Technical Note: Combining Quantile Forecasts and Predictive Distributions of Stream-flows

Konrad Bogner[1], Katharina Liechti[1], and Massimiliano Zappa[1]

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

Correspondence to: K. Bogner (konrad.bogner@wsl.ch)

Abstract. The enhanced availability of many different hydro-meteorological modelling and forecasting systems raises the issue of how to optimally combine this great deal of information. Especially the usage of deterministic and probabilistic forecasts with sometimes widely divergent predicted future stream-flow values makes it even more complicated for decision makers to sift out the relevant information. In this study multiple stream-flow forecast information will be aggregated based on several different predictive distributions, resp. quantile forecasts. For this combination the Bayesian Model Averaging (BMA) approach, the Nonhomogeneous Gaussian Regression (NGR), also known as Ensemble Model Output Statistic (EMOS) model and a novel method called Beta transformed Linear Pooling (BLP) will be applied. By the help of the Quantile Score (QS) and the Continuous Ranked Probability Score (CRPS), the combination results for the Sihl river in Switzerland with about five years of forecast data will be compared and the differences between the raw and the optimally combined forecasts will be highlighted. The results demonstrate the importance of applying proper forecast combination methods for decision makers in the field of flood and water resources management.

1 Introduction

The combination, or aggregation, of differing probability distributions into a single one could result in beneficial effects, since the differences between various forecast systems provide a better understanding of the uncertainty about the target quantities and the aggregates may reflect more accurately the information. However, the biggest advantage of aggregation is that the forecaster is not forced to decide a priori which forecast system is the most reliable at the actual point of issuing a forecast, because the combination method will be optimized at each forecast run by taking into consideration the quality of the forecast from previous time steps. Thus, the data itself will automatically lead to the optimal decision incorporating all available information about the different deficiencies and strengths of the individual forecast systems.

In econometrics and related disciplines, the combination of forecasts has a long tradition starting with Bates and Granger (1969) suggesting the use of empirical weights derived from 'out of sample' forecast variances. An overview over the last forty years of forecast combination in the economic field can be found in Wallis (2011). Thompson (1977) was one of the first who outlined the advantages of forecast combinations in meteorology and Shamseldin et al. (1997) showed different methods of combining the output of different hydrological models. In Abrahart and See (2002) different combination methods for hydrological forecast models are compared. Diks and Vrugt (2010) compare different model averaging approaches, showing that a

simple regression method could result in improvements comparable to more sophisticated methods.

In general the challenge of model combination is that, apart from the simple model averaging methodologies, different weights need to be assigned according to the quality of the forecast of the preceding days and periods. A frequently used method for model averaging and forecast combination is the method of Bayesian Model Averaging (BMA) introduced by Min and Zellner (1993) and Raftery et al. (1997), where the weights are based on posterior model probabilities within a Bayesian framework. The BMA method has been applied in the field of ensemble forecast calibration (Raftery et al. (2005); Fraley et al. (2010)) and for flood forecasting purposes, e.g. in Ajami et al. (2007), Vrugt and Robinson (2007), Todini (2008) and Hemri et al. (2013).

In Gneiting et al. (2005) and Gneiting et al. (2007) the term calibration is used to describe the statistical consistency between the distributional forecasts and the observations and is a joint property of the predictions and the events that materialise. A state of the art calibration and bias correction method is the Non-homogeneous Gaussian Regression (NGR), also known as Ensemble Model Output Statistic (EMOS) technique of Gneiting et al. (2005). It fits a single parametric predictive probability density function (pdf) using summary statistics from the (multi-model) ensemble and corrects simultaneously for biases and dispersion errors. Also NGR has been applied many times successfully for calibrating and combining hydro-meteorological ensemble forecasts (see for example Hemri et al. (2014)).

The Beta transformed Linear Pooling (BLP) approach, which has been developed recently by Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013) for combining predictive distributions, will be tested and compared with the NGR and the BMA in this study. To the author's knowledge the BLP and the associated estimation of weights, which assign relative importance to the individual predictive distributions, has not been applied to hydrological forecasts so far.

Before the combination methods are applied, the errors of the hydrological model are corrected in order to minimize the difference between the last available observation and the predictions at the time of initialization of the forecast. This process of error correction is later on called post-processing, since it starts after completing the hydrological simulations and predictions given meteorological observations or forecasts. Depending on the post-processing method, quantiles or pdf's for future streamflows will be derived for each single forecast time-step. Whereas Quantile Regression (QR) methods (Koenker (2005)) and modifications of it will lead to predictions of quantiles, a predictive pdf can be derived for example by the recently developed waveVARX method (Bogner and Pappenberger (2011) directly. For more details of these post-processing methods the reader is referred to Bogner et al. (2016), whereas the objective of this paper will be the analysis of combination methods of forecasts. In the next section the three combination methods and the applied verification measures will be described. After the presentation of the data and the results, the outcome of the comparison will be discussed and summarized in the conclusions.

2 Methods

Three different combination methods have been applied to the flood forecasting system for the river Sihl at the station Zurich (Switzerland), where two meteorological forecasts, the 16 member ensemble system COSMO-LEPS (Montani et al. (2011)) and the deterministic C7 system (produced at MeteoSwiss with $\approx$ 7 km resolution) are implemented (a detailed description can be found in Addor et al. (2011); Ronco et al. (2015); Liechti et al. (2016)).

In a first step the hydrological modelling errors of all these forecasts will be minimized, using a QR method in combination with Neural Networks (QRNN, Taylor (2000); Cannon (2011)). This will result in direct estimates of the inverse cumulative density function (i.e. the quantile function), which in turn allows the derivation of the predictive uncertainty (see for example (Weerts et al., 2011; López López et al., 2014; Dogulu et al., 2015), where the application of the QR in order to estimate Predictive Uncertainties (PU's) is outlined). If the number of estimated quantiles within the domain $\{0 < \tau < 1\}$ is sufficiently large the resulting distribution could be considered as continuous. In this study the number of quantiles is set to nine with probability levels $\tau = 0.01, 0.1, 0.2, 0.25, 0.5, 0.75, 0.8, 0.9, 0.99$. In Quiñonero Candela et al. (2006) the cdf, respectively pdf is constructed by combining step-interpolation of probability densities for specified τ-quantiles with exponential lower and upper tails, which will be called the empirical method (EMP). Alternatively the pdf could be constructed by monotone re-arranging the τ-quantiles and estimating a log-normal distribution (LN) to these quantiles for each lead-time Δt. The advantage of the quantile re-arranging and the distribution fitting is twofold and efficiently prevents known problems occurring with QR: firstly it eliminates the problem of crossing of different quantiles (i.e. the unrealistic, but possible outcome of the non-linear optimization problem yielding lower quantiles for higher stream-flow values Chernozhukov et al. (2010), e.g. the value of the 0.90 quantile is higher than the value of the 0.95 quantile) and secondly it permits the extrapolation to extremes not included in the training sample (Bowden et al. (2012)).

This QRNN method will be applied to each ensemble member of the COSMO-LEPS forecasts resulting in 16 forecasts of quantiles and to the C7 forecasts. Lichtendahl et al. (2013) have examined averaging quantiles of continuous distributions given by multiple information sources rather than averaging probabilities. Both approaches of probability and quantile averaging have been applied in this paper for averaging the post-processed Ensemble Prediction System (EPS) based stream-flow forecasts in order to get one predictive pdf, resp. quantile forecast. Before applying the probability averaging approach, a pdf has been constructed by the LN method, i.e. a log-normal distribution has been fitted to the re-arranged τ - quantiles.

Thus, in total there are 5 different forecasts available after post-processing, two based on the application of the QRNN method for the COSMO-LEPS with probability averaging (p.aver.), resp. quantile averaging (q.aver.), two post-processed C7 forecasts based on QRNN with the EMP and the LN aproach, and one forecast based on the waveVARX method. Additionally the raw COSMO-LEPS forecast will be included in the following combination procedures as well (see Fig. 1).

Three different methods will be tested for optimally combining these six forecast models $(M1, \ldots, M6)$, which allow to assign different weights to the raw and the five post-processed forecasts. For the application of the first two methods, BMA and NGR, the stream-flow values have been transformed to the Normal Space by the help of the Normal Quantile Transformation (Van der Waerden (1952), Van der Waerden (1953a, b)).

2.1 Bayesian Model Averaging (BMA)

If the combination is calculated within a Bayesian Framework by using weights corresponding to the posterior model probabilities, it is usually referred to as BMA and follows from direct application of Bayes' theorem as explained in e.g. Min and Zellner (1993) and Raftery et al. (1997).

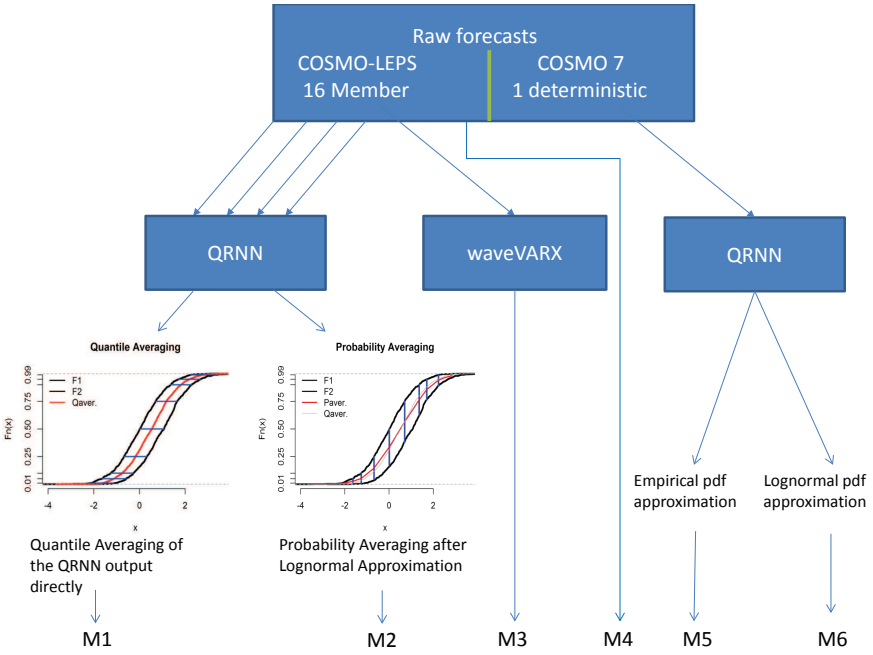

Figure 1. Set of six different forecast models available for combination, five post-processed plus one raw forecast. For the quantile averaging (M1) and the probability averaging (M2) method an example of averaging two ensemble members is indicated.

In Raftery et al. (2005) the statistical BMA model is extended to dynamical forecast models, where each forecast and/or ensemble member is represented by a probabilistic distribution for which a weight is assigned based on the past performance of each individual forecast. These weights are used to combine all distributions into one single mixture distribution. Therefore the BMA predictive model of the quantity of interest y is given by

$$p(y|k_1,\ldots,k_M) = \sum_{m=1}^{M} h_m g_m(y|k_m), \tag{1}$$

where h_m is the posterior probability (i.e. weight) of forecast k_m being the best forecast derived from its performance in the training period and the conditional pdf of y on k_m, $g_m(y|k_m)$, given that k_m is the best forecast in the ensemble with $m = 1,\ldots,M$ members, resp. models. The transformation of the stream-flow values to the Normal Space beforehand allows the application of the BMA method based on mixtures of univariate normal distributions. In the work of Wang et al. (2012) and Schepen and Wang (2015) variants of the BMA method have been applied, which allow the direct usage of the cdf's for estimating the weighting parameters. However, in this study these BMA approaches have not been implemented and the estimated medians ($\tau = 0.5$) from the five post-processing methods and from the raw COSMO-LEPS are taken as input only in order to allow better comparison with the following NGR approach.

2.2 Non-homogeneous Gaussian Regression (NGR)

Another possibility to address under-dispersion and forecast bias is the use of the Non-homogeneous Gaussian Regression (NGR) method, also known as Ensemble Model Output Statistics (EMOS) and is based on multiple linear regression for linear variables, such as temperature or stream-flows, and logistic regression for binary variables, such as precipitation occurrence or freezing. More information about the MOS technique can be found for example in Glahn and Lowry (1972) and Wilks (1995). Its extension for ensembles is explained in Gneiting et al. (2005) and a brief summary of this method is given hereafter. Let y denote again the variable of interest (e.g. stream flow) and let $k_1, \ldots, k_M$ be the corresponding forecast of the M ensemble members or models. If $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian density with mean μ and variance σ^2, the NGR predictive distribution is given by

$$y|k_1, \ldots, k_M \sim \mathcal{N}(a_0 + a_1 k_1 + \cdots + a_M k_M, b_0 + b_1 s^2), \tag{2}$$
$$\text{where } s^2 = \frac{1}{M} \sum_{m=1}^{M} \left(k_m - \frac{1}{M} \sum_{m=1}^{M} k_m \right)^2 .$$

Thus the predictive mean is equal to the regression estimates with coefficients $a_0, \ldots, a_m, b_0$, and b_1 and forms a bias-corrected weighted average of the different forecasts (ensemble members), whereas the predictive variance depends linearly on the variance of the forecast models (ensemble members). Although modifications for the NGR exists for non-normal distributed variates (see for example Baran (2014), Baran and Lerch (2015)), the stream-flow values have been transformed to the Normal Space for comparison reasons and the medians ($\tau = 0.5$) from the five post-processing methods and from the raw COSMO-LEPS are taken as input as in the BMA method.

2.3 Beta transformed Linear Pool (BLP)

In Ranjan and Gneiting (2010) it has been stated that any non-trivially weighted average of distinct probability forecasts will be uncalibrated and lack sharpness, even when the individual forecasts have been calibrated. Hence they suggested a composite of the traditional linear pool with a beta transform. The aggregation method introduced by Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013) considers the Beta transformed Linear Pool (BLP) for a set of predictive cdfs $F_1, \ldots, F_M$ as

$$F(y) = B_{\alpha,\beta} \left(\sum_{m=1}^{M} \omega_m F_m(y) \right) \tag{3}$$

for $y \in R$, where $B_{\alpha,\beta}$ denotes the cdf of the standard Beta distribution with parameters $\alpha > 0$ and $\beta > 0$ and $\omega_1, \ldots, \omega_M$ being nonnegative weights that sum to 1. The BLP density forecast for the component densities $f_i, \ldots, f_M$ then is

$$f(y) = \left(\sum_{m=1}^{M} \omega_m f_m(y) \right) b_{\alpha,\beta} \left(\sum_{m=1}^{M} \omega_m F_m(y) \right) \tag{4}$$

with parameters $\alpha > 0$ and $\beta > 0$ of the Beta density function $b_{\alpha,\beta}$. For $\alpha = \beta = 1$ the BLP corresponds to the traditional linear opinion pool.

Thus $B_{\alpha,\beta}$ can be interpreted as a parametric calibration function for combining $F_1,\ldots,F_M$ with mixture weights $\omega\in\Delta_M$, which assign relative importance to the individual predictive distributions. The parameters $\alpha>0$ and $\beta>0$ and the weights $\omega_1,\ldots,\omega_M$ are estimated with the maximum likelihood method. The log likelihood function for the BLP model (4) is

$$\begin{aligned}\ell(\omega_1,\ldots,\omega_M;\alpha,\beta) &= \sum_{j=1}^{J}\log(f(y_j)) \\ &= \sum_{j=1}^{J}\log\left(\sum_{m=1}^{M}\omega_m f_{mj}\left(y_j\right)\right)+\sum_{j=1}^{J}\log\left(b_{\alpha,\beta}\left(\sum_{m=1}^{M}\omega_m F_{mj}\left(y_j\right)\right)\right) \\ &= \sum_{j=1}^{J}\log\left(\sum_{m=1}^{M}\omega_m f_{mj}\left(y_j\right)\right) \\ &\quad+\sum_{j=1}^{J}\left((\alpha-1)\log\left(\sum_{m=1}^{M}\omega_m F_{mj}\left(y_j\right)\right)\right. \\ &\quad\left.+(\beta-1)\log\left(1-\sum_{m=1}^{M}\omega_m F_{mj}\left(y_j\right)\right)\right)+J\log B(\alpha,\beta)\end{aligned} \tag{5}$$

where B is the classical Beta function.

This BLP approach has been applied now to combine the different forecast systems. The quantiles resulting from the QRNN method (models M1, M4, M5) forecasts have been converted to pdfs applying the LN method (by fitting a log-normal distribution to the re-arranged τ quantiles).

2.4 Verification

Although probability and quantile forecasts are both probabilistic products, the former is expressed in terms of a probability (e.g. that a certain threshold will be exceeded) and the latter is given by a quantile for a particular probability level of interest (Bouallègue et al. (2015)). Since the output of the QRNN model are quantiles, it is reasonable to evaluate the performance with a skill score which has been developed for predictive quantiles (Koenker and Machado (1999); Friederichs and Hense (2007)), known as the Quantile Score (QS). It is based on an asymmetric piecewise linear function, the so called check function,

$\rho_\tau\left(y_i-q_{\tau,i}\right)$, which is a function of the probability level τ $(0<\tau<1)$ and the error between the observation y_i and the quantile forecast $q_{\tau,i}$ for $i=1,\ldots,N$, where N is the sample size. The check function is defined as:

$$\rho_\tau\left(y_i-q_{\tau,i}\right)=\begin{cases}\tau\left(y_i-q_{\tau,i}\right) & \forall y_i\geq q_{\tau,i}\\ (\tau-1)\left(y_i-q_{\tau,i}\right) & \forall y_i<q_{\tau,i}\end{cases} \tag{6}$$

and the QS results as the mean of the check function with penalties $1-\tau$ and τ for under- and over-forecasting (see Bouallègue et al. (2015)):

$$QS=\frac{1-\tau}{N}\sum_{i:y_i<q_{\tau,i}}\left(q_{\tau,i}-y_i\right)+\frac{\tau}{N}\sum_{i:y_i\geq q_{\tau,i}}\left(y_i-q_{\tau,i}\right) \tag{7}$$

The CRPS compares the forecast probability distribution with the observation and both are represented as cdfs. If F is the predictive cdf and y is the verifying observation, Gneiting and Ranjan (2011) showed that the CRPS can be defined equivalently as standard form,

$$CRPS(F,y) = \int_{-\infty}^{\infty} \left(F(t) - I\{y \leq t\}\right)^2 dt, \quad \text{and as} \tag{8}$$

$$= 2\int_{0}^{1} \left(I\left\{y < F^{-1}(\tau)\right\} - \tau\right)\left(F^{-1}(\tau) - y\right) d\tau \tag{9}$$

Thus, in the standard form (Equ. 8) an ensemble of predictions can be converted into a piecewise constant cdf with jumps at the different models (ensemble members), and $I\{.\}$ is a Heaviside step function, with a single step from 0 to 1 at the observed value of the variable. The equivalence of Equ. 8 to Equ. 9 was noted by Laio and Tamea (2007). For the quantile forecast $q_\tau = F^{-1}(\tau)$, the integrand in Equ. 9 equals the quantile score, i.e. the mean of the check function (Equ. 6). That means the CRPS corresponds to the integral of the QS over all thresholds, or likewise the integral of the QS over all probability levels (Laio and Tamea (2007) and Gneiting and Ranjan (2011)). Hence, the CRPS averages over the complete range of forecast thresholds and probability levels, whereas the QS looks at specific τ-quantiles; thus, it is more efficient in revealing deficiencies in different parts of the distributions, especially with respect to the tails of the distribution. Both verification measures are negatively oriented, meaning the smaller the better.

3 Results

COSMO-LEPS and C7 forecasts are available from 2010-02-24 to 2016-04-27 once a day with hourly time resolution, which have been post-processed in order to derive predictive distributions and quantile forecasts. For calibrating and validating the post-processing parameters (QRNN and waveVARX) the data set of available hourly observations and corresponding simulations have been split into two half (calibration period: 2010-2012; validation period: 2013-2016). The results of the validation, which are not shown due to lack of space, highlight the improvements of the QRNN method (similar to the results shown in Bogner et al. (2016)).

The weighting parameters of the combination methods are estimated by applying a moving window with a size of 7 days (168 hours) for optimization. Different window sizes have been tested as well, but 7 days was chosen finally as a trade-off between computing time and efficiency. In Fig. 2 an example of the temporal evolution of the hourly weights for a lead-time of 48 hours for the three combination methods is shown.

Before the forecast skill of the three combination methods are compared, the statistical consistency between the predictive cdf and the observations are analysed with the help of the Probability Integral Transform (PIT) as proposed by Dawid (1984) (see Fig. 3). In case of well calibrated forecasts, the sequence of PIT values will follow a uniform distribution $U(0,1)$. U-shaped

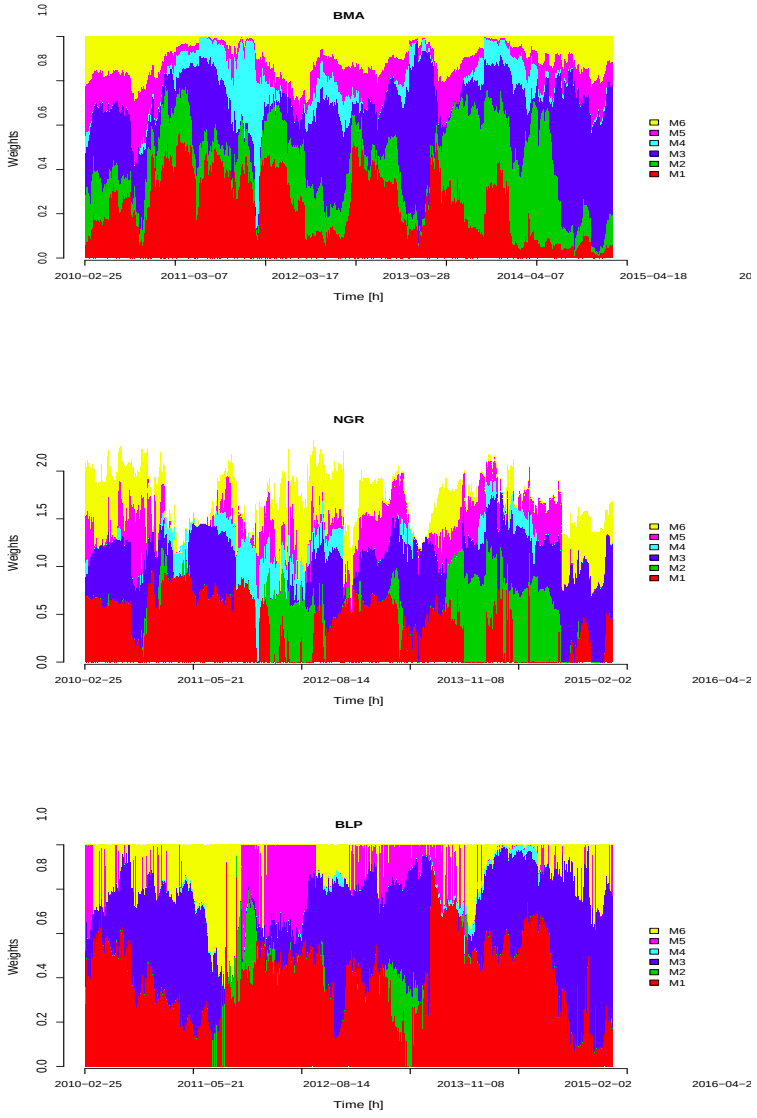

Figure 2. Hourly weights of the BMA (top), NGR (middle), BLP (bottom) method estimated for a lead-time of 48 hours. The 6 forecasts are the QRNN method for the COSMO-LEPS with quantile averaging (QRNN-CL-q.) - **M1**, probability averaging (QRNN-CL-p.) - **M2**, the waveVARX(-CL) method - **M3**, the raw COSMO-LEPS (CL) forecast - **M4**, the two post-processed C7 forecasts based on QRNN with the EMP - **M5**, resp. the LN aproach - **M6**

PIT histograms indicate underdispersed forecasts with too little spread on average, inverse U-shaped histograms correspond to overdispersed forecasts (see for example Gneiting et al. (2007), Laio and Tamea (2007)).

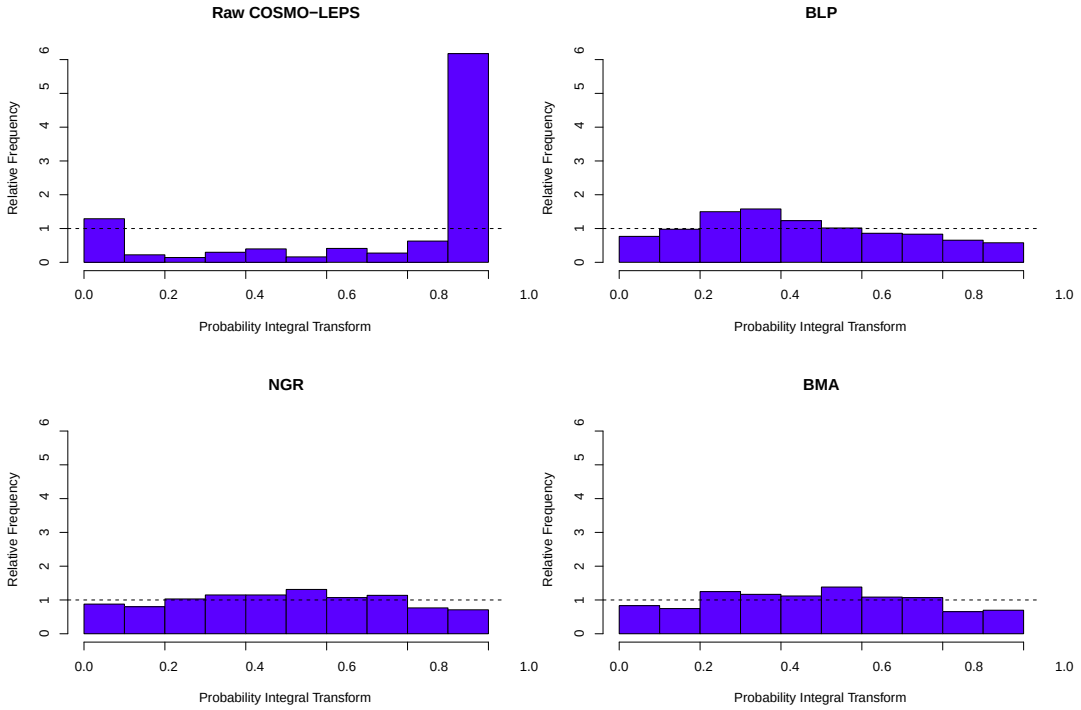

Figure 3. Probability Integral Transform (PIT) of the raw and the three combined forecast at a lead-time of 48 hours

The question now is whether there are significant differences between the three combination methods. Therefore the QS has been applied at first to highlight possible differences between the combination methods in more detail.

In Fig. 4 the results of the QS at four lead-times for the raw COSMO-LEPS (C-L, black line) and for the three combination methods BLP (red line), NGR (green line), BMA (blue line) are shown and compared to the QS results of the raw C-L (black circles). Additionally, a simple Quantile Mapping (QM) is applied (cyan diamonds) to the raw C-L forecasts in order to evaluate the positive effect of using more complex methods. Thereby the cdf of the raw forecast is matched to the cdf of the observations. As mentioned in Zhao et al. (2017) QM is highly effective for bias correction, but ensemble spread reliability problems cannot be solved properly.

In Figure 5 the CRPS results of the 6 forecast models are shown in comparison to the BLP in order to demonstrate the motivation of aggregating these systems. As can be seen clearly, the combined forecast outperforms each of the individual forecasts in view of the CRPS.

The CRPS for the raw C-L, the QM approach and the three combination methods is shown in Fig. 6.

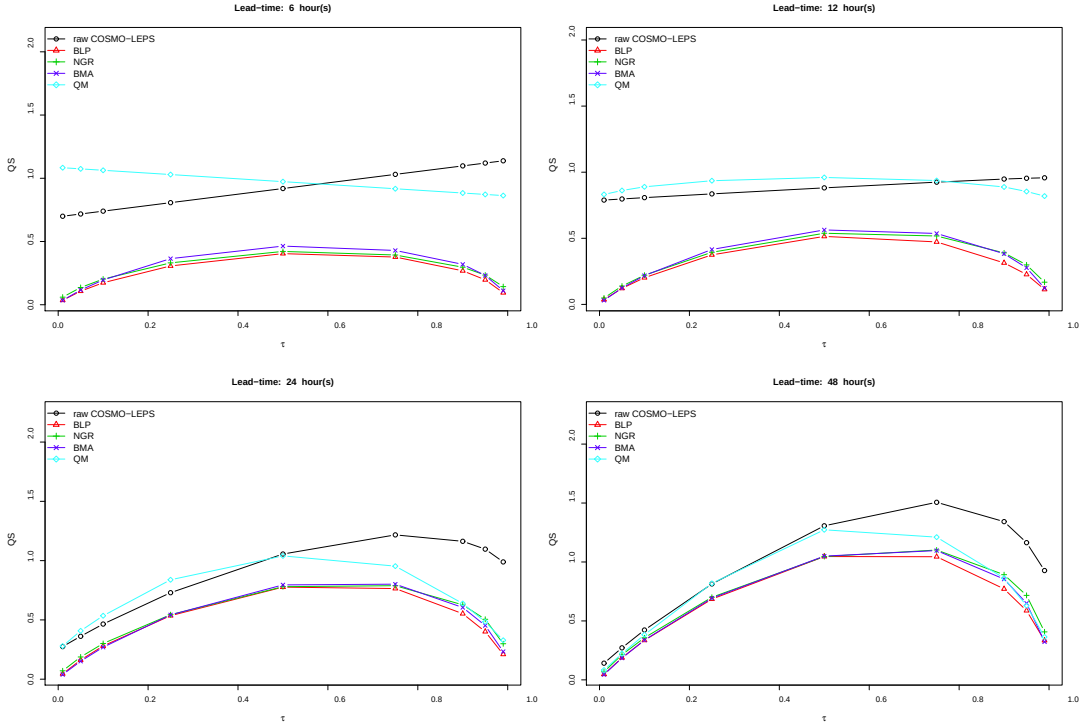

Figure 4. Quantile Score (QS) for various lead-times and the three combination methods in comparison to the raw COMSO-LEPS and a simple Quantile Mapping (QM) approach

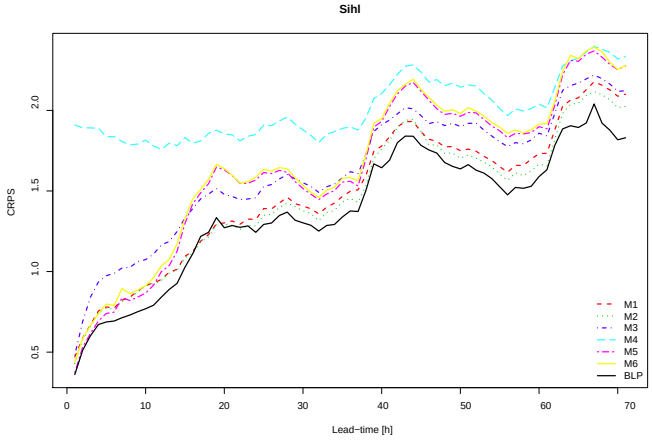

Figure 5. CRPS of the six forecast models: COSMO-LEPS with quantile averaging (QRNN-CL-q.) - **M1**, probability averaging (QRNN-CL-p.) - **M2**, the waveVARX(-CL) method - **M3**, the raw COSMO-LEPS (CL) forecast - **M4**, the two post-processed C7 forecasts based on QRNN with the EMP - **M5**, resp. the LN aproach - **M6**. Additionally the CRPS of the BLP combined forecast is shown.

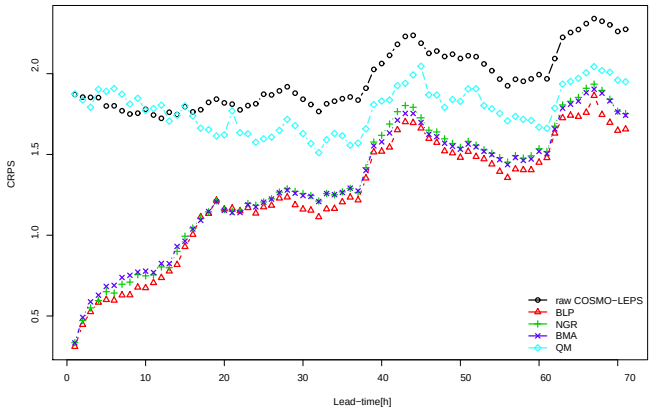

Figure 6. CRPS of the raw and the combined forecast

4 Discussion

So far most of the studies comparing the results of the BMA and the NGR approach did not find any preference (see for example Williams et al. (2014)). In this paper these two methods are checked against the BLP, which has not been used for hydrological purposes until now. In a first step the weights derived for each individual, raw and post-processed, forecast system are compared. The pattern of these optimized weights in Fig. 2 show rather vague similarities between the three combination methods. The BLP and the NGR are in general more spiky with rapid changes between consecutive hours. This could result from problems on convergence from the optimization algorithm applied for estimating the parameters ("constrOptim" in R (R Core Team (2015))).

In general the weights show some periodicity, which indicates that some models are more appropriate to be used at certain seasons and for certain flow conditions during a year. However, the limited amount of data does not allow to draw clear conclusions.

The results of the PIT clearly indicate that all three combination result in well-calibrated forecasts with close to uniform histograms. In Fig. 3 the examples for the 48h forecast are given, highlighting the heavy underdispersiveness of the raw forecasts. The same behaviour is visible for almost all lead-times, however the raw COSMO-LEPS forecasts are getting less underdisperse with increasing lead-time, since the spread and the uncertainty in the ensemble increases.

The analysis of the QS (Fig. 4) show slightly better results for the BLP followed by the NGR and BMA. The raw COSMO-LEPS (C-L) and the QM are much worse, especially for smaller lead-times. It is interesting to see that the QS of the raw C-L follows a straight line for smaller lead-times (6 and 12 hours) in the same manner as one would expect from deterministic forecasts, because of the under-dispersiveness of the C-L at the beginning of the forecast horizon. The slope of this line is an

indicator of the size of the (positive) bias. The QM at a lead-time of 6 hours is also a straight line, however with an opposite, but much smaller and negative slope (bias) in comparison to the raw C-L. With increasing lead-times the QS of the raw C-L and the QM forecasts come closer to the combined forecasts for probability levels between 0.1 and 0.5. This is most probably caused by the increased spread of the ensemble. However, for a lead-time of 24 and 48 hours the raw C-L forecasts still show the worst behaviour at higher flows, whereas the QM method performs at a lead-time of 48 hour almost as well as the combination methods, apart from the forecasts around the median.

As already stated previously, the comparison of the CRPS of the different post-processed methods and the aggregated ones (e.g. BLP) clearly identifies the advantage of combination (Fig. 5). The CRPS, i.e. the integral of the QS, for the different combination methods (Fig. 6) confirm the results of the QS. In general the results of the BLP are slightly better than the NGR and the BMA results. It seems that for those periods of lead-times, where the BLP is not superior (e.g. around 20 hours), the optimization routines had problems on convergence. However further analysis will be necessary. The comparison with the QM approach confirmed the results of Zhao et al. (2017), since the forecast quality did not show any improvements at the first lead-times because of the underdispersiveness of the raw C-L. Thus, the more complex combination by far outperforms the QM method.

5 Conclusions

Combination is an essential tool for improving the forecast quality. The different methods are all more or less equally suited. Although the BLP showed slightly better results, the straight forward application and the low computational costs of the NGR make this method an equally good alternative, at least for this case study. The parameter estimation of the BMA and the BLP could get quite time consuming and sometimes results in suboptimal solutions, which could degrade the gain of applying combination methods.

Competing interests. The authors declare that no competing interests are present

Acknowledgements. The real-time operational system for the Sihl basin is financed by the Office of Waste, Water, Energy and Air of the Canton of Zurich. This study was conducted in the framework of the Swiss Competence Center for Energy Research - Supply of Electricity (SCCER-SoE) with funding from the Commission for Technology and Innovation CTI (grant 2013.0288). MeteoSwiss is greatly acknowledged for providing all used meteorological data. The Swiss Federal Office for Environment (FOEN) provided the observed discharge data. The authors would like to thank especially Vanessa Round for proofreading.

References

Abrahart, R. J. and See, L.: Multi-model data fusion for river flow forecasting: An evaluation of six alternative methods based on two contrasting catchments, Hydrology and Earth System Sciences, 6, 655–670, 2002.

Addor, N., Jaun, S., Fundel, F., and Zappa, M.: An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): Skill, case studies and scenarios, Hydrology and Earth System Sciences, 15, 2327–2347, 2011.

Ajami, N. K., Duan, Q., and Sorooshian, S.: An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, Water Resources Research, 43, W01403, doi:doi:10.1029/2005WR004745, 2007.

Baran, S.: Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components, Computational Statistics & Data Analysis, 75, 227 – 238, doi:https://doi.org/10.1016/j.csda.2014.02.013, 2014.

Baran, S. and Lerch, S.: Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting, Quarterly Journal of the Royal Meteorological Society, 141, 2289–2299, http://dx.doi.org/10.1002/qj.2521, 2015.

Bates, J. and Granger, C.: The combination of forecasts, Operations Research Quarterly, 20, 451–468, 1969.

Bogner, K. and Pappenberger, F.: Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system, Water Resources Research, 47, W07524, doi:10.1029/2010WR009137, 2011.

Bogner, K., Liechti, K., and Zappa, M.: Post-Processing of Stream Flows in Switzerland with an Emphasis on Low Flows and Floods, Water, 8, 115, doi:10.3390/w8040115, 2016.

Bouallègue, Z. B., Pinson, P., and Friederichs, P.: Quantile forecast discrimination ability and value, Quarterly Journal of the Royal Meteorological Society, 141, 3415–3424, doi:10.1002/qj.2624, http://dx.doi.org/10.1002/qj.2624, 2015.

Bowden, G. J., Maier, H. R., and Dandy, G. C.: Real-time deployment of artificial neural network forecasting models: Understanding the range of applicability, Water Resources Research, 48, n/a–n/a, doi:10.1029/2012WR011984, w10549, 2012.

Cannon, A. J.: Quantile regression neural networks: Implementation in R and application to precipitation downscaling, Computers & Geosciences, 37, 1277 – 1284, doi:http://dx.doi.org/10.1016/j.cageo.2010.07.005, 2011.

Chernozhukov, V., Fernández-Val, I., and Galichon, A.: Quantile and Probability Curves Without Crossing, Econometrica, 78, 1093–1125, doi:10.3982/ECTA7880, 2010.

Dawid, A.: Statistical theory: The prequential approach, J. Roy. Statist. Soc. Ser. A, 147, 278–292, 1984.

Diks, C. G. H. and Vrugt, J. A.: Comparison of point forecast accuracy of model averaging methods in hydrologic applications, Stochastic Environmental Research and Risk Assessment, 24, 809–820, doi:10.1007/s00477-010-0378-z, 2010.

Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H., and Shrestha, D. L.: Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments, Hydrology and Earth System Sciences, 19, 3181–3201, doi:10.5194/hess-19-3181-2015, 2015.

Fraley, C., Raftery, A., and Gneiting, T.: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging, Monthly Weather Review, 138, 190–202, 2010.

Friederichs, P. and Hense, A.: Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression, Monthly Weather Review, 135, 2365–2378, doi:10.1175/MWR3403.1, 2007.

Glahn, H. and Lowry, D.: The use of model output statistics (MOS) in objective weather forecasting, J. Appl. Meteor., 11, 1203–1211, 1972.

Gneiting, T. and Ranjan, R.: Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules, Journal of Business & Economic Statistics, 29, 411–422, 2011.

Gneiting, T. and Ranjan, R.: Combining predictive distributions, Electron. J. Statist., 7, 1747–1782, doi:10.1214/13-EJS823, 2013.

Gneiting, T., Raftery, A., Westveld III, A., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Monthly Weather Review, 133, 1098–1118, 2005.

Gneiting, T., Balabdaoui, F., and Raftery, A.: Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society. Series B: Statistical Methodology, 69, 243–268, 2007.

Hemri, S., Fundel, F., and Zappa, M.: Simultaneous calibration of ensemble river flow predictions over an entire range of lead times, Water Resources Research, 49, 6744–6755, doi:10.1002/wrcr.20542, http://dx.doi.org/10.1002/wrcr.20542, 2013.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., and Haiden, T.: Trends in the predictive performance of raw ensemble weather forecasts, Geophys. Res. Lett., doi:10.1002/2014GL062472, 2014.

Koenker, R.: Quantile Regression, Econometric Society Monographs, Cambridge University Press, 2005.

Koenker, R. and Machado, J. A. F.: Goodness of Fit and Related Inference Processes for Quantile Regression, Journal of the American Statistical Association, 94, 1296–1310, doi:10.1080/01621459.1999.10473882, 1999.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrology and Earth System Sciences, 11, 1267–1277, 2007.

Lichtendahl, K. C. J., Grushka-Cockayne, Y., and Winkler, R. L.: Is It Better to Average Probabilities or Quantiles?, Management Science, 59, 1594–1611, doi:10.1287/mnsc.1120.1667, 2013.

Liechti, K., Oplatka, M., Eisenhut, N., and Zappa, M.: Early Flood Warning for the City of Zurich: Evaluation of real-time Operations since 2010, in: 13th Congress Interpraevent 2016, Living with natural risks, 2016.

López López, P., Verkade, J. S., Weerts, A. H., and Solomatine, D. P.: Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the upper Severn River: a comparison, Hydrology and Earth System Sciences, 18, 3411–3428, doi:10.5194/hess-18-3411-2014, 2014.

Min, C.-K. and Zellner, A.: Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates, Journal of Econometrics, 56, 89 – 118, doi:http://dx.doi.org/10.1016/0304-4076(93)90102-B, http://www.sciencedirect.com/science/article/pii/030440769390102B, 1993.

Montani, A., Cesari, D., Marsigli, C., and Paccagnella, T.: Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges, Tellus Series A - Dynamic Meteorology and Oceanography, 63, 605–624, 2011.

Quiñonero Candela, J., Rasmussen, C., Sinz, F., Bousquet, O., and Schölkopf, B.: Evaluating Predictive Uncertainty Challenge, in: Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment, edited by Quiñonero Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., vol. 3944 of *Lecture Notes in Computer Science*, pp. 1–27, Springer Berlin Heidelberg, doi:10.1007/11736790_1, 2006.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2015.

Raftery, A., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Monthly Weather Review, pp. 1155–1174, 2005.

Raftery, A. E., Madigan, D., and Hoeting, J. A.: Bayesian Model Averaging for Linear Regression Models, Journal of the American Statistical Association, 92, 179–191, doi:10.1080/01621459.1997.10473615, 1997.

Ranjan, R. and Gneiting, T.: Combining probability forecasts, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72, 71–91, doi:10.1111/j.1467-9868.2009.00726.x, 2010.

Ronco, P., Bullo, M., Torresan, S., Critto, A., Olschewski, R., Zappa, M., and Marcomini, A.: KULTURisk regional risk assessment methodology for water-related natural hazards - Part 2: Application to the Zurich case study, HYDROLOGY AND EARTH SYSTEM SCIENCES, 19, 1561–1576, 2015.

Schepen, A. and Wang, Q. J.: Model averaging methods to merge operational statistical and dynamic seasonal streamflow forecasts in Australia, Water Resources Research, 51, 1797–1812, doi:10.1002/2014WR016163, 2015.

Shamseldin, A., O'Connor, K., and Liang, G.: Methods for combining the outputs of different rainfall–runoff models, Journal of Hydrology, 197, 203–229, 1997.

Taylor, J. W.: A quantile regression neural network approach to estimating the conditional density of multiperiod returns, Journal of Forecasting, 19, 299–311, 2000.

Thompson, P. D.: How to Improve Accuracy by Combining Independent Forecasts, Monthly Weather Review, 105, 228–229, doi:10.1175/1520-0493(1977)105<0228:HTIABC>2.0.CO;2, 1977.

Todini, E.: A model conditional processor to assess predictive uncertainty in flood forecasting, International Journal of River Basin Management, 6, 123–137, 2008.

Van der Waerden, B. L.: Order tests for two-sample problem and their power I, Indagat. Math., 14, 453 – 458, 1952.

Van der Waerden, B. L.: Order tests for two-sample problem and their power II, Indagat. Math., 15, 303 – 310, 1953a.

Van der Waerden, B. L.: Order tests for two-sample problem and their power III, Indagat. Math., 15, 311 – 316, 1953b.

Vrugt, J. A. and Robinson, B. A.: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, Water Resources Research, 43, doi:10.1029/2005WR004838, 2007.

Wallis, K. F.: Combining forecasts - forty years later, Applied Financial Economics, 21, 33–41, 2011.

Wang, Q. J., Schepen, A., and Robertson, D. E.: Merging Seasonal Rainfall Forecasts from Multiple Statistical Models through Bayesian Model Averaging, Journal of Climate, 25, 5524–5537, doi:10.1175/JCLI-D-11-00386.1, 2012.

Weerts, A. H., Winsemius, H. C., and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrology and Earth System Sciences, 15, 255–265, doi:10.5194/hess-15-255-2011, 2011.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences: An Introduction, Academic Press, New York, 1995.

Williams, R. M., Ferro, C. A. T., and Kwasniok, F.: A comparison of ensemble post-processing methods for extreme events, Quarterly Journal of the Royal Meteorological Society, 140, 1112–1120, doi:10.1002/qj.2198, http://dx.doi.org/10.1002/qj.2198, 2014.

Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E., and Ramos, M.-H.: How Suitable is Quantile Mapping For Postprocessing GCM Precipitation Forecasts?, Journal of Climate, 30, 3185–3196, doi:10.1175/JCLI-D-16-0652.1, 2017.