quad \boldsymbol{\Sigma}_{i,g2}^0 \quad \cdots \quad \boldsymbol{\Sigma}_{i,g(g-1)}^0 \right) \mathbf{C}^{-1} \left(\tilde{\boldsymbol{\varphi}}_{1:l_{g-1}} - \mathbf{f}_{i,j,1:l_{g-1}} \right) \\ \mathbf{B} &= \boldsymbol{\Sigma}_{i,gg}^0 - \left(\boldsymbol{\Sigma}_{i,g1}^0 \quad \boldsymbol{\Sigma}_{i,g2}^0 \quad \cdots \quad \boldsymbol{\Sigma}_{i,g(g-1)}^0 \right) \mathbf{C}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}_{i,1g}^0 \\ \boldsymbol{\Sigma}_{i,2g}^0 \\ \vdots \\ \boldsymbol{\Sigma}_{i,(g-1)g}^0 \end{pmatrix} \\ \mathbf{C} &= \begin{pmatrix} \boldsymbol{\Sigma}_{i,11}^0 & \boldsymbol{\Sigma}_{i,12}^0 & \cdots & \boldsymbol{\Sigma}_{i,1(g-1)}^0 \\ \boldsymbol{\Sigma}_{i,21}^0 & \boldsymbol{\Sigma}_{i,22}^0 & \cdots & \boldsymbol{\Sigma}_{i,2(g-1)}^0 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{i,(g-1)1}^0 & \boldsymbol{\Sigma}_{i,(g-1)2}^0 & \cdots & \boldsymbol{\Sigma}_{i,(g-1)(g-1)}^0 \end{pmatrix} \end{aligned}$$

[34] According to Berrocal *et al.* [2007] each iteration can be interpreted as an ensemble member from the multivariate BMA model. Hence, after many iterations, the theoretical quantiles for any lead time can be approximated by the empirical quantiles of the simulated ensemble.

4. Results

4.1. Verification Tools

[35] As stated in section 1, the goal of statistical postprocessing are well calibrated and sharp predictive density functions of the quantity to be forecast [Raftery *et al.*, 2005; Gneiting *et al.*, 2007]. In the following some methods for assessing calibration and sharpness will be introduced. Calibration can be examined by the Probability Integral Transform (PIT). The Continuous Ranked Probability Score (CRPS) addresses both, calibration and sharpness. The widths of the prediction intervals are a simple measure for sharpness.

4.1.1. Calibration

[36] As stated above the PIT is a measure for calibration. The PIT is given by:

$$z_i = \int_{-\infty}^{y_i} p(y|x_i) dy, \quad (12)$$

where $p(y_i|x_i)$ is the conditional predictive probability density of the response variable y_i , i.e., observed runoff in this work, given the forecast ensemble $\mathbf{x}_i$. Rosenblatt [1952] showed that $z_i \stackrel{\text{iid}}{\sim} U(0, 1)$ for a probabilistic forecast with perfect calibration. The PIT can be visualized by a histogram of the relative frequencies into which quantile

interval of the forecast density the observations fall. In case of discrete predictive distributions, like ensemble forecasts or quantiles of continuous predictive distributions, the deviation from uniformity can be estimated by the reliability index [Delle Monache et al., 2006; Berrocal et al., 2007]. To this end the relative frequencies of ranks of the observations among the discrete forecasts in the training period is determined. From this, the negatively oriented reliability index can be calculated by:

$$\Delta = \sum_{j=1}^{m+1} \left| f_j - \frac{1}{m+1} \right|, \quad (13)$$

where f_j denotes the relative frequency of rank j .

4.1.2. Continuous Ranked Probability Score

[37] Hersbach [2000] discusses the CRPS and its decomposition in detail. Nevertheless a short summary is provided here. The CRPS assesses how much the predicted and the observed cumulative distribution differ from each other. Denote runoff, or any other quantity of interest, by y and the predictive density by $\rho(y)$. And let again y_i be the observed runoff, then the CRPS is:

$$\text{CRPS} = \text{CRPS}(P, y_i) = \int_{-\infty}^{\infty} (P(y) - P_i(y))^2 dy, \quad (14)$$

where the cumulative distributions P and P_i are given by:

$$P(y) = \int_{-\infty}^y \rho(u) du, \quad (15)$$

$$P_i(y) = H(y - y_i), \quad (16)$$

where $H(y)$ is the Heaviside function:

$$H(y) = \begin{cases} 0 & \text{if } y < y_i \\ 1 & \text{if } y \geq y_i. \end{cases} \quad (17)$$

[38] The best possible CRPS value is zero for a deterministic forecast that equals the observation, i.e., $P = P_i$. The lower the sharpness and the higher the bias, the higher is also the CRPS. Hence, CRPS should be as small as possible. Usually one is interested in $\overline{\text{CRPS}} := \frac{1}{T} \sum_{t=1}^T \text{CRPS}_t$, i.e., the average CRPS over the verification period. The CRPSS, which is its associated skill score, can be calculated by

$$\text{CRPSS} = \frac{\overline{U} - \overline{\text{CRPS}}}{\overline{U}}, \quad (18)$$

where the uncertainty, $\overline{U}$ equals the $\overline{\text{CRPS}}$ of the climatological forecast, i.e., the empirical density of the climatology. Note that the CRPSS of a perfect forecast equals 1. CRPSS values below zero indicate that the forecast is worse than the climatological forecast in terms of CRPS. We have calculated $\overline{\text{CRPS}}$ and $\overline{U}$ using the R-package verification.

4.1.3. Multivariate Generalizations of PIT Histogram and CRPS

[39] The scoring rules introduced above assess only the univariate properties of the marginals of the BMA fore-

casts. In order to examine the dependency structure among different lead times multivariate generalizations of the PIT histogram and the CRPS are needed.

[40] Gneiting et al. [2008] introduced the multivariate rank histogram as a multivariate generalization of the PIT histogram. In this study multivariate ranks are assigned in a similar way, but with an adjustment in order to account for the high dimensionality of the data at hand. As proposed by Pinson and Girard [2012] dimensions are reduced by an orthogonal transformation. The most obvious way to do this is to consider only the first few components of the principal component transform in the following estimation procedure:

[41] 1. Sample randomly m realizations of the predictive distribution at hand. Collect the realizations $\mathbf{x}_j \in \mathbb{R}^L$, $j = 1, \dots, m$ and the corresponding observation $\mathbf{x}_0 \in \mathbb{R}^L$ in the set $\{\mathbf{x}_j : j = 0, 1, \dots, m\}$, where L is the number of lead times considered.

[42] 2. Reduce the dimension by applying a principal component transform to the set $\{\mathbf{x}_j : j = 0, 1, \dots, m\}$. Since only the first d principal components are used, the transformed set $\{\mathbf{x}_j^* \in \mathbb{R}^d : j = 0, 1, \dots, m\}$ consists of vectors of dimension d .

[43] 3. Calculate preranks $\rho_j = \sum_{k=0}^m \mathbf{1}_{[\mathbf{x}_k^* \preceq \mathbf{x}_j^*]}$, where $\mathbf{1}$ denotes the indicator function, and $\mathbf{x}_k^* \preceq \mathbf{x}_j^*$ if and only if $[\mathbf{x}_k^*]_i \leq [\mathbf{x}_j^*]_i$ for $i = 1, \dots, d$. A prerank of n for vector j indicates, for instance, that there exist $n - 1$ other vectors with all elements smaller or equal to the corresponding elements of vector j .

[44] 4. The multivariate ranks correspond then to their corresponding preranks, but with ties, i.e., multiple principal components with equal preranks, resolved at random.

[45] Without dimension reduction preranks are very likely to be equal to one due to the high dimensionality of 240 lead times considered. Hence, any multivariate ranks would be assigned completely at random leading to a flat multivariate rank histogram in almost every case. The downside of the dimension reduction approach is that the number of principal components used has to be selected deliberately.

[46] Gneiting and Raftery [2007] and Gneiting et al. [2008] proposed the energy score (ES) as a multivariate generalization of the CRPS. For univariate quantities the energy score equals the CRPS, with which it shares its negative orientation. It is given by

$$\text{es}(P, \mathbf{x}) = \mathbb{E}_P \|\mathbf{X} - \mathbf{x}\| - \frac{1}{2} \mathbb{E}_P \|\mathbf{X} - \mathbf{X}'\|, \quad (19)$$

where $\|\cdot\|$ denotes the Euclidian norm. $\mathbf{X}$ and $\mathbf{X}'$ are both independent random vectors following the same distribution P . For an ensemble forecasts P_{ens} of size M the energy score can be calculated by:

$$\text{es}(P_{\text{ens}}, \mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \|\mathbf{x}_m - \mathbf{x}\| - \frac{1}{2M^2} \sum_{n=1}^M \sum_{m=1}^M \|\mathbf{x}_n - \mathbf{x}_m\|, \quad (20)$$

where $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^L$ denote the members of P_{ens} and $\mathbf{x} \in \mathbb{R}^L$ is the vector of verifying observations. $\ell \in (1, \dots, L)$ denotes the prediction lags. If P is a density forecast, the

Figure 4. CRPSS of the BMA fits. The horizontal dashed line corresponds to the CRPSS of the climatology, which is zero by definition. The vertical dotted lines indicate the lead times at which COSMO-2, COSMO-7, or COSMO-LEPS drop out.

energy score can be approximated by sampling randomly $\mathbf{x}_1, \dots, \mathbf{x}_k$ from P :

$$\hat{e}_s = \frac{1}{k} \sum_{i=1}^k \|\mathbf{x}_i - \mathbf{x}\| - \frac{1}{2(k-1)} \sum_{i=1}^{k-1} \|\mathbf{x}_i - \mathbf{x}_{i+1}\|, \quad (21)$$

where k is the size of the random sample.

4.2. Summary Results

[47] In the following, the univariate and the multivariate BMA approach are compared by means of summary results over the entire training period.

4.2.1. CRPSS

[48] First, we analyze CRPSS over the entire range of lead times up to 240 h. As shown in Figure 4, both BMA approaches, univariate and multivariate, lead to an increase in CRPSS compared to the raw ensemble for lead times up to 170 h. Note that any CRPSS values below zero indicate that using the climatology as a prediction would outperform the actual forecast. For lead times up to 72 h univariate and multivariate BMA are equally good. From 73 to 120 h the multivariate model outperforms its univariate

counterpart. While dropping out of COSMO-2 and COSMO-7 at lead-times 72 h and 120 h, respectively, does not lead to a sudden decrease in CRPSS, losing the COSMO-LEPS forecasts at 120 h has a big impact. The multivariate approach dampens this effect resulting in better CRPSS scores for lead times between 120 and 150 h.

4.2.2. Calibration and Sharpness

[49] Now we investigate calibration by means of a modified, i.e., time dependent, 3-D PIT histogram. Figure 5 compares the PIT of the raw ensemble with that of the BMA models. The raw ensemble is underdispersed over the entire range of lead times. The BMA models are much better calibrated and do not show the cyclic calibration pattern of the raw ensemble. Differences between the univariate and the multivariate model are very small and do not seem to be systematic except for a drop in the $(0.9, 1.0]$ interval right after lead time 120 h in case of multivariate BMA. Note that both models exhibit a frequency shift from interval $(0.0, 0.1]$ to the intervals $(0.2, 0.3]$ and $(0.3, 0.4]$ at lead time 120 h.

[50] After having assessed calibration in more detail, we now take a closer look at the sharpness. This is done by comparing the mean width of the 95% prediction intervals

Figure 5. 3-D PIT histograms for lead times 1–240 h for the raw ensemble and both, the univariate and multivariate, BMA models. Each lead time slice corresponds to the marginal PIT histogram for a particular time lag. Perfect calibration would result in a horizontal plane at 0.1

Figure 6. Mean width of 95% prediction intervals against lead time. The dashed line corresponds to the 95% prediction interval widths of a climatology, which is dependent on the time of day. The vertical dotted lines indicate the lead times at which COSMO-2, COSMO-7, or COSMO-LEPS dropout.

of the BMA models with the 95% width of a lead time dependent climatological forecast. Lead time dependent climatology stands for the empirical density of observations at a particular time of day. That is, we have 24 climatologies instead of 1, because the raw and BMA forecasts are provided on an hourly basis. For lead times up to 120 h sharpness of the BMA models decreases linearly with increasing lead time as shown in Figure 6. At lead time 120 h the BMA models are still clearly sharper than the climatology. The univariate and the multivariate BMA models are quite similar in terms of sharpness up to lead time 120 h. The multivariate model is slightly sharper for lead times from 72 to 120 h. Interestingly, the multivariate model exhibits a prominent decrease in sharpness at lead time 120 h, while this is not the case for the univariate model. For lead times beyond 150 h the BMA models are not sharper than the corresponding climatologies.

4.2.3. Multivariate Evaluation

[51] As shown above both, univariate and multivariate, BMA methods lead to marginally well calibrated forecasts, which are also of similar sharpness. Considering multivariate evaluation methods multivariate BMA performs much better than its univariate counterpart. Figure 7 compares multivariate rank histograms that have been obtained by using only the first few principal components of the principal component transform. Multivariate BMA is likely to be well calibrated in multivariate terms, whereas the bell-shaped histograms of univariate BMA indicate poor calibration. Table 2 shows that in case of multivariate BMA the first six principal components are able to explain 86% of the variance, while also leading to a low reliability index. For univariate BMA 40 principal components are needed in order to explain only 79% of the variance. The reliability indices of univariate BMA are comparable with

Figure 7. Multivariate rank histograms of univariate (left) and multivariate BMA (right) forecasts. The histograms are shown for the first 2, 4, and 6 principal components.

Table 2. Proportion of Variances and Corresponding Reliability Indices^a

n pc	Proportion of Variance		Reliability Index	
	uni BMA	mult BMA	uni BMA	mult BMA
2	0.06	0.58	0.727	0.285
4	0.12	0.78	0.727	0.245
6	0.18	0.86	0.462	0.264
8	0.23	0.90	0.297	0.216
10	0.28	0.93	0.236	0.262
15	0.40	0.96	0.245	0.236
20	0.50	0.98	0.255	0.252
40	0.79	0.99	0.238	0.243
69	1.00	1.00	0.214	0.243

^aProportion of variance explained by the first n principal components and the corresponding reliability indices for univariate and multivariate BMA.

those of multivariate BMA only when using ≥ 10 principal components. Using more than 10 principal component means also calculating a multivariate rank histogram on ≥ 10 dimensions. This typically leads to flat rank histograms even for badly calibrated ensembles, because the observations and most of the ensemble members are very likely to have prerank 1 which leads to randomly distributed observation ranks.

[52] As stated above the ES is a multivariate generalization of the CRPS, which assesses sharpness and multivariate calibration simultaneously. Table 3 shows the mean and median ES's over all verification dates. The ES values confirm the multivariate rank histograms, which indicate that multivariate BMA performs better than univariate BMA. Note that even the ensemble consisting of the ordered $\{\frac{0.5}{69}, \frac{1.5}{69}, \dots, \frac{68.5}{69}\}$ quantiles exhibits a higher mean and median ES than a random sample of size $k=1000$ from the multivariate BMA forecast distributions.

5. Discussion

[53] We first consider univariate verification tools only. Both BMA methods, the univariate and the multivariate, are able to increase calibration and sharpness compared to the raw forecast for lead times up to 120 h. Beyond lead time 150 h both BMA methods do not perform better than the raw forecasts. As stated in the results section 4, multivariate BMA is able to improve CRPSS significantly between lead time 120 and 150 h. Since COSMO-LEPS, which is a regional ensemble model with high spatial resolution, forecasts only up to 120 h, the pronounced drop in CRPSS from lead time 120–121 h in case of the raw ensemble and the univariate BMA approach is not surprising. Note that the high resolution deterministic models COSMO-2 and COSMO-7 forecast only up to 24 and 72 h, respectively. Hence, it is probable that multivariate BMA is able to make use of the more valuable COSMO-LEPS forecasts during the first 10–20 h after the dropping out of COSMO-LEPS. As this is reflected in CRPSS, but not in sharpness, multivariate BMA seems to mainly improve reliability. This might be due to the higher spread of the COSMO-LEPS ensemble compared to the ECMWF ensemble. Consequently, the multivariate BMA is less sharp for

Table 3. Mean and Median Energy Scores Over the Training Period^a

Method	Mean	Median
uni BMA 69 quant	0.865	0.631
uni BMA mc	0.975	0.709
mult BMA mc	0.831	0.598

^aIn case of univariate BMA, ES has been calculated once from the exact $\{\frac{0.5}{69}, \frac{1.5}{69}, \dots, \frac{68.5}{69}\}$ quantiles of the forecast distribution and once using the random sampling (Monte Carlo) approach with sample size $k=1000$. For multivariate BMA, ES has been obtained by using a random sample of size $k=1000$.

lead times right after the dropping out of COSMO-LEPS. From the above considerations, it follows that even in terms of univariate verification multivariate BMA is to be preferred in cases where the most valuable models drop out at an early lead time. Taking multivariate verification tools like the energy score or the multivariate rank histogram into account, multivariate BMA performs much better than its univariate counterpart. Additionally, we could show that the multivariate rank histogram can be used to assess high dimensional data. To this end, dimension reduction by adequate selection of the first few components of the principal component transform has to be performed. Further investigations on multivariate verification might be valuable to overcome the loss of information due to principal component selection.

[54] As stated in section 3.2, the dependencies among lead times of the multivariate BMA model may be modeled by the more general Matérn variogram model instead of the exponential variogram model. This would on the one hand result in more accurate variogram fits, but on the other hand also in a more complex and more time consuming fitting process. As a test, of which we do not show any further details here, we have fitted both models to the average empirical variogram over all training periods used for this study. The exponential variogram fits proved to be only slightly worse in this case. This may be different in case of other hydrologic settings and needs to be checked for any other catchment separately.

6. Conclusions

[55] We have shown that multivariate BMA provides a valuable tool for postprocessing of probabilistic forecasts for Thur River over an entire range of lead times simultaneously. It performs much better than its univariate counterpart in terms of multivariate verification measures, and more surprisingly it outmatches univariate BMA also in terms of univariate verification. Multivariate BMA especially improves the reliability of the prediction when the forecast system is changing, i.e., fewer models are available due to dropping out at particular lead times. In order to generalize those findings, the strength of multivariate BMA should be verified using other catchments of different sizes and flow regimes. Additionally, multivariate BMA needs to be contrasted with other multivariate postprocessing methods like Ensemble Copula Coupling (Scheffzik et al., Uncertainty quantification in complex simulation models using ensemble copula coupling, to be published in *Statistical Science*, 2013). Due to its inherent correlation structure,

sampling randomly from the multivariate BMA forecast distribution leads to more realistic predicted trajectories than univariate approaches. Furthermore, in spite of its complexity multivariate BMA comes along with a rather low computing time, since one single BMA model typically applies to a whole range of lead times. However, multivariate BMA features also some shortcomings: first, it is much more complex than the univariate BMA approach. Second, multivariate BMA would systematically fail in cases of sudden changes in variance structure among different lead times. Note that changes in variance structures from the training period to the verification day would be even more problematic. Further, improvements of multivariate BMA may go in the direction of applying different models depending on the river flow regimes like winter low flow, summer low flow et cetera [Tongal et al., 2013]. One of the main tasks would be to find sound flow regime differentiation criteria.

[56] **Acknowledgments.** We are grateful to H. R. Künsch of the Seminar for Statistics at ETH Zürich for advising the master thesis of S. Hemri, on which this article is based. T. Gneiting, M. Scheuerer and other colleagues of the Institute for Applied Mathematics at Heidelberg University is thanked for helpful discussions and inputs. We like to thank three anonymous reviewers for their helpful comments. We are indebted to the Swiss Federal Office of Meteorology and Climatology, MeteoSwiss and ECMWF for granting access to their products. This research was partly funded from grants of the EU FP7 Project IMPRINTS (grant 226555/FP7-ENV-2008-1-226555) and by the Department of Waste, Water, Energy and Air (AWEL) of Canton Zurich.

References

- Addor, N., S. Jaun, F. Fundel, and M. Zappa (2011), An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): Skill, case studies and scenarios, *Hydrol. Earth Syst. Sci.*, **15**, 2327–2347, doi:10.5194/hess-15-2327-2011.
- Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resour. Res.*, **43**, W01403, doi:10.1029/2005WR004745.
- Berrocal, V. J., A. E. Raftery, and T. Gneiting (2007), Combining spatial statistical and ensemble information in probabilistic weather forecasts, *Mon. Weather Rev.*, **135**, 1386–1402, doi:10.1175/MWR3341.1.
- Bogner, K., F. Pappenberger, and H. L. Cloke (2013), Model combination and weighting methods in operational flood forecasting, *Geophys. Res. Abstr.*, **15**, 13629.
- Box, G., and D. Cox (1964), An analysis of transformations, *J. R. Stat. Soc., Ser. B*, **26**, 211–252.
- Chilès, J.-P., and P. Delfiner (1999), *Geostatistic: Modeling Spatial Uncertainty*, John Wiley, New York.
- Consortium for Small-scale Modeling (2011), Operational applications within COSMO. MeteoSwiss, Zürich. [Available at <http://www.cosmo-model.org/content/tasks/operational/meteoSwiss/default.htm>, last accessed 23 Apr. 2013.]
- Cressie, N. A. C. (1993), *Statistics for Spatial Data*, rev. ed., John Wiley, New York.
- Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull (2006), Probabilistic aspects of meteorological and ozone regional ensemble forecasts, *J. Geophys. Res.*, **111**, D24307, doi:10.1029/2005JD003917.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc., Ser. B*, **39**, 1–39.
- Diks, C. G. H., and J. A. Vrugt (2010), Comparison of point forecast accuracy of model averaging methods in hydrologic applications, *Stochastic Environ. Res. Risk Assess.*, **24**, 809–820, doi:10.1007/s00477-010-0378-z.
- Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian (2007), Multi-model ensemble hydrologic prediction using Bayesian model averaging, *Adv. Water Resour.*, **30**(5), 1371–1386, doi:10.1016/j.advwatres.2006.11.014.
- Fraley, C., A. E. Raftery, and T. Gneiting (2010), Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging, *Mon. Weather Rev.*, **138**(1), 190–202, doi:10.1175/2009MWR3046.1.
- Fundel, F., and M. Zappa (2011), Hydrological ensemble forecasting in mesoscale catchments: Sensitivity to initial conditions and value of reforecasts, *Water Resour. Res.*, **47**, W09520, doi:10.1029/2010WR009996.
- Fundel, F., S. Jörg-Hess, and M. Zappa (2013), Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, *Hydrol. Earth Syst. Sci.*, **17**, 395–407, doi:10.5194/hess-17-395-2013.
- Gel, Y., A. E. Raftery, and T. Gneiting (2004), Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method, *J. Am. Stat. Assoc.*, **99**(467), 575–583, doi:10.1198/016214504000000872.
- Gneiting, T., and A. E. Raftery (2007), Strictly proper scoring rules, prediction, and estimation, *J. Am. Stat. Assoc.*, **102**, 359–378, doi:10.1198/016214506000001437.
- Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and sharpness, *J. R. Stat. Soc., Ser. B*, **69**, 243–268, doi:10.1111/j.1467-9868.2007.00587.x.
- Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson (2008), Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds, *Test*, **17**(2), 211–235, doi:10.1007/s11749-008-0114-x.
- Hersbach, H. (2000), Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecasting*, **15**(5), 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.
- Krzanowski, W. J. (2000), *Principles of Multivariate Analysis: A User's Perspective*, vol. 23, Oxford Univ. Press, Oxford, U. K.
- Marsigli, C., A. Montani, F. Nerozzi, and T. Paccagnella (2004), Probabilistic high-resolution forecast of heavy precipitation over Central Europe, *Nat. Hazards Earth Syst. Sci.*, **4**, 315–322.
- Marsigli, C., F. Boccanera, A. Montani, and T. Paccagnella (2005), The COSMO-LEPS mesoscale ensemble system: Validation of the methodology and verification, *Nonlinear Processes Geophys.*, **12**, 527–536.
- McLachlan, G. J., and T. Krishnan (1997), *The EM Algorithm and Extensions*, p. 274, Wiley, New York.
- Molteni, F., R. Buizza, T. Palmer, and T. Petroligis (1996), The ECMWF ensemble prediction system: Methodology and validation, *Q. J. R. Meteorol. Soc.*, **122**, 73–119, doi:10.1002/qj.49712252905.
- Montani, A., D. Cesari, C. Marsigli, and T. Paccagnella (2011), Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: Main achievements and open challenges, *Tellus, Ser. A*, **63**(3), 605–624, doi:10.1111/j.1600-0870.2010.00499.x.
- Palmer, T. (2000), Predicting uncertainty in forecasts of weather and climate, *Rep. Prog. Phys.*, **63**, 71–116, doi:10.1088/0034-4885/63/2/201.
- Parrish, M. A., H. Moradkhani, and C. M. DeChant (2012), Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation, *Water Resour. Res.*, **48**, W03519, doi:10.1029/2011WR011116.
- Pinson, P., and R. Girard (2012), Evaluating the quality of scenarios of short-term wind power generation, *Appl. Energy*, **96**, 12–20, doi:10.1016/j.apenergy.2011.11.004.
- Pinson, P., H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klöckl (2008), From probabilistic forecasts to statistical scenarios of short-term wind power production, *Wind Energy*, **12**(1), 51–62, doi:10.1002/we.284.
- R Development Core Team (2011), *R: A language and environment for statistical computing*, R Found. for Stat. Comput., Vienna, Austria. [Available at <http://www.R-project.org/>, last accessed 6 Oct. 2013.]
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, **133**(2), 1155–1174, doi:10.1175/MWR2906.1.
- Rosenblatt, M. (1952), Remarks on multivariate transformation, *Ann. Math. Stat.*, **23**, 470–472.
- Tibaldi, S., T. Paccagnella, C. Marsigli, A. Montani, and F. Nerozzi (2006), Limited-area ensemble forecasting: The COSMO-LEPS system, in *Predictability of Weather and Climate*, edited by T. Palmer and R. Hagedorn, pp. 489–513, Cambridge Univ. Press, Cambridge, U. K., doi:10.1017/CBO9780511617652.019.
- Tongal, H., M. C. Demirel, and M. J. Booij (2013), Seasonality of low flows and dominant processes in the Rhine River, *Stochastic Environ. Res. Risk Assess.*, **27**(2), 489–503, doi:10.1007/s00477-012-0594-9.

- Viviroli, D., M. Zappa, J. Schwanbeck, J. Gurtz, and R. Weingartner (2007a), Continuous simulation for flood estimation in ungauged meso-scale catchments of Switzerland. Part I: Modelling framework and calibration results, *J. Hydrol.*, *377*(1–2), 191–207, doi:10.1016/j.jhydrol.2009.08.023.
- Viviroli, D., H. Mittelbach, J. Gurtz, and R. Weingartner (2007b), Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland. Part II: Parameter regionalisation and flood estimation results, *J. Hydrol.*, *377*(1–2), 208–225, doi:10.1016/j.jhydrol.2009.08.022.
- Viviroli, D., M. Zappa, F. J. Gurtz, and R. Weingartner (2009), An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools, *Environ. Modell. Software*, *24*, 1209–1222, doi:10.1016/j.envsoft.2009.04.001.
- Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*, W01411, doi:10.1029/2005WR004838.
- Yuan, X., and E. F. Wood (2012), Downscaling precipitation or bias-correcting steamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast, *Water Resour. Res.*, *48*, W12519, doi:10.1029/2012WR012256.
- Zappa, M., S. Jaun, U. Germann, A. Walser, and F. Fundel (2011), Superposition of three sources of uncertainties in operational flood forecasting chains, *Atmos. Res.*, *100*, 246–262, doi:10.1016/j.atmosres.2010.12.005.