Manuscript prepared for Hydrol. Earth Syst. Sci.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 26 April 2016

Simultaneous calibration of hydrological models in geographical space

A. Bárdossy[1], Y. Huang[1], and T. Wagener[2]

[1]Institute for Modelling Hydraulic and Environmental Engineering, University of Stuttgart, Stuttgart, Germany
[2]Department of Civil Engineering, Queen's School of Engineering, University of Bristol, Bristol, UK

Correspondence to: A. Bárdossy (andras.bardossy@iws.uni-stuttgart.de)

Abstract. Hydrological models are usually calibrated for selected catchments individually using specific performance criteria. This procedure assumes that the catchments show individual behavior. As a consequence, the transfer of model parameters to other ungauged catchments is problematic. In this paper, the possibility of transferring part of the model parameters was investigated. Three different conceptual hydrological models were considered. The models were restructured by introducing a new parameter η which exclusively controls water balances. This parameter was considered as individual to each catchment. All other parameters, which mainly control the dynamics of the discharge (dynamical parameters), were considered for spatial transfer. Three hydrological models combined with three different performance measures were used in three different numerical experiments to investigate this transferability. The first numerical experiment, individual calibration of the models for 15 selected MOPEX catchments, showed that it is difficult to identify which catchments share common dynamical parameters. Parameters of one catchment might be good for another catchment but not reversed. In the second numerical experiment, a common spatial calibration strategy was used. It was explicitly assumed that the catchments share common dynamical parameters. This strategy leads to parameters which perform well on all catchments. A leave one out common calibration showed that in this case a good parameter transfer to ungauged catchments can be achieved. In the third numerical experiment, the common calibration methodology was applied for 96 catchments. Another set of 96 catchments were used to test the transfer of common dynamical parameters. The results show that even a large number of catchments share similar dynamical parameters. The performance is worse than those obtained by individual calibration, but the transfer to ungauged catchments remains possible. The performance of the common parameters in the second experiment was better than in the third, indicating that the selection of the catchments for common calibration is important.

1 Introduction

Hydrological models are widely used to describe catchment behavior, and for subsequent use for water management, flood forecasting and other purposes. Hydrological modeling is usually done for catchments with observed precipitation and discharge data. The unknown (and partly not measurable) parameters of a conceptual or to some extent physics-based model are adjusted in a calibration procedure to reproduce the measured discharge from the observed weather and catchment properties. Due to the high variability of catchment properties and hydrological behavior (Beven, 2000), this modeling procedure is usually performed individually for each catchment. Different catchments are often modeled using different models. This great variety of models and catchments makes a generalization of the description of the hydrological processes very challenging (Sivapalan, 2003). Additionally, even for a selected model applied for a specific catchment, the parameter identification is not unique. A great number of parameter vectors might lead to a very similar performance (Beven and Freer, 2001).

Moreover, due to over-reliance on measured discharge for model calibration, estimation of model parameters for ungauged basins is a big challenge. Instead of model calibration, parameters have to be estimated on the basis of other information (Sivapalan, 2003). A decade of world-wide research efforts have been carried out for the runoff prediction in ungauged basins(PUB) (Hrachowitz et al., 2013). The PUB synthesis book (Blöschl et al., 2013) takes a comparative approach to learning from similarities between catchments and summarizes a great number of interesting methods that are being used for predicting runoff regimes in ungauged basins. Many attempts have been made to develop catchment classification schemes to identify groups of catchments which behave similarly (Grigg, 1965; Sawicz et al., 2011; Ali et al., 2012; Sivakumar and Singh, 2012; Toth, 2013). However, the task is of great importance, McDonnell and Woods (2004) discussed the need for a widely accepted classification system and Wagener et al. (2007) pointed out that a good classification would help to model the rainfall–runoff process for ungauged catchments.

Razavi and Coulibaly (2012) give a comprehensive review of regionalization methods for predicting streamflow in ungauged basins. Catchment similarity can be determined by comparing their corresponding discharge series using correlation (Archfield and Vogel, 2010) or copulas (Samaniego et al., 2010). Much of the variability in discharge time series is controlled by the weather patterns. Therefore, it is likely that similarity in discharge is higher for catchments with well correlated weather, which often requires geographical closeness (Archfield and Vogel, 2010). However, discharge series produced by catchments can be very different under different meteorological conditions. Even the same catchment behaves differently in a dry and in a wet year. Due to the different weather forcing, the above methods would consider the same catchment in one time period as dissimilar to itself in another time period.

One can also define catchment similarity using hydrological models (McIntyre et al., 2005; Oudin et al., 2010; Razavi and Coulibaly, 2012). Catchments are similar if they can be modeled reasonably well by the same model using the same model parameters (Bárdossy, 2007). Due to observational errors and specific features in the calibration period, the adjustment of the model can be very specific to the observation period leading to an overcalibration (Andréassian et al., 2012). To overcome such limitations, regional calibration (Fernandez et al., 2000) approach is suggested to identify single parameter set that perform well for all catchments within the modeled domain. Parajka et al. (2007) indicate that the iterative regional calibration indeed reduced the uncertainty of most parameters. Regional calibration can result in a better temporal robustness than normal individual calibration (Gaborit et al., 2015) and it provides effective approach in large-scale hydrological assessments (Ricard et al., 2012).

The focus of this paper is to investigate if the transformation of precipitation to discharge is possible independently of the weather. For this purpose, the hydrological model parameters are separated into two groups:

- Parameters describing the water balances, which are strongly related to climate.

- Parameters describing the dynamics of the runoff triggered by weather.

The second group of parameters is supposed to be weather independent and represent the focus of this paper. To simplify the problem, a single new parameter η was introduced to describe water balance. This parameter is conditional on the other model parameters and adjusts the long term water balances.

The purpose of this paper is to investigate to what extent do different catchments share a similar dynamical rainfall–runoff behavior and can be modeled using the same model parameters with exception of the newly introduced individualized water balance parameter η.

Hydrological models are usually judged according to the degree of reproducing discharge dynamics and water balances. While water balances are mainly driven by weather in terms of precipitation, temperature, radiation and wind. Dynamics is controlled by catchment properties in terms of size, terrain, slopes, soils etc. Formation of landscapes as a result of long time climate is a quasi equilibrium process. The hypothesis of this paper is that this equilibrium is mirrored in a similar dynamic behavior. Thus, a large number of catchments can be modelled by using the same dynamic parameters.

Three simple conceptual hydrological models combined with three different performance measures are used to describe the rainfall–runoff behavior on the daily time scale for a large number of catchments.

The following three different numerical experiments, including calibration and validation procedures, are carried out for different sets of selected catchments:

1. The usual catchment by catchment calibration is carried out. In order to test if dynamical model parameters are shared, the parameters are directly transferred to all of other catchments.

2. Instead of the traditional catchment by catchment calibration, it is assumed that the model parameters are similar for a set of catchments in a close geometrical setting. Thus a simultaneous calibration of the models is carried out and tested both in a gauged and an ungauged version.

3. The geographical extent of the catchments used for simultaneous calibration is expanded. A great number of assumed ungauged catchments are used for testing the hypothesis.

The hypothesis is that the rainfall–runoff process can be described using the same dynamical hydrological model parameters for a number of catchments. The very different climatic conditions and water balances of the catchments are considered by the newly introduced specific parameter η controlling the long term water balance of each catchment individually. The other model parameters control the discharge dynamics on both short and long time scales. These dynamical parameters are supposed to be shared despite the great heterogeneity of the catchments. This procedure simplifies the hydrological model parameter estimation for ungauged catchments, namely the procedure is reduced to the estimation of a single parameter η, which can be related to long term water balances.

The paper is structured as follows: after the introduction, the investigation area is described. This is followed by a description of the three conceptual hydrological models and the three performance criteria used for calibration and validation. In section four, the new model parameter η controlling the water balance is introduced. In sections five to seven, three numerical experiments are described and the results are presented, starting with the individual calibration of the models and ending with a transfer of the model parameters to randomly selected catchments. The paper concludes with a discussion of the results.

2 Investigation area and available data

The study area is the eastern United States. Locations of the 196 catchments used in this study are shown in Fig. 1. The catchments for a subset used for the international Model Parameter Estimation Experiment (MOPEX) project. Catchments range in size from 134 to 9889 km^2 and exhibit aridity indices (long-term potential evapotranspiration to precipitation rates) between 0.41 and 3.3, hence representing a heterogeneous dataset. Time-series data of daily streamflow, precipitation, and temperature for all catchments were provided by the MOPEX project (Duan et al., 2006). Catchments within this dataset are minimally impacted by human influences. Streamflow information within this dataset was originally provided by the United States Geological Survey (USGS) gauges, while precipitation and temperature was supplied by the National Climate Data Center (NCDC). The MOPEX dataset has been used widely for hydrological model comparison studies (see references in Duan et al., 2006).

3 Hydrological models and performance criteria

Three simple conceptual hydrological models were applied in this study. The reason for this is that the great number of calibration and validation experiments could only be performed with relatively simple model structures. It is important to see if the results are similar for different models and performance measures. In a subsequent study, spatially distributed models will be considered.

3.1 HYMOD

The HYMOD (Boyle et al., 2001) is a conceptual rainfall–runoff model derived from the Probability Distributed Model (Moore, 1985). The soil moisture accounting module of HYMOD utilizes a Pareto distribution function of storage elements of varying sizes. The storage elements of the catchment are distributed according to a probability density function defined by the maximum soil moisture storage CMAX and the distribution of soil moisture stores b (Wagener et al., 2001). Evaporation from the soil moisture store occurs at the rate of the potential evaporation estimates using the Hamon approach. After evaporation, the remaining rainfall and snowmelt is used to fill the soil moisture stores. A routing module divides the excess rainfall using a split parameter α which separates fluxes amongst two parallel conceptual linear reservoirs meant to simulate the quick and slow flow response of the system (defined by residence times k_{q} and k_{s}).

3.2 HBV

The HBV model is a conceptual model and was originally developed at the Swedish Meteorological and Hydrological Institute (SMHI) (Bergström and Forsman, 1973). Snow accumulation and melt, actual soil moisture and runoff generation are calculated using conceptual routines. The snow accumulation and melt is based on the degree-day approach. Actual soil moisture is calculated by considering precipitation and evapotranspiration. Runoff generation is estimated by a non-linear function of actual soil moisture and precipitation. The dynamics of the different flow components at the subcatchment scale are conceptually represented by two linear reservoirs. The upper reservoir simulates the near surface and interflow in the sub-surface layer, while the lower reservoir represents the base flow. They are connected through a linear percolation rate. Finally, there is a transformation function consisting of a triangular weighting function with one free parameter for smoothing the generated flow.

3.3 Xinanjiang model (XAJ)

The XAJ model was established in the early 1970s in China. This conceptual rainfall–runoff model has been applied to a large number of basins in the humid and semi-humid regions in China. The lumped version of XAJ model consisted of four main components (Zhao, 1995). The evapotranspiration is represented by a 3-layer soil moisture module which differentiates upper, lower and deeper soil layers. Runoff production is calculated based on rainfall and soil storage deficit, tension water capacity curve is introduced to provide for a non-uniform distribution of tension water capacity throughout the whole catchment. The runoff separation module separates the determined runoff into three parts, namely surface runoff, interflow and groundwater. The flow routing module transfers the local runoff to the outlet of the basin. In order to account for the precipitation that is contributed from snowmelt, the degree-day snowmelt approach is added in this model. In this study, the model has 16 parameters which can be adjusted using calibration.

3.4 Performance criteria

Model calibration depends strongly on the performance criteria used. In order to obtain reasonably general results, three different criteria were selected to evaluate model performance.

The Nash–Sutcliffe Efficiency (Nash and Sutcliffe, 1970) between the observed and modeled flow is most frequently taken as the first evaluation criterion:

$$O^{(1)}: \mathrm{NS} = 1 - \frac{\sum_{T=1}^{T} \left(Q_{\mathrm{o}}(t) - Q_{\mathrm{m}}(t)\right)^2}{\sum_{T=1}^{T} \left(Q_{\mathrm{o}}(t) - \overline{Q_{\mathrm{o}}}\right)^2} \tag{1}$$

Here $Q_{\mathrm{o}}(t)$ is the observed discharge and $Q_{\mathrm{m}}(t)$ is the modeled discharge on a given day *t*. The abbreviation NS is used subsequently for this performance measure.

The NS model performance criterion was often criticized (for example in Schaefli and Gupta, 2007), and several modifications and other criteria were suggested. One interesting suggestion was published in Gupta et al. (2009), the authors suggest using a performance measure which accounts for the water balances and the correlation of the observed and modeled time series separately. Their approach was slightly modified and the following performance criterion was introduced:

$$O^{(2)}: \mathrm{GK} = 1 - \beta \left(\frac{\sum_{T=1}^{T} \left(Q_{\mathrm{o}}(t) - Q_{\mathrm{m}}(t)\right)}{\sum_{T=1}^{T} Q_{\mathrm{o}}(t)} \right)^2 - \left(1 - r(Q_{\mathrm{o}}, Q_{\mathrm{m}})\right)^2 \tag{2}$$

Here $r(Q_{\mathrm{o}}, Q_{\mathrm{m}})$ is the correlation coefficient between the observed and modeled time series of discharge. β is a weight to express the importance of the water balance. In our study, $\beta = 5$ was selected. The reason for selecting this version of the coefficient is that a model should produce good water balances and appropriate discharge dynamics simultaneously. The quadratic form in Eq. (2) assures that both aspects are considered, and the worse of them is dominating. The abbreviation GK is used subsequently for this performance measure.

The Nash–Sutcliffe coefficient of the logarithm of the discharges is focusing on the low flow conditions more than the traditional NS coefficient:

$$L_{\mathrm{NS}} = 1 - \frac{\sum_{T=1}^{T} \left(\log(Q_{\mathrm{o}}(t)) - \log(Q_{\mathrm{m}}(t))\right)^2}{\sum_{T=1}^{T} \left(\log(Q_{\mathrm{o}}(t)) - \overline{\log(Q_{\mathrm{o}})}\right)^2} \tag{3}$$

To equally concentrate on high and low flows, a combination of the original NS and the logarithmic NS is used as a third measure:

$$O^{(3)}:\mathrm{NS}+\mathrm{LNS}=\frac{\mathrm{NS}+L_{\mathrm{NS}}}{2} \tag{4}$$

The abbreviation $\mathrm{NS}+\mathrm{LNS}$ is used subsequently for this performance measure.

The three performance criteria were modified, hence the higher the value the better the model. Further the best value for the criteria is 1.

4 Method

4.1 Model parameter to control water balance

Climatic conditions are of central importance for water balances. The relationship of potential to actual evapotranspiration can differ strongly due to water or energy limitations. This suggests that catchments might have similar dynamical behavior but with different water balances. In order to account for this, the model parameters could be separated to form two groups, one group with parameters controlling the water balances and another controlling the discharge dynamics. This separation of existing model parameters is difficult, as they often influence simultaneously both components. Instead of an artificial model specific separation, a new parameter η was introduced to all three models. This parameter controls the ratio between daily potential and actual evapotranspiration depending on the available water and depends on the long term water balance only. This parameter η gives:

$$E_{\mathrm{ta}}=\begin{cases} E_{\mathrm{tp}} & \text{if } \frac{\mathrm{SM}}{\mathrm{CMAX}}>\eta \\ \min\left(\frac{\mathrm{SM}}{\eta\cdot\mathrm{CMAX}}E_{\mathrm{tp}},\mathrm{SM}\right) & \text{else} \end{cases} \tag{5}$$

Here SM is the actual soil water available for evapotranspiration. CMAX is the maximum possible soil moisture. E_{tp} stands for the potential and E_{ta} for the actual evapotranspiration, respectively.

The parameter η regulates the water balances in accordance with the dynamical parameters. It can be calculated directly for each parameter vector $\boldsymbol{\theta}$. This is necessary as it is thought to establish correct water balances. Thus it is a catchment and parameter vector dependent parameter. $f(\eta)=V_{iM}(\eta,\theta)$ is a monotonically decreasing function of η. If the model can provide correct long term water balances then:

$$V_{iM}(1,\theta)<V_{iO}<V_{iM}(0,\theta) \tag{6}$$

As $f(\eta)=V_{iM}(\eta,\theta)$ is continuous, there is a unique $\eta(\theta)$ for which:

$$V_{iM}(\eta(\theta),\theta)=V_{iO} \tag{7}$$

If Eq. (6) is not fulfilled, then the parameter vector $\boldsymbol{\theta}$ is not appropriate for the model.

The parameter η is fitted individually for each $\boldsymbol{\theta}$, in this way a correct water balance is assured for the calibration period.

4.2 Experimental design

In this study, the ROPE algorithm (Bárdossy and Singh, 2008) was applied for model parameter optimization. This parameter optimization method could obtain pre-determined number of optimal parameter sets that perform very similar to the models, although the parameter sets are very heterogeneous. In this study, each calibration yielded 10 000 convex sets of good parameter vectors. Three numerical experiments on a large number of catchments were carried out to investigate the transferability of the model parameters under different calibration strategies. For a clear explanation and understanding of the methods, the procedure and results for these three experiments are presented in the following three sections.

5 Numerical experiment 1: individual calibration and parameter transfer

The first experiment is thought to test the transferability of the model parameters under the usual individual calibration for each catchment.

As a first step, 15 catchments with reliable data and slightly varying catchment properties in the eastern United States were selected. Locations of the selected gauges are marked as red plus on Fig. 1. Table 1 lists the basic catchment properties and Table 2 summarizes the meteorological conditions for the selected 15 catchments, respectively (Falcone et al., 2010). The tables show that despite their geographical proximity, these catchments have quite different climate and hydrographic properties.

For the 15 selected catchments, an individual calibration was performed using all three models and all three performance measures. Data series from year 1951 to 2000 were split up into 5 sub-periods. This leads to 45 calibrations for each catchment. Each calibration yielded convex sets $\mathcal{G}_i$ of good parameters for each catchment i. 10 000 parameter vectors from each of these sets were generated. (Note that the corresponding parameter η was estimated for each element of the parameter set separately.)

Let $O_i^{(j)}(\theta)$ denote the value of the objective function j for a parameter vector $\boldsymbol{\theta}$ in catchment i. The best objective function value for each individual catchment is denoted with $O_i^{(j)*}$. Although the parameter sets are very heterogeneous, all of them perform very similar. For simplicity, we used the average value of the 10 000 performances to represent the simulation result for each catchment.

The left part of Fig. 2 shows the mean values of the objective function NS for the 10 000 parameter vectors for the calibration period 1971–1980 for the three selected models (denoted as individual calibration). As expected, the model performance varies across catchments. The reasons for this are observation errors both in input and output as well as a possible inability of the model to reasonably well represent the main hydrological processes.

The ranges of the model parameters are relatively large. As a first step, we checked if the catchments have common parameter vectors. For each pair of catchments (i, j), for the same performance measure and time period, the intersection of the convex hull of the good parameter sets $\mathcal{G}_i \cap \mathcal{G}_j$ is empty showing that there are no common best parameters. From the result, seemingly none of the catchments are similar.

As a next step, the 10 000 generated best dynamical parameter vectors for a given time period and hydrological model obtained for catchment i were applied to model all other catchments using the same hydrological model and time period. Note that the value of η is not transferred but adjusted to the true long term water balance. Figure 3 shows the color coded matrices for the mean NS performance and GK performance of the three hydrological models using transferred parameters for all 15 catchments for a calibration period (1971–1980).

The performance of the transferred parameter vectors displays a strongly varying picture. While in some cases the catchments seem to share parameter vectors with reasonably good performance, in other cases the transfer lead to weak performances. A further surprising fact is that none of the matrices is symmetrical. One can see that some catchments are good donors as their parameters are good for nearly all catchments, while others have parameters which are hardly transferable.

The asymmetry of the parameter transition matrices cannot be explained by catchment properties. Two different catchments seem to share well performing parameters if calibrated on one catchment and no common good parameters if calibrated on the other one. Take the catchments 1 and 12 with the NS performance as an example. For all three models, parameters calibrated for catchment 1 are not suitable for catchment 12, but parameters of catchment 12 perform reasonably well for catchment 1. From the observation data, we found that catchment 12 is under relatively dry climate conditions during the calibration period. We also found from the simulated hydrographs that the parameter sets calibrated on catchment 1 could not well capture the dynamic behavior of catchment 12 as the low flows were underestimated for most of the time and the peak flows were obviously overestimated. The matrices for NS show different performances with different models. In general, HBV model performs the best. The average value of the matrix is 0.62 for HBV, 0.55 for HYMOD and 0.54 for XAJ model. Furthermore, the correlation of transferred model performance between different models are all greater than 0.7. From the viewpoint of parameter transferability, the three models perform similarly, if a parameter transfer is reasonable from catchment i to j for one model then it is also reasonable for the other models. The results for the GK performance differ from those of the NS performance. Here the XAJ model seems to give the generally best transferable parameters. Parameter vectors from other catchments generally fail to perform on catchment 15 across all three models.

The difference of the transferability for these two performance measures could be explained by different focuses – while NS is mainly focusing on the squared difference between the observed and modeled discharge, GK focuses on water balances and good timing and NS+LNS is strongly influenced by low flow events. It is interesting to observe that catchment 12 is a very bad receiver for model parameters for NS, while it is an excellent receiver for GK. This means that different events have different influence on the performance. A possible explanation for the asymmetry is the fact that the catchments have different weather forcing in the calibration period. It could be that runoff events which are most important for a performance measure occur in the calibration period frequently in one catchment leading to good transferability, and seldom in the other causing weak transferability of the parameters from one catchment to another.

The transferability of the model parameters was also tested for an independent validation period between 1991 and 2000. Figure 4 shows the corresponding color coded results for NS as performance measure. The matrices are similar to those obtained for calibration. Catchment 12 remained a bad receiver but a good donor indicating that the bad performance is unlikely to be caused by observation errors. Further, for some columns the off diagonal elements are larger than the diagonal ones which is a sign of a possible overcalibration of models.

To investigate the influence of climate on calibration, the hydrological models calibrated for different time periods using the same model and performance measure were compared. As the different time periods represent different climate conditions, the calibrations lead to different parameter sets. As a comparison, the differences in calibrated model parameters using the same model and performance measure for different catchments were compared. As an example, the left part of Fig. 5 shows two calibrated parameters of the HYMOD model for catchment 13 on three different 10 years time periods. The right part of Fig. 5 shows the same parameters obtained by calibration for three different catchments 7, 8 and 13 during time period 1951–1960. The structural similarity of the two scatterplots suggests that the difference between the different catchments is comparable to the difference between the different time periods. In hydrological modeling, it is usually assumed that model parameters are constant over time assuming no significant change in climate or other characteristics. The results however show the assumption that parameters are the same over space is not completely unrealistic. The figures even suggest that there might be parameter vectors which perform reasonably well for all 15 catchments. As a next step, an experiment to test this assumption was devised.

6 Numerical experiment 2: simultaneous calibration

Since for many pairs of catchments, the parameter transfer worked reasonably well. As a next step, we investigated if there are parameters which perform reasonably well for all catchments. As seen in the previous section, none of the catchments share optimal parameters. Therefore common sub-optimal parameters have to be found.

In order to identify parameter vectors which perform simultaneously well for each catchment, the hydrological models were calibrated for all 15 catchments simultaneously. The simultaneous calibration of the model for all catchments is a multi-objective optimization problem. The goal is to find parameter vectors which are almost equally good for all catchments with no exception. As the models perform differently for the different catchments due to data quality and catchment particularities, the performance was measured through the loss in performance compared to the usual individual calibration. Thus the objective function was formulated using the formulation of the compromise programming method (Zeleny, 1981):

$$R^{(j)}(\theta) = \sum_{i=1}^{n} \left(O_i^{(j)*} - O_i^{(j)}(\theta) \right)^p \tag{8}$$

Here index i indicates the catchment number, index j indicates the type of the individual performance measure specified in Eqs. (1), (2) and (4). The goal in this objective function is to minimize $R^{(j)}$. Here p is the so called balancing factor, the larger p is the more the biggest loss in performance contributes to the common performance. In order to obtain parameters which are good for all catchments, a relatively high $p = 4$ was selected for all three performance measures.

As same as individual calibration, the ROPE algorithm was used for the simultaneous calibration. The optimized parameter sets $\mathcal{H}^{(j)}$ are simultaneously well performed for each model and time period. The left part of Fig. 2 compares the performance of the individually calibrated and the common calibration for the 15 selected catchments using NS as performance criterion. As expected, the results show that the individual calibrations lead to better performances, but the joint parameter vectors perform reasonably well for all catchments.

As the goal of modeling is not the reconstruction of already observed data, the performances on a different validation period (1991–2000) were also compared. The right part of Fig. 2 shows the mean model performances for the 15 individually calibrated and the common calibrated datasets. The observation that parameter vectors obtained through common calibration may outperform individual on-site calibration may also indicate the weakness of the calibration process for an individual catchment, which should ideally be able to identify the ‘best’ parameter set.

These results indicate that instead of transferring model parameters from a single catchment, a parameter transfer might perform better if the parameters obtained through common calibration on all other catchments are used. In order to test this kind of parameter transfer, a set of simple “leave one out” calibrations were performed. This means that for a catchment *i*, the hydrological models were simultaneously calibrated for the remaining 14 catchments. Each time another catchment *i* was not considered for calibration, leading to 15 simultaneous calibrations. These common model parameters were then applied for the catchment which was left out. The performance of the models on these catchments in the calibration period is reasonably good for all catchments. Figure 6 shows the result of HBV and HYMOD using the NS performance measure. It compares the performance of the parameters obtained via individual calibrations (red x-mark), parameter transfers from other catchments individually (blue plus) and the transfer of the common parameters obtained by leave one out procedure (green diamond). The performance of common parameters is obviously weaker than that of the individual calibration, but better than many parameter transfer obtained using individual parameter transfer. To test the potential of the transferability of the common parameters, a validation period was used. Figure 7 shows the results for the validation time period 1991–2000. In this case, the common calibration performs very well. For HYMOD, it outperforms the parameter vectors obtained by individual calibration for 6 out of the 15 catchments. For the other catchments, the loss in performance is relatively small. Note that this good performance of the common models was obtained without using any information of the target catchment. The transfer of parameters obtained from individual calibrations on other catchments shows a highly inhomogeneous picture as described in experiment 1. The transferred common calibration is better than most of these performances. Further note that the results of experiment 1 show that there is no explanation why certain transfer work well and others do not. Thus for the transfer of model parameters to ungauged catchments, common calibration seems to be a reasonable method.

In order to illustrate how model parameters of the leave one out common calibration perform in validation, two hydrographs are presented. Figures 8 and 9 show a part of the observed, the modeled and the common calibration transferred hydrographs for a randomly selected parameter set obtained by individual calibration and leave one out common calibration of HBV for catchments 5 and 14. While for catchment 5, the common calibration leads to a hydrograph which is slightly better than that obtained by individual calibration, in the second case for catchment 14 the performance is reversed. However, in both cases the common parameters, which were obtained without using any observations of the catchment perform surprisingly well.

7 Numerical experiment 3: extension to other catchments

The results of the previous experiment suggest that even more catchments might share parameters which perform well on all. The 15 catchments used in experiments 1 and 2 are however to some extent similar and can thus not necessarily be considered as representative for a great number of other catchments. Thus, for the third experiment, 192 catchments of the MOPEX dataset were considered. 96 of them were randomly selected for common calibration (marked as blue circle on Fig. 1), the other 96 catchments were used as receivers to test the performance of the common parameters (marked as green triangle on Fig. 1). HBV model using three selected performance measures were considered in this experiment.

For each of the 192 catchments, an individual model calibration was carried out using 1971–1980 as calibration period. Common calibration was performed for the selected 96 catchments the same way as in experiment 2, for HBV model using all performance measures.

As a first step, the model performances for the individual and common calibration were compared. As expected and already seen in experiment 2, the performance for the common calibration is lower than the individual one for HBV using all performance measures. For example, the mean performance NS over all 96 catchments drops from 0.69 to 0.50. When one applies the models for the validation period 1991–2000, the individually calibrated model mean performance is 0.65, while for the common calibration the mean increases to 0.51. Figure 10 shows the histograms of the performance NS for the calibration and validation periods for the individual and the common calibrations. Results indicate the robustness of the common calibration. The transfer to the 96 assumed ungauged catchments shows very similar performance for the common parameters as for the catchments selected for common calibration. Figure 11 shows the histograms of the performance NS for the individual calibration and the transfer for the assumed ungauged catchments. It can be seen clearly from the histogram that there is very little difference between the performance for the gauged and the ungauged catchments. In 90 % of catchments, the common calibration works reasonably well even for the ungauged cases. The common parameters describing runoff dynamics of all 192 catchments indicate that there is a high degree of similarity of these catchments.

Comparing the results of the common calibration using the 96 catchments to that obtained using the 15 catchments, one can observe that the increase of catchments considered for the common calibration lead to a decrease of the performance. The common parameter sets calibrated by 15 catchments in a reasonable geographic proximity perform better than the parameter sets calibrated by 96 catchments. Thus the parameters obtained through common calibration can be regarded to describe the common dynamical behavior of many very different catchments over a large geographical area. If one is interested to find model parameters for a specific ungauged catchment, the common calibration using a more careful selection of the donor set of catchments is likely to lead to good parameter transfers.

The water balances of the 192 catchments are very different leading to very different η parameters. Figure 12 shows the distribution of η values for three randomly selected common good parameter sets for HBV model using NS as performance measure for the calibration time period. It can be seen clearly from the curve that for the same catchment, η is specific for different dynamical parameter sets. And due to the differences in water balance, different catchments requires very different η-s to control actual evapotranspiration. Furthermore, for all 192 catchments, parameter η present very similar tendency for different dynamical parameter sets. Figure 13 plots the mean η value against the ratio of the long term actual evapotranspiration to potential evapotranspiration ($E_{\mathrm{ta}}/E_{\mathrm{tp}}$) for each catchment. It shows strong negative correlation (-0.72) between η and $E_{\mathrm{ta}}/E_{\mathrm{tp}}$.

8 Discussion

8.1 Robust parameter sets

The three experiments were carried out in way that a set of parameters (usually represented by 10 000 individual parameter sets) was used. This leads to a considerable fluctuation of the results. Modelers often prefer to use single parameter vector. If a single parameter vector is desired, then according to Bárdossy and Singh (2008), the deepest parameter set (which represents the most central point in the whole parameter vectors) is the most likely candidate to be robust. This study also indicates the deepest parameter set perform slightly better than the mean of the parameter sets considered.

8.2 Variability and estimation of η

As defined, the water balance related parameter η is specific for each catchment and each model parameter vector. Therefore, each individual catchment has a large variation in η for the calibrated 10000 parameter sets. And for the same set of good parameters that matching different water balances, different catchments always require very different η-s to control actual evapotranspiration. Parameter η was not transferred, only the other parameters controlling flow dynamics and short term water balances were assumed to be shared by many catchments. However, regionalization of η directly is not feasible and η remains different after regionalization. In the numerical experiments, in order to estimate water balance parameter η, the long term discharge volumes were treated as known variables for both gauged and ungauged catchments. For application in practical system, the long term discharge volumes have to be estimated for ungauged catchments. This problem is not explicitly treated in this paper. For the study area, the discharge coefficients which relate discharge volumes to (known) precipitation show a quite smooth spatial behavior as shown on Fig. 14. Thus the regionalization of this parameter seems to be a not extremely complicated task. According to the previous analysis of η, for each common dynamical parameter set, one can have a possible estimator of η for a certain catchment based on the regionalization of discharge coefficients.

8.3 Prediction in ungauged basins

The results of this study supported the general finding of Ricard et al. (2012) and Gaborit et al. (2015), where the simultaneous calibration lead to weaker model performance than the individual one for both calibration and validation time period. The loss of model performance in validation is smaller than that in calibration. When applied to ungauged catchments, the simultaneous calibration shows more robustness than the individual one. Simultaneous calibration of models in geographical space offers a good possibility for the runoff prediction in ungauged basins. Compared with traditional regionalization method, only the water balance parameter η has to be estimated based on the regionalization of discharge coefficients.

It was examined from the hydrographs that high flows are often underestimated and low flows are probably overestimated. This kind of phenomenon has also been detected in previous regional calibration studies Ricard et al. (2012); Gaborit et al. (2015). This behavior mainly due to the uncertainty of model structure and the low spatial and temporal resolutions of both models and input variables Gaborit et al. (2015).

9 Conclusions

In this paper, the transfer of the dynamical parameters of hydrological models was investigated. A new model parameter η controlling the actual evapotranspiration was introduced to cope with the clear differences in water balances due to water or energy limitations. Three hydrological models were used in combination with three different performance measures in three numerical experiments on a large number of catchments.

The individual calibration and transfer results indicate that models are often overfitted during calibration. The parameters are sometimes more specific for the calibration time period and their relation to catchment properties seems to be unclear. This makes parameter transfers or parameter regionalization based on individual calibration difficult. The common spatial calibration strategy, which explicitly assumed that catchments share dynamical parameters, was tested on a number of 15 catchments and 96 catchments, respectively. The common calibration provides an effective way to identify parameter sets which work reasonably for all catchments within the modeled domain. Testing the parameters on an independent time period shows that common parameters perform comparably well as those obtained using individual calibration. The transfer of the common parameters to model ungauged catchments works well. The performance of common parameters on a small number(15) of catchments was better than on a big number (96) of catchments covering a large spatial scale. It indicates that the performance of the common parameters depends strongly on the selection of the catchments used to assess them and a reasonable geographic proximity of the catchments might be a good choice for common calibration. The results of the experiments were similar for all three hydrological models applied independently of the choice of the performance measures. Note however that the common parameters corresponding to the different performance measures differ considerably. Common behavior is dependent on how one evaluates the performance of the models.

The fact that many catchments share common parameters which describe their dynamical behavior does not mean that they have the same dynamical behavior. The model output highly depends on the parameter η which varies from catchment to catchment and also as a function of the other model parameters describing dynamical behavior. Common parameters offer a good possibility for the prediction of ungauged catchments, only the parameter η which controls the long term water balances has to be estimated individually. This however can be done using other modelling approaches including regionalization methods.

In this study, all the models were tested on the daily time scale. The results show that many catchments behave similar as the same dynamical parameter sets could perform reasonable for all of them. This means that hydrological behavior on the daily scale is mainly dominated by precipitation characteristics and actual evapotranspiration and we believe that differences in catchment properties rather have significant effects on smaller temporal scales like hourly. Results also indicate that the differences in catchment properties cannot be captured well by simple lumped model parameters.

Acknowledgements. The research of the second author (Yingchun Huang) was supported by China Scholarship Council. The authors wish to thank two anonymous reviewers for their helpful comments and suggestions.

References

Ali, G., Tetzlaff, D., Soulsby, C., McDonnell, J. J., and Capell, R.: A comparison of similarity indices for catchment classification using a cross-regional dataset, Advances in Water Resources, 40, 11–22, 2012.

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.-H., Oudin, L., Mathevet, T., Lerat, J., and Berthet, L.: All that glitters is not gold: the case of calibrating hydrological models, Hydrological Processes, 26, 2206–2210, 2012.

Archfield, S. A. and Vogel, R. M.: Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments., Water Resources Research, 46, W10 513, 2010.

Bárdossy, A.: Calibration of hydrological model parameters for ungauged catchments., Hydrol. Earth Syst. Sci., 11, 703–710, 2007.

Bárdossy, A. and Singh, S. K.: Robust estimation of hydrological model parameters., Hydrology and Earth System Sciences, 12, 1273–1283, 2008.

Bergström, S. and Forsman, A.: Development of a conceptual deterministic rainfall-runoff model., Nordic Hydrology, 4, 174–190, 1973.

Beven, K. and Freer, J.: Equifinality, data assimilation, and data uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of Hydrology, 249, 11–29, 2001.

Beven, K. J.: Uniqueness of place and process representations in hydrological modelling, Hydrology and Earth System-Sciences, 4, 203–213, 2000.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H. e.: Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales., Cambridge University Press,, Cambridge, 2013.

Boyle, D. P., Gupta, H. V., Sorooshian, S., Koren, V., Zhang, Z., and Smith, M.: Toward Improved Streamflow Forecasts: Value of Semidistributed Modeling., Water Resources Research, 37 (11), 2749–2759, 2001.

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Goteti, G., Gupta, H., Gusev, Y., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, Journal of Hydrology, 320, 3–17, 2006.

Falcone, J. A., Carlisle, D. M., Wolock, D. M., and Meador, M. R.: GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States: Ecological Archives E091-045, Ecology, 91, 621–621, 2010.

Fernandez, W., Vogel, R., and Sankarasubramanian, A.: Regional calibration of a watershed model, Hydrological Sciences Journal, 45, 689–707, 2000.

Gaborit, É., Ricard, S., Lachance-Cloutier, S., Anctil, F., Turcotte, R., and Polat, A.: Comparing global and local calibration schemes from a differential split-sample test perspective, Canadian Journal of Earth Sciences, 52, 990–999, 2015.

Grigg, D.: THE LOGIC OF REGIONAL SYSTEMS 1, Annals of the Association of American Geographers, 55, 465–491, 1965.

Gupta, H., Kling, H., Yilmaz, K., and Martinez, G.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, 2009.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., et al.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological sciences journal, 58, 1198–1255, 2013.

McDonnell, J. and Woods, R.: On the need for catchment classification, Journal of Hydrology, 299, 2–3, 2004.

McIntyre, N., Lee, H., Wheater, H., Young, A., and Wagener, T.: Ensemble predictions of runoff in ungauged catchments, Water Resources Research, 41, 2005.

Moore, R. J.: The probability-distributed principle and runoff production at point and basin scales., Hydrological Sciences Journal, 30(2), 273–297, 1985.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models. 1. A discussion of principles, Journal of Hydrology, 10, 282–290, 1970.

Oudin, L., Kay, A., Andréassian, V., and Perrin, C.: Are seemingly physically similar catchments truly hydrologically similar?, Water Resources Research, 46, 2010.

Parajka, J., Blöschl, G., and Merz, R.: Regional calibration of catchment models: Potential for ungauged catchments, Water Resources Research, 43, 2007.

Razavi, T. and Coulibaly, P.: Streamflow prediction in ungauged basins: review of regionalization methods, Journal of Hydrologic Engineering, 18, 958–975, 2012.

Ricard, S., Bourdillon, R., Roussel, D., and Turcotte, R.: Global calibration of distributed hydrological models for large-scale applications, Journal of Hydrologic Engineering, 18, 719–721, 2012.

Samaniego, L., Bárdossy, A., and Kumar, R.: Streamflow prediction in ungauged catchments using copula-based dissimilarity measures., Water Resources Research, 46, W02 506, 2010.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P., and Carrillo, G.: Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA., Hydrology & Earth System Sciences, 15, 2011.

Schaefli, B. and Gupta, H.: Do Nash values have value?, Hydrological Processes, 21, 2075–2080, 2007.

Sivakumar, B. and Singh, V.: Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework, Hydrology and Earth System Sciences, 16, 4119–4131, 2012.

Sivapalan, M.: Prediction in ungauged basins: a grand challenge for theoretical hydrology, Hydrological Processes, 17, 3163–3170, 2003.

Toth, E.: Catchment classification based on characterisation of streamflow and precipitation time series, Hydrology and Earth System Sciences, 17, 1149–1159, 2013.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrology and Earth System Sciences, 5, 13–26, 2001.

Wagener, T., Sivapalan, M., Troch, P., and Woods, R.: Catchment classification and hydrologic similarity, Geography Compass, 1, 901–931, 2007.

Zeleny, M.: Multiple Criteria Decision Making., McGraw-Hill, New York, USA, 1981.

Zhao, R.J.and Liu, X.: The Xinanjiang model. In: Computer Models of Watershed Hydrology, Water Resources Publications, Littleton, Colorado, USA, 1995.

Table 1. Catchment properties for the selected 15 catchments

Streamgauge ID	Streamgauge name	Drainage area (km^2)	Shape factor	Field capacity	Average porosity	Base flow index	Snow proportion (%)
01548500	Pine Creek at Cedar Run, PA	1564	0.14	0.32	0.42	0.44	26.6
01606500	So. Branch Potomac River near Petersburg, WA	1663	0.15	0.31	0.28	0.45	19.5
01611500	Cacapon River near Great Cacapon, WV	1753	0.17	0.269	0.27	0.41	15.6
01663500	Hazel River at Rixeyville at Rixeyville, VA	743	0.16	0.30	0.39	0.51	12.1
01664000	Pappahannock River at Remington, VA	1606	0.11	0.294	0.40	0.50	11.8
01667500	Rapidan River near Culpeper, VA	1222	0.13	0.32	0.40	0.51	10.6
02016000	Cowpasture River near Clifton Forge, VA	1194	0.18	0.28	0.27	0.43	16.0
02018000	Craig Creek at Parr, VA	852	0.24	0.27	0.30	0.44	11.3
02030500	Slate River near Arvonia, VA	585	0.20	0.30	0.46	0.48	8.5
03114500	Middle Island Creek at Little, WV	1186	0.14	0.36	0.27	0.21	15.6
03155500	Hughes River at Cisco, WV	1171	0.14	0.36	0.27	0.22	14.9
03164000	New River near Galax, VA	2929	0.09	0.29	0.43	0.64	13.3
03173000	Walker Creek at Bane, VA	790	0.24	0.32	0.37	0.46	13.5
03180500	Greenbrier River at Durbin, WV	344	0.26	0.36	0.27	0.37	25.3
03186500	Williams River at Dyer, WV	332	0.33	0.36	0.28	0.36	24.3

Table 2. Climate variables of the 15 selected catchments.

No	Streamgauge ID	Annual precipitation (mm)	Average temperature (°C)	Annual potential evapotranspiration (mm)	Annual runoff (mm)
1	01548500	951.7	7.2	727.0	495.1
2	01606500	948.6	10.3	716.3	378.3
3	01611500	905.6	10.8	800.0	310.5
4	01663500	1049.9	11.7	897.2	402.6
5	01664000	1027.7	12.0	906.1	367.5
6	01667500	1087.4	12.3	915.2	380.4
7	02016000	1029.5	11.0	746.0	402.9
8	02018000	1010.6	11.4	764.6	406.3
9	02030500	1075.9	13.5	918.2	350.3
10	03114500	1089.7	11.4	737.4	483.9
11	03155500	1057.8	11.6	740.0	443.7
12	03164000	1247.9	10.6	807.4	593.3
13	03173000	958.6	11.1	762.7	371.9
14	03180500	1224.2	8.3	710.9	543.2
15	03186500	1401.5	9.1	710.9	945.0

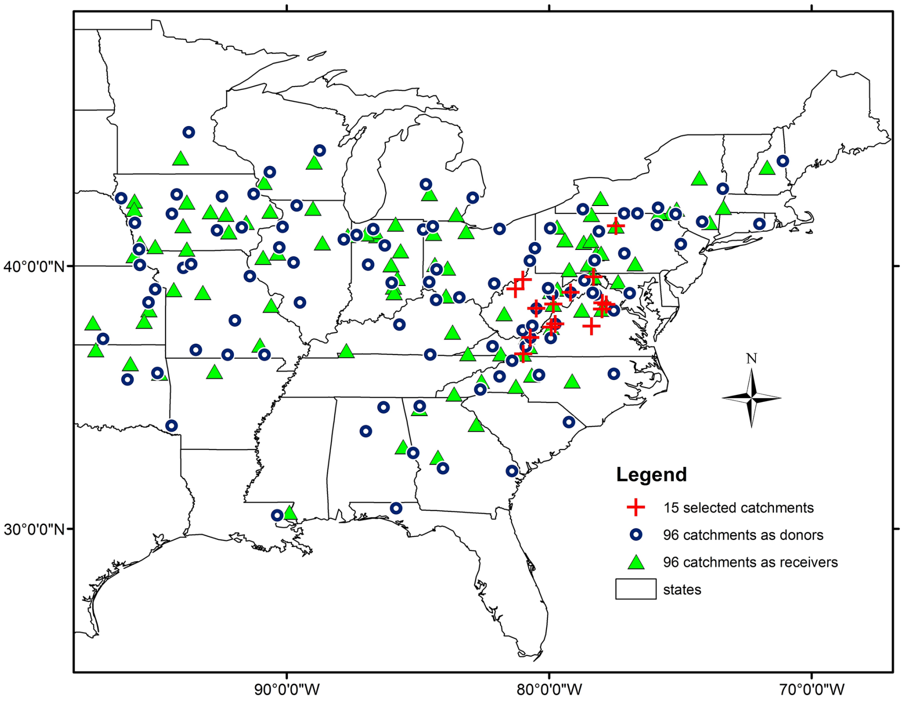

Figure 1. Location of the catchments selected for the experiments.

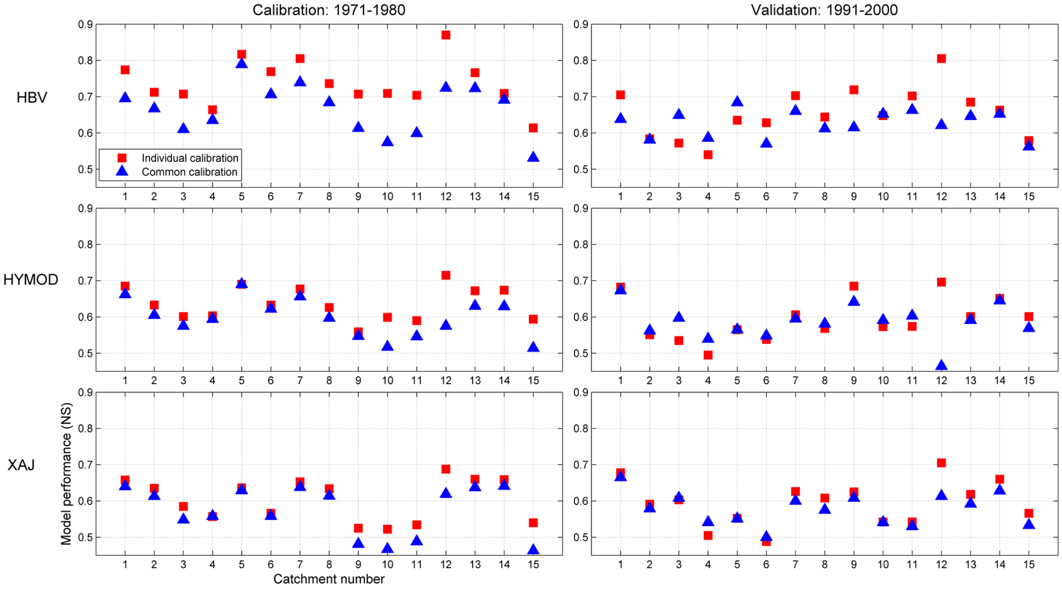

Figure 2. Performance of the individually calibrated and the common calibrated models using NS as performance criterion.

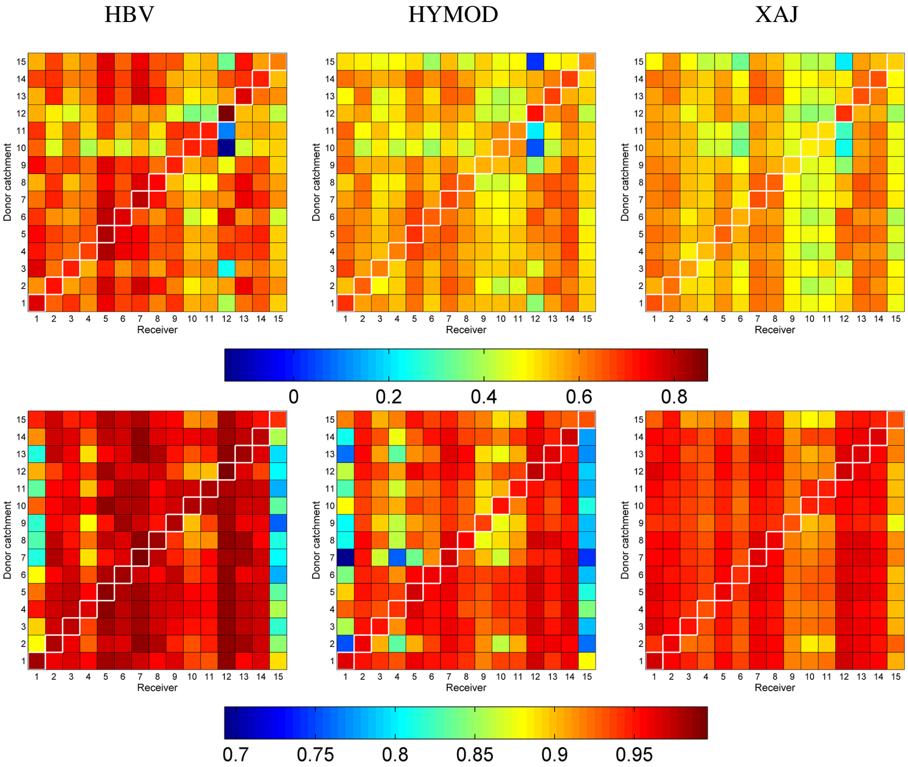

Figure 3. Color coded matrices for the mean model performance of the parameter transfer for the selected 15 catchments. The upper panel used NS as performance measure, the lower panel used GK as performance measure.

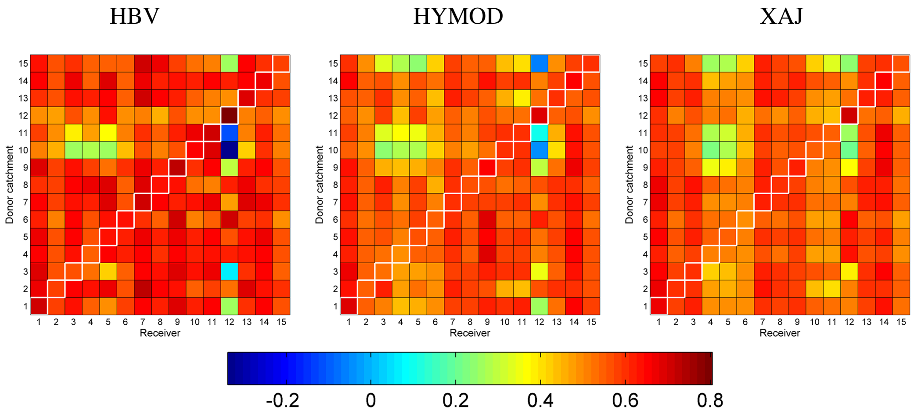

Figure 4. Color coded matrices for the mean NS model performance of the parameter transfer for the validation period for the selected 15 catchments.

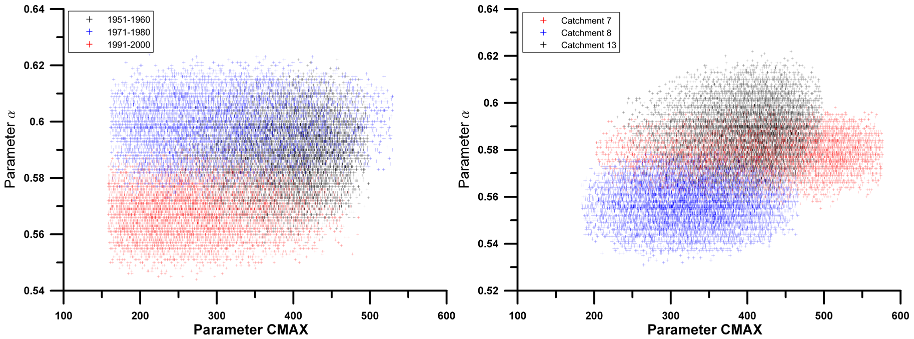

Figure 5. Scatterplots for two selected HYMOD parameters CMAX and α obtained via model calibration using NS as performance measures. Left: for catchment 13 (black: 1951–1960, blue: 1971–1980 and red: 1991–2000); Right: for catchments 7 (red), 8 (blue) and 13 (black) for 1951–1960

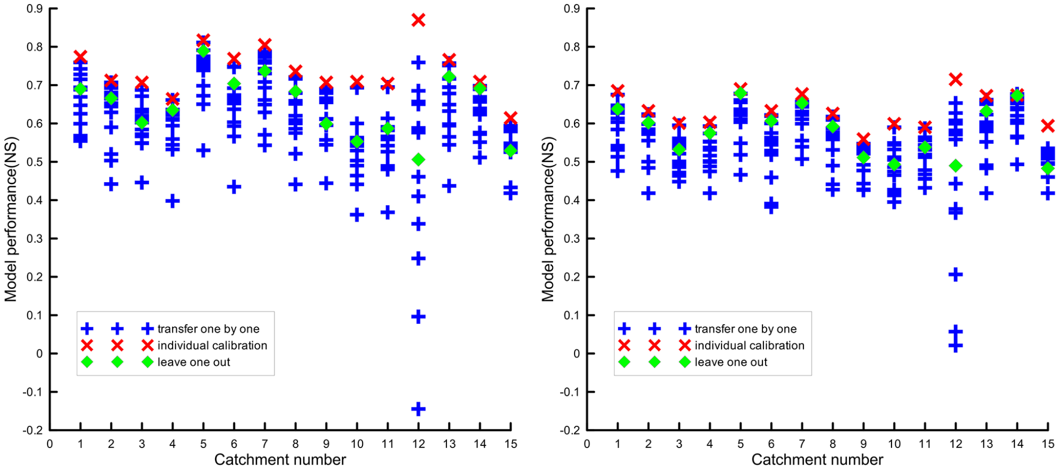

Figure 6. Mean NS model performance of the calibration, individual parameter transfer and for the leave one out transfer for the selected 15 catchments for the calibration time period 1971–1980. Left panel: HBV, right panel: HYMOD.

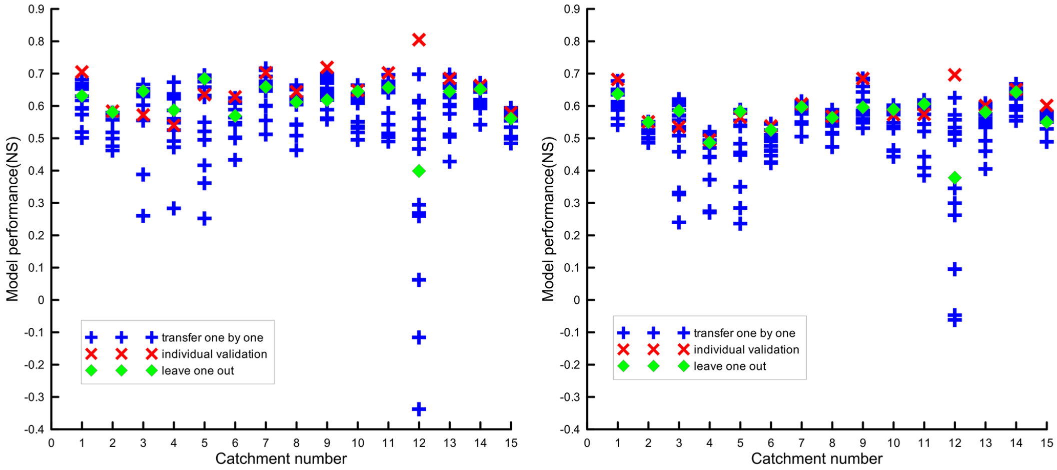

Figure 7. Mean NS model performance of the calibration, individual parameter transfer and for the leave one out transfer for the selected 15 catchments for the validation time period 1991–2000. Left panel: HBV, right panel: HYMOD.

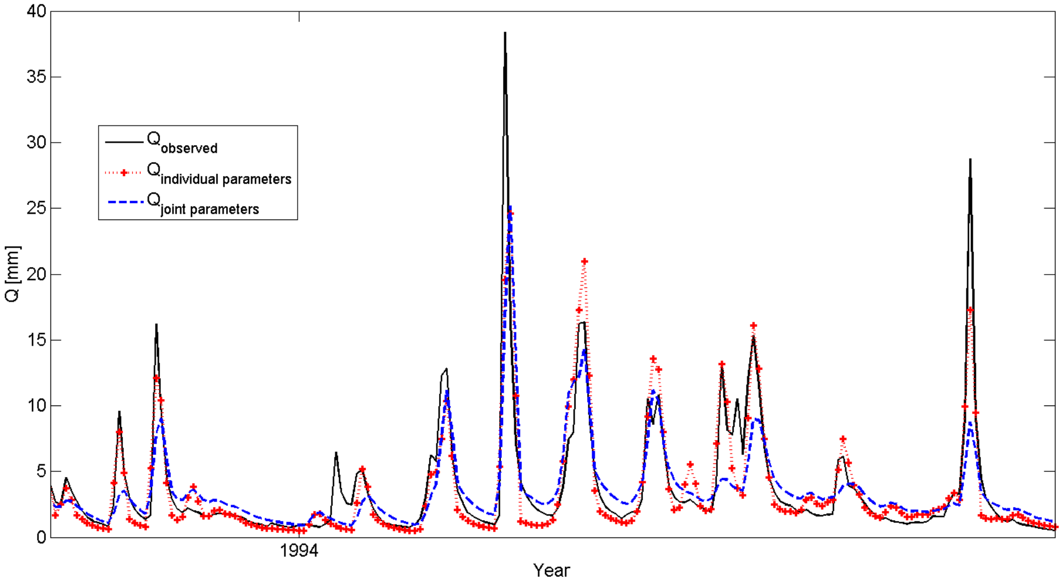

Figure 8. Runoff hydrographs for catchment 14 obtained using individual and leave one out common calibrations of HBV using the GK performance measure.

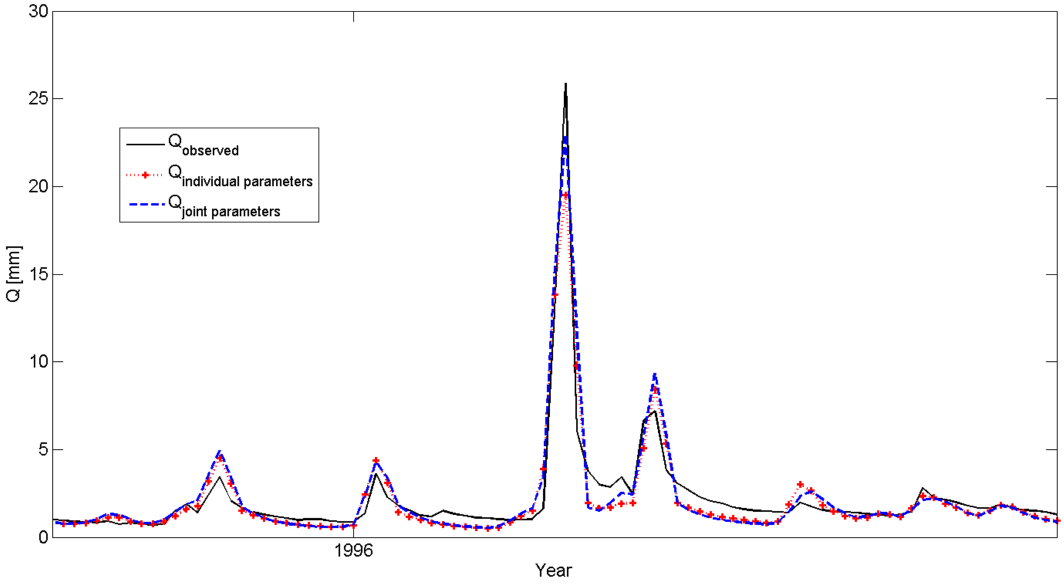

Figure 9. Runoff hydrographs for catchment 5 obtained using individual and leave one out common calibrations of HBV using the NS performance measure.

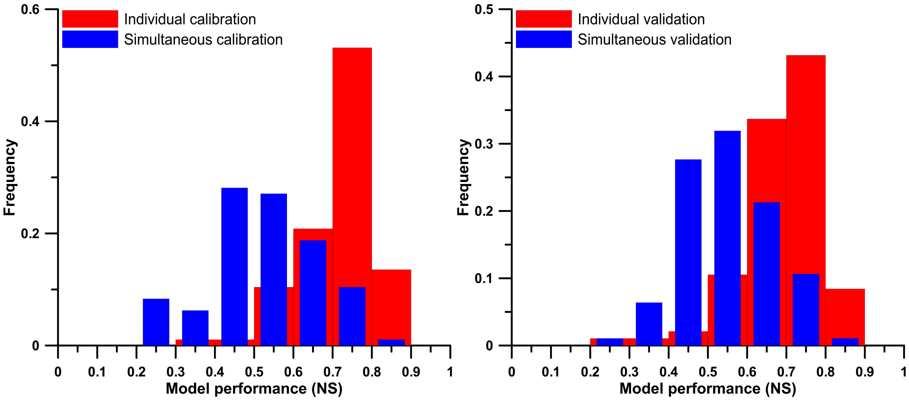

Figure 10. Histograms of the NS model performance of HBV for the 96 selected (donor) catchments. Left: calibration period (1971–1980), right: validation period (1991–2000).

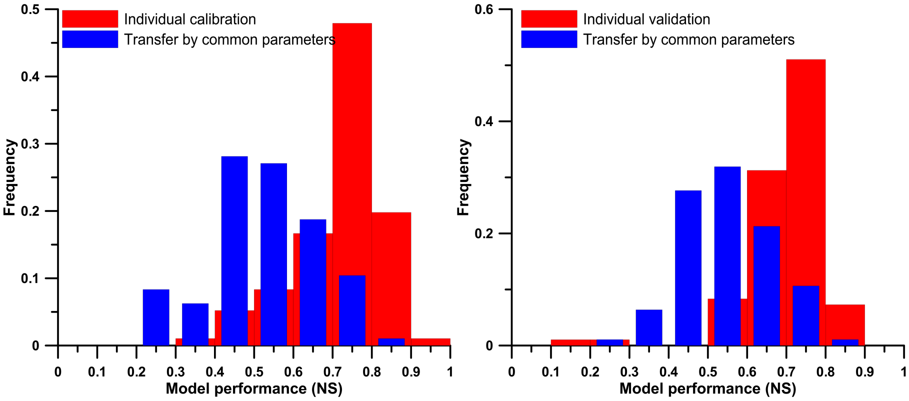

Figure 11. Histograms of the NS model performance of HBV for the 96 test (ungauged) catchments. Left: calibration period (1971–1980), right: validation period (1991–2000).

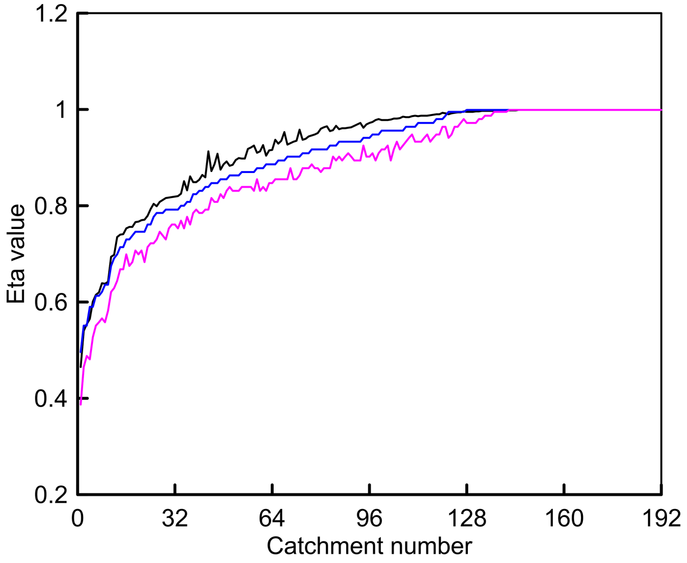

Figure 12. Distribution of water balance parameter η for three randomly selected common parameter vectors obtained via HBV using the NS performance measure for 192 selected catchments.

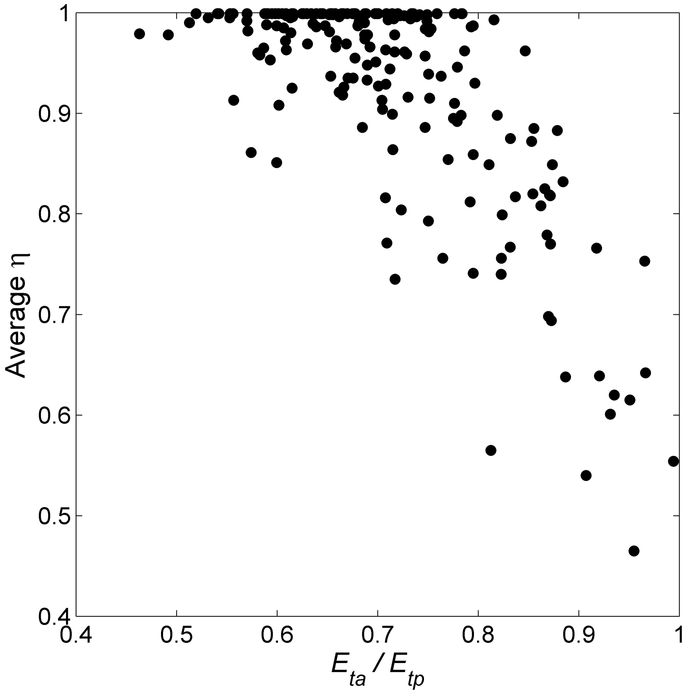

Figure 13. Scatterplots of mean η value and ratio of actual evapotranspiration to potential evapotranspiration for 192 selected catchments.

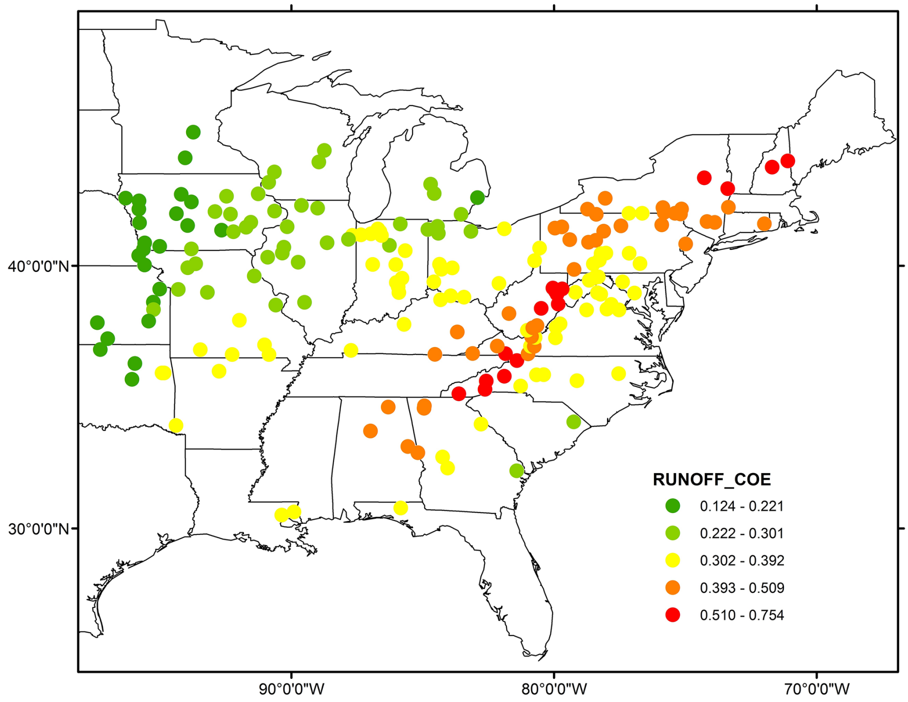

Figure 14. The discharge coefficient of the catchments selected for the experiments.