Adv. Geosci., 29, 109–118, 2011
www.adv-geosci.net/29/109/2011/
doi:10.5194/adgeo-29-109-2011

Advances in Geosciences

Reliability of autoregressive error models as post-processors for probabilistic streamflow forecasts

M. Morawietz, C.-Y. Xu, and L. Gottschalk

Department of Geosciences, University of Oslo, P.O. Box 1047 Blindern, 0316 Oslo, Norway

Received: 20 July 2010 – Revised: 26 January 2011 – Accepted: 2 April 2011 – Published: 27 April 2011

Abstract. In this study, the reliability of different versions of autoregressive error models as post-processors for probabilistic streamflow forecasts is evaluated. Rank histograms and reliability indices are used as performance measures. An algorithm for the construction of confidence intervals to indicate ranges of reliable forecasts within the rank histograms is presented. To analyse differences in performance of the post-processors, scatter plots of the standardized residuals of the error models are generated to assess the homoscedacity of the residuals with respect to streamflow. A problem of distorted impressions may appear when such plots are generated with a regular x-scale. The problem is analysed with both synthetic and real data, and a rank scaled x-axis is proposed to remedy the problem. The results of the study reveal large differences in the reliability of the post-processors. Versions with empirical distribution functions are clearly superior to those with standard normal distribution, but for validations with independent data their rank histograms still lie outside of the confidence bands for reliable forecasts.

1 Introduction

Many studies that aim at a description of the uncertainties of a streamflow forecast do this through an explicit description of the main input uncertainties, i.e. the meteorological uncertainties (e.g. Roulin, 2007; Thielen et al., 2009; He et al. 2010; see also review on ensemble flood forecasting by Cloke and Pappenberger, 2009). Distributions of the meteorological variables, represented through a finite number of values, are transformed through a deterministic precipitation-runoff model into a distribution of simulated streamflow values that represent the forecast. However, the simulation process itself introduces further sources of uncertainty such as uncertainty in model structure, model parameters and initial states of the precipitation-runoff model (Seo et al., 2006). These other uncertainties, labelled as hydrologic uncertainties, should also be accounted for in order to obtain a well-calibrated probabilistic forecast. A treatment of the hydrologic uncertainties in a lumped form can be achieved through a hydrologic uncertainty processor (Krzysztofowicz, 1999). In the context of the forecast chain, such a hydrologic uncertainty processor can be labelled as post-processor (Seo et al., 2006).

In this study, different versions of autoregressive error models are applied as post-processors to the HBV model (Bergström, 1992). The aspects addressed with the different versions are the type of transformation applied to the original streamflow values, the formulation of the parameters of the post-processor as either constant or state dependent, and the type of distribution function used for description of the residuals. In Morawietz et al. (2011) the performance of these post-processors was evaluated with the discrete ranked probability score, an evaluation measure that mainly characterizes the sharpness of a probabilistic forecast (Carney and Cunningham, 2006). In this study, the post-processors are evaluated with respect to their reliability. Rank histograms (Talagrand et al., 1997) and reliability indices (Delle Monache et al., 2006) are used as evaluation measures for the reliability. Furthermore, the homoscedacity (i.e. equal scattering) of the residuals of the post-processors was investigated using scatter plots of standardized residuals versus simulated streamflow. Within the evaluation, two aspects are especially addressed. (1) For the rank histograms it needs to be tested if deviations from the uniform distribution are significant. As an alternative to the chi-square and similar tests (Hamill and Colucci, 1997; Jolliffe and Primo, 2008), we propose a construction of confidence intervals for the rank histograms that not only allow the detection of significant deviations but also their visualization in the plots of the rank histogram.

Correspondence to: M. Morawietz
(martin.morawietz@geo.uio.no)

Published by Copernicus Publications on behalf of the European Geosciences Union.

(2) The scaling of the x-axis in scatter plots for assessing homoscedacity has a strong influence on the visual impression and may lead to distortions and misinterpretations. We illustrate the problems with both real and synthetic data, and propose a scaling that remedies the problem.

2 Methods

2.1 The probabilistic streamflow forecast

A probabilistic streamflow forecast that accounts for both the input uncertainties and the hydrologic uncertainties can be described through the law of total probability (Krzysztofowicz, 1999):

$$\psi(o_t|y) = \int_{-\infty}^{\infty} \varphi(o_t|s_t, y)\pi(s_t|y)ds_t \quad (1)$$

ψ is the probability density function (pdf) of the observed streamflow o_t for a future time t, conditioned on a variable y (y may also stand for several variables; the specific implementation of y for this study is given in Sect. 2.2). π is the pdf of the simulated streamflow s_t that is obtained, when the distributions of forecast temperature and forecast precipitation are transformed through a deterministic precipitation-runoff model. Thus π incorporates the input uncertainties (meteorological uncertainties) but not the hydrologic uncertainties. The hydrologic uncertainties are accounted for through φ, the pdf of the observed streamflow o_t conditioned on the simulated streamflow s_t and the variable y. The density φ thus constitutes the hydrologic uncertainty processor or post-processor.

In this study, the focus is on the hydrologic uncertainty. Therefore, to remove the influence of the meteorological uncertainties, observed values of precipitation and temperature are used as input to the precipitation-runoff model as a representation of a perfect meteorological forecast. The density π of the simulated streamflow s_t will thus become a Dirac delta function $\delta(s_t)$ and the probabilistic forecast density ψ becomes equivalent to the post-processor density φ:

$$\psi(o_t|y) = \varphi(o_t|s_t, y) \quad (2)$$

2.2 Autoregressive error model as post-processor

The simulation errors of the deterministic precipitation-runoff model can be described through an autoregressive error model:

$$d_t = \alpha_t d_{t-1} + \sigma_t \varepsilon_t \quad (3)$$

The simulation error d_t is defined as difference between the transformed observed streamflow, o_t, and transformed simulated streamflow, s_t:

$$d_t = o_t - s_t \quad (4)$$

where α_t and σ_t are the parameters of the error model, and ε_t is the standardized residual error described through a random variable.

Solving the error model for the observed streamflow at a future time t yields:

$$o_t = s_t + \alpha_t(o_{t-1} - s_{t-1}) + \sigma_t \varepsilon_t \quad (5)$$

With κ being the pdf of ε_t, the density of an observed streamflow o_t conditioned on s_t, o_{t-1} and s_{t-1} is equal to the density of the value ε_t that corresponds to o_t through Eq. (5):

$$\varphi(o_t|s_t, o_{t-1}, s_{t-1}) = \kappa(\varepsilon_t) \text{ with } \varepsilon_t = \frac{(o_t - s_t) - \alpha_t(o_{t-1} - s_{t-1})}{\sigma_t} \quad (6)$$

i.e.

$$\varphi(o_t|s_t, o_{t-1}, s_{t-1}) = \kappa\left(\frac{(o_t - s_t) - \alpha_t(o_{t-1} - s_{t-1})}{\sigma_t}\right) \quad (7)$$

Equation (7) constitutes a post-processor as defined in Sect. 2.1. The variable y from Eqs. (1) and (2) is now realized through the two variables observed streamflow o_{t-1} and simulated streamflow s_{t-1} at the time $t-1$ where the forecast is generated. With Φ and K being the cumulative distribution functions (cdfs) corresponding to φ and κ respectively, it follows equivalent to Eq. (7):

$$\Phi(o_t|s_t, o_{t-1}, s_{t-1}) = \mathrm{K}\left(\frac{(o_t - s_t) - \alpha_t(o_{t-1} - s_{t-1})}{\sigma_t}\right) \quad (8)$$

The following three aspects of the post-processor are investigated:

1. Parameters: state dependent (SD) formulation of the parameters α_t and σ_t versus state independent parameters (SI).

2. Transformation: logarithmic transformation (Log) of the original values of observed streamflow, Q_{obs}, and simulated streamflow, Q_{sim},

   $$o_t = \ln(Q_{\mathrm{obs}}(t)) \quad (9)$$

   $$s_t = \ln(Q_{\mathrm{sim}}(t)) \quad (10)$$

   versus square root transformation (Sqrt)

   $$o_t = \sqrt{Q_{\mathrm{obs}}(t)} \quad (11)$$

   $$s_t = \sqrt{Q_{\mathrm{sim}}(t)} \quad (12)$$

3. Distribution: standard normal distribution (Norm) for the density κ of the standardized residuals ε_t versus an empirical distribution (Emp).

Each of the three aspect has two possible realizations, and by forming all possible combinations of the three aspects, eight versions of post-processors are generated (Table 1).

Table 1. Model versions investigated in this study.

Version	Label: Parameters.Transformation.Distribution
1	SD.Log.Norm
2	SI.Log.Norm
3	SD.Sqrt.Norm
4	SI.Sqrt.Norm
5	SD.Log.Emp
6	SI.Log.Emp
7	SD.Sqrt.Emp
8	Si.Sqrt.Emp

Table 2. Five classes of meteorological and snow states (modified from Morawietz et al., 2011).

Meteorological and snow states	$i(t)$
$T_t \leq 0\,^{\circ}\mathrm{C}$	1
$T_t > 0\,^{\circ}\mathrm{C}$ and $P_t = 0\,\mathrm{mm}$ and $\mathrm{SWE}_t \leq \mathrm{swe}_{\mathrm{tresh}}$	2
$T_t > 0\,^{\circ}\mathrm{C}$ and $P_t = 0\,\mathrm{mm}$ and $\mathrm{SWE}_t > \mathrm{swe}_{\mathrm{tresh}}$	3
$T_t > 0\,^{\circ}\mathrm{C}$ and $P_t > 0\,\mathrm{mm}$ and $\mathrm{SWE}_t \leq \mathrm{swe}_{\mathrm{tresh}}$	4
$T_t > 0\,^{\circ}\mathrm{C}$ and $P_t > 0\,\mathrm{mm}$ and $\mathrm{SWE}_t > \mathrm{swe}_{\mathrm{tresh}}$	5

2.2.1 State dependent parameter formulation

As state dependent parameter formulation, a parameter description used at the Norwegian Water Resources and Energy Directorate (NVE) is applied (Langsrud et al., 1998). State dependence of the parameters is realized in three ways:

1. Firstly, the parameters α_t and σ_t of the autoregressive error model are formulated to be linearly dependent on the transformed simulated streamflow s_t:

$$\alpha_t = a_{i(t)} + b s_t \tag{13}$$

$$\ln \sigma_t = A_{i(t)} + B s_t \tag{14}$$

2. Secondly, the parameters $a_{i(t)}$ and $A_{i(t)}$ of the linear relations can assume different values, depending on the states defined through the variables temperature T_t, precipitation P_t and simulated snow water equivalent SWE_t at time t. It is distinguished if the temperature is below or above 0 °C, if precipitation occurs or not, and if the snow water equivalent is above or below a certain threshold value $\mathrm{swe}_{\mathrm{thresh}}$; if the amount of snow is below $\mathrm{swe}_{\mathrm{thresh}}$, the catchment is assumed to behave as a snow free catchment. Through combination of the three variables, five different states $i(t)$ are distinguished (Table 2).

3. Thirdly, two different sets of parameters a_j, b, A_j, B, j=1, ..., 5, are used, depending on if the simulated streamflow at time t is above or below a flow threshold q_{thresh}.

As threshold $\mathrm{swe}_{\mathrm{thresh}}$, the average simulated snow water equivalent that corresponds to a snow cover of 10% is used. As threshold q_{thresh}, the 75-percentile of the observed streamflow of the calibration period is used.

2.2.2 Empirical distribution function

The empirical distribution function is based on the empirical standardized residuals

$$\hat{\varepsilon}_t = \frac{d_t - \alpha_t d_{t-1}}{\sigma_t} \tag{15}$$

of the error model. A set of standardized empirical residuals $\hat{\varepsilon}_g$, $g \in \{1, 2, ..., G\}$, that constitutes the empirical distribution function, is calculated from all days $g = 1, \ldots, G$ of the calibration period after the parameters of the error model have been estimated.

2.3 Evaluation

2.3.1 Rank histograms and reliability indices

The property of reliability, also known as statistical consistency (Talagrand et al., 1997), empirical validity (Carney and Cunningham, 2006) or calibration (Carney and Cunningham, 2006), describes the property of the forecast being correct in a statistical sense. That means that the predicted probabilities are in agreement with the verifying observations (Talagrand, 1997). In terms of this definition, reliability is a property that can only be evaluated for a probabilistic forecast and not for a deterministic forecast. A measure to evaluate the reliability is the rank histogram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand, 1997). For a continuous probabilistic forecast, a rank histogram is constructed as follows. The probability space [0, 1] is divided into a number of N bins, $1 \ldots N$, with equal width $1/N$. For a probability forecast with forecast distribution F, the nonexceedance probability $F(x)$ for the verifying observation x is determined. It is checked, into which bin the probability $F(x)$ is falling. Repeating this over a number of forecasts and counting the number of occurrences of $F(x)$ in each of the bins generates a discrete distribution which represents the rank histogram. For a well-calibrated forecast, the probability of $F(x)$ falling into a certain bin is equal for all bins, i.e. $1/N$. Thus the rank histogram from such forecasts constitutes a sample from a discrete uniform distribution with the categories $1 \ldots N$; apart from some random variation, a rank histogram from a calibrated forecast should be approximately uniform. For the current study, the forecast distribution F is given through the post-processor cdf Φ (Eq. 8) with the verifying observation x being equivalent to the observed streamflow o_t.

To condense the rank histogram into one numerical measure, the reliability index (Delle Monache et al., 2006) can be calculated as a summary measure for the flatness of the rank histogram. As the number of bins is the same for all rank histograms in this study (N=10), the correction factor

that accounts for different numbers of bins becomes equal to one and the reliability index is calculated as

$$\mathrm{RI} = \frac{\text{mean distance from ideal bin count}}{\text{ideal bin count}} \times 100 = \frac{\frac{1}{N}\sum_{i=1}^{N} |\mathrm{count}_i - \frac{r}{N}|}{\frac{r}{N}} \times 100 \quad (16)$$

where N is the number of bins of the rank histogram, count_i is the number of times the probability $F(x)$ of the observed variable x is falling into the i-th bin, and r is the sum of the count_i for $i = 1,\ldots,N$, i.e. the sample size or number of forecasts.

Finally, to summarize the overall performance of the different post-processors, the average reliability index over all catchments, $\overline{\mathrm{RI}}$, is calculated as:

$$\overline{\mathrm{RI}} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{RI}_c \quad (17)$$

where RI_c is the reliability index of catchment c, and C is the number of catchments.

An overall reliability might also be assessed through an average rank histogram with the bin counts pooled over all catchments. However, if the individual rank histograms have different shapes, an overlay of such histograms may mask deficiencies in the individual histograms. Therefore, an overall evaluation through averaged reliability indices seemed preferable.

2.3.2 Confidence intervals for rank histograms

To check for a given rank histogram if deviations from the uniform distribution are significant, chi square goodness of fit tests (Hamill and Colucci, 1997) or similar tests (Jolliffe and Primo, 2008) can be applied. However, in order to get a visual impression, we found it desirable to not only characterize the significance by a single test statistic and the corresponding p-value, but to visualize it through confidence intervals in the plots of the rank histograms. Therefore, an algorithm for the construction of such confidence intervals based on Monte Carlo simulations was developed. The algorithm is described as follows.

The confidence intervals are constructed for the null-hypothesis that the rank histogram belongs to a well calibrated forecast, i.e. that the sample rank histogram comes from a uniform distribution. For a rank histogram with N bins which is derived from a number of r forecasts, a confidence interval for a significance level of α is constructed as follows.

1. Generate a number of M sample histograms. Each histogram is generated by sampling r-times from a uniform distribution with the possible outcomes $1\ldots N$.

2. Determine the highest count h_m and the lowest count l_m in each histogram $m = 1\ldots M$.

3. Sort the values of h_m in descending order and the values of l_m in ascending order. Set the lower boundary of the confidence interval b_l equal to the first element of the sorted series of l_m, and the upper boundary b_u of the confidence interval equal to the first element of the sorted series of h_m.

4. Set the lower boundary b_l equal to the next element of the sorted series of l_m and the upper boundary b_u equal to the next element of the sorted series of h_m.

5. Repeat step 4 until the number of histograms for which $\mathrm{h}_m \geq \mathrm{b}_u$ or $\mathrm{l}_m \leq \mathrm{b}_l$ is equal to $\alpha^* M$.

The final values of the lower boundary b_l and upper boundary b_u constitute a confidence interval which encloses the fraction $(1-\alpha)$ of sample histograms that come from a uniform distribution. A fraction α of the histograms will have at least one bar that lies outside the confidence interval. The confidence intervals are constructed so that the fraction of histograms that exceed the upper boundary is equal to the fraction of histograms that go below the lower boundary. However, these fractions will usually not be equal to $\alpha/2$ but rather larger as a number of histograms that exceed the upper bound will also go below the lower bound.

2.3.3 Choice of x-scale in scatter plots for assessing homoscedacity

Plots and their visual evaluation are important tools for analysis of data and results in hydrology and other sciences. Often they reveal aspects and nuances much clearer than numerical measures. However, some plots may be misleading when not generated or interpreted in the right way. One case of such plots that may give a distorted picture appeared in this study when producing scatter plots to assess the homoscedacity of the standardized residuals ε_t. The problem will be illustrated in the following paragraphs with both real data and a synthetic example, and a solution to eliminate the distorting influences is given.

The term homoscedacity is defined as having the same variance. Two random variables are homoscedastic if they have the same variance, and a variable y can be described as homoscedastic with respect to another variable x, if the variance of the variable y does not vary with varying values of x. A common way for checking or illustrating this is to plot the variable y versus x and to evaluate if the scatter of y is constant over the range of x values. In this study, the homoscedacity of the standardized residuals ε_t of the post-processors with respect to the variable simulated streamflow Q_{sim} was checked using scatter plots of standardized residuals versus simulated streamflow. Figure 1a shows an example plot for the catchment Bulken for the SD.Log model. The visual impression is that the residuals are strongly non-homoscedastic and that the variance of ε_t decreases with increasing values of Q_{sim}. As the actual calculations were performed with the log-transformed values of streamflow, the

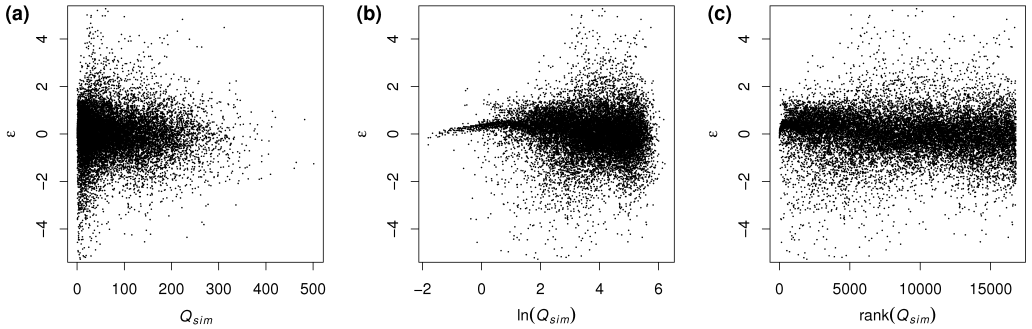

Fig. 1. Plots of standardized residuals ε_t versus simulated streamflow with different scales on the x-axis for the SD.Log model for the catchment Bulken, period 1 January 1962–31 December 2005; **(a)** original scale; **(b)** logarithmic scale; **(c)** rank scale.

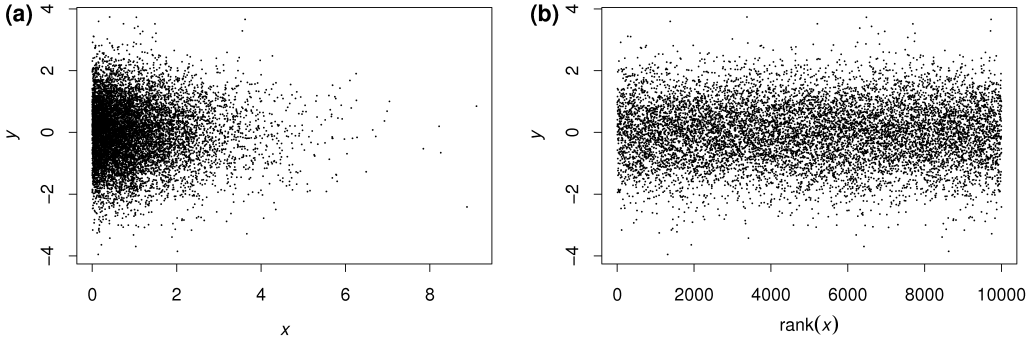

Fig. 2. Plots of the standard normally distributed variable y versus the variable x, which has a non-uniform density, with different scales on the x-axis; **(a)** original scale; **(b)** rank scale.

same plot also was generated using the log-transformed values of simulated streamflow $\ln(Q_{\text{sim}})$ on the x-axis (Fig. 1b). Surprisingly, this plot gives a very contradictory impression of the behaviour of the variance of ε_t compared to the plot in the original scale. While the variable still appears non-homoscedastic, the variance now seems to rather increase with increasing values of streamflow. The explanation for these seemingly contradictory impressions is found in the varying density of the points in x-direction in the two plots. A non-constant density of the points in x-direction distorts the picture and makes an evaluation of the variance of the variable ε_t difficult.

The effect can be illustrated with an example of synthetic data. A data set (x, y) is defined so that the variable y is homoscedastic with respect to the variable x, but the data points have a non-constant density in x-direction. The data is generated as follows.

1. A vector of y-values is generated by random sampling 10 000 times from a standard normal distribution.

2. A vector of x-values is generated by random sampling 10 000 times from a non-uniform distribution. The distribution used in this example is a gamma distribution with shape and scale parameters equal to one.

As the data in the vector of y-values are unordered random values coming all from the same distribution, any association of this data with a vector of another variable x results in a data set where y is by definition homoscedastic with respect to x. However, in a plot of y versus x (Fig. 2a), the visual impression is distorted through the decreasing density of points in x-direction, and the data appears to be strongly non-homoscedastic with seemingly decreasing variance of the variable y with increasing x-values.

In order to remedy the distorting effect of a non-constant density of the points in x-direction, a plot to check for homoscedacity should be made against the rank of the variable x instead of the original values. This generates a transformed x-scale with a constant density of the points in x-direction. Figure 2b shows such a plot for the example of synthetic data. With a constant point density in x-direction, the variance of the variable y appears as constant over the whole range of x-values, as it would be expected from the generation of the data as a homoscedastic data set. Figure 1c shows the analogue plot for the example of real-world data. The impression is very different from both the plot with the original scale (Fig. 1a) and the plot with the log-transformed scale (Fig. 1b). Though the data might not be totally homoscedastic, the non-homoscedastic behaviour seems much less pronounced than any of the other two plots would suggest.

Table 3. Overview of the five validations (from Morawietz et al., 2011).

Label	Validation type	Period for validation		Period for parameter estimation	
Cal. 1 + 2	Dependent	1 + 2	(1962–2005)	1 + 2	(1962–2005)
Cal. 1	Dependent	1	(1962–1983)	1	(1962–1983)
Cal. 2	Dependent	2	(1984–2005)	2	(1984–2005)
Val. 1	Independent	1	(1962–1983)	2	(1984–2005)
Val. 2	Independent	2	(1984–2005)	1	(1962–1983)

We may conclude that, while different scales of the x-axis may reveal different aspects of the data, for scatter plots to evaluate homoscedacity, an x-scale using the rank of the x variable is preferable in order to avoid distortions through non-constant data density in x-direction

2.4 Practical implementation of the study

The underlying deterministic precipitation-runoff model used to generate the series of simulated streamflow s_t is the HBV model (Bergström 1976, Bergström 1992). The model version used in this study is the "Nordic" HBV model (Sælthun, 1996). The model is run with daily time steps with catchment averages of mean daily air temperature T_t and accumulated daily precipitation P_t as model inputs and mean daily streamflow $Q_{\mathrm{sim}}(t)$ as model output.

Fifty-five catchments throughout Norway have been selected for the study. Basis for the selection was a common data period from 1 September 1961–31 December 2005. Catchment sizes vary from 6 to 15450 km^2, with the majority of the catchments (45) being smaller than 1000 km^2.

In each catchment, the HBV model was run for the complete period of data. The first four months of each model run were discarded as warm up period and the remaining period 1 January 1962–31 December 2005 was kept to investigate the eight post-processors.

For each of the eight post-processors, three different parameter sets were estimated from three different periods of data in each of the 55 catchments:

- Period 1: 1 January 1962–31 December 1983
- Period 2: 1 January 1984–31 December 2005
- Period 1 + 2: 1 January 1962–31 December 2005

Then the post-processors were evaluated in three dependent and two independent validations (Table 3). In the dependent validations, the post-processors were applied to the same data that was used for estimation of the parameters of the post-processors. In the independent validations, the post-processors were applied to independent data which had not been used in the parameter estimation.

For each day of a validation period, a probabilistic forecast was generated using Eq. (8). A time step of one day was used as interval between $t-1$ and t, corresponding to the time step of the precipitation-runoff model. Based on the forecasts distributions $\Phi(o_t|s_t, o_{t-1}, s_{t-1})$ and the actual observations o_t, rank histograms were generated and the corresponding reliability indices were calculated according to Eq. (16). This was done for each of the eight post-processors, version 1–8, in each of the 55 catchments and for all five validations (Table 3), i.e. altogether 2200 rank histograms and reliability indices were derived.

3 Results

Figures 3 and 4 show the rank histograms for a selected catchment (Polmak). Figure 3 shows the results for the independent validation for Period 2 (Val. 2) and Fig. 4 shows the corresponding results for the dependent validation for the same period (Cal. 2). The aspects described for the selected catchment in the following paragraphs are representative for the results received in the other catchments (either in all of them or in the majority).

The most apparent aspect is the superior reliability of models with an empirical distribution function (Figs. 3e–f and 4e–f) over models with normal distribution. For the dependent validation (Fig. 4e–f) the rank histograms are absolutely flat and thus show a perfect reliability. This is of course expected through the definition of the distribution function by the empirical residuals of the calibration period. But also for the independent validation (Fig. 3e–f) the rank histograms have a relatively flat appearance; the histograms of the different model versions 5–8 look thereby relatively similar. However, despite the much better performance over the models with normal distribution, all rank histograms in Fig. 3e–h contain bars that lie outside the 95% confidence intervals. That means that, assuming a significance level of 5%, the null-hypothesis of being well calibrated in a statistical sense has to be rejected also for models using an empirical distribution function when they are applied in an independent validation.

For post-processors with normal distribution, the results for the dependent (Fig. 4a–d) and independent (Fig. 3a–d) validation look similar for each of the model versions 1–4. All models show significant deviations from a flat histogram. The worst performance has the state independent (SI) model with square root (Sqrt) transformation (Figs. 3d

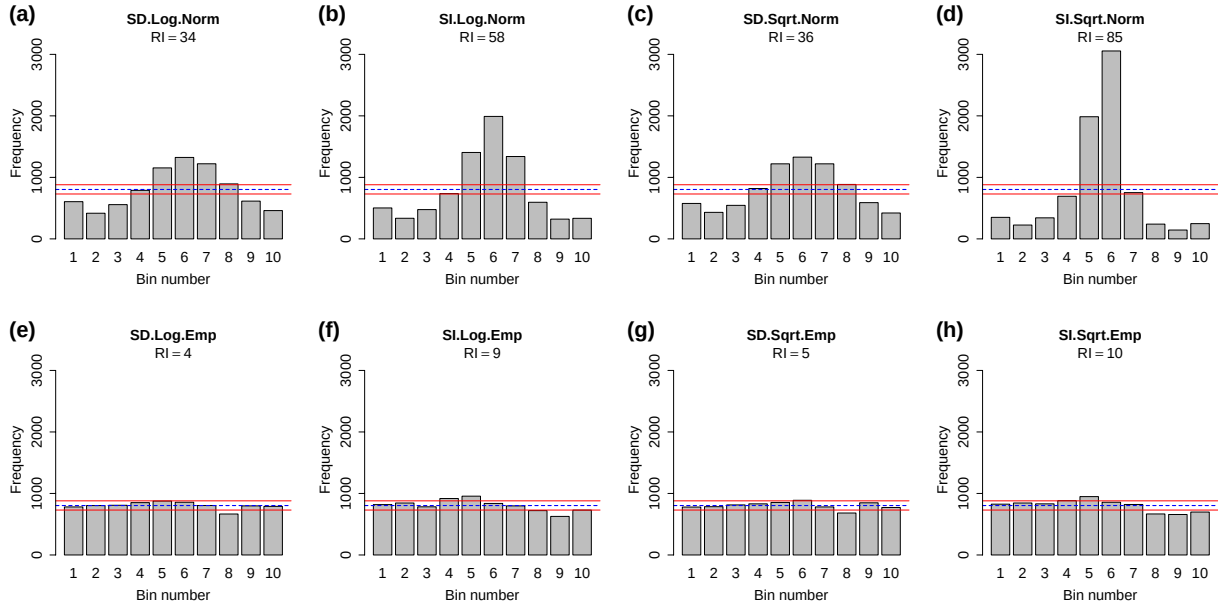

Fig. 3. Rank histograms for the independent validation in Period 2 (Val. 2) for the catchment Polmak. Red lines indicate 95% confidence intervals; dashed blue lines indicate the ideal bin count, i.e. a perfectly flat histogram.

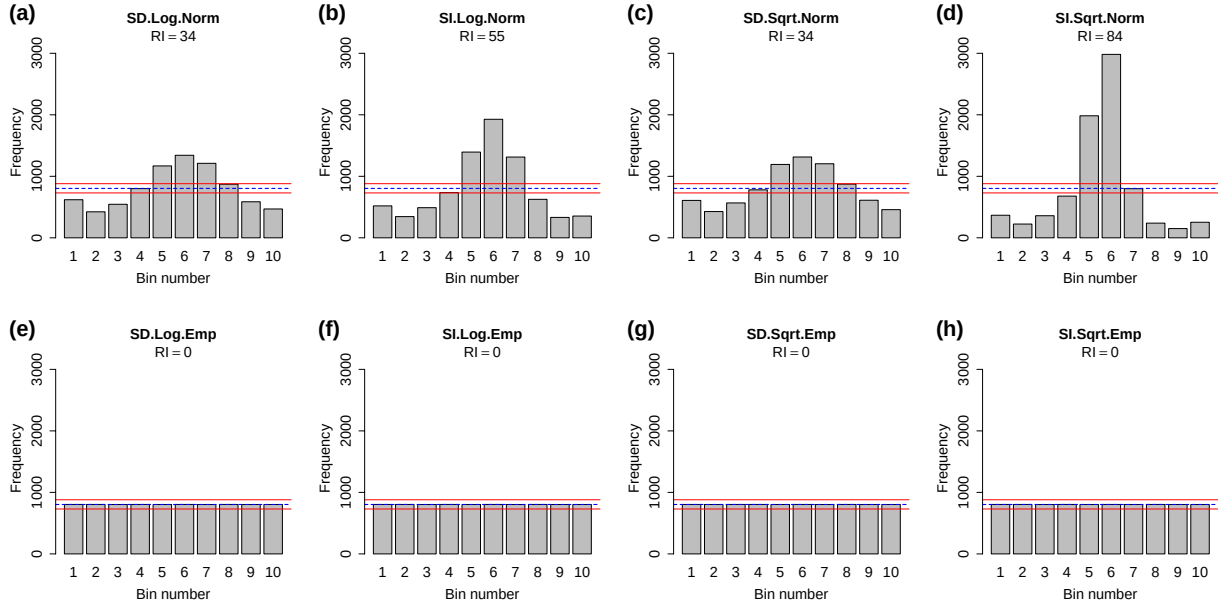

Fig. 4. Rank histograms for the dependent validation in Period 2 (Cal. 2) for the catchment Polmak. Red lines indicate 95% confidence intervals; dashed blue lines indicate the ideal bin count, i.e. a perfectly flat histogram.

and 4d), followed by the state independent (SI) model with logarithmic (Log) transformation (Figs. 3b and 4b). The state dependent (SD) models (Figs. 3a, c and 4a, c) have a better performance. The differences between SD-models with logarithmic and SD-models with square root transformation are in general small. In the example shown, the reliability index is slightly better for the model with logarithmic transformation (Figs. 3a and 4a), as it is in the majority of catchments. However for some catchments, models with square root transformation may perform slightly better instead.

Table 4 gives an overview of the number of catchments for which the null-hypothesis of a reliable forecast is rejected. The overview is for all eight post-processors and all five validations. The results are given for confidence intervals constructed according to the algorithm of Sect. 2.3.2 (numbers without brackets) and for traditional chi square goodness of

Table 4. Number of catchments for which the null-hypothesis that the rank histogram comes from a uniform distribution is rejected (significance level: 5%); the numbers without brackets are for confidence intervals constructed according to the algorithm of Sect. 2.3.2; the numbers in brackets are for traditional chi-square goodness of fit tests.

Validation	SD.Log.Norm	SI.Log.Norm	SD.Sqrt.Norm	SI.Sqrt.Norm	SD.Log.Emp	SI.Log.Emp	SD.Sqrt.Emp	Si.Sqrt.Emp
Cal. 1 + 2	55 (55)	55 (55)	55 (55)	55 (55)	0 (0)	0 (0)	0 (0)	0 (0)
Cal. 1	55 (55)	55 (55)	55 (55)	55 (55)	0 (0)	0 (0)	0 (0)	0 (0)
Cal. 2	55 (55)	55 (55)	55 (55)	55 (55)	0 (0)	0 (0)	0 (0)	0 (0)
Val. 1	55 (55)	55 (55)	55 (55)	55 (55)	54 (54)	55 (55)	53 (54)	54 (55)
Val. 2	55 (55)	55 (55)	55 (55)	55 (55)	55 (55)	55 (55)	55 (55)	54 (55)

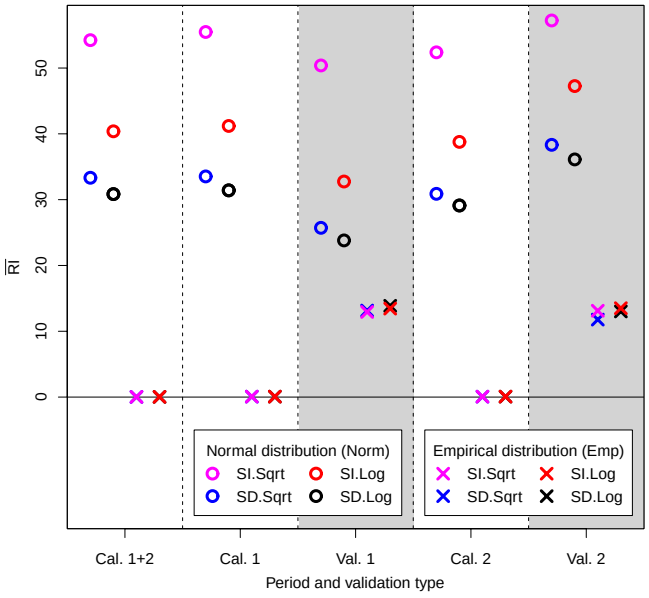

Fig. 5. Values of the average reliability indices $\overline{\text{RI}}$ (Eq. 17) for the eight post-processors for the different periods with dependent (white background) and independent (grey background) validation.

fit tests (Hamill and Colucci, 1997; numbers in brackets). The numbers for the algorithm of Sect. 2.3.2 are either identical or almost identical to the corresponding numbers for the chi square goodness of fit test. This underpins the usefulness of the new approach of Sect. 2.3.2; the approach gives equivalent results as the traditional chi square goodness of fit test while in addition it allows a visualization of confidence intervals in the rank histograms as presented in Figs. 3 and 4.

Figure 5 shows the average reliability indices $\overline{\text{RI}}$ over all catchments (Eq. 17) for the eight post-processors in the three dependent validations (white backgrounds) and two independent validations (grey backgrounds). The average reliability indices confirm the description of the model performances given for the example catchment in the paragraphs above. In addition they illustrate if and how average model performances change when going from dependent to independent validations. For the models with normal distribution (circles), there is no consistent change. Period 2 shows a certain decrease in model performance when going from the dependent (Cal. 2) to the independent (Val. 2) validation, whereas for Period 1 there is an increase in model performances. This shows that these post-processors do not have an over-fitting in the sense that the model performances would consistently deteriorate in periods with independent data. In contrast, for the models with empirical distribution (crosses) there is a consistent decrease of the performance from the dependent validation with average reliability indices of 0 to the independent validations with average reliability indices in the order of 12–14.

To investigate the differences in performance between models with logarithmic and models with square root transformation, scatter plots of the standardized residuals ε_t versus the simulated streamflow Q_{sim} were generated. A rank-scale was used for the x-axis to avoid distortion through a non-constant data density in x-direction. Figure 6 gives an example with plots for the catchment Knappom that is representative for the behaviour in most of the catchments. In Fig. 6a the standardized residuals for the state independent model with logarithmic transformation show a fairly homoscedastic behaviour, while in Fig. 6b the residuals for the corresponding model with square root transformation show an increasing variance with increasing streamflow values. For the state independent models with square root transformation the assumption of a constant standard deviation σ is therefore less justified and this is reflected in their inferior reliability indices compared to corresponding models with logarithmic transformation. For models with state dependent parameters, the formulation of σ as being dependent on the simulated streamflow (Eq. 14) and other flexibilities introduced with the state dependent formulation, can account for the more non-homoscedastic behaviour and other deficiencies that the state independent models with square root transformation might have compared to the corresponding models with logarithmic transformation. Thus the differences between reliability indices of models with logarithmic and models with square root transformation are strongly diminished for the models with state dependent parameter formulation.

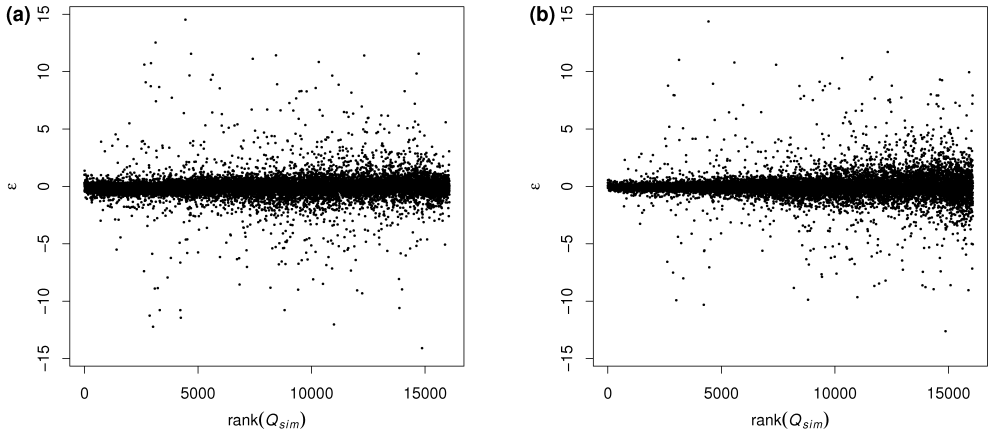

Fig. 6. Plots of the standardized residuals ε_t versus rank of the simulated streamflow $Q_{\mathrm{sim}}(t)$ for models with state independent (SI) parameters for the catchment Knappom; **(a)** model with logarithmic transformation (Log) of the original streamflow values; **(b)** model with square root transformation (Sqrt) of the original streamflow values.

4 Discussion

When using autoregressive error models as post-processors for probabilistic streamflow forecasts, the assumption of a normal distribution of the residuals leads to a relatively weak performance of the post-processors with respect to reliability. Use of an empirical distribution function instead strongly increases the reliability. However for the independent validations, the rank histograms, though relatively flat, are still significantly different from a uniform distribution. This reflects the fact that the distribution functions of the empirical residuals of the calibration period are significantly different from the distribution functions of periods with independent data. This in turn is an indication that the implicit assumption that the distribution of the residuals would be the same for all days, is not true. Rather, the distribution of the empirical residuals from the calibration period can be thought of as an overlay of different distributions for different days. As we may not find identical conditions in the period of the independent validation, the composition of the distributions of the individual days is different for the period with independent data. Thus, the resulting "pooled" distribution of the empirical residuals in the period of independent data becomes significantly different from the calibration period, resulting in rank histograms that are significantly different from the uniform distribution.

In principle, two approaches can be thought of to further improve the reliability of the post-processors with empirical distribution function for periods with independent data so that the reliability diagrams would not be significantly different from a uniform distribution. The first approach would be to modify the model structure of the error model (or possibly the precipitation-runoff model) so that the distribution function of the residuals of the autoregressive error model becomes *identical* for all days. The second approach would be to identify *different* empirical distribution functions for different conditions similar to the estimation of different model parameters for different conditions. While the first approach may be more satisfying from a scientific viewpoint, it seems more difficult to assess, especially for the modification of the precipitation-runoff model, how this approach can be realized in practice. The second approach however, though less satisfying in terms of its conciseness, may be more straightforward to investigate.

The findings of this study for the performances of the eight post-processors with respect to the average reliability indices are compatible with the performances with respect to the average ranked probability score described in Morawietz et al. (2011). More specific, the findings for corresponding model versions, i.e. model versions differing in one of the three aspects transformation, parameters or distribution function, do not contradict each other. If one model is superior over a corresponding model with respect to one score, for the other score either the same superiority is found or the models do not differ significantly, but no opposite ranking is found. Within this compatibility, the most notable differences are two fold. (1) All four post-processors with empirical distribution function, versions 5–8, give equally good performances with respect to the average reliability index, whereas some differences between these versions can be found for the ranked probability score. (2) With respect to the average reliability index, the four post-processors with empirical distribution function, versions 5–8, are superior over all of the four post-processors with normal distribution. However with respect to the discrete ranked probability score, the four post-processors with empirical distribution function, versions 5–8, are superior over the corresponding post-processor with normal distribution but not necessarily over all four post-processors with normal distribution.

5 Summary and conclusions

The reliability of eight different versions of autoregressive error models as post-processors for probabilistic streamflow forecasts was evaluated using rank histograms and reliability indices. The reliability is best for models using an empirical distribution function for description of the standardized residuals, and all models using this approach are equally reliable irrespective of the other two aspect of transformation type and parameter formulation. The confidence intervals that were constructed to indicate ranges of reliable forecasts in the rank histograms indicate however, that for independent data these post-processors still show deviations from a reliable forecast. Two approaches that might further improve the reliability of the post-processors are indicated.

The findings for the performance of the different post-processors with respect to rank histograms are consistent with the findings for performances with respect to the discrete ranked probability score. The post-processor that performs best with respect to both performance measures is a model with state dependent parameter formulation (SD) and an empirical distribution function (Emp) irrespective of the transformation type as either logarithmic or square root.

Acknowledgements. We thank the Norwegian Water Resources and Energy Directorate (NVE) for the provision of the data, the "Nordic" HBV model program and computing facilities. We further thank Thomas Skaugen and Elin Langsholt for their information on the "Nordic" HBV model code and the uncertainty procedures operational at NVE. We are also very grateful to Lukas Gudmundsson and Lena Tallaksen for their discussions and comments, and we want to thank two anonymous reviewers for their valuable comments, which helped to improve the paper.

Edited by: M. Bruen
Reviewed by: two anonymous referees

References

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Climate, 9, 1518–1530, 1996.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, SMHI RHO, 7, 1976.

Bergström, S.: The HBV model – its structure and applications, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, SMHI Reports Hydrology, 4, 35 pp., 1992.

Carney, M. and Cunningham, P.: Evaluating density forecasting models, Trinity College Dublin, Department of Computer Science, Dublin, Ireland, Computer Science Technical Report TCD-CS-2006-21, 12 pp., available at: https://www.cs.tcd.ie/publications/tech-reports/reports.06/TCD-CS-2006-21.pdf, 2006.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, J. Hydrol. 375, 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009.

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X., and Stull, R. B.: Probabilistic aspects of meteorological and ozone regional ensemble forecasts, J. Geophys. Res., 111, D24307, doi:10.1029/2005JD006917, 2006.

Hamill, T. M. and Colucci, S. J.: Verification of Eta-RSM short-range ensemble forecasts, Mon. Weather Rev., 125, 1312–1327, 1997.

He, Y, Wetterhall, F., Bao, H, Cloke, H., Li, Z., Pappenberger, F., Hu, Y., Manful, D., and Huang, Y: Ensemble forecasting using TIGGE for the July-September 2008 floods in the Upper Huai catchment: a case study, Atmos. Sci. Lett., 11, doi:10.1002/asl.270, 132–138, 2010.

Jolliffe, I. T. and Primo, C.: Evaluating rank histograms using decompositions of the chi-square test statistic, Mon. Weather Rev., 136, 2133–2139, doi:10.1175/2007MWR2219.1, 2008.

Krzysztofowicz, R.: Bayesian theory of probabilistic forecasting via deterministic hydrological model, Water Resour. Res., 35, 2739–2750, doi:10.1029/1999WR900099, 1999.

Langsrud, Ø., Frigessi, A., and Høst, G.: Pure model error of the HBV model, Norwegian Water Resources and Energy Directorate, Oslo, Norway, HYDRA note no. 4, 28 pp., 1998.

Morawietz, M., Xu, C.-Y., Gottschalk, L., and Tallaksen, L. M.: Systematic evaluation of autoregressive error models as post-processors for a probabilistic streamflow forecast system, J. Hydrol., in revision, 2011.

Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions, Hydrol. Earth Syst. Sci., 11, 725–737, doi:10.5194/hess-11-725-2007, 2007.

Seo, D.-J., Herr, H. D., and Schaake, J. C.: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction, Hydrol. Earth Syst. Sci. Discuss., 3, 1987–2035, doi:10.5194/hessd-3-1987-2006, 2006.

Sælthun, N. R.: The Nordic HBV Model, Norwegian Water Resources and Energy Administration, Oslo, NVE Publication no. 7, 26 pp., 1996.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Proceedings of the ECMWF Workshop on Predictability, Reading, UK, 20–22 October 1997, 1–25, available at: http://www.ecmwf.int/publications/library/ecpublications/_pdf/workshop/1997/predictability/ws_predictability_talagrand.pdf, 1997.

Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System – Part 1: Concept and development, Hydrol. Earth Syst. Sci., 13, 125–140, doi:10.5194/hess-13-125-2009, 2009.