Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: www.tandfonline.com/journals/thsj20

Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency

Sandra Pool, Marc Vis & Jan Seibert

To cite this article: Sandra Pool, Marc Vis & Jan Seibert (2018) Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, Hydrological Sciences Journal, 63:13-14, 1941-1953, DOI: 10.1080/02626667.2018.1552002

To link to this article: https://doi.org/10.1080/02626667.2018.1552002

HYDROLOGICAL SCIENCES JOURNAL
2018, VOL. 63, NOS. 13–14, 1941–1953
https://doi.org/10.1080/02626667.2018.1552002

Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency

Sandra Pool [a], Marc Vis [a] and Jan Seibert [a,b]

[a]Department of Geography, University of Zurich, Zurich, Switzerland; [b]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

ABSTRACT

Goodness-of-fit measures are important for an objective evaluation of runoff model performance. The Kling-Gupta efficiency (R_{KG}), which has been introduced as an improvement of the widely used Nash-Sutcliffe efficiency, considers different types of model errors, namely the error in the mean, the variability, and the dynamics. The calculation of R_{KG} is implicitly based on the assumptions of data linearity, data normality, and the absence of outliers. In this study, we propose a modification of R_{KG} as an efficiency measure comprising non-parametric components, i.e. the Spearman rank correlation and the normalized flow–duration curve. The performances of model simulations for 100 catchments using the new measure were compared to those obtained using R_{KG} based on a number of statistical metrics and hydrological signatures. The new measure resulted overall in better or comparable model performances, and thus it was concluded that efficiency measures with non-parametric components provide a suitable alternative to commonly used measures.

ARTICLE HISTORY
Received 8 March 2018
Accepted 13 November 2018

EDITOR
S. Archfield

ASSOCIATE EDITOR
A. Efstratiadis

KEYWORDS
runoff modelling; calibration; non-parametric; multi-objective; Kling-Gupta efficiency

Introduction

Runoff models are important tools in hydrology. Their application requires some form of parameter estimation to ensure reliable discharge simulations for the catchment of interest. Parameter estimation is often based on comparing simulated and observed discharge using a goodness-of-fit measure, also called an objective function. The most widely used objective function in hydrological modelling is the model efficiency (Nash and Sutcliffe 1970), which is based on the mean squared error. The mean squared error between simulated and observed discharge can be decomposed into the three components mean, variability, and dynamics (Murphy 1988, Gupta *et al.* 2009). Estimating model parameters by optimizing the mean squared error is critical in two ways. Gupta *et al.* (2009) demonstrated that a high model performance for discharge dynamics is inevitably related to an underestimation of discharge variability and that the importance of discharge volume in model calibration depends on a catchment's discharge variability. This motivated them to suggest an objective function (the so called Kling-Gupta model efficiency, R_{KG}), which is based on an improved combination of the three diagnostically meaningful components of the mean squared error.

The Kling-Gupta model efficiency is in line with the paradigm of using multiple objectives for model calibration with the aim of preventing an overfitting of model parameters to a particular hydrograph aspect (some early studies are Lindström 1997, Gupta *et al.* 1998, Boyle *et al.* 2000, Madsen 2003). Taking into account multiple objectives can reduce simulation uncertainties and provides more reliable predictions given that the individual objectives are uncorrelated (Efstratiadis and Koutsoyiannis 2010). Multi-objective functions were originally composed mostly of purely statistical metrics, such as the root mean squared error of low, high or peak flows (see review of Efstratiadis and Koutsoyiannis 2010). In more recent years, hydrological signatures were applied as multi-objective functions (Yilmaz *et al.* 2008, Hingray *et al.* 2010, Euser *et al.* 2013, Zhang *et al.* 2016, Kiesel *et al.* 2017, Shafii *et al.* 2017) with the aim of focusing model calibration on relevant hydrograph aspects or major catchment functions. The term multi-objective function can also refer to multiple variables or multiple sites within a catchment (Madsen 2003). In this study, however, we used only discharge time series for calibration.

The calculation of R_{KG} is implicitly based on the assumptions of data linearity and normality, as well as

CONTACT Sandra Pool sandra.pool@geo.uzh.ch
Supplementary data for this article can be accessed here.

the absence of outliers. However, discharge time series and model simulation errors are known to be highly skewed, which violates the implicit assumptions underlying R_{KG}. The aim of this study was therefore to make a step towards using non-parametric efficiency measures by reformulating the variability and the correlation term of R_{KG} in a non-parametric form. For a non-parametric alternative to the standard deviation, we decided to use the flow–duration curve (FDC). The FDC describes the relationship between the frequency and magnitude of streamflow and is an indicator of flow variability across all flow magnitudes of a catchment (Vogel and Fennessey 1995), whereas the standard deviation is, in cases of non-normally distributed data, only a metric for the variability of flows around the mean flow. Since catchment characteristics such as flashiness or baseflow can be linked to specific segments of the FDC, it has become a widely used signature for model calibration (Yilmaz *et al.* 2008, Westerberg *et al.* 2011, Pokhrel *et al.* 2012, Euser *et al.* 2013, Pfannerstill *et al.* 2014, Garcia *et al.* 2017). As proposed by Legates and McCabe (1999), we used the Spearman rank correlation to describe discharge dynamics instead of the Pearson correlation coefficient as is used by R_{KG}. Spearman rank correlation is less sensitive to extreme values in a time series than Pearson correlation and is therefore less prone to artificially high correlation values, leading to a more robust characterization of the correlation (Legates and McCabe 1999, Krause *et al.* 2005). Spearman rank correlation is, just as the Pearson correlation, insensitive to additive and proportional differences between simulated and observed discharge (Legates and McCabe 1999), which stresses the importance of the volume term in R_{KG}. To our knowledge, the Spearman rank correlation has only been used in a limited number of calibration studies (Vis *et al.* 2015, Seibert and Vis 2016).

In this study, we propose a modification of R_{KG} towards a non-parametric calibration criterion (R_{NP}) that is composed of the mean discharge, the FDC, and the Spearman rank correlation. Model calibrations with R_{KG}, R_{NP}, and different combinations of their mean, variability, and dynamic components were evaluated by comparing the model performance for a number of selected hydrograph aspects. The goal was to evaluate the potential of non-parametric formulations of goodness-of-fit measures for runoff model calibrations aiming at multiple hydrograph aspects.

Data and methods

Study area

This study was based on model applications in 100 catchments located across the contiguous United States (Fig. 1). The catchments are a subset of the Newman *et al.* (2015) dataset and were selected by stratified random sampling from the drainage area of the major river regions proportional to the number of gauged catchments in these river regions. The Newman *et al.* (2015) dataset provides daily temperature, precipitation, and discharge data along with catchment outline information for over 600 catchments in the United States with minimal human disturbance. Monthly potential evaporation was estimated using

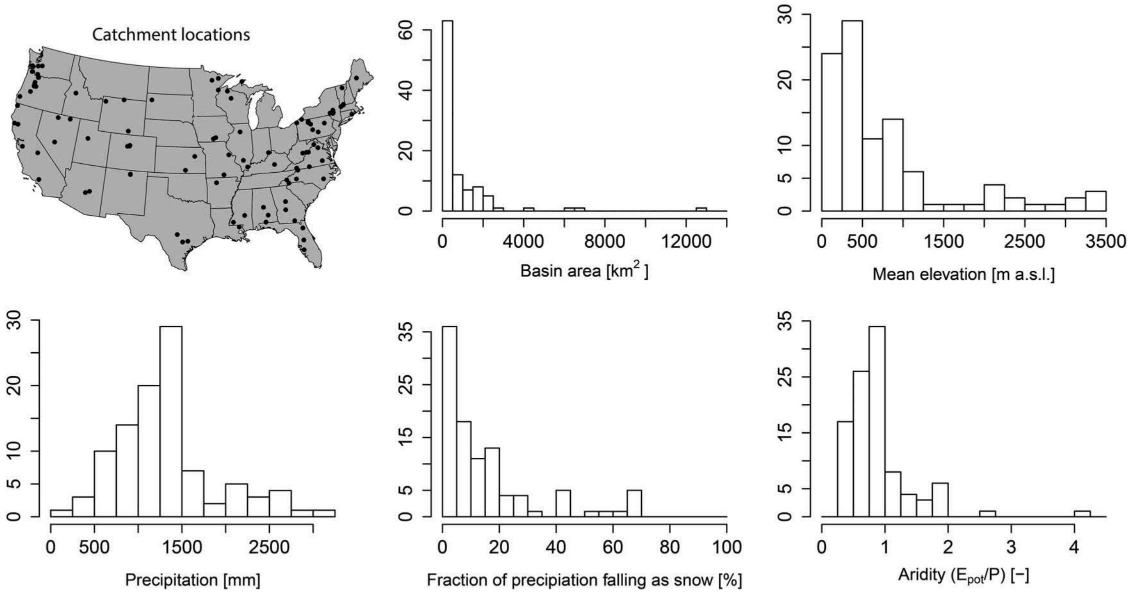


Figure 1. Locations and hydroclimatic characteristics of the 100 study catchments.

the Priestley-Taylor equation, for which the required input data were also extracted from the Newman *et al.* (2015) dataset. The catchment areas range from 10 to 12 630 km^2 with a median of 340 km^2. Mean catchment elevations are between 25 and 3355 m a.s.l. Annual precipitation sums vary from 240 to 3070 mm, of which more than 15% falls as snow in a third of the catchments (based on the time period from 1990 to 2009). Of the study catchments, 43% can be classified as humid, 40% as temperate, and 17% as arid (classification according to Coopersmith *et al.* 2014). The wide range of catchment areas and hydroclimatic conditions of the selected catchments ensured that a large variability in runoff processes were represented among the study catchments.

The HBV runoff model

The HBV runoff model (Hydrologiska Byråns Vattenbalansavdelning, Bergström 1976, Lindström *et al.* 1997) in the version HBV-light (Seibert and Vis 2012) was used to test the influence of the different objective functions on simulated runoff. The runoff model has been successfully applied in many different hydroclimates (e.g. Häggström *et al.* 1990, Lidén and Harlin 2000, Perrin *et al.* 2001, Beck *et al.* 2016, Seibert and Vis 2016). The HBV model is a bucket-type runoff model with a conceptual representation of runoff processes at the catchment scale. The model consists of four routines representing snow, soil water, groundwater, and stream network routing. Daily temperature and precipitation are input to the snow routine, where snow accumulation and melt are calculated with a degree-day method. Snowmelt and rainfall supply the soil moisture storage from which, together with monthly potential evaporation, the actual evaporation and the groundwater recharge are computed. Groundwater storage is represented by a shallow and a deep reservoir from which the fast runoff response, intermediate runoff response, and baseflow are calculated. These three runoff components are summed and transformed by a triangular weighting function to simulate the hydrograph at the catchment outlet.

The HBV model was applied in a semi-distributed configuration by dividing the catchment into elevation bands of 200 m with separate computations for the snow and soil routines. Temperature and precipitation inputs to the elevation bands were calculated with a lapse rate of −0.6°C per 100 m and 10% per 100 m, respectively. Potential evapotranspiration was assumed to be uniform over the entire catchment. Elevation bands were determined using SRTM elevation data (Shuttle Radar Topography Mission, Jarvis *et al.* 2008).

Model calibration criteria

In this study, multiple model calibration criteria were defined based on the decomposition of the mean squared error into the three aspects, mean (β), variability (α), and dynamics (i.e. correlation r; Murphy 1988, Gupta *et al.* 2009).

The three terms β, α, and r were first calculated as originally proposed by Gupta *et al.* (2009), (Equations (1)–(3)). The bias between simulated (sim) and observed (obs) mean discharge μ and the bias between simulated and observed standard deviation σ was used to compute β and α_{KG}, respectively. The Pearson correlation between observed and simulated discharge time series Q with length n was used as the indicator for discharge dynamics (r_{p}). Together the three parametric components β, α_{KG}, and r_{p} were input to the Kling-Gupta efficiency R_{KG} (Equation (4)):

$$\beta = \frac{\mu_{\mathrm{sim}}}{\mu_{\mathrm{obs}}} \tag{1}$$

$$\alpha_{\mathrm{KG}} = \frac{\sigma_{\mathrm{sim}}}{\sigma_{\mathrm{obs}}} \tag{2}$$

$$r_{\mathrm{p}} = \frac{\sum_{i=1}^{n}\left(Q_{\mathrm{obs}}(i)-\mu_{\mathrm{obs}}\right)\left(Q_{\mathrm{sim}}(i)-\mu_{\mathrm{sim}}\right)}{\sqrt{\left(\sum_{i=1}^{n}\left(Q_{\mathrm{obs}}(i)-\mu_{\mathrm{obs}}\right)^2\right)\left(\sum_{i=1}^{n}\left(Q_{\mathrm{sim}}(i)-\mu_{\mathrm{sim}}\right)^2\right)}} \tag{3}$$

$$R_{\mathrm{KG}} = 1-\sqrt{(\beta-1)^2+(\alpha_{\mathrm{KG}}-1)^2+(r_{\mathrm{p}}-1)^2} \tag{4}$$

To make a step towards a non-parametric variant of R_{KG}, the terms α and r were furthermore expressed in a non-parametric way. The non-parametric form of the discharge variability (α_{NP}) was built on the FDC. The FDC was normalized to remove the volume information and keep only the distribution signal. The absolute error was then computed between all ranked simulated and observed discharge values (Equation (5); where $I(k)$ and $J(k)$ are the time steps when the kth largest flow occurs within the simulated and observed time series, respectively). For a non-parametric alternative to the correlation term, the Spearman rank correlation (r_{s}) was calculated on the ranks of the observed (R_{obs}) and simulated (R_{sim}) discharge time series (Equation (6)). The combination of β, α_{NP}, and r_{s} into a single metric resulted in the partly non-parametric objective function R_{NP} (Equation (7)). An R-script with the calculation for R_{NP} is provided in the supplementary material.

$$\alpha_{\mathrm{NP}} = 1 - \frac{1}{2}\sum_{k=1}^{n}\left|\frac{Q_{\mathrm{sim}}(I(k))}{n\bar{Q}_{\mathrm{sim}}} - \frac{Q_{\mathrm{obs}}(J(k))}{n\bar{Q}_{\mathrm{obs}}}\right| \tag{5}$$

$$r_s = \frac{\sum_{i=1}^{n}(R_{\mathrm{obs}}(i) - \bar{R}_{\mathrm{obs}})(R_{\mathrm{sim}}(i) - \bar{R}_{\mathrm{sim}})}{\sqrt{\left(\sum_{i=1}^{n}(R_{\mathrm{obs}}(i) - \bar{R}_{\mathrm{obs}})^2\right)\left(\sum_{i=1}^{n}(R_{\mathrm{sim}}(i) - \bar{R}_{\mathrm{sim}})^2\right)}} \tag{6}$$

$$R_{\mathrm{NP}} = 1 - \sqrt{(\beta - 1)^2 + (\alpha_{\mathrm{NP}} - 1)^2 + (r_s - 1)^2} \tag{7}$$

Overall, the components β, α, and r used in their parametric and non-parametric variants built the foundation for the various one-, two-, and three-component objective functions used in this study (Fig. 2):

(1) One-component objective functions were defined so that each consisted of a single variable from R_{NP} (R_β, R_α, and R_r).

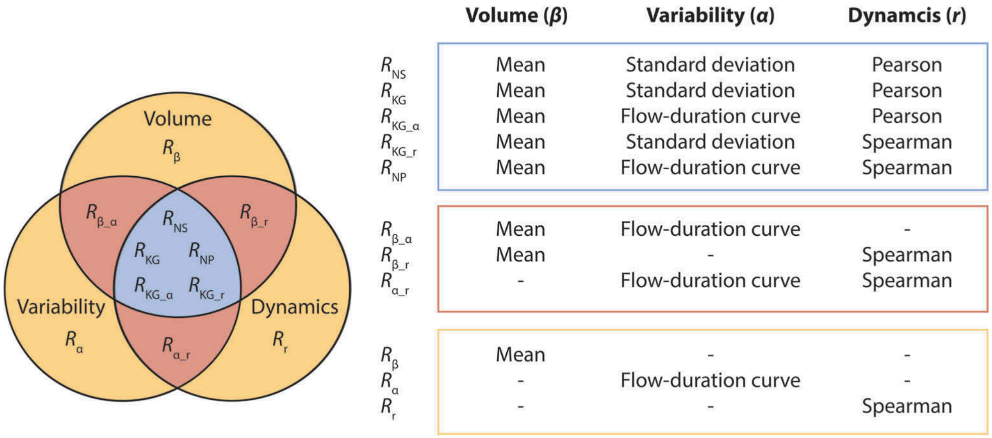
(2) Two-component objective functions consisted of two equally weighted variables from R_{NP} (R_{β_α}, R_{β_r}, and R_{α_r}).

(3) For the three-component objective functions we used β and both parametric and non-parametric variants of α and r. The Nash-Sutcliffe model efficiency (R_{NS}), the Kling-Gupta model efficiency (R_{KG}) and its non-parametric modification (R_{NP}) were assigned to this third group of objective functions. To complement R_{KG} and R_{NP}, two further objective functions were introduced where either α (R_{KG_α}) or r (R_{KG_r}) was modified to be non-parametric. These two versions were used to analyse the effect of each of the individual modifications that were made to R_{KG} in order to generate R_{NP}. Similar to R_{KG} and R_{NP}, the multiple components of R_{KG_α} and R_{KG_r} were combined using the Euclidean distance measure (Equations (4) and (7)).

Model calibration and evaluation

The HBV model was calibrated against the continuous daily discharge time series of the hydrological years 1990 to 1999 for each of the 100 study catchments. Model parameters were optimized within predefined parameter ranges using a genetic algorithm (Seibert 2000) and each of the 11 objective functions (Fig. 2). To consider parameter uncertainty, the parameter optimization was performed 100 times. The model calibration process resulted in an ensemble of 100 calibrated parameter sets for each catchment and objective function. These parameter sets were additionally used to simulate discharge for a validation period (1 October 2000 to 30 September 2009). For both calibration and validation a two-year warming-up period was used to ensure suitable initial values for the state variables.

Model simulations in calibration and validation were evaluated in three ways. First, we evaluated hydrograph uncertainty related to the use of different objective functions. The spread between the 100 simulated hydrographs of each catchment was used as information on how well an objective function constrains model parameters for a particular catchment. To evaluate this spread in simulated hydrographs, we computed the difference between the 0.05 and 0.95 quantiles of the 100 simulated hydrographs at each time step in the calibration time period. The difference was then normalized by the observed discharge and evaluated for different discharge quantiles to see if simulation uncertainty differed for different flow conditions. Hydrograph uncertainty was evaluated for simulations based on the objective functions R_{KG}, R_{KG_β}, R_{KG_α}, and R_{NP}.



	Volume (*β*)	Variability (*α*)	Dynamcis (*r*)
R_{NS}	Mean	Standard deviation	Pearson
R_{KG}	Mean	Standard deviation	Pearson
R_{KG_α}	Mean	Flow-duration curve	Pearson
R_{KG_r}	Mean	Standard deviation	Spearman
R_{NP}	Mean	Flow-duration curve	Spearman
R_{β_α}	Mean	Flow-duration curve	-
R_{β_r}	Mean	-	Spearman
R_{α_r}	-	Flow-duration curve	Spearman
R_β	Mean	-	-
R_α	-	Flow-duration curve	-
R_r	-	-	Spearman

Figure 2. Objective functions used for model calibration. The basic components describing discharge volume (β), variability (α), and dynamics (r) are combined into 11 one-, two- or three-component objective functions.

Second, the 100 simulated hydrographs of each catchment were evaluated in terms of R_{KG}, R_{NP}, and the (non-)parametric β, α, and r components. The further analysis was based on the median of the 100 efficiencies of each catchment. The median efficiencies from all catchments were used to compute cumulative distribution functions for each evaluation metric. Furthermore, we were interested in to what extent R_{KG}, R_{NP}, and their (non-)parametric β, α, and r components are correlated with each other. Therefore, the Spearman rank correlation was calculated for different pairs among the three components (β, α, and r), R_{KG}, and R_{NP} for simulations in the calibration period.

Lastly, model performance for simulations with each objective function was evaluated for three commonly used statistical metrics and five hydrograph signatures that were not explicitly considered in calibration. These statistical metrics were the model efficiency calculated for peak flows (R_{NS_peak}), model efficiency calculated on logarithmic flow (R_{NS_logQ}), and R_{MARE}, a measure for low flows (1 minus the mean absolute relative error between observed and simulated flow). The chosen hydrograph signatures provide information on the major catchment functions by linking rainfall input to the flow response of a catchment (Yilmaz *et al.* 2008). The five signatures are the percent bias in runoff ratio (B_{rr}), the difference in watershed lag time (B_t), the percent bias in the high-flow segment of the FDC (B_{hf}), the slope of the mid-flow segment of the FDC (B_{FDC}), and the low-flow segment of the FDC (B_{lf}). The signatures were calculated according to Yilmaz *et al.* (2008), except for the watershed lag time, where only the difference in observed and simulated lag time, and not the percent bias, was calculated. Throughout this study, we always evaluated the absolute values of the percent bias or the absolute values of the difference between signatures. To statistically quantify the different effect of R_{KG} and R_{NP} on statistical metrics and signatures, we conducted a Wilcoxon signed-rank test (Wilcoxon 1945) using the median efficiency of each of the 100 study catchments.

Results

Evaluation of model simulations for R_{KG}, R_{NP} and their components

The calibrated model simulations in general reproduced the observed hydrographs reasonably well. The 100 simulated hydrographs resulting from calibration with R_{KG} and R_{NP} for two example catchments, one snow and one winter-rain dominated, indicated that, independent of a catchment's runoff regime, the range of model simulations (0.05 to 0.95 quantile) is generally wider for simulations based on R_{KG} than on R_{NP} (Fig. 3). This observation was confirmed by the results from all 100 catchments (Fig. 4). The difference in the range of simulated hydrographs was especially pronounced during recession periods and low-flow conditions. At exceptionally high peak flows (0.95 flow quantile), however, simulation uncertainty for calibrations with R_{NP} exceeded those of calibrations with R_{KG}. Simulation uncertainty resulted from the interplay between both the variability and the dynamics measure of R_{KG} and R_{NP} (Fig. 4). While variability and dynamics comparably influenced the simulation uncertainty for mean-flow conditions, their individual effect varied for low and high flows. During low-flow conditions, simulation uncertainty was most strongly influenced by the dynamics component of R_{KG} and R_{NP}, whereby the sensitivity of the Pearson correlation coefficient for high discharge values resulted in less confined simulations during low flows. At high-flow conditions, it was the described sensitivity of the Pearson correlation coefficient and the use of the FDC that reduced the range in simulated hydrographs.

The median model efficiencies of the 100 hydrograph simulations for each study catchment are presented in Figure 5. As expected, model efficiencies for R_{KG} and R_{NP} decreased when moving from calibration (median R_{KG} 0.86 and median R_{NP} 0.85) to validation (median R_{KG} 0.77 and median R_{NP} 0.80). This decrease was more pronounced for the objective function the model was calibrated on. Interestingly, the discharge variability measured in terms of the standard deviation was underestimated for calibrations based on R_{NP} in 80% of the catchments, as opposed to an almost equal fraction of catchments being under- and overestimated for calibrations with R_{KG}. Hydrograph dynamics, measured in terms of the Pearson correlation coefficient, were well represented by simulations calibrated with R_{KG}. However, the same model calibrations performed relatively poorly in terms of the Spearman rank correlation coefficient, stressing the stronger sensitivity of the Pearson correlation to discharge extremes than to discharge dynamics.

The individual components of a multi-objective function should ideally be uncorrelated to have a high information value for model calibration (Efstratiadis and Koutsoyiannis 2010). Table 1 shows that this requirement is only partly met for R_{KG} and R_{NP}. The rank correlation between r and β could be considered as weak, whereas it was strong between the r and α components and moderate to strong between α and β. The correlation between the multi-objective function (R_{KG} or R_{NP}) and its individual components (α, β, and r) is an

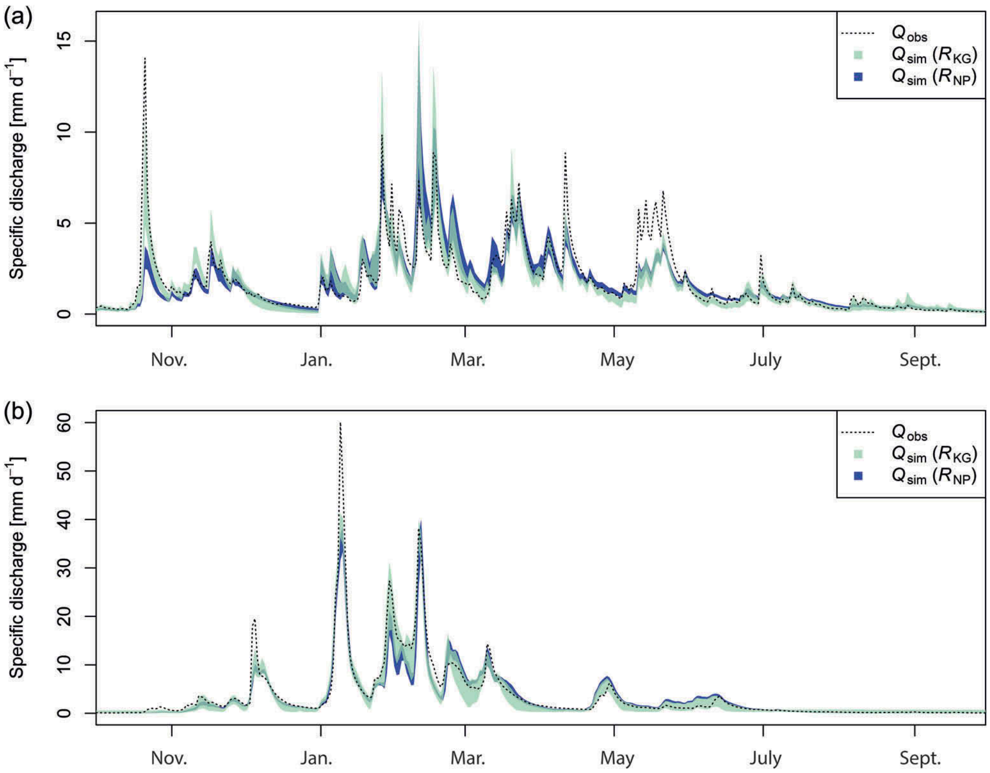


Figure 3. Observed and simulated hydrographs from model calibrations with R_{KG} and R_{NP} for (a) a snow-dominated catchment in the northeast (USGS gauge id 01423000) and (b) a winter-rain-dominated catchment in the northwest (USGS gauge id 14301000) of the United States. The range in hydrograph simulations indicates the 0.05 to 0.95 quantile of all 100 simulations.

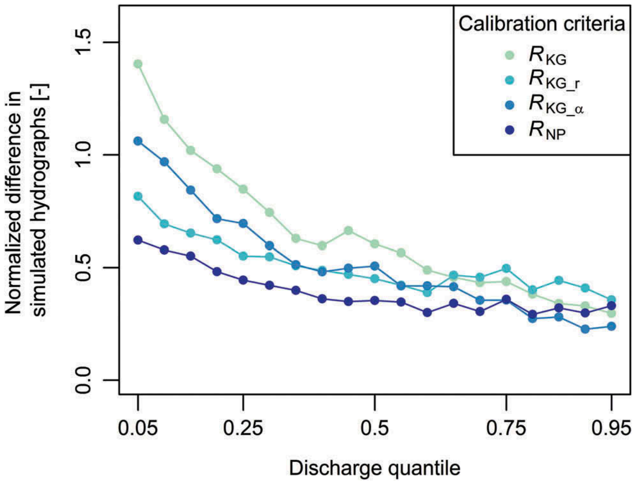


Figure 4. Hydrograph uncertainty for model calibrations with R_{KG}, R_{KG_r}, R_{KG_α}, and R_{NP}. For each catchment, uncertainty was calculated as the difference between the 0.05 and 0.95 quantiles of the 100 hydrograph simulations at a particular point in time normalized by the observed discharge. Uncertainty was computed for various discharge quantiles. Here, the median uncertainty over all 100 study catchments is presented.

indicator for the strength of their relation. Hydrograph dynamics were strongly related to the efficiency score of the multi-objective function, followed by the discharge variability component and the volume component, which was not necessarily well simulated when model efficiencies R_{KG} and R_{NP} were good.

Effect of a stepwise modification of R_{KG} to R_{NP} on statistical metrics and hydrological signatures

The stepwise modification of the variability and correlation components of R_{KG} gives an indication of their individual effect on the model calibration with R_{NP} (Fig. 6). Statistical metrics (Fig. 6(a)) measuring model performance related to the magnitude and timing of high flows (R_{NS} and R_{NS_peak}) were better simulated with R_{KG} than R_{NP}. Adapting the variability component of R_{KG} by introducing the FDC led to negligible changes in model performance, whereas the replacement of the Pearson correlation by the Spearman rank correlation clearly impaired the timing and magnitudes of high flows. In contrast, the non-parametric variants of the variability and correlation components strongly improved the model performance for low-flow measures (R_{NS_logQ} and R_{MARE}), with the highest positive effect when changing both components simultaneously (R_{NP}). Similar effects as described for the statistical metrics could be

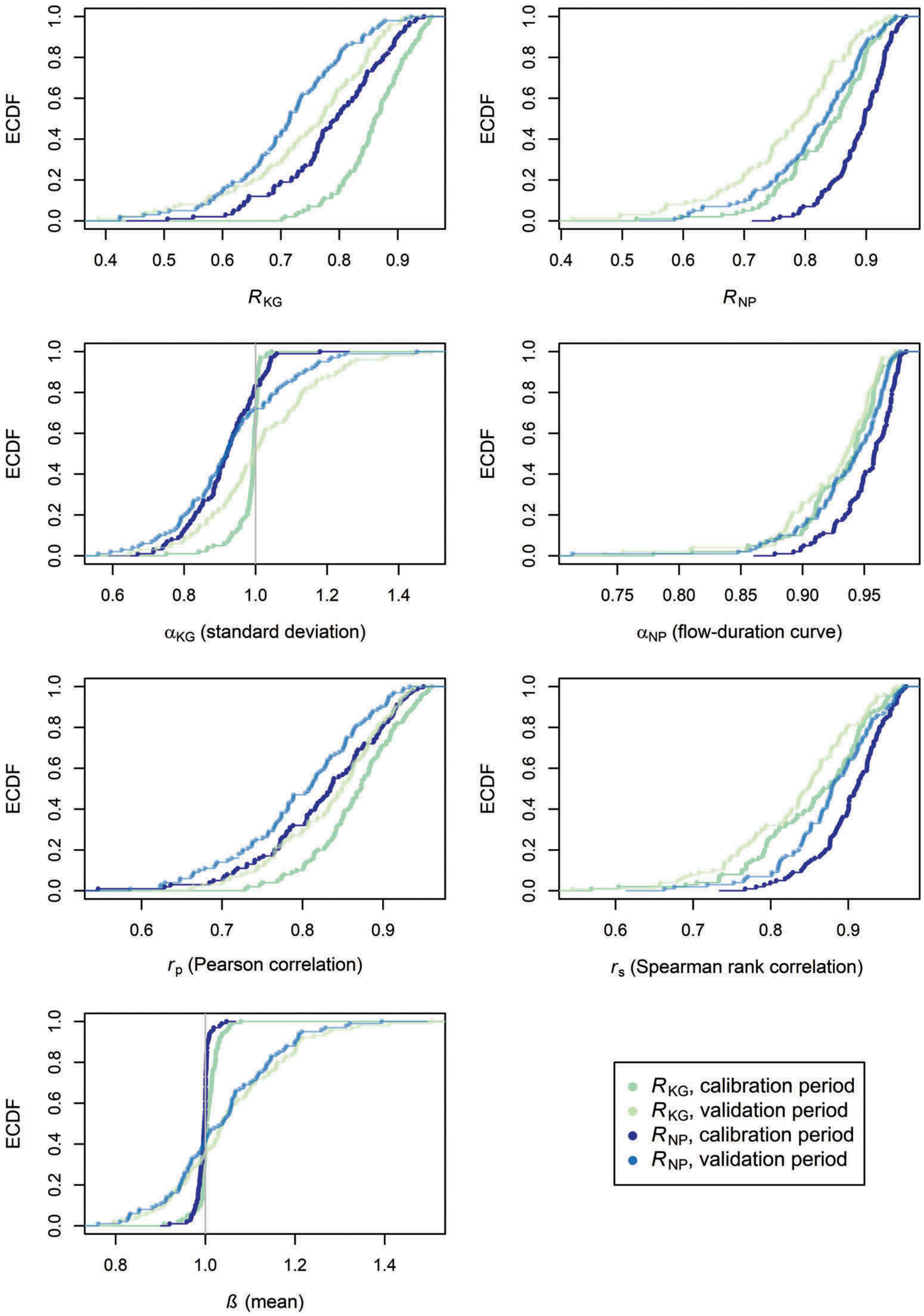


Figure 5. Model efficiencies (R_{KG}, R_{NP}, α_{KG}, α_{NP}, r_p, r_s, and β) in calibration and validation for model calibrations with R_{KG} and R_{NP}. Empirical cumulative distribution curves (ECDF) consist of the median model efficiency for each of the 100 study catchments.

observed for the high- and low-flow related hydrograph signatures (Fig. 6(b)). The two signatures runoff ratio and watershed lag time were not much affected by changes in the variability and correlation components.

Ranking the objective functions R_{KG}, R_{KG_r}, R_{KG_α}, and R_{NP} (Fig. 6(c)) according to their model performance provides a generalized picture of their effect on various hydrograph characteristics. The ranking highlights that the introduction of the non-parametric variant of the variability component (R_{KG_α}) often resulted in better simulations in comparison to R_{KG}. The non-parametric variant of R_{KG} could be considered as a valuable alternative for R_{KG}, unless timing and magnitude of high flows were of major importance.

Table 1. Spearman rank correlation coefficients for R_{KG} and R_{NP} and their three components in calibration. The correlation coefficients were calculated using the median values of all 100 catchments.

	R_{KG}				R_{NP}			
	R_{KG}	β	α_{KG}	r_p	R_{NP}	β	α_{NP}	r_s
$R_{KG/NP}$	1	0.42	0.58	0.97	1	0.27	0.71	0.97
β		1	0.61	0.3		1	0.34	0.21
$\alpha_{KG/NP}$			1	0.48			1	0.59
$r_{p/s}$				1				1

The Wilcoxon signed-rank test revealed that the model efficiency for statistical metrics and signatures (Fig. 6) differed significantly for calibrations with R_{KG} and R_{NP}, except for the signatures B_{rr}, B_{hf}, and B_{lf}.

Effect of the number of components on statistical metrics and signatures

Figures 7 and 8 present the results for model calibrations with nine objective functions consisting of a varying number of components for all catchments. For most statistical metrics and hydrograph signatures, performance increased with an increasing number of components. Especially the loss of information on dynamics (by excluding the correlation component) negatively affected model performance in the two-component objective function. In the case of the one-component objective functions, hydrograph dynamics were most important for model calibration, followed by the information on discharge variability. Model calibrations on volume only resulted in the poorest model simulations consistently throughout all evaluation metrics. Altogether, these results indicate that capturing all three components, i.e. discharge volume, variability, and dynamics of a catchment, is important for simulations aiming at multiple aspects of the hydrograph.

There are, of course, exceptions that do not follow the general observation made above. For example, the slope of the FDC (B_{FDC}) was best simulated when the α component, expressed in terms of the FDC, had a relatively high weight in calibration, which was not the case for calibrations with a three-component objective function. Another exception is the percent bias in

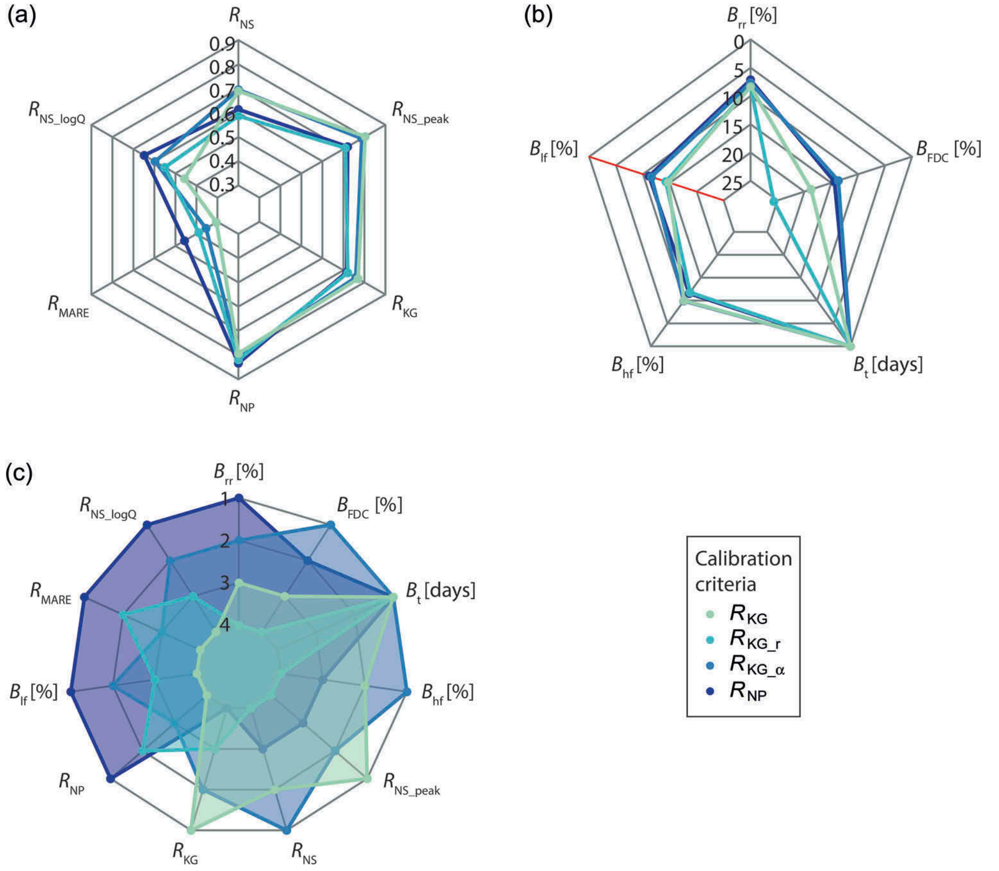


Figure 6. Model efficiencies in validation for model calibrations with R_{KG}, R_{KG_r}, R_{KG_α}, and R_{NP}. Calibration criteria are evaluated in terms of (a) statistical metrics and (b) hydrological signatures (note that the axis for B_{lf} is scaled by a factor of five, meaning that percent bias is five times higher than indicated). Each calibration criterion is ranked according to its performance for statistical metrics and hydrological signatures in (c). Results are presented for the median efficiency of all 100 study catchments.

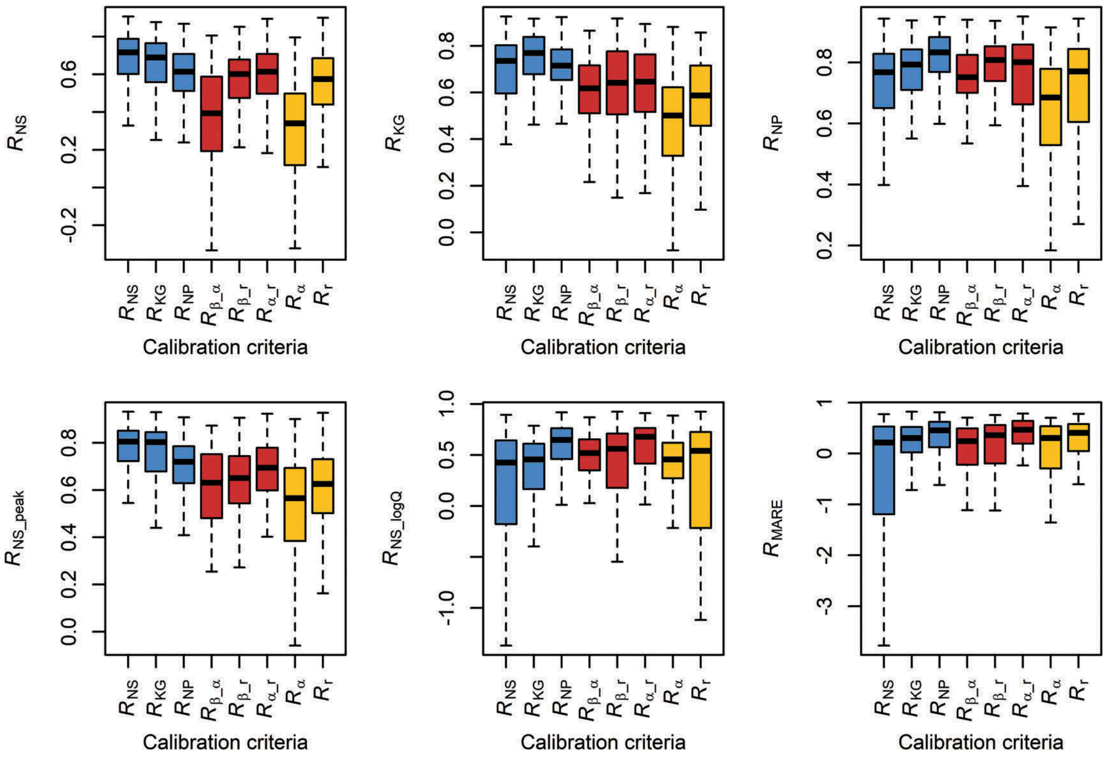


Figure 7. Model efficiencies (R_{NS}, R_{KG}, R_{NP}, R_{NS_peak}, R_{NS_logQ}, and R_{MARE}) in validation for model calibrations with three-, two- or one-component calibration criteria. Box plots consist of the median model efficiency for each of the 100 study catchments. Note the different scales of the *y*-axis. Results for calibrations based on R_β are not displayed as they were much poorer than for all other calibrations. Median efficiencies for calibration with R_β were −0.50, −0.28, −0.14, 0.43, −1.13, and −2.55 for $R_{NS,}$ R_{KG}, R_{NP}, R_{NS_peak}, R_{NS_logQ}, and R_{MARE}, respectively.

runoff ratio (B_{rr}) for which it was more essential to include a volume metric in the multi-objective function than to consider discharge dynamics. Lastly, discharge dynamics were less important than flow variability for simulating the high-flow segment of the FDC (B_{hf}) most likely because the timing is not of major relevance for that signature.

Discussion

The use of non-parametric goodness-of-fit measures is still a relatively new approach to model calibration. A comparison of various hydrograph characteristics resulting from calibrations with a partly non-parametric formulation of the popular Kling-Gupta efficiency (R_{NP}) and its original formulation (R_{KG}) demonstrated the potential of calibration criteria with non-parametric components. Overall, R_{NP} proved to be a valuable alternative for R_{KG}. It resulted in more confined hydrograph simulations (Figs. 3 and 4) and, except for high-flow metrics, in comparable or improved model performance for many of the statistical metrics and signatures (Fig. 6). Altogether, the flow–duration curve positively affected parameter selection, whereas the Spearman rank correlation had a varied effect on hydrograph simulations (Figs. 4 and 6).

More specifically, the use of the normalized FDC instead of the standard deviation had a positive effect on hydrograph simulations for all evaluated performance criteria. This favourable effect is encouraging although it may not be surprising. Unlike the standard deviation, which is a measure of discharge variability around the mean flow, the FDC contains information about the distribution of discharge over the full range of magnitudes (Vogel and Fennessey 1995). It therefore supports the model calibration with more information on discharge variability than the standard deviation.

A non-parametric formulation for discharge dynamics was especially valuable for simulating mean and low-flow conditions of a catchment as opposed to exceptionally high flow volumes or the timing of peak flows, which were better simulated when the Pearson correlation coefficient was used for model calibration. The sensitivity of the Pearson correlation coefficient to high discharge magnitudes (Legates and McCabe 1999, Krause *et al.* 2005) might seem beneficial for certain hydrograph aspects, but makes calibration sensitive to potential rating curve uncertainties at high flows. The results that model calibrations with R_{KG} (and therefore

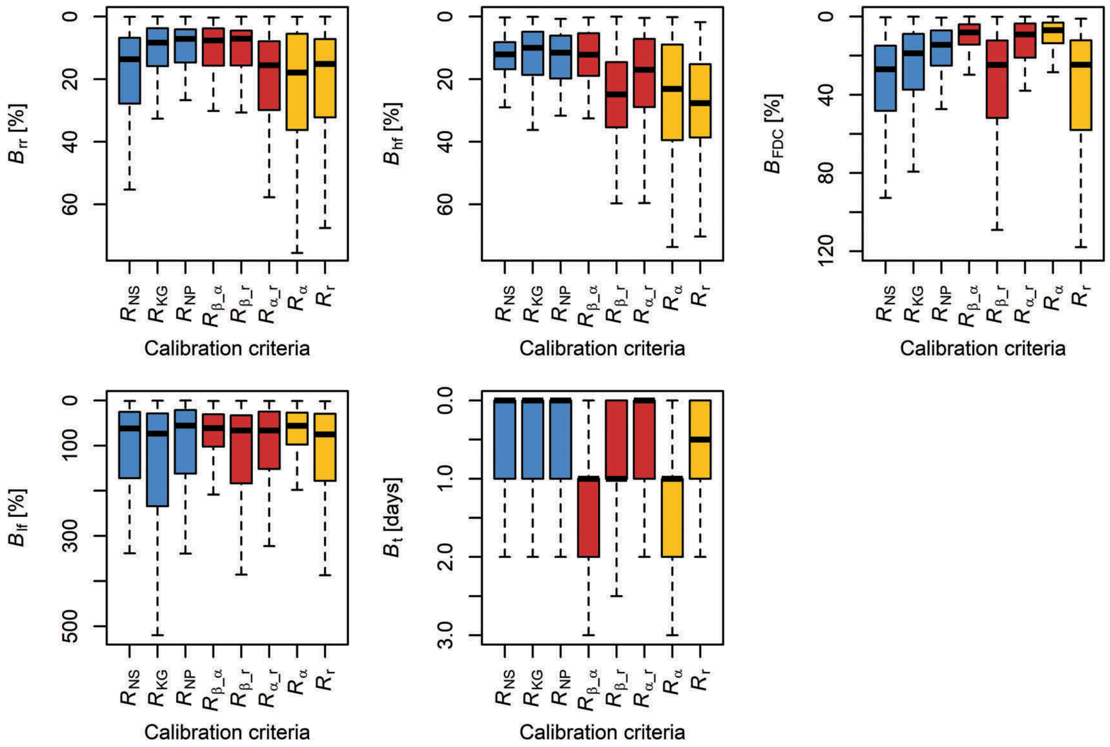


Figure 8. Model efficiencies (B_{rr}, B_{hf}, B_{FDC}, B_{lf}, and B_t) in validation for model calibrations with three-, two- or one-component calibration criteria. Box plots consist of the median model efficiency for each of the 100 study catchments. Note the different scales of the *y*-axis. Results for calibrations based on R_β are not displayed as they were much poorer than for all other calibrations. Median efficiencies for calibration with R_β were 96.0%, 28.2%, 62.7%, 87.9%, and 1.8 days for B_{rr}, B_{hf}, B_{FDC}, B_{lf}, and B_t respectively.

with the Pearson correlation coefficient) did not necessarily end in high Spearman rank correlations, questions the current predominant use of the Pearson correlation for describing discharge dynamics. The loss of information usually attributed to the use of Spearman rank correlation can therefore be a desirable effect when aiming at evaluating dynamics aspects. As a consequence, for many modelling applications Spearman rank correlation probably results in a more realistic representation of the overall dynamics and magnitudes of a catchment's runoff response than the Pearson correlation.

Although the goal of the proposed modification of R_{KG} was to make a step towards non-parametric calibration criteria, we decided to use the mean instead of the median, which would have been the non-parametric alternative, to describe discharge volumes for two main reasons. First, information on the total discharge volume in a -hydrological year is essential to close the water balance during model calibration, i.e. to constrain model parameters and ensure a correct simulation of actual evapotranspiration. For skewed distributions, such as those of discharge time series, the median can deviate largely from the mean and a good model fit could have been achieved without closing the water balance. Second, the median discharge of semi-arid and arid catchments with prolonged dry periods might be zero, which would result in numerical problems when computing the ratios of simulated and observed values.

Mean discharge, normalized FDC, and Spearman rank correlation all provide unique information for model calibration that is not represented by one of the other criteria. As a consequence, using all three components for model calibration (R_{NP}) resulted in a better overall model performance than using a subset of the three components. These results are consistent with the observation that more robust hydrograph simulations are achieved with multi-objective model calibration (e.g. Lindström 1997, Gupta *et al.* 1998, Boyle *et al.* 2000, Madsen 2003, Efstratiadis and Koutsoyiannis 2010). Since mean discharge, normalized FDC, and Spearman rank correlation represent different hydrograph characteristics, it was to some extent surprising to see the moderate to strong correlation between them. One explanation for this observation is that discharge is often closely related to precipitation input. Especially in humid catchments, discharge volume and variability are reasonably modelled as long as hydrograph dynamics are well captured by the runoff model (Seibert and Vis 2016). A correlation between efficiency criteria can to some degree be desirable as it inhibits solutions where only a single hydrograph aspect is well simulated while others are poorly represented.

By selecting objective functions for the evaluation of runoff models, we implicitly make assumptions about the statistical nature of discharge data and model simulation errors. These assumptions can be that a discharge time series is normally distributed or does not include any outliers. However, such assumptions are often violated when working with real data. We therefore argue that from a conceptual point of view it is desirable to use non-parametric formulations of objective functions, requiring weaker assumptions that are more likely met by observed and simulated discharge data. From a results perspective, we demonstrated that good results can be achieved when using a multi-objective function with non-parametric components to calibrate a model for multiple hydrograph aspects. Our results can be considered as relatively robust given that modelling results were based on 100 catchments with long modelling time series and which represent a large variety of hydroclimates. In practice, modellers often use log-transformed discharge to put less emphasis on high flows. This approach should be avoided for computing R_{KG}, because using log-transformed discharge would result in R_{KG} values that are sensitive to discharge close to zero and that are dependent on the chosen flow unit (Santos *et al.* 2018). Therefore, using R_{NP} could provide a valuable alternative in cases where one would otherwise use log-transformed flows.

Conclusions

In this study, we propose a modified variant of the Kling-Gupta efficiency towards a non-parametric calibration criterion for hydrological models. In this modified formulation, discharge volume is described by the mean discharge, discharge variability is represented by the FDC, and discharge dynamics are expressed in terms of Spearman rank correlation. Given the conceptual advantages of non-parametric calibration criteria, the goal was to evaluate the potential and limits of such a goodness-of-fit measure for simulating multiple hydrograph aspects simultaneously. The proposed calibration approach was tested on 100 catchments across the contiguous United States, which span a large range of hydroclimatic conditions. From the evaluation of the simulated hydrographs on commonly used statistical metrics and signatures that represent various hydrograph aspects, the following conclusions can be drawn:

1. The non-parametric modification of the Kling-Gupta efficiency generally resulted in better agreement between simulated and observed discharge than the original formulation, except when evaluating the magnitude and timing of high flows. The proposed non-parametric based multi-objective function can therefore be seen as a useful alternative to existing performance measures when aiming at acceptable simulations of multiple hydrograph aspects.
2. The use of the FDC instead of the standard deviation to describe discharge variability positively affected all evaluated hydrograph aspects, which is likely due to the complete information on the discharge distribution contained in the FDC.
3. The Spearman rank correlation generally improved simulations during mean and low-flow conditions compared to the Pearson correlation, which can be attributed to the insensitivity of the Spearman rank correlation to extreme values, strengthening its characterization of discharge dynamics.
4. The combination of all three components of the mean squared error, namely discharge volume, variability, and dynamics, in a single objective function generally resulted in simulations that represent multiple hydrograph aspects well. In contrast, model calibrations with a subset of the three components put emphasis on rather specific hydrograph aspects at the expense of a realistic representation of several hydrograph characteristics simultaneously.

Acknowledgements

We thank Gilles Kratzer for discussion about the definitions of parametric and non-parametric metrics. Alexander Gelfan and two anonymous reviewers are acknowledged for their valuable comments on an earlier draft of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the University of Zurich. Hydro-meteorological data and catchment shapefiles were made available by Newman *et al.* (2015). SRTM elevation data were provided by Jarvis *et al.* (2008).

ORCID

Sandra Pool http://orcid.org/0000-0001-9399-9199
Marc Vis http://orcid.org/0000-0002-5589-2611
Jan Seibert http://orcid.org/0000-0002-6314-2124

References

Beck, H.E., *et al.*, 2016. Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, 52 (5), 3599–3622. doi:10.1002/2015WR018247

Bergström, S., 1976. *Development and application of a conceptual runoff model for Scandinavian catchments*. Norrköping, Sweden: Swedish Meteorological and Hydrological Institute. No. RHO 7 134.

Boyle, D.P., Gupta, H.V., and Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. *Water Resources Research*, 36 (12), 3663–3674. doi:10.1029/2000WR900207

Coopersmith, E.J., Minsker, B.S., and Sivapalan, M., 2014. Patterns of regional hydroclimatic shifts: an analysis of changing hydrologic regimes. *Water Resources Research*, 50, 1960–1983. doi:10.1002/2012WR013320

Efstratiadis, A. and Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. *Hydrological Sciences Journal*, 55 (1), 58–78. doi:10.1080/02626660903526292

Euser, T., *et al.*, 2013. A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, 17 (5), 1893–1912. doi:10.5194/hess-17-1893-2013

Garcia, F., Folton, N., and Oudin, L., 2017. Which objective function to calibrate rainfall–runoff models for low-flow index simulations? *Hydrological Sciences Journal*, 62 (7), 1149–1166.

Gupta, H.V., Sorooshian, S., and Yapo, P.O., 1998. Toward improved calibration of hydrologic models: multiple and non-commensurable measures of information. *Water Resources Research*, 34 (4), 751–763. doi:10.1029/97WR03495

Gupta, H.V., *et al.*, 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377 (1–2), 80–91. doi:10.1016/j.jhydrol.2009.08.003

Häggström, M., *et al.*, 1990. *Application of the HBV model for flood forecasting in six central American rivers*. Norrköping, Sweden: Swedish Meteorological and Hydrological Institute. No. RHO 27 14.

Hingray, B., *et al.*, 2010. Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments. *Hydrological Sciences Journal*, 55 (6), 1002–1016. doi:10.1080/02626667.2010.505572

Jarvis, A., *et al.*, 2008. *Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m*. Available from: http://srtm.csi.cgiar.org [accessed January 2016].

Kiesel, J., *et al.*, 2017. Improving hydrological model optimization for riverine species. *Ecological Indicators*, 80, 376–385. doi:10.1016/j.ecolind.2017.04.032

Krause, P., Boyle, D.P., and Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5, 89–97. doi:10.5194/adgeo-5-89-2005

Legates, D.R. and McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35 (1), 233–241. doi:10.1029/1998WR900018

Lidén, R. and Harlin, J., 2000. Analysis of conceptual rainfall–runoff modelling performance in different climates. *Journal of Hydrology*, 238 (3–4), 231–247. doi:10.1016/S0022-1694(00)00330-9

Lindström, G., 1997. A simple automatic calibration routine for the HBV model. *Hydrology Research*, 28 (3), 153–168. doi:10.2166/nh.1997.0009

Lindström, G., *et al.*, 1997. Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, 201, 272–288. doi:10.1016/S0022-1694(97)00041-3

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Advances in Water Resources*, 26 (2), 205–216. doi:10.1016/S0309-1708(02)00092-1

Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116 (12), 2417–2424. doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I - a discussion of principles. *Journal of Hydrology*, 10 (3), 282–290. doi:10.1016/0022-1694(70)90255-6

Newman, A.J., *et al.*, 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19, 209–223. doi:10.5194/hess-19-209-2015

Perrin, C., Michel, C., and Andréassian, V., 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, 242 (3–4), 275–301. doi:10.1016/S0022-1694(00)00393-0

Pfannerstill, M., Guse, B., and Fohrer, N., 2014. Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. *Journal of Hydrology*, 510, 447–458. doi:10.1016/j.jhydrol.2013.12.044

Pokhrel, P., Yilmaz, K.K., and Gupta, H.V., 2012. Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *Journal of Hydrology*, 418, 49–60. doi:10.1016/j.jhydrol.2008.12.004

Santos, L., Thirel, G., and Perrin, C., 2018. Technical note: pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*, 22, 4583–4591. doi:10.5194/hess-22-4583-2018

Seibert, J., 2000. Multi-Criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, 4, 215–224. doi:10.5194/hess-4-215-2000

Seibert, J. and Vis, M.J., 2016. How informative are stream level observations in different geographic regions? *Hydrological Processes*, 30 (14), 2498–2508. doi:10.1002/hyp.10887

Seibert, J. and Vis, M.J.P., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, 16, 3315–3325. doi:10.5194/hess-16-3315-2012

Shafii, M., *et al.*, 2017. A diagnostic approach to constraining flow partitioning in hydrologic models using a multiobjective optimization framework. *Water Resources Research*, 53 (4), 3279–3301. doi:10.1002/2016WR019736

Vis, M., *et al.*, 2015. Model calibration criteria for estimating ecological flow characteristics. *Water*, 7 (5), 2358–2381. doi:10.3390/w7052358

Vogel, R.M. and Fennessey, N.M., 1995. Flow duration curves II: a review of applications in water resources planning. *JAWRA Journal of the American Water Resources Association*, 31 (6), 1029–1039. doi:10.1111/jawr.1995.31.issue-6

Westerberg, I.K., *et al.*, 2011. Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, 15 (7), 2205. doi:10.5194/hess-15-2205-2011

Wilcoxon, F., 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1 (6), 80–83. doi:10.2307/3001968

Yilmaz, K.K., Gupta, H.V., and Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. *Water Resources Research*, 44, 9. doi:10.1029/2007WR006716

Zhang, Y., *et al.*, 2016. Multi-metric calibration of hydrological model to capture overall flow regimes. *Journal of Hydrology*, 539, 525–538. doi:10.1016/j.jhydrol.2016.05.053