Hydrol. Earth Syst. Sci., 17, 149–161, 2013
www.hydrol-earth-syst-sci.net/17/149/2013/
doi:10.5194/hess-17-149-2013

Hydrology and Earth System Sciences

An approach to identify time consistent model parameters: sub-period calibration

S. Gharari[1,2], **M. Hrachowitz**[1], **F. Fenicia**[1,2], **and H. H. G. Savenije**[1]

[1]Delft University of Technology, Faculty of Civil Engineering and Geosciences, Water Resources Section, Delft, The Netherlands
[2]Public Research Center-Gabriel Lippmann, Belvaux, Luxembourg

Correspondence to: S. Gharari (s.gharari@tudelft.nl)

Received: 2 February 2012 – Published in Hydrol. Earth Syst. Sci. Discuss.: 13 February 2012
Revised: 28 November 2012 – Accepted: 5 December 2012 – Published: 17 January 2013

Abstract. Conceptual hydrological models rely on calibration for the identification of their parameters. As these models are typically designed to reflect real catchment processes, a key objective of an appropriate calibration strategy is the determination of parameter sets that reflect a "realistic" model behavior. Previous studies have shown that parameter estimates for different calibration periods can be significantly different. This questions model transposability in time, which is one of the key conditions for the set-up of a "realistic" model. This paper presents a new approach that selects parameter sets that provide a consistent model performance in time. The approach consists of testing model performance in different periods, and selecting parameter sets that are as close as possible to the optimum of each individual sub-period. While aiding model calibration, the approach is also useful as a diagnostic tool, illustrating tradeoffs in the identification of time-consistent parameter sets. The approach is applied to a case study in Luxembourg using the HyMod hydrological model as an example.

1 Introduction

Conceptual hydrological models represent an abstraction of real world processes, and are typically constituted of a number of interconnected reservoirs which are supposed to represent the main catchment compartments and dominant processes (Wagener et al., 2003). Typically, several of these model parameters are not measurable, even if they are supposed to represent physical catchment characteristics, and as a result they have to be determined by calibration (Wheater et al., 1993). Different approaches to infer parameter values and their distributions have been developed, for example single or multi-objective calibration (Gupta et al., 1998), generalized likelihood uncertainty estimation (GLUE, Beven and Binley, 1992), dynamic identifiability analysis (DYNIA, Wagener et al., 2003) and Bayesian inference (Wood and Rodríguez-Iturbe, 1975).

A key objective for hydrological modeling is the development of "realistic" models, that is, models which are able to reflect real catchment processes (Wagener, 2003). The set-up of a realistic model requires the determination of a realistic model structure and a suitable parameterization. While the determination of a suitable model structure is a theoretical development in its own right (e.g. Wagener et al., 2002; Fenicia et al., 2007, 2011; Clark et al., 2008; Savenije, 2009), we focus here on the determination of realistic parameter sets, and in particular, on parameter sets that reflect a consistent model behavior in time.

Model transposability in time is in fact recognized as one of the main requirements to a successful "validation" of model performance (Klemeš, 1986). Hartmann and Bárdossy (2005) advocate that "if a model is to be used under non-stationary conditions, its parameters and process descriptions should be transferable".

The calibration–validation approach (or the split-sample test proposed by Klemeš, 1986) has become standard in hydrological practice (Andréassian et al., 2009). A model is calibrated for a period of time and the parameter sets which are selected as behavioral in the calibration period

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

are subsequently evaluated for a different validation period. Several combinations of calibration and validation for multiple-response data were suggested by Mroczkowski et al. (1997). Calibration and validation is proposed as a crucial step in the comprehensive model developing scheme proposed by Refsgaard et al. (2005).

Seibert (2003) pointed out that the success of identifying the best parameter set (or model structure) relies on the selection of time periods with similar characteristics. He argued that the reason for the scarce literature on models which perform well in time periods with characteristics different from the calibration period is due to the fact that they most probably fail this test (i.e. the differential split sample test proposed by Klemeš, 1986). Kirchner (2006) criticized commonly used model evaluation methods. He argued that "such models are often good mathematical marionettes; they often can dance to the tune of the calibration data. However, their predictive validity is often in doubt". This shortcoming was repeatedly addressed in the literature (Anderson and Woessner, 1992; Hassan, 2004; Gupta et al., 2008; Refsgaard and Hansen, 2010).

The failure of validation tests has its counterpart in the fact that calibrated model parameters are inherently linked to the calibration time period, and may be inadequate to represent other periods. Wagener et al. (2003) developed a method to screen across the time series of model prediction in order to investigate the identifiability of model parameters. They show that uncertainties associated to model parameters can vary substantially in different time periods. Coron et al. (2012) used a similar concept to investigate the performance of the three models in contrasting climate conditions. They questioned the validity of parameter transferability in time due to varying climate conditions.

Previously, Freer et al. (2003) evaluated the dynamic TOPMODEL using GLUE with different objective functions based on the rising or falling limbs of the hydrograph. They showed that it may be difficult to propose a consistently parameterized model structure due to the significant variability of the observed responses. They concluded that the model fails to meet even relaxed acceptability thresholds. Hartmann and Bárdossy (2005) investigated parameter transferability in different climatic conditions ("warm", "cold", "wet" and "dry") and for different time scales (from days to years). They designed a calibration method that allows a good performance on different time scales simultaneously. Li et al. (2012) investigated the transferability of model parameters for dry and wet conditions. They showed that the dry period contains more information for model calibration than the wet period. Bárdossy and Singh (2008), using the depth function (Tukey, 1975), concluded "that equally performing parameters are not necessarily equally transferable or equally sensitive".

While the decrease of model performance in the validation period can have many causes, we focus here on how it is affected by the parameter selection approach. Various approaches have been proposed to extract meaningful hydrological information from the observed time series. Boyle et al. (2000, 2001) used the multi-objective calibration approach proposed by Gupta et al. (1998) to calibrate a model for different flow segments of the hydrograph. The multi-objective approach makes it possible to identify optimal parameter sets for a set of objective functions. This approach was extensively used in several applications (for a review see Efstratiadis and Koutsoyiannis, 2010). Incorporating multiple calibration-criteria, for instance tracer data or remotely sensed evaporation, into model calibration helps in identifying a more realistic model structure and parameter sets (e.g. Weiler et al., 2003; Freer et al., 2004; Uhlenbrook and Sieber, 2005; Vaché and McDonnell, 2006; Son and Sivapalan, 2007; Winsemius et al., 2008; Dunn et al., 2008; Birkel et al., 2010; Fenicia et al., 2010; Hrachowitz et al., 2012).

Both multi-objective and multi-criteria optimizations constrain the feasible parameter space and facilitate parameter selection on the basis of performance trade-offs, i.e. Pareto fronts. However, as argued by Beven (2006), the mere mappings of optimum parameter sets after calibration are: "too simplistic, since they arbitrarily exclude many models that are very nearly as good as the *optima*". As argued by Andréassian et al. (2012), mathematically optimum parameter sets may be far different from hydrologically optimum parameter sets. These arguments simply imply that the parameter realization should include "sub-optimal" parameter sets as well.

Hence the question of how to retain model parameters that have a consistent model behavior in time deserves further investigations. A related challenge is how to establish the tradeoff between behavioral and non-behavioral parameters in a meaningful way.

With the attempt to address this question, we introduce a new approach for parameter identification including optimal and sub-optimal parameter sets which are more time consistent. The method is based on the calibration on different periods, and determines the parameter sets which perform best for all these sub-periods. As the selected parameter sets are evaluated in different periods, only the time consistent parameter sets are selected. The new method is applied to a case study in the Wark catchment in Luxembourg, using the lumped conceptual model HyMod, and compared with a calibration–validation approach with respect to parameter identifiability and performance.

2 Sub-period calibration

The aim of the sub-period calibration is to identify a time consistent parameterization for a certain model structure and data set. The approach involves two steps. First, the available input and output data sets are split into (ideally equal-length) k sub-periods. These sub-periods and their lengths

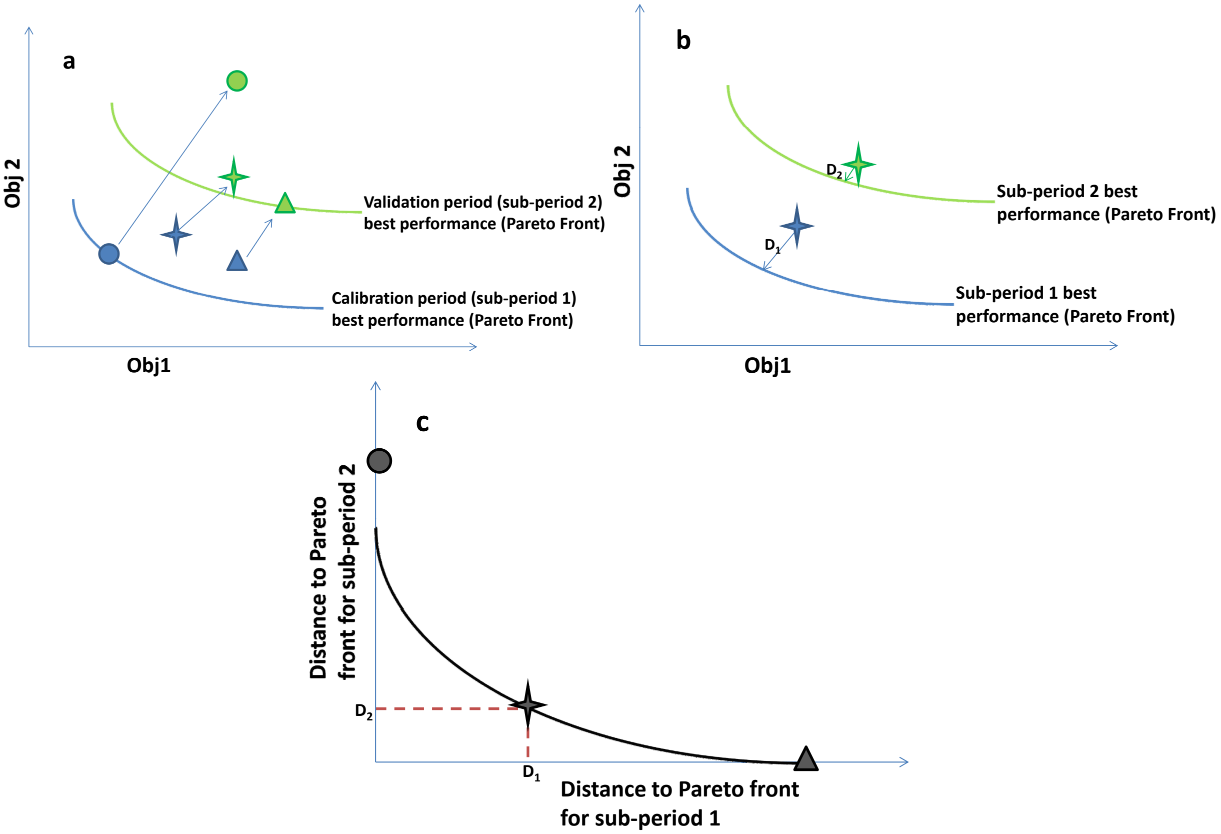
a
Obj 2
Obj1
Validation period (sub-period 2) best performance (Pareto Front)
Calibration period (sub-period 1) best performance (Pareto Front)
b
Obj 2
Obj1
D_2
D_1
Sub-period 2 best performance (Pareto Front)
Sub-period 1 best performance (Pareto Front)
c
Distance to Pareto front for sub-period 2
Distance to Pareto front for sub-period 1
D_2
D_1

Fig. 1. Schematic illustration of the sub-period calibration approach (circles, stars and triangles represent the performance of different parameter sets in the 2 periods): **(a)** Calibration–validation of a two dimensional optimization problem; the lines represent the Pareto fronts in 2 periods (CPF_1 and CPF_2 for sub-period calibration, respectively). **(b)** Proposed method of calibration that aims at minimizing the distance to the 2 Pareto fronts (CPF_1 and CPF_2) of each sub-period. **(c)** Minimum distance Pareto front (MDPF). Performance of parameter sets in every sub-period is depicted by the same color as the calibration Pareto front (CPF) of that specific sub-period.

can be arbitrarily chosen. They can, for example, be months, seasons, years, or wetness conditions (e.g. Hartmann and Bárdossy, 2005; Seiller et al., 2012). Additionally, a number n of objective functions is defined.

Each sub-period is then calibrated individually by sampling the parameter space and identifying the n-dimensional Pareto front for each sub-period. Therefore k n-dimensional calibration Pareto fronts (CPF) are obtained.

Subsequently, the parameter space is sampled to find parameter sets which minimize the distance to the k Pareto fronts. Distance measures can, for example, be the Euclidian distance to the Pareto front or any other measure which evaluates the performance of a parameter set relative to the Pareto front. This leads, for each parameter set, to k distances for each of the k sub-periods.

The goal is to find parameter sets that minimize the distances to all Pareto fronts. In order to achieve this, in a k-dimensional space, we represent each parameter set by its distance to each of the k Pareto fronts. The Pareto front of this cloud of points represents the parameter sets with minimum distance to all Pareto fronts. We call it the minimum distance Pareto front (MDPF). It contains the parameter sets that have the most consistent performance in each sub-period.

The concept is illustrated in Fig. 1 with a schematic 2-objective function, 2-sub-period example. The CPFs for the two sub-periods are shown in Fig. 1a. The circle represents a parameter set that is a Pareto member of the first sub-period (zero distance to the CPF_1); however, it does not perform well compared to the optimum in the second sub-period (large distance to the CPF_2). The parameter set represented by the triangle, although sub-optimal in the first sub-period, is a Pareto member in the second sub-period. The parameter set represented by the star, on the other hand, although not a Pareto member in both sub-periods, performs rather well overall (small distance to both the CPF_1 and CPF_2).

Figure 1c plots the distance of each parameter set to the Pareto fronts. The circle has zero distance to CPF_1, and large distance to CPF_2. It does not belong to the MDPF. The triangle has zero distance to CPF_2, and small distance to CPF_1, indicating the edge of MDPF. The star has small distances to both CPF_1 and CPF_2, and it belongs to the MDPF at some intermediate position.

Table 1. Rainfall, runoff and potential evaporation for year 1998 to 2009 for the Wark catchment.

year	Rainfall ($\mathrm{mm\,yr^{-1}}$)	Runoff ($\mathrm{mm\,yr^{-1}}$)	Potential evaporation ($\mathrm{mm\,yr^{-1}}$)
1998	997	370	709
1999	1065	381	738
2000	1062	419	729
2001	1198	397	725
2002	1025	406	744
2003	788	225	797
2004	865	247	713
2005	738	154	741
2006	830	244	774
2007	983	410	750
2008	966	418	727
2009	886	397	749

3 Case study

3.1 Study area and data

The outlined methodology will, in the following, be illustrated with a case study using data from the Wark catchment in the Grand Duchy of Luxembourg. The catchment has an area of 82 $\mathrm{km^2}$ with the catchment outlet located downstream of the town of Ettelbrück at the confluence with the Alzette River (49.85° N, 6.10° E). With an average precipitation of 850 $\mathrm{mm\,yr^{-1}}$ and an average potential evaporation of 650 $\mathrm{mm\,yr^{-1}}$ the average runoff is approximately 250 $\mathrm{mm\,yr^{-1}}$. The geology in the northern part is dominated by schist while the southern part of the catchment is mostly underlain by sandstone and conglomerate. The dominant land uses are forest on hillslopes, agricultural land on plateaus and pastures in the valley bottoms. The elevation varies between 195 to 532 m a.s.l. with an average of 380 m a.s.l. The slope of the catchment varies between 0–200 %, with an average of 17 % (Gharari et al., 2011). The hydrological data include: discharge at the outlet of the Wark catchment, potential evaporation estimated by the Hamon equation (Hamon, 1961) with data measured at Findel (Luxembourg airport; Fenicia et al., 2008), and precipitation by three tipping bucket rain gauges. The data series has been discretized at 12-h resolution. For model evaluation, the period 1998–2009 was used. The meteorological conditions of each year are summarized in Table 1.

3.2 Hydrological model

The rainfall-runoff model applied to the Wark catchment is the lumped conceptual HyMod model (Wagener et al., 2001). HyMod was chosen for its low number of parameters while still maintaining adequate process representation including slow and fast responses together with a non-linear soil moisture component.

HyMod is characterized by five reservoirs, including the soil moisture reservoir (S_M[L]), three linear reservoirs in series (S_{F_1}[L], S_{F_2}[L], S_{F_3}[L]) mimicking the fast runoff component, and one slow reservoir (S_{S_1}[L]). It has five parameters representing the maximum soil moisture storage capacity ($S_\mathrm{M,max}$(L)), the spatial variability of soil moisture (β[-]), the partitioning between fast reservoirs and slow reservoir (α[-]), as well as the timescales of the fast and slow reservoirs ($R_\mathrm{F}[\mathrm{T}^{-1}]$, $R_\mathrm{S}[\mathrm{T}^{-1}]$). Model equations were solved using the forward explicit Euler method using 12-h resolution time series.

$P[\mathrm{LT}^{-1}]$, $E_\mathrm{a}[\mathrm{LT}^{-1}]$, $E_\mathrm{p}[\mathrm{LT}^{-1}]$ and $Q_\mathrm{m}[\mathrm{LT}^{-1}]$ represent precipitation, actual evaporation, potential evaporation and modeled runoff, respectively. The simulated runoff by the model is the summation of slow and fast components ($Q_\mathrm{m} = Q_{\mathrm{S}_1} + Q_{\mathrm{F}_3}$). The water balance equations and constitutive relations are listed in Table 2 and the HyMod schematic illustration is depicted in Fig. 2.

3.3 Implementation of sub-period calibration

In the following, two case studies are presented where we compare performance and selected parameter sets by two approaches: (1) calibration over the entire length of a (sub-)period, which for sake of simplicity thereafter is referred to as standard calibration; and (2) calibration over decomposed sub-periods which is referred to as SuPer (sub-period) calibration. The case studies are designed to show the performance of SuPer calibration for parameter identification extracting information from sub-periods. The first case study intends to make the best use of limited available data by decomposing it into different sub-periods. The second case study intends to investigate how standard calibration might average out the characteristics of the sub-periods over the long time series.

3.3.1 Case study 1 – Short data series

The 3 consecutive years 2001–2003 are used for model evaluation, with the year 2001 selected as the warm-up period.

1. The model is calibrated using standard calibration on the year 2002 and Pareto front members (CPF_{2002}) are validated for the year 2003.

2. The model is calibrated using standard calibration on the year 2003 and Pareto front members (CPF_{2003}) are validated for the year 2002.

3. The model is calibrated using standard calibration on the years 2002–2003 and Pareto front members ($\mathrm{CPF}_{2002-2003}$) are validated for the individual years of 2002 and 2003.

Table 2. Equations used in HyMod.

Reservoir	Water balance equations	Constitutive relations	
Soil moisture (S_{M})	$\mathrm{d}S_{\mathrm{M}}/\mathrm{d}t = P - P_{\mathrm{e}} - E_{\mathrm{a}}$	$P_{\mathrm{e}} = FP$ $E_{\mathrm{a}} = WE_{\mathrm{p}}$	$F = 1 - (1 - S_{\mathrm{M}}/S_{\mathrm{M,max}})^{\beta}$ $W = \lceil \frac{S_{\mathrm{M}}}{S_{\mathrm{M,max}}} \rceil$
First fast reservoir ($S_{\mathrm{F_1}}$)	$\mathrm{d}S_{\mathrm{F_1}}/\mathrm{d}t = \alpha P_{\mathrm{e}} - Q_{\mathrm{F_1}}$	$Q_{\mathrm{F_1}} = S_{\mathrm{F_1}} R_{\mathrm{F}}$	
Second fast reservoir ($S_{\mathrm{F_2}}$)	$\mathrm{d}S_{\mathrm{F_2}}/\mathrm{d}t = Q_{\mathrm{F_1}} - Q_{\mathrm{F_2}}$	$Q_{\mathrm{F_2}} = S_{\mathrm{F_2}} R_{\mathrm{F}}$	
Third fast reservoir ($S_{\mathrm{F_3}}$)	$\mathrm{d}S_{\mathrm{F_3}}/\mathrm{d}t = Q_{\mathrm{F_2}} - Q_{\mathrm{F_3}}$	$Q_{\mathrm{F_3}} = S_{\mathrm{F_3}} R_{\mathrm{F}}$	
Slow reservoir ($S_{\mathrm{S_1}}$)	$\mathrm{d}S_{\mathrm{S_1}}/\mathrm{d}t = (1 - \alpha) P_{\mathrm{e}} - Q_{\mathrm{S_1}}$	$Q_{\mathrm{S_1}} = S_{\mathrm{S_1}} R_{\mathrm{S}}$	

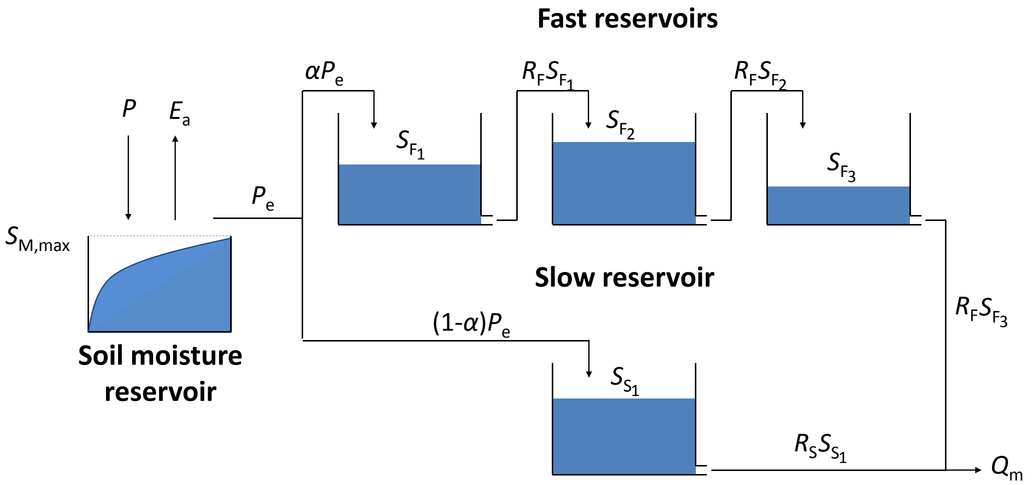

Fig. 2. Schematic illustration of HyMod rainfall/runoff conceptual model.

4. The model is calibrated using SuPer calibration using the years 2002 and 2003 as sub-periods. The performances of the obtained parameter sets ($\mathrm{MDPF}_{2002-2003}$) are then validated in each sub-period (2002 and 2003).

Note that the years 2002 and 2003 are hydrologically very different. Rainfall, runoff and potential evaporation are presented in Table 1 for the two years. Year 2002 is wet compared to 2003.

3.3.2 Case study 2 – Long data series

The available time series of the Wark catchment are divided into three parts. The years 1996–1997 are used as warm up period. The years 1998–2005 are used for parameter identification. The years 2006–2009 are retained for validation to compare the performance of parameter sets selected by the different calibration approaches. Two parameter identification approaches are compared:

1. The model is calibrated using standard calibration for the eight-year period of 1998–2005.

2. The model is calibrated using SuPer calibration considering each individual year of the period 1998–2005 as a sub-period. This requires the determination of the 8 sub-period calibration Pareto fronts CPF_{1998}, ..., CPF_{2005}. Therefore parameter set identification is based on an 8-dimensional $\mathrm{MDPF}_{1998-2005}$.

The two approaches are compared both with respect to performance and parameter distributions. The performance of the different parameter sets retained by each calibration approach is compared relative to the calibration Pareto front of each individual year (CPF_{2006}, ..., CPF_{2009}) and of the entire validation period ($\mathrm{CPF}_{2006-2009}$).

The sensitivity of model parameters by standard calibration (1998–2005) is assessed with 3 different approaches (graphically illustrated in Fig. 3):

1. Pareto optimal parameter sets ($\mathrm{CPF}_{1998-2005}$).

2. Parameter sets within a pre-defined distance to the origin. In this case study, the parameter sets with a distance smaller than 1.05 times of the closest Pareto member to the origin.

3. Parameter sets contained within the quadrant determined by the single objective optima.

The parameter distributions of both standard calibration and SuPer calibration (MDPF) are compared with the optimal

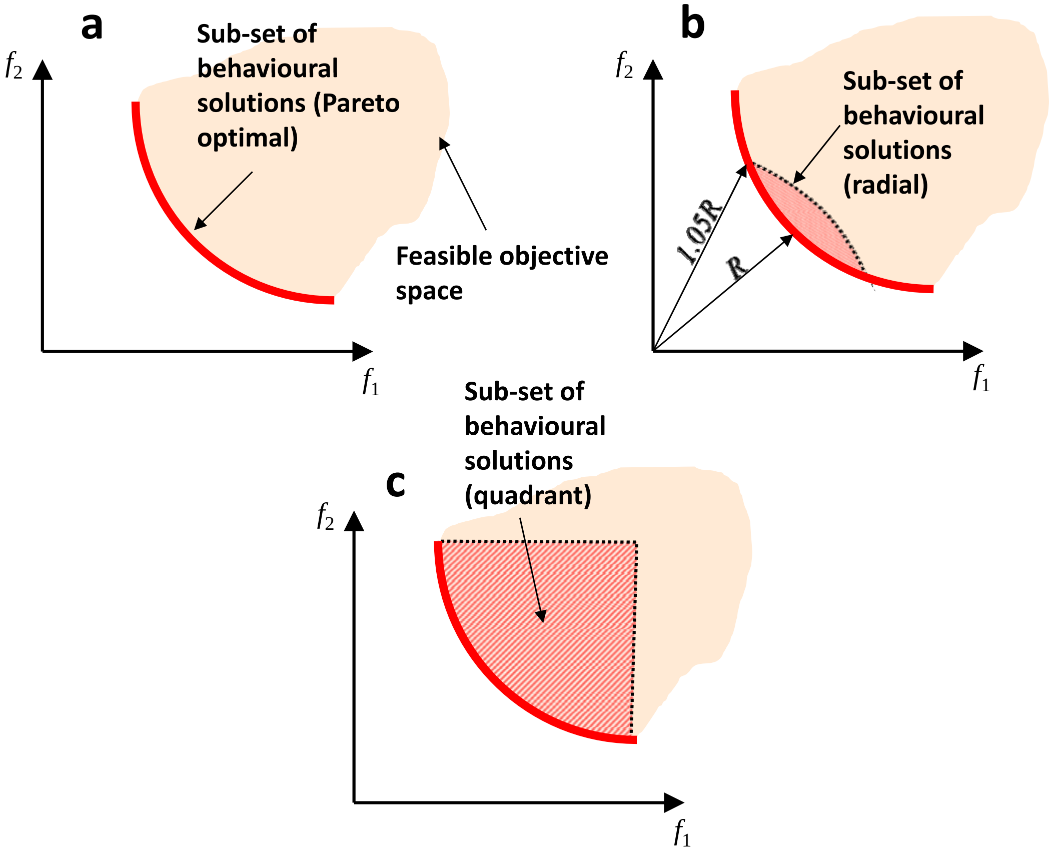
a
Sub-set of behavioural solutions (Pareto optimal)
Feasible objective space
f_2
f_1
b
Sub-set of behavioural solutions (radial)
1.05R
R
f_2
f_1
c
Sub-set of behavioural solutions (quadrant)
f_2
f_1

Fig. 3. Different approaches for the selection of behavioral parameter sets for a two-dimensional ($[f_1\ f_2]$) multi-objective problem; behavioral parameter sets are selected as **(a)** Pareto optimal parameter sets, **(b)** parameter sets which perform closer than 1.05 of minimum distance of Pareto front to origin (radial), and **(c)** parameter sets which perform simultaneously better than the lowest performance of any dimension of Pareto front (quadrant).

parameter sets of each individual year of the entire calibration and validation periods (CPFs).

In the two case studies presented, HyMod was evaluated by two objective functions. These are the root mean square error of flows (I_{RMSE}) and the root mean square error of the logarithm of flows (I_{LRMSE}), which emphasize high flow and low flow respectively:

$$I_{\mathrm{RMSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_{\mathrm{m},i} - Q_{\mathrm{o},i})^2}, \qquad (1)$$

$$I_{\mathrm{LRMSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(Q_{\mathrm{m},i}) - \log(Q_{\mathrm{o},i}))^2}, \qquad (2)$$

where $Q_{\mathrm{m},i}$ and $Q_{\mathrm{o},i}$ are the modeled and observed flow for time step i, and N is the number of time steps. I_{RMSE} was used instead of the Nash–Sutcliffe efficiency (I_{NSE}), as I_{NSE} depends on the average of the observations, which may be different in different sub-periods (Schaefli and Gupta, 2007).

The relative performance of a parameter set is presented by calculating the Euclidian distance to the calibration Pareto front (CPF) for every individual sub-period. We assume that the two objective functions in this case study are in the same order of magnitude and therefore do not need normalization.

Parameter search was performed using the MOSCEM-UA algorithm (Vrugt et al., 2003) for both calibration Pareto fronts (CPFs) and minimum distance Pareto front (MDPF). SuPer calibration selects parameter sets with the best performance relative to CPFs; therefore MOSCEM-UA was chosen as it uses Zitzler strength Pareto ranking (Zitzler and Thiele, 1999), which allows robust estimation of CPF.

4 Results

4.1 Case study 1 – Short data series

The calibration Pareto fronts, CPF_{2002}, CPF_{2003} and $\mathrm{CPF}_{2002-2003}$ are shown in Fig. 4. CPF_{2003} and $\mathrm{CPF}_{2002-2003}$ show a large tradeoff between I_{RMSE} and I_{LRMSE}. In Fig. 4, model performance in periods outside the calibration period

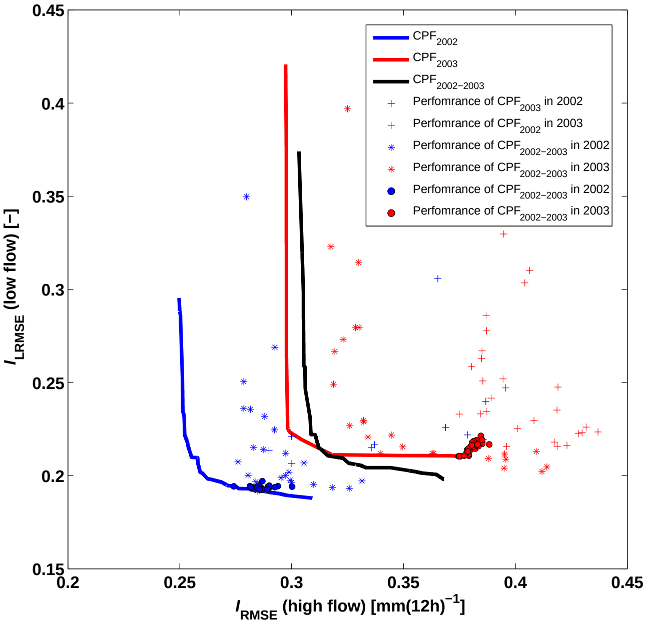

Fig. 4. The calibration Pareto fronts based on 2002, 2003 and 2002–2003 (CPF$_{2002}$, CPF$_{2003}$, CPF$_{2002-2003}$) are illustrated by blue, red and black respectively. The blue and red crosses show the performance of CPF$_{2002}$ members in 2003 and the performance of CPF$_{2003}$ members in 2002 respectively. The blue and red asterisks illustrate performance of CPF$_{2002-2003}$ in 2002 and 2003 respectively. The blue and red dots illustrate the performance of MDPF$_{2002-2003}$ in year 2002 and 2003 respectively.

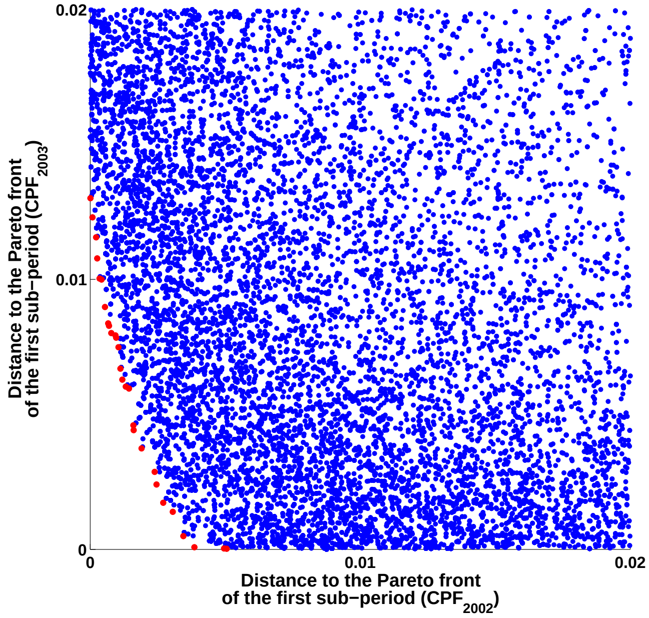

Fig. 5. The two-dimensional minimum distance Pareto front (red dots) based on year 2002–2003 (MDPF$_{2002-2003}$).

is indicated by crosses of the same color as the sub-period CPF. Figure 4 shows the performance of CPF$_{2002}$ members in 2003 and the performance of CPF$_{2003}$ in 2002. Moreover the performance of CPF$_{2002-2003}$ in 2002 and 2003 are illustrated with stars of the same color as the sub-period CPF of the same year. It can be observed that model performance in periods outside the calibration period may differ significantly from the optimal performance. Even the standard calibration based on the entire time period (2002–2003) deviates significantly from the optimal performance in each sub-period.

The parameter sets as identified with the SuPer calibration approach are shown by dots in Fig. 4 for the sub-periods 2002 and 2003 with the same color as the sub-period CPFs. As shown in Fig. 4, SuPer calibration picks parameter sets with relatively good performance in both sub-periods, excluding parameter sets that work well in one period, but very poorly in another. Moreover, Fig. 4 shows SuPer calibration emphasizes on the parameter sets with better performance regarding I_{LRMSE}, indicating low flow can be modeled more consistent over time. The relative performance of parameter sets selected with SuPer calibration to CPFs of every sub-period (2002 and 2003) are illustrated in Fig. 5 (MDPF$_{2002-2003}$).

Figure 6 illustrates the distribution of the parameters of the fast and slow reservoirs (R_F, R_S) selected by different approaches. We can see that the parameter distributions associated to CPF$_{2002}$ and CPF$_{2003}$ are very different from each other. However, the parameter distributions associated to MDPF$_{2002-2003}$ are close to the intersection of the distributions in the two individual periods. The parameter distributions identified by SuPer calibration indicate a narrower range compared to calibration over the entire time series (CPF$_{2002-2003}$).

4.2 Case study 2 – Long data series

The comparison between standard calibration and SuPer calibration using each year as an individual sub-period over the period of 1998–2005 is illustrated in Fig. 7. The parameter sets obtained by SuPer calibration are different from those identified by the different selection rules (Pareto optimal, radial and quadrant see Sect. 3.3), but similarly to the previous case study, SuPer calibration tends to select parameter sets towards the I_{LRMSE} objective function, indicating that low flow parameters are more consistent in time.

The distance to Pareto front (relative performance) of parameter sets obtained by different selection methods for behavioral parameters using standard calibration (Pareto, radial and quadrant rules) and those obtained by SuPer calibration are illustrated in Fig. 8 for the entire validation period (2006–2009) as well as for every individual year (2006, 2007, 2008, and 2009). The Pareto front members (CPF$_{1998-2005}$) perform differently and the 25/75th interquartile ranges of their performance only have limited overlap for individual validation sub-periods (2006, ..., 2009). For parameter sets retained

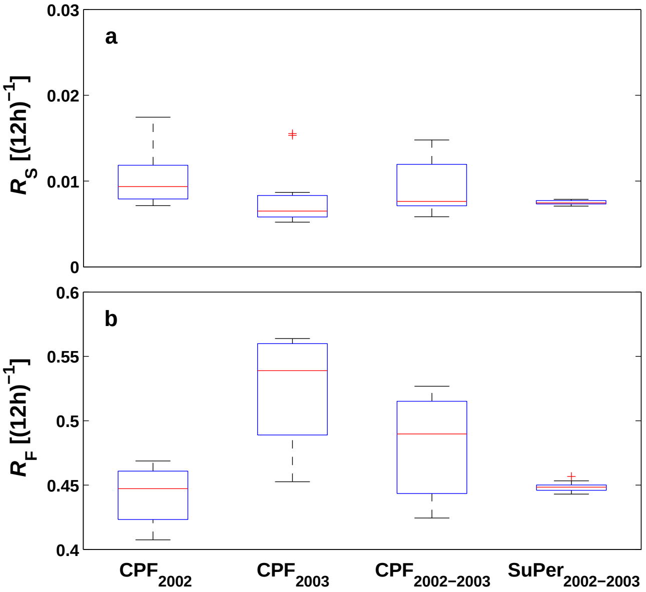

Fig. 6. Parameter distributions of Pareto members of 2002, 2003 and 2002–2003 (CPF2002, CPF2003, CPF2002–2003) and SuPer calibration (MDPF2002–2003) for **(a)** slow reservoir coefficient (R_S) and **(b)** fast reservoir coefficient (R_F). Whiskers represent the 1.5 times the interquartile range (IQR) and the red crosses show outliers.

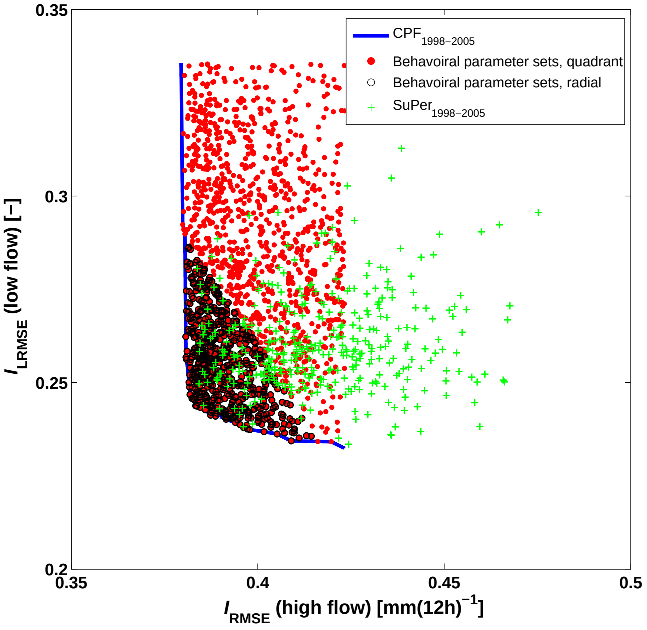

Fig. 7. CPF$_{1998-2005}$ is shown by the blue line. Black circles and red dots illustrate the parameter set retained by radial and quadrant rules (see Sect. 3.3, Fig. 3). The green crosses indicate the performance of parameter sets identified by MDPF$_{1998-2005}$ over the period of 1998–2005.

by quadrant or radial rules, similar to Pareto front members (CPF$_{1998-2005}$), the distance to Pareto fronts during the validation period varies significantly for sub-periods. However, for every individual year as well as for the entire validation periods, the 25/75th interquartile ranges of the parameter sets retained by SuPer calibration show significant overlap. Overall, the parameter sets selected by SuPer calibration tend to show more consistency over individual years (CPF$_{2006}$,..,CPF$_{2009}$), as well as over the entire validation period (CPF$_{2006-2009}$) compared to parameter sets retained by calibration over the entire period.

The distributions of two characteristic parameters for the calibration (1998–2005) and validation (2006–2009) periods (over the entire time series and every individual year) are shown for different parameter identification approaches in Fig. 9. The comparison between parameter distributions of sub-period Pareto members for the slow reservoir coefficients (R_S) shows that SuPer calibration is less affected by an anomaly of one sub-period (2001). As can be seen in Fig. 9a, the parameter distribution of standard calibration retained by the quadrant rule, emphasizes also on the values which are not optimal in sub-periods (1998, ..., 2005). Comparing in Fig. 9b, the distribution of the fast reservoir coefficient (R_F) obtained by standard calibration and retained by the quadrant rule with SuPer calibration, indicates that SuPer calibration selects parameter sets which overlap for every sub-period, while standard calibration over the entire calibration period (1998–2005) may cover values which do not have any overlap with sub-periods (1998, ..., 2005). Figure 9a, b, indicating Pareto optimal members identified by standard calibration over the entire calibration period (1998–2005), shows a narrower range compared to parameter sets retained by SuPer calibration. However the Pareto optimal distribution does not have any intersection with the parameter distribution of Pareto members of sub-period 2003 for the slow reservoir coefficient (R_S), meaning that they cannot perform optimally in that specific sub-period, while the distribution of the parameter set selected by SuPer calibration covers the distribution range of every sub-period. As was also illustrated in Fig. 8, the performance of Pareto optimal members, although confined to narrow ranges, may not perform optimally in every sub-period.

5 Discussion

SuPer calibration focuses on different parts of sub-period calibration Pareto fronts (CPFs), and helps to identify parameter sets with a time consistent behavior. These parameter sets may therefore be regarded as more "realistic" (Figs. 6 and 9). We attribute this to the fact that the processes identified by some objective functions (in the present case low flows) may have a more time consistent behavior than the processes represented by other objective functions (e.g. high flows).

SuPer calibration identifies parameter sets which perform optimally in sub-periods. The corresponding parameter

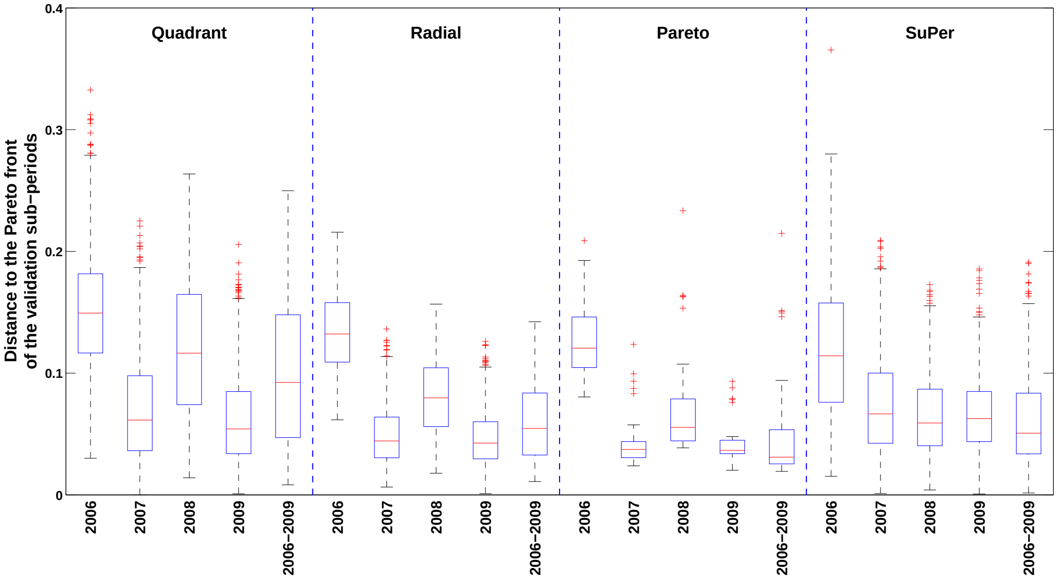

Fig. 8. Distribution of Euclidian distance or relative performance of behavioral parameter sets obtained by calibration, and parameter sets retained by $\mathrm{MDPF}_{1998-2005}$ to calibration Pareto fronts of the individual year and the entire validation period (CPF_{2006}, CPF_{2007}, CPF_{2008}, CPF_{2009}, $\mathrm{CPF}_{2006-2009}$). Whiskers represent the 1.5 times the interquartile range (IQR) and the red crosses show outliers.

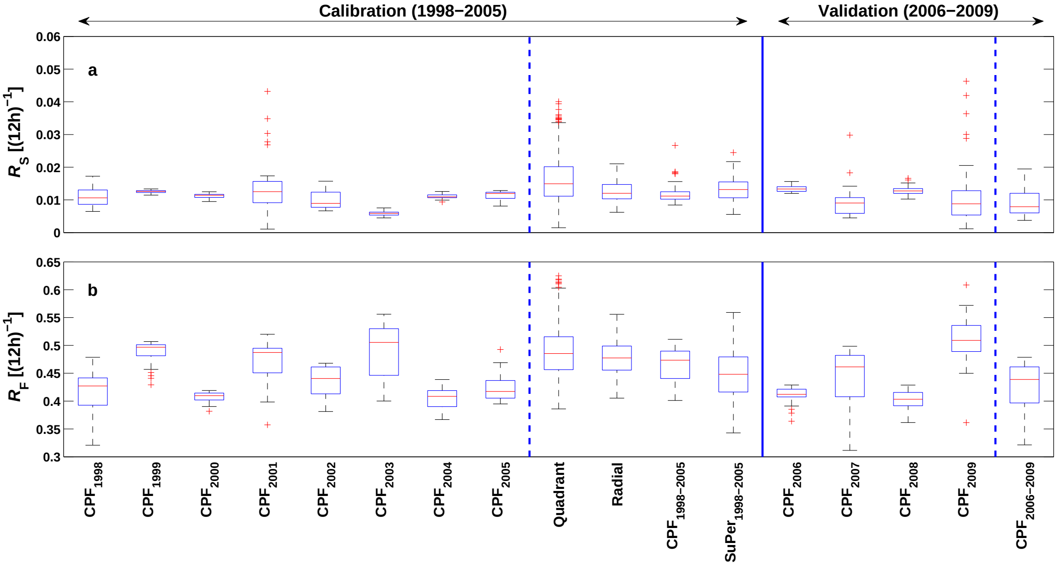

Fig. 9. Parameter distributions of fast reservoir coefficient (R_F) and slow reservoir coefficient (R_S) for the calibration Pareto front of every individual year and the entire calibration and validation periods. Whiskers represent the 1.5 times the interquartile range (IQR) and the red crosses show outliers.

ranges, although maybe not optimal over the entire time series, are the narrowest ranges considering optimal behavior in every sub-period (Figs. 8 and 9); therefore making it possible to obtain parameter distributions that are just dependent on data quality, sub-period characteristics and the selected hydrological model. Moreover, unlike common selection methods of behavioral parameter sets, which as highlighted by Efstratiadis and Koutsoyiannis (2010) require the specification of a subjective threshold for identifying behavioral parameter sets, SuPer calibration does not require this. The difference between parameter sets selected by calibration and SuPer calibration is illustrated graphically in Fig. 10. Our results have indicated that parameter sets selected with this approach may be grouped towards one or more objective function at the expense of others (Figs. 4 and 7).

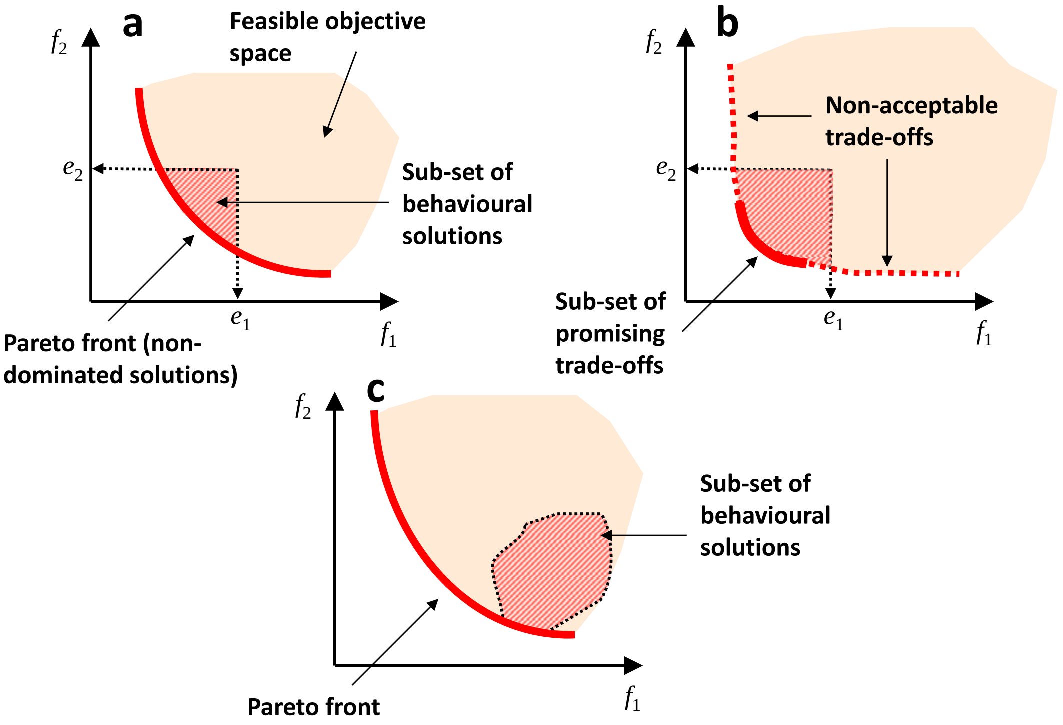
a
Feasible objective space
Sub-set of behavioural solutions
Pareto front (non-dominated solutions)
f_2
e_2
e_1
f_1
b
Non-acceptable trade-offs
Sub-set of promising trade-offs
f_2
e_2
e_1
f_1
c
Sub-set of behavioural solutions
Pareto front
f_2
f_1

Fig. 10. Graphical examples illustrate Pareto optimal and behavioral solutions in the objective space for a two-dimensional ([f_1 f_2]) multi-objective problem, with e=[e_1 e_2] indicating limits of acceptability, i.e. cut-off thresholds for distinguishing behavioral and non-behavioral solutions for **(a)** smooth and **(b)** steep trade-off Pareto fronts. **(c)** The position of parameter sets identified by SuPer calibration (MDPF) in the objective space (After: Efstratiadis and Koutsoyiannis, 2010, with permission of the first author and the publisher).

One might argue that SuPer calibration can be achieved by applying multi-objective calibration to different objective functions of different sub-periods in a single step. As an example, the first case study can be presented by introducing the 2 objective functions (I_{RMSE} and I_{LRMSE}) for different sub-periods (2002 and 2003), therefore parameter identification will be formulated as a four-dimensional optimization practice with $I_{\mathrm{RMSE}_{2002}}$, $I_{\mathrm{RMSE}_{2003}}$, $I_{\mathrm{LRMSE}_{2002}}$ and $I_{\mathrm{LRMSE}_{2003}}$ as objective functions. This approach would determine the trade-off of the model performance in different sub-periods. However, parameter identification is still based on the selection of Pareto front members and therefore the challenge of selecting behavioral parameter sets, or in this case time consistent parameter sets, from the Pareto front members remains the same as mentioned by Efstratiadis and Koutsoyiannis (2010) (Fig. 10).

SuPer calibration can also be used as a tool to analyze parameter time consistency in different sub-periods. By identifying non-time consistent parameters, SuPer calibration can be used as a diagnostic tool for identifying model structural deficiencies (see Clark et al., 2008). This approach can also provide information about the behavior of each parameter with respect to the hydrological condition of that period. As an example, the fast reservoir coefficient (R_{F}) shows higher values for the sub-period 2003 than for 2002. The years 2002 and 2003 are hydrologically distinct years (Table 1). This analysis, similar to the DYNIA (Wagener et al., 2003), can help the modeler to identify a model deficiency and guide towards model improvements.

Although in this work we used hydrological years as the basis for sub-period analysis, periods can be selected in different ways. For example, they can be applied to storm events with different magnitude and return period to retain their characteristics during the calibration process. Sub-periods can also be defined as different parts of the flow duration curve (Westerberg et al., 2011) or can be used for calibration based on unusual events (Singh and Bárdossy, 2012; Krauße and Cullmann, 2012). Building on previous studies (e.g. Wagener, 2003; Seiller et al., 2012), we support the conclusion that looking individually at different periods is an approach to extract more information from the data, rather than considering the data series as a whole.

Sampling strategies for the parameter space were not discussed and in principle different approaches can be used. As the method requires the identification of Pareto fronts, methods that sample the vicinity of the optimal parameter sets are preferable. The uncertainty in Pareto front identification may introduce uncertainty in the final selected parameter set selected by SuPer calibration. In this study MOSCEM-UA (Vrugt et al., 2003) was used to generate Pareto fronts in both steps of the procedure (creating CPFs and MDPFs).

Limitations of the presented SuPer calibration approach include, at least in its current implementation, that it cannot be applied to represent meaningful uncertainty estimates; the potential application of this approach in a Bayesian framework remains to be investigated.

6 Conclusions

In this paper a calibration approach based on splitting the available data sets into sub-periods has been proposed. The sub-period calibration approach makes use of calibration in individual sub-periods, and extracts parameter sets with a time consistent performance. Although this comes at the cost of potentially reduced performance during the calibration of each individual period, model parameterizations obtained by SuPer calibration perform consistently better in the validation period, which is what modelers actually should look for. The design of SuPer calibration is such that acceptable parameterizations have to perform consistently well when predicting any of the defined sub-periods, which is implicitly enforced in SuPer calibration, thus avoiding the need for explicit model validation. Furthermore, by the transformation of the traditional objective-space into a minimum Euclidean distance space, the need for subjective choices of parameter acceptance thresholds is avoided.

It should be again emphasized here that SuPer calibration is not a calibration algorithm, nor is it explicitly addressing parameter uncertainty. It is rather a more advanced method of model testing, building on traditional split sample tests and making more efficient use of available data. SuPer calibration can in principle be done with any number and type of objective functions (e.g. I_{NSE} or I_{RMSE}) but also with any number and type of calibration criteria (e.g. only using runoff or using runoff and tracer dynamics). A Matlab function of the SuPer calibration approach can be obtained by personal communication with the lead author.

Acknowledgements. The authors would like to thank the people who facilitated this research at Gabriel Lippmann Research Institute especially Laurent Pfister and Lucien Hoffmann. The author would like to thank Vazken Andréassian, Abdolreza Bahremand, Keith Beven, Andreas Efstratiadis, Uwe Ehret, Jim Freer, Hoshin Gupta, Jean-François Iffly, Demetris Koutsoyiannis, Rohini Kumar, Patrick Matgen, Thibault Mathevet, Saket Pande, Elena Ridolfi, Gerrit Schoups, Dimitri Solomatine, Jasper Vrugt, Thorsten Wagener, Ling Wang, Erwin Zehe, and two anonymous reviewers for their constructive comments on this work. Shervan Gharari is funded during his PhD program by Fond National de la Recherche (AFR) of Luxembourg with project reference number of 1383201.

Edited by: N. Verhoest

References

Anderson, M. P. and Woessner, W. W.: The role of the postaudit in model validation, Adv. Water Resour., 15, 167–173, 1992.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H., and Valéry, A.: HESS Opinions "Crash tests for a standardized evaluation of hydrological models", Hydrol. Earth Syst. Sci., 13, 1757–1764, doi:10.5194/hess-13-1757-2009, 2009.

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, Hydrol. Process., 26, 2206–2210, doi:10.1002/hyp.9264, 2012.

Bárdossy, A. and Singh, S. K.: Robust estimation of hydrological model parameters, Hydrol. Earth Syst. Sci., 12, 1273–1283, doi:10.5194/hess-12-1273-2008, 2008.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320, 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K. J. and Binley, A. M.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6, 279–298, 1992.

Birkel, C., Dunn, S. M., Tetzlaff, D., and Soulsby, C.: Assessing the value of high-resolution isotope tracer data in the stepwise development of a lumped conceptual rainfall-runoff model, Hydrol. Process., 24, 2335–2348, doi:10.1002/hyp.7763, 2010.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resour. Res., 36, 3663–3674, doi:10.1029/2000WR900207, 2000.

Boyle, D. P., Gupta, H. V., Sorooshian, S., Koren, V., Zhang, Z., and Smith, M.: Toward improved streamflow forecasts: value of semidistributed modeling, Water Resour. Res., 37, 2749–2759, doi:10.1029/2000WR000207, 2001.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007WR006735, 2008.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resour. Res., 48, W05552, doi:10.1029/2011WR011721, 2012.

Dunn, S. M., Bacon, J. R., Soulsby, C., Tetzlaff, D., Stutter, M. I., Waldron, S., and Malcolm, I. A.: Interpretation of homogeneity in 18O signatures of stream water in a nested sub-catchment system in north-east Scotland, Hydrol. Process., 22, 4767–4782, doi:10.1002/hyp.7088, 2008.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrol. Sci. J., 55, 58–78, doi:10.1080/02626660903526292, 2010.

Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: A comparison of alternative multiobjective calibration strategies for hydrological modeling, Water Resour. Res., 43, W03434, doi:10.1029/2006WR005098, 2007.

Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: Understanding catchment behavior through stepwise model concept improvement, Water Resour. Res., 44, W01402, doi:10.1029/2006WR005563, 2008.

Fenicia, F., Wrede, S., Kavetski, D., Pfister, L., Hoffmann, L., Savenije, H. H. G., and McDonnell, J. J.: Assessing the impact of mixing assumptions on the estimation of streamwater mean residence time, Hydrol. Process., 24, 1730–1741, doi:10.1002/hyp.7595, 2010.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resour. Res., 47, W11510, doi:10.1029/2010WR010174, 2011.

Freer, J., Beven, K., and Peters, N.: Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure, Water Sci. Appl., 6, 69–87, doi:10.1029/WS006p0069, available at: http://www.agu.org/books/ws/v006/WS006p0069/WS006p0069.shtml, 2003.

Freer, J., McMillan, H., McDonnell, J., and Beven, K.: Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures, J. Hydrol., 291, 254–277, doi:10.1016/j.jhydrol.2003.12.037, 2004.

Gharari, S., Hrachowitz, M., Fenicia, F., and Savenije, H. H. G.: Hydrological landscape classification: investigating the performance of HAND based landscape classifications in a central European meso-scale catchment, Hydrol. Earth Syst. Sci., 15, 3275–3291, doi:10.5194/hess-15-3275-2011, 2011.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34, 751–763, doi:10.1029/97WR03495, 1998.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802–3813, doi:10.1002/hyp.6989, 2008.

Hamon, W. R.: Estimating potential evapotranspiration, Journal Hydraulic Division, 87, 107–120, 1961.

Hartmann, G. and Bárdossy, A.: Investigation of the transferability of hydrological models and a method to improve model calibration, Adv. Geosci., 5, 83–87, doi:10.5194/adgeo-5-83-2005, 2005.

Hassan, A. E.: Validation of numerical ground water models used to guide decision making, Ground Water, 42, 277–290, doi:10.1111/j.1745-6584.2004.tb02674.x, 2004.

Hrachowitz, M., Savenije, H., Bogaard, T. A., Tetzlaff, D., and Soulsby, C.: What can flux tracking teach us about water age distributions and their temporal dynamics?, Hydrol. Earth Syst. Sci. Discuss., 9, 11363–11435, doi:10.5194/hessd-9-11363-2012, 2012.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42, W03S04, doi:10.1029/2005WR004362, 2006.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrolog. Sci. J., 31, 13–24, doi:10.1080/02626668609491024, 1986.

Krauße, T. and Cullmann, J.: Towards a more representative parametrisation of hydrologic models via synthesizing the strengths of Particle Swarm Optimisation and Robust Parameter Estimation, Hydrol. Earth Syst. Sci., 16, 603–629, doi:10.5194/hess-16-603-2012, 2012.

Li, C. Z., Zhang, L., Wang, H., Zhang, Y. Q., Yu, F. L., and Yan, D. H.: The transferability of hydrological models under nonstationary climatic conditions, Hydrol. Earth Syst. Sci., 16, 1239–1254, doi:10.5194/hess-16-1239-2012, 2012.

Mroczkowski, M., Raper, P. G., and Kuczera, G.: The quest for more powerful validation of conceptual catchment models, Water Resour. Res., 33, 2325–2335, doi:10.1029/97WR01922, 1997.

Refsgaard, J. C. and Hansen, J. R.: A good-looking catchment can turn into a modeller's nightmare, Hydrolog. Sci. J., 55, 899–912, doi:10.1080/02626667.2010.505571, 2010.

Refsgaard, J. C., Henriksen, H. J., Harrar, W. G., Scholten, H., and Kassahun, A.: Quality assurance in model based water management – review of existing practice and outline of new approaches, Environ. Modell. Softw., 20, 1201–1215, doi:10.1016/j.envsoft.2004.07.006, 2005.

Savenije, H. H. G.: HESS Opinions "The art of hydrology"*, Hydrol. Earth Syst. Sci., 13, 157–161, doi:10.5194/hess-13-157-2009, 2009.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrol. Process., 21, 2075–2080, doi:10.1002/hyp.6825, 2007.

Seibert, J.: Reliability of model predictions outside calibration conditions, Nord. Hydrol., 34, 477–492, 2003.

Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrol. Earth Syst. Sci., 16, 1171–1189, doi:10.5194/hess-16-1171-2012, 2012.

Singh, S. K. and Bárdossy, A.: Calibration of hydrological models on hydrologically unusual events, Adv. Water Resour., 38, 81–91, doi:10.1016/j.advwatres.2011.12.006, 2012.

Son, K. and Sivapalan, M.: Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, Water Resour. Res., 43, W01415, doi:10.1029/2006WR005032, 2007.

Tukey, J. W.: Mathematics and the Picturing of Data, 2, pp. 523–531, in Proceedings of the 1975 international 17 congress of mathematics, 1975.

Uhlenbrook, S. and Sieber, A.: On the value of experimental data to reduce the prediction uncertainty of a process-oriented catchment model, Environ. Modell. Softw., 20, 19–32, doi:10.1016/j.envsoft.2003.12.006, 2005.

Vaché, K. and McDonnell, J.: A process-based rejectionist framework for evaluating catchment runoff model structure, Water Resour. Res., 42, W02409, doi:10.1029/2005WR004247, 2006.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S.: Effective and efficient algorithm for multiobjective optimization of hydrologic models, Water Resour. Res., 39, 1214, doi:10.1029/2002WR001746, 2003.

Wagener, T.: Evaluation of catchment models, Hydrol. Process., 17, 3375–3378, doi:10.1002/hyp.5158, 2003.

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, Hydrol. Earth Syst. Sci., 5, 13–26, doi:10.5194/hess-5-13-2001, 2001.

Wagener, T., Lees, M. J., and Wheater, H. S.: Mathematical Models of Large Watershed Hydrology, chap. A framework for the development and application of parsimonios hydrological models, Water Resources Publications, 91–140, 2002.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, Hydrol. Process., 17, 455–476, doi:10.1002/hyp.1135, 2003.

Weiler, M., McGlynn, B., McGuire, K., and McDonnell, J.: How does rainfall become runoff? A combined tracer and runoff transfer function approach, Water Resour. Res., 39, 1315, doi:10.1029/2003WR002331, 2003.

Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C.-Y.: Calibration of hydrological models using flow-duration curves, Hydrol. Earth Syst. Sci., 15, 2205–2227, doi:10.5194/hess-15-2205-2011, 2011.

Wheater, H. S., Jakeman, A. J., and Beven, K. J.: Progress and directions in rainfall-runoff modeling, John Wiley & Sons, 1993.

Winsemius, H. C., Savenije, H. H. G., and Bastiaanssen, W. G. M.: Constraining model parameters on remotely sensed evaporation: justification for distribution in ungauged basins?, Hydrol. Earth Syst. Sci., 12, 1403–1413, doi:10.5194/hess-12-1403-2008, 2008.

Wood, E. F. and Rodríguez-Iturbe, I.: A Bayesian approach to analyzing uncertainty among flood frequency models, Water Resour. Res., 11, 839–843, doi:10.1029/WR011i006p00839, 1975.

Zitzler, E. and Thiele, L.: Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach, IEEE T. Evolut. Comput., 3, 257–271, doi:10.1109/4235.797969, 1999.