Journal of Hydrology 377 (2009) 191–207

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland – Part I: Modelling framework and calibration results ☆

Daniel Viviroli [a,b,*], Massimiliano Zappa [c], Jan Schwanbeck [a,b], Joachim Gurtz [d], Rolf Weingartner [a,b]

[a] Institute of Geography, University of Bern, Hallerstrasse 12, CH-3012 Bern, Switzerland
[b] Oeschger Centre for Climate Change Research, University of Bern, Zähringerstrasse 25, CH-3012 Bern, Switzerland
[c] Swiss Federal Research Institute WSL, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland
[d] Institute for Atmospheric and Climate Science, ETH Zürich, Universitätsstrasse 16, CH-8092 Zürich, Switzerland

ARTICLE INFO

Article history:
Received 10 April 2009
Received in revised form 8 July 2009
Accepted 17 August 2009

This manuscript was handled by K. Georgakakos, Editor-in-Chief, with the assistance of Emmanouil N. Anagnostou, Associate Editor.

Keywords:
Model calibration
Flood estimation
Precipitation–runoff model
Efficiency scores
Switzerland

SUMMARY

With the aim of calibrating a large number of catchments for a semi-distributed, process-based conceptual hydrological model, we introduce a straightforward yet robust automatic calibration procedure. Since identification of a global parameter optimum is not feasible in practical terms, the procedure presents a trade-off between computational time and algorithm complexity to identify, with reasonable effort, a parameter set that is well representative of the catchment's dynamics. In its standard mode, the calibration combines three efficiency scores which are evaluated both for the entire calibration period and in terms of their annual and monthly variations. These scores are furthermore assessed only in their relevant value range, producing a comprehensive overall acceptability score which is used to guide an iterative parameter search algorithm. An additional set of four flood-sensitive scores is added, thereby improving results in the peak-flow range.

Calibration was done for 140 mesoscale (roughly 10–1000 km^2) catchments in Switzerland, using the hydrological modelling system PREVAH (Precipitation–Runoff-EVApotranspiration-HRU related model) in hourly time steps. For 49 representative catchments with long gauge records, a median Nash–Sutcliffe efficiency (*NSE*) of 0.75 was achieved for the calibration period in standard mode. The limited loss in efficiency when moving to the validation period (median *NSE*: 0.72) proves the stability and representativity of the parameter sets identified, while a Monte-Carlo analysis underscores the effectiveness of our procedure. The Nash–Sutcliffe efficiencies for the additional flood calibration are slightly lower, but again almost equally high for the calibration (0.69) and the validation (0.67) period. Despite the concessions made to improve peak-flow results, the simulation's hydrological plausibility was not compromised.

The ultimate goal of our study is flood estimation in ungauged Swiss catchments through continuous simulation using PREVAH. With the extensive calibration task presented in this article, the foundation is laid for regionalisation of the tuneable model parameters, which will be addressed in the companion paper (Viviroli et al., 2009a), along with detailed flood estimation results.

Introduction

Calibration is a key prerequisite for hydrological modelling; it is necessary for model parameters which either do not have a direct physical interpretation or cannot be measured at the appropriate scale. Adjustment of these parameters is typically regarded as successful when a measurable system output, e.g. runoff, and the model output show an acceptable level of agreement. A major difficulty is the fact that the data available for calibration are usually limited to observed time series of runoff (Jakeman and Hornberger, 1993; Kuczera and Mroczkowski, 1998). From a theoretical point of view, this information is rather sparse for identification of all model parameters (Wagener et al., 2003). Therefore, fitting a model on the basis of a single efficiency score may lead to unsatisfactory calibration results (Eckhardt, 2002). Additional measurements – e.g. snow cover, soil moisture or groundwater stages – would increase parameter identifiability (Kirchner, 2006), but they are usually not available for exhaustive applications such as the present one. The parameter set achieving the best performance is commonly

☆ This is the companion paper of "Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland – Part II: Parameter regionalisation and flood estimation results" by Viviroli, Mittelbach, Gurtz and Weingartner (2009a).

* Corresponding author. Address: Institute of Geography, University of Bern, Hallerstrasse 12, CH-3012 Bern, Switzerland. Tel.: +41 31 631 80 17; fax: +41 31 631 85 11.
E-mail address: viviroli@giub.unibe.ch (D. Viviroli).

0022-1694/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2009.08.023

assumed to be representative of the natural system under investigation. Perfect agreement of simulation and observation is not feasible due to various error and uncertainty sources involved e.g. in model structure or field measurements (see e.g. Beven, 1993; Vrugt et al., 2005); especially the influence of parameter uncertainty and equifinality has been discussed in dozens of papers since the mid-1990s and still finds a large response in the hydrological literature (see e.g. Beven, 2006a; Mantovan and Todini, 2006; Pappenberger and Beven, 2006).

Several calibration philosophies are reflected in the scientific literature. Following the advances in mathematical and hydrological knowledge, the formerly popular manual calibration has been gradually replaced by a series of more or less sophisticated search algorithms and objective functions (for a comprehensive overview see Duan et al., 2003). But along with these high-level discussions of model calibration issues, solutions are also required for practical hydrological questions. In the present article and its companion paper (Viviroli et al., 2009a), we focus on flood estimation in ungauged mesoscale basins, a subject which has high relevance in view of the large damage frequently caused by floods, particularly in mountain areas (see e.g. Weingartner et al., 2003; Kron, 2006). For the ultimate goal of a process-based flood estimation system based on continuous simulation (Viviroli et al., 2009a; see also Viviroli, 2007), an extensive calibration effort (140 basins using a process-based model at hourly resolution) is necessary which in turn forms the basis for regionalisation of the model parameters for ungauged basins.

Calibration of such a large number of catchments rules out any manual procedure as it would simply be too time-consuming and too subjective (Botterweg, 1995; Seibert, 1997; Madsen et al., 2002). On the other hand, comprehensive objective calibration schemes based on Monte-Carlo search procedures are too costly in terms of CPU time for extensive applications of process-based models. Therefore, we propose a cost-efficient and yet robust automatic calibration procedure which is uniformly suitable for treating a large sample of catchments and requires no user intervention. The procedure combines multiple measures of goodness-of-fit which are assessed over different time segments, and a fuzzy approach which considers the most sensitive range of the respective measures. Additionally, extending the work of Lamb (1999), a suite of flood-sensitive efficiency scores serves to adjust the model's performance under high flow conditions without compromising overall hydrological plausibility.

In this first paper, the modelling framework (“Model description” and “Test catchments and data”) and the calibration strategy (“Standard calibration” and “Flood calibration”) are presented. In “Results” and “Discussion”, we show that despite its pragmatic architecture, the calibration procedure adopted here very effectively provides representative parameter sets which yield good validation results. The subsequent companion paper (Viviroli et al., 2009a) will then deal with regionalisation of the large set of tuned model parameters thus obtained and will discuss the flood estimation results.

Model description

We use the distributed hydrological model PREVAH (Precipitation–Runoff-EVApotranspiration-HRU related model; for definition of HRU see below) (Viviroli et al., 2009b), which has been developed with the intent of improving the understanding of the spatial and temporal variability of hydrological processes in catchments with complex topography (Gurtz et al., 1999; Gurtz et al., 2003). The suitability of PREVAH for such challenging environments has been proven since the mid-1990s in a wide range of applications (for an overview see Viviroli et al., 2007 and Viviroli et al., 2009b).

The spatial discretisation of PREVAH relies on the aggregation of gridded spatial information into hydrological response units (HRUs) (Ross et al., 1979; Gurtz et al., 1999). HRUs are clusters representing areas of a basin where similar hydrological behaviour is expected. In mountainous environments, it is most advisable to assign to an HRU all the grid cells located in the same meteorological sub-unit (e.g. the same range of elevation) and showing similar aspect, land-use type and soil properties (Gurtz et al., 1999). The HRU size is smaller where the ensemble of soil, land surface and topographic characteristics shows higher spatial variability. For the present project, the HRUs were generated from 0.5×0.5 km^2 raster cells.

The model is forced by interpolated values of observed climatic variables; six meteorological input variables at time steps of 1 h were used: precipitation (mm h^{-1}), air temperature (°C), global radiation (W m^{-2}), relative sunshine duration (–), wind speed (m s^{-1}) and relative humidity (%). For spatial and temporal interpolation, procedures based on Detrended Inverse Distance Weighting (e.g. Garen and Marks, 2001) and Ordinary Kriging (e.g. Isaaks and Srivastava, 1989) were adopted.

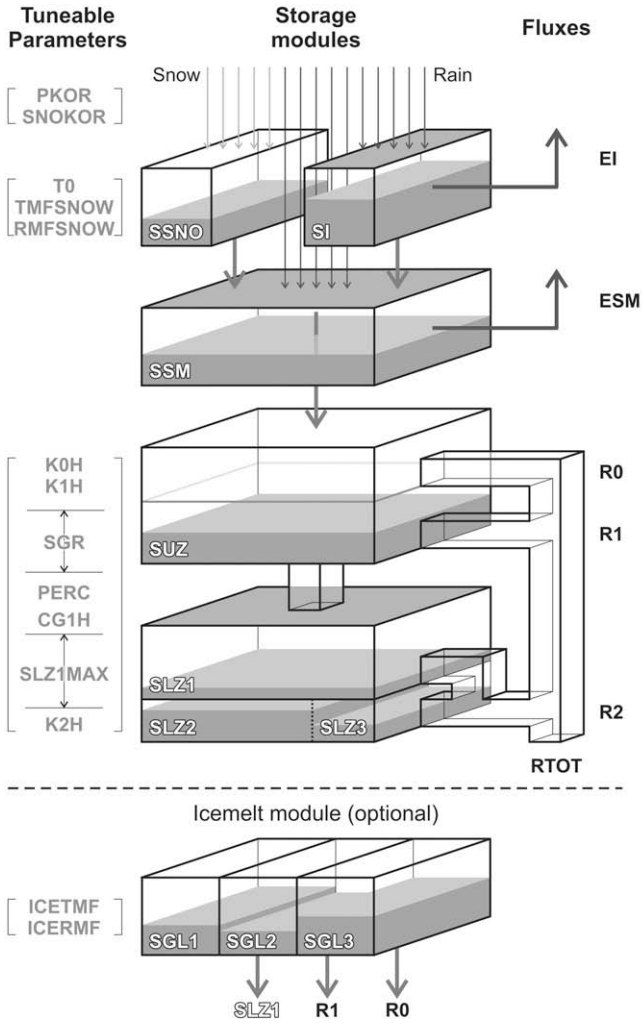
The pre-processing and basic parameterisation of PREVAH includes the topographic analysis of the investigated catchments based on a Digital Elevation Model (DEM), on digital representations of land cover characteristics and on digital maps of soil types. Each HRU has to be provided with a set of parameters based on information derived from the DEM (elevation, aspect and slope) and the soil maps (plant-available soil field capacity, soil depth, hydraulic conductivity). The land-use information allows additional values required for determining evapotranspiration to be parameterised (albedo, root depth, interception storage capacity, vegetation height, leaf area index and minimum stomatal resistance of the various vegetation classes). Non-vegetated surfaces (snowpack, glaciers, rock, large water bodies and urban areas) have to be parameterised separately (Gurtz et al., 1999). Both meteorological and geophysical pre-processing are handled using a suite of comprehensive tools (Viviroli et al., 2007,2009b).

The most sensitive catchment-specific tuneable parameters to be calibrated are then (see Fig. 1): water balance adjustment factors for rainfall (PKOR) and snowfall (SNOKOR); parameters for snowmelt, i.e. threshold temperature (T0), temperature melt factor (TMFSNOW) and radiation melt factor (RMFSNOW); parameters for ice melt, i.e. temperature melt factor (ICETMF) and radiation melt factor (ICERMF); storage times governing the process of runoff generation, i.e. surface runoff (K0H), interflow (K1H), quick baseflow (CG1H) and slow baseflow (K2H); threshold storage parameter for the generation of surface runoff (SGR); maximum storage available for fast baseflow (SLZ1MAX); percolation rate (PERC). In the model version used here, the non-linearity exponent for soil moisture recharge (BETA) is parameterised as a function of soil depth and altitude (Viviroli, 2007).

Details of the model physics, structure and parameterisation are reported in the comprehensive description by Viviroli et al. (2007), further information is found in the publications by Gurtz et al. (1999, 2003), Zappa and Gurtz (2003), Zappa et al. (2003) and Viviroli et al. (2009b).

Test catchments and data

The study area encompasses the Swiss Northern Alpine area, which is a challenging environment for modelling tasks due to its large heterogeneity of geophysical conditions (see Frei and Schär, 1998; Gurtz et al., 2003; Weingartner et al., 2003). Our focus is on mesoscale catchments with an area of roughly 10–1000 km^2,



TUNEABLE PARAMETERS

Precipitation

PKOR	Precipitation adjustment for rain [%]
SNOKOR	Precipitation adjustment for snow [%]

Snowmelt

T0	Threshold temperature snowmelt [°C]
TMFSNOW	Temperature melt factor for snow [mm d^{-1} K^{-1}]
RMFSNOW	Radiation melt factor snow [mm h^{-1} K^{-1} W^{-1} m^2]

Runoff formation

K0H	Storage time for surface runoff [h]
K1H	Storage time for interflow [h]
SGR	Threshold for quick runoff formation [mm]
PERC	Percolation rate [mm h^{-1}]
CG1H	Storage time for quick baseflow [h]
SLZ1MAX	Maximal content of the quick baseflow storage [mm]
K2H	Storage time for slow baseflow [h]

Icemelt (optional)

ICETMF	Temperature melt factor for ice [mm d^{-1} K^{-1}]
ICERMF	Radiation melt factor for ice [mm h^{-1} K^{-1} W^{-1} m^2]

STORAGE MODULES

SSNO	Snow storage [mm]
SI	Interception storage [mm]
SSM	Soil moisture storage [mm]
SUZ	Upper zone (unsaturated) runoff storage [mm]
SLZ1–3	Lower zone (saturated) runoff storages [mm]
SGL1–3	Glacier module: firn, snow and ice (optional) [mm]

FLUXES

EI	Evaporation from interception storage [mm h^{-1}]
ESM	Evapotranspiration from soil moisture storage [mm h^{-1}]
R0	Quick runoff (surface runoff) [mm h^{-1}]
R1	Delayed runoff (interflow) [mm h^{-1}]
R2	Slow runoff (baseflow) [mm h^{-1}]
RTOT	Total runoff [mm h^{-1}]

Fig. 1. Schematic of the PREVAH model structure with most sensitive tuneable parameters, storage modules and hydrological fluxes.

while only catchments without major influence of lake regulation and hydropower are examined (Fig. 2).

All catchments studied have at least 5 years of continuous gauged discharge data in hourly resolution within the 1984–2003 period (FOEN, 2008); for 49 representative catchments, complete time series are available for these years (also indicated in Fig. 2). The need for uninfluenced records restricts the availability of study catchments in the eastern Alpine part of Switzerland, which has a high hydropower density (Margot et al., 1992). Model forcing is derived from station data provided by MeteoSwiss (2008) in hourly to daily resolution (see Viviroli et al., 2007, 2009b). Geophysical information required for model parameterisation, such as soil properties and land use, is available for all of Switzerland (SFSO, 2003).

Standard calibration

Manual calibration of tuneable model parameters is often surprisingly effective (Boyle et al., 2000), particularly if the nonlinear nature of models and the complex nature of parameter interactions are borne in mind (Gupta et al., 2005). However, manual calibration has a number of relevant disadvantages: It requires a large amount of specific expert knowledge, it may be very time-consuming, and it has noticeable subjective components (Botterweg, 1995; Seibert, 1997; Madsen et al., 2002). For the present study, the large number of catchments (n = 140) already rules out any manual procedure as it would simply be too time-consuming. Furthermore, in view of the regionalisation of the tuneable model parameters envisaged by Viviroli et al. (2009a), the parameter sets should be as representative as possible for the physical and meteorological characteristics of the individual catchments (see e.g. Bronstert et al., 2003). This is more likely achieved by applying a uniform procedure with maximum objectivity. It should be noted that automatic procedures, too, contain subjective components (e. g. parameter range allowed), but since such user-defined boundary conditions can be kept constant for the calibration of multiple basins, the degree of subjectivity is certainly less pronounced. Obviously, the results from an automatic calibration also have to be verified by an expert.

For the aforementioned purposes, a monitored automatic procedure is proposed below. While being pragmatic and cost-effective, it yields robust and representative results and is therefore suitable for calibrating a large number of catchments and achieving reproducible parameter values. While we admit that the factor of computational power has become less limiting in the past years, e.g. through cluster programming, a major benefit of our method is that an extensive number of basins can be calibrated simultaneously by a single user within a reasonable timeframe. This is of particular relevance if regionalisation is to include information from as many calibrated basins as possible. The extensive calibrated and subsequently regionalised data may thus

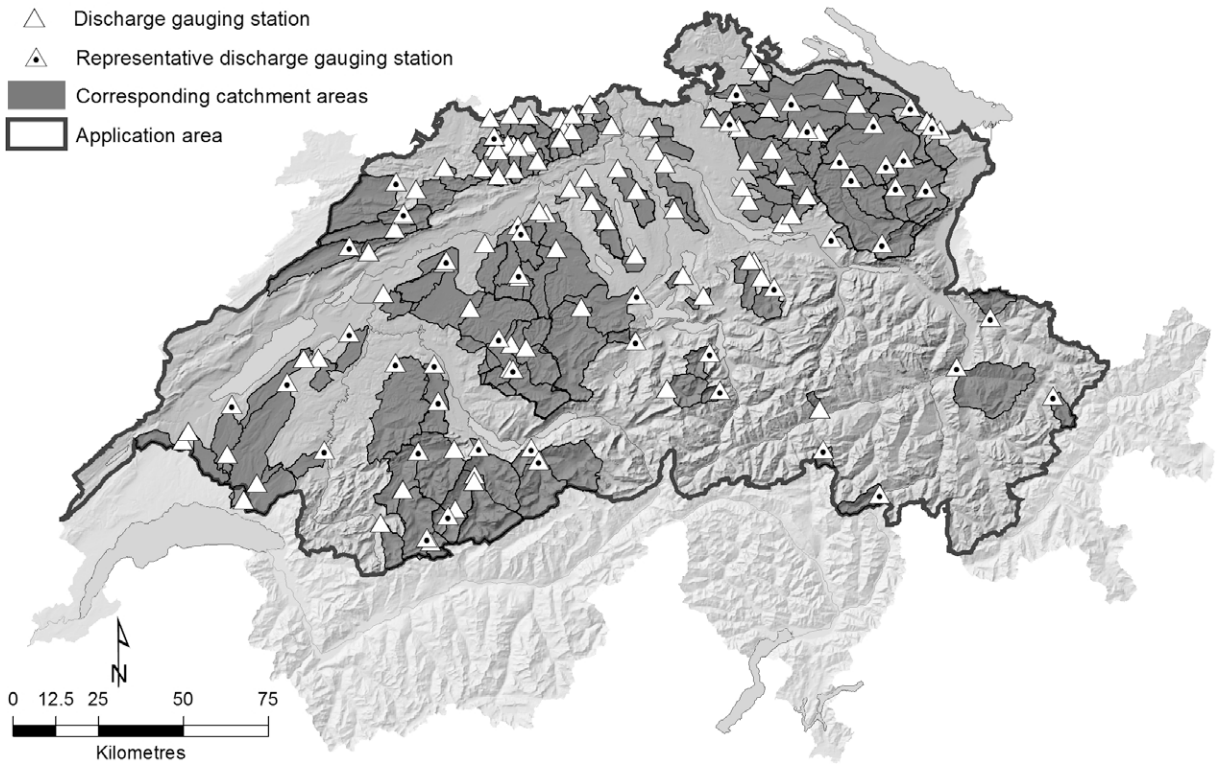


Fig. 2. Study area, with a total of 140 examined study catchments and 49 representative test catchments.

lay the foundations for a national flood estimation system, as in the present work (see also Viviroli et al., 2009a).

Efficiency score system

Elementary scores used

As discussed above, discharge data are usually the only observations available for calibration; consequently, the amount of information for tuning the model parameters is very restricted. To improve utilisation of these sparse data, it was proposed to combine different efficiency scores (e.g. Gupta et al., 1998; Seibert and McDonnell, 2002; Madsen, 2000, 2003; Merz, 2002). Therefore, an objective scoring system has been designed to assess the goodness-of-fit between simulated and observed discharge from different points of view. The system develops recommendations by Seibert and McDonnell (2002) and is essentially based on the standard efficiency score by Nash and Sutcliffe (1970) (*NSE*), its logarithmic derivate (NSE_{ln}) and the volumetric deviation between observed and simulated runoff (*VD*).

The Nash–Sutcliffe efficiency (*NSE*) score is defined as follows:

$$NSE = \frac{\sum_{t=1}^{n}(Q_t - \overline{Q})^2 - \sum_{t=1}^{n}(Q_t - q_t)^2}{\sum_{t=1}^{n}(Q_t - \overline{Q})^2}, \quad NSE \in]-\infty, 1] \tag{1}$$

where Q_t is the observed runoff at time step t, $\bar{Q}$ the average of observed runoff, q_t the simulated runoff at time step t and n the number of time steps. *NSE* quantifies the improvement of the model relative to the mean of the observations; it tends towards 1 when q_t tends towards Q_t. Although criticised for being particularly sensitive to high flows, runoff variance and meteorological model input (see Legates and McCabe, 1999; Eckhardt, 2002; Schaefli and Gupta, 2007), it has remained a popular benchmark and is well suited to comparison and evaluation of model runs using different model parameters.

The particular sensitivity of *NSE* to higher flow values can be amended with a simple logarithmic transformation which emphasises differences in medium and low flow periods:

$$NSE_{ln} = \frac{\sum_{t=1}^{n}(\ln(Q_t) - \ln(\overline{Q}))^2 - \sum_{t=1}^{n}(\ln(Q_t) - \ln(q_t))^2}{\sum_{t=1}^{n}(\ln(Q_t) - \ln(\overline{Q}))^2}, \quad NSE_{ln} \in]-\infty, 1] \tag{2}$$

Finally, the volumetric deviation (*VD*) between observed and simulated runoff is assessed, this being an important measure to ensure the overall hydrological plausibility of the simulation. For use in calibration, we compute the average of absolute volumetric deviations relative to the observed value. This measure is well suited to guiding a parameter search algorithm since it is dynamic relative to observed discharge and has a convenient scale range.

$$VD = \frac{\left|\sum_{t=1}^{n}\frac{q_t}{Q_t} - 1\right|}{n}, \quad VD \in [0, \infty[\tag{3}$$

Since *VD* is based on the rate of simulated to observed runoff, it is difficult to interpret. Therefore, we introduce the annual average of the sum of volumetric deviations (where k is the number of years that are simulated), which is far more straightforward to interpret:

$$SVD_a = \frac{\sum_{t=1}^{n}(q_t - Q_t)}{k}, \quad SVD_a \in]-\infty, \infty[\tag{4}$$

SVD_a is given in mm yr^{-1} and sums up the amount of runoff by which the model overestimates (positive sign) or underestimates (negative sign) observed runoff per year. Its advantages include that it can be directly related to the water balance of the corresponding catchment or region and is independent of the modelling time step. With an observed annual runoff sum of 1500 mm, for example, a SVD_a of + 60 mm yr^{-1} means that the model

overestimates average annual runoff by 4%. The bias estimated by SVD_a is usually largely attributable to errors in the precipitation input and water balance adjustment factors PKOR and SNOKOR.

Assessment periods

Each of the three scores introduced above has its particular sensitivity: *NSE* for high flows, NSE_{log} for low flows and *VD* for the water balance. In order to further increase the information content available for a monitored automatic model calibration, each of these scores is additionally assessed with regard to three time aspects:

- Entire time range (*full*): The complete hydrograph output is considered at model time-step resolution (1984–2003 at hourly resolution in the present case). This time aspect is identical to the initial definition of the score. *full* is to be maximised in calibration.
- Year-to-year standard deviation (*ann*): The score is computed at model time-step resolution for each year, and subsequently, the standard deviation is computed for these yearly scores. This is to differentiate parameter sets with good scores over the entire period from scores which show high scores in some years and low scores in other years. When results over the entire time period (*full*) are similar, parameter sets with more constant year-to-year results (*ann*) are preferred. *ann* is to be minimised in calibration.
- Month-to-month standard deviation (*mon*): The score is computed at model time-step resolution for all data in the entire time period which belong to the same month (in the present case e.g. complete hydrograph output for January 1984, January 1985, . . . , January 2003). Subsequently, the standard deviation is computed for these monthly scores. Parameter sets with similar performance in all seasons are preferred. *mon* is to be minimised in calibration.

Combining the three scores (*NSE*, NSE_{log} and *VD*) with the three time aspects, a total of nine objective elementary scores are defined: NSE_{full}, NSE_{ann}, NSE_{mon}, $NSE_{ln,full}$, $NSE_{ln,ann}$, $NSE_{ln,mon}$, VD_{full}, VD_{ann} and VD_{mon}. The year-to-year (*ann*) and month-to-month (*mon*) scoring systems give priority to parameter sets with constant performance and eliminate the effects of high frequency variability on the efficiency estimate. At the cost of somewhat lower overall efficiencies, it can be assumed that this procedure limits losses in performance when moving from the calibration to the validation period (Merz and Blöschl, 2004) and thus yields stable and representative parameter sets because the model produces 'the right answers for the right reasons' (Kirchner, 2006).

Combination into a total score

In order to combine the nine elementary scores into a total score, they have to be transformed to a common value range between 0 and 1. If the value of a score *S* is lower than the respective user-defined lower threshold s_{low}, its transformed value *T* is 0 (worst); if *S* is higher than the defined upper threshold s_{high}, *T* is 1 (best); in-between, the transformation function is linear (see Fig. 3). The score-specific values for s_{low} and s_{high} are provided in Table 1. Using this transformation procedure, it is further possible to assess only the most sensitive – and therefore most relevant – value range of the scores (see grey area in Fig. 3). Very low and very high scores, respectively, are not further differentiated (see Seibert, 1997).

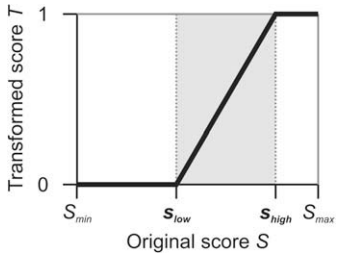


Fig. 3. Remapping an original score *S* (range: S_{min}–S_{max}) to a transformed score *T* with uniform value range from 0 to 1, using thresholds s_{low} and s_{high}. Using *T*, *S* is only assessed in the value range between s_{low} and s_{high} (grey area). Specific values for s_{low} and s_{high} are provided in Table 1.

Using individual weights *w* as indicated in Table 1, the nine elementary scores NSE^*_{full}, NSE^*_{ann}, NSE^*_{mon}, $NSE^*_{ln,full}$, $NSE^*_{ln,ann}$, $NSE^*_{ln,mon}$, VD^*_{full}, VD^*_{ann} and VD^*_{mon} (the asterisk indicating that the scores have been re-mapped to a common range of values) are then combined into three intermediate acceptability scores *ALIN*, *ALOG* and *AVOL*, which summarise the model's overall performance (*full*) as well as the respective year-to-year (*ann*) and month-to-month variability (*mon*), each for NSE^*, NSE^*_{ln} and VD^*. Ultimately, the total acceptability score *ATOT* is calculated from *ALIN*, *ALOG* and *AVOL* as a weighted geometric mean (following Seibert and McDonnell, 2002). The weights proposed here attach particularly high importance to the overall volume difference (VD_{full}) and to the overall Nash–Sutcliffe efficiency (NSE_{full}); for these scores, good performance is absolutely required in order to attain hydrologically plausible simulations.

$$ALIN = \left(NSE^*_{full} \cdot w^{lin}_1 + NSE^*_{ann} \cdot w^{lin}_2 + NSE^*_{mon} \cdot w^{lin}_3\right)/w^{lin} \quad (5)$$

$$ALOG = \left(NSE^*_{ln,full} \cdot w^{log}_1 + NSE^*_{ln,ann} \cdot w^{log}_2 + NSE^*_{ln,mon} \cdot w^{log}_3\right)/w^{log} \quad (6)$$

$$AVOL = \left(VD^*_{full} \cdot w^{vol}_1 + VD^*_{ann} \cdot w^{vol}_2 + VD^*_{mon} \cdot w^{vol}_3\right)/w^{vol} \quad (7)$$

$$ATOT = ALIN^{w^{lin}/w^{tot}} \cdot ALOG^{w^{log}/w^{tot}} \cdot AVOL^{w^{vol}/w^{tot}} \quad (8)$$

with

$$w^{lin} = \sum_{i=1}^{3} w^{lin}_i, \quad w^{log} = \sum_{i=1}^{3} w^{log}_i, \quad w^{vol} = \sum_{i=1}^{3} w^{vol}_i,$$
$$w^{tot} = w^{lin} + w^{log} + w^{vol}$$

Fig. 4 summarises the calculation of the scores introduced above for a real-world catchment. This is also to illustrate how a relatively indistinct pattern of the popular Nash–Sutcliffe efficiency (NSE_{full}, top left) is further differentiated through logarithmic and volumetric scores, resulting in a clearly more distinctive pattern in the overall acceptability score *ATOT*. This additional differentia-

Table 1
User-defined lower (s_{low}, worst, =0) and upper (s_{high}, best, =1) threshold values for transformation of the provisional efficiency scores; in-between the thresholds, a linear decrease applies (see Fig. 3). Weights used for calculation of summary score *ATOT* are indicated in the last column.

Score	Lower threshold (s_{low})	Upper threshold (s_{high})	Weight
NSE_{full}	0.25	0.95	$w_{1,1} = 6$
NSE_{ann}	0.10[a]	0.01[a]	$w_{1,2} = 4$
NSE_{mon}	0.50[a]	0.02[a]	$w_{1,3} = 2$
$NSE_{ln,full}$	0.25	0.95	$w_{2,1} = 3$
$NSE_{ln,ann}$	0.10[a]	0.01[a]	$w_{2,2} = 2$
$NSE_{ln,mon}$	0.50[a]	0.02[a]	$w_{2,3} = 1$
VD_{full}	0.10[a]	0.01[a]	$w_{3,1} = 10$
VD_{ann}	0.15[a]	0.02[a]	$w_{3,2} = 4$
VD_{mon}	0.10[a]	0.02[a]	$w_{3,3} = 3$

[a] These scores actually need to be minimised in calibration. To keep the overall score consistent in using a function value of 1 as best result and calibration goal, s_{high} (where the function value is 1) is lower than s_{low} (where the function value is 0).

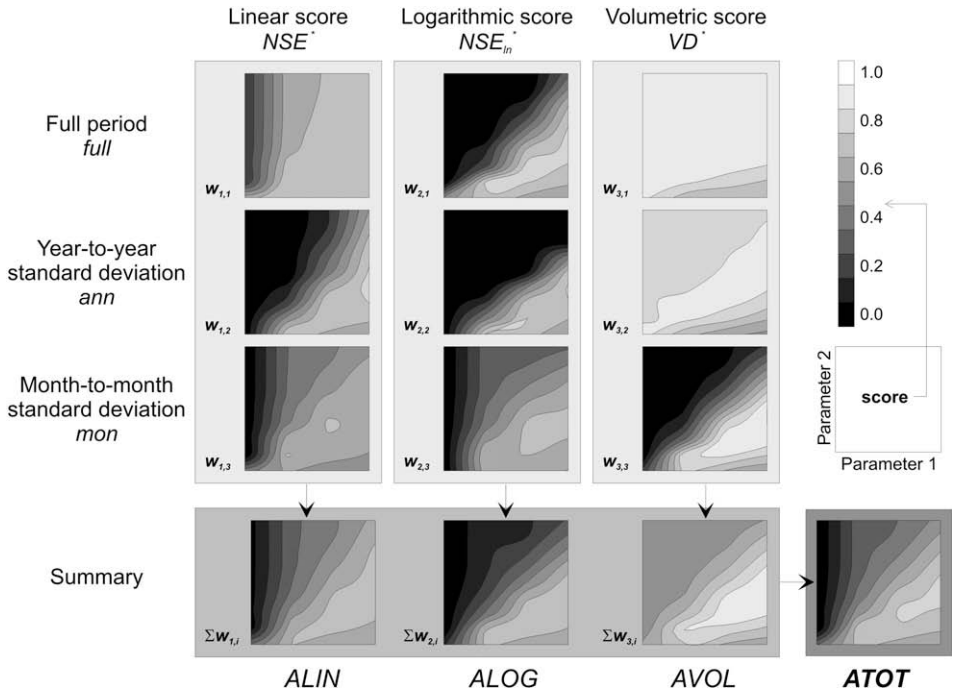


Fig. 4. Set-up of PREVAH's calibration score system consisting of nine subscores. The asterisks indicate that the scores have been mapped to a common value range from 0 to 1. Individual weights *w* serve to calculate the intermediate scores *ALIN*, *ALOG* and *AVOL* and, ultimately, the summary score *ATOT* (see Eqs. (5)–(8)). The score system is illustrated taking the example of the comparative assessment of model parameters CG1H (abscissa each) and SLZ1MAX (ordinate each), catchment of Bibere@Kerzers, 1994–1997.

tion is also useful for tackling the problems induced by frequent overparameterisation of hydrological models (see e.g. Jakeman and Hornberger, 1993; Perrin et al., 2001; for HBV-type models see particularly Bergström, 1995) since the limited calibration information (see Introduction) is exploited more thoroughly.

Search algorithm

A major issue in tuning the parameters of a hydrological model is that several different sets of parameter values can result in similarly good results (Seibert, 1997; Madsen et al., 2002). This implies the presence of a number of local optima in parameter space, and typical calibration procedures will identify different local optima depending on the efficiency scores in use and the design and sophistication of the search algorithm (Wilby, 1996). In practical terms, it is usually impossible to find a single 'optimal' parameter set (Duan et al., 1992; Wagener et al., 2001; Beven, 2006b). Our aim, therefore, is to identify a parameter set which yields good results and is representative for the catchment's hydrological behaviour. The latter is particularly important with respect to our overall goal of parameter regionalisation (Viviroli et al., 2009a) and will be discussed in more depth in "Discussion".

We here propose an iterative method which sequentially treats the parameters pair-wise and narrows down the considered parameter space step by step (Zappa and Kan, 2007). As Fig. 5 shows, the user-defined acceptable parameter space is first of all divided into 3×3 tiles, and the model is run for each of the four intersection points of the dividing lines. Then, the four tiles surrounding the intersection point with the best overall acceptability score *ATOT* are kept, while the remaining five tiles are discarded (Fig. 5a). With such a single iteration step, the parameter space is reduced by a factor of five ninths. The same procedure is then repeated, each time using the respective remaining parameter space for a maximum number of iterations as defined by the user; 4–8 iterations are recommended for each parameter pair (Fig. 5b). If there is no significant improvement of the overall acceptability score *ATOT*, the iteration is stopped for the current parameter pair, and the procedure continues with the next pair.

The parameter arrangement scheme chosen (see Table 2) groups parameters with reference to similar processes and assumes that best results are obtained by treating these common sensitivities. Furthermore, the grouping follows the model schematic from input treatment and melt processes to fast and then slow components; hence, the most sensitive parameters are treated first. Since the parameters are treated pair-wise and not at once, multiple sequential runs of the above parameter search algorithm are recommended in order to allow all parameters to adjust to each other. For the extensive sample used in this study, best results were achieved with three sequential calibration runs. Further runs did not lead to significant changes in parameter values.

Calibration period

A split-sample approach was adopted to divide the simulation into a calibration period and a validation period.

Recommendations concerning how many years to use for calibration differ from author to author. A basic requirement is that all hydrological conditions relevant for the respective catchment should occur in the selected time-span. Statistically, 2–3 years are sufficient for this according to Sorooshian and Gupta (1995), Merz (2002) recommends 5 years for a lumped HBV-type model with 11 tuneable parameters running in daily time steps. On the basis of previous experience with PREVAH, 4 years were chosen for calibration in the present study. A preceding warm-up year is used to estimate the storage fill levels, particularly the levels of snow and baseflow storages, which are very difficult to determine a priori. A warm-up year also precedes model validation and all other simulations presented here; in all cases, it is discarded and not used for evaluation of model performance. The suitability of this model initialisation is discussed in "Model warm-up".

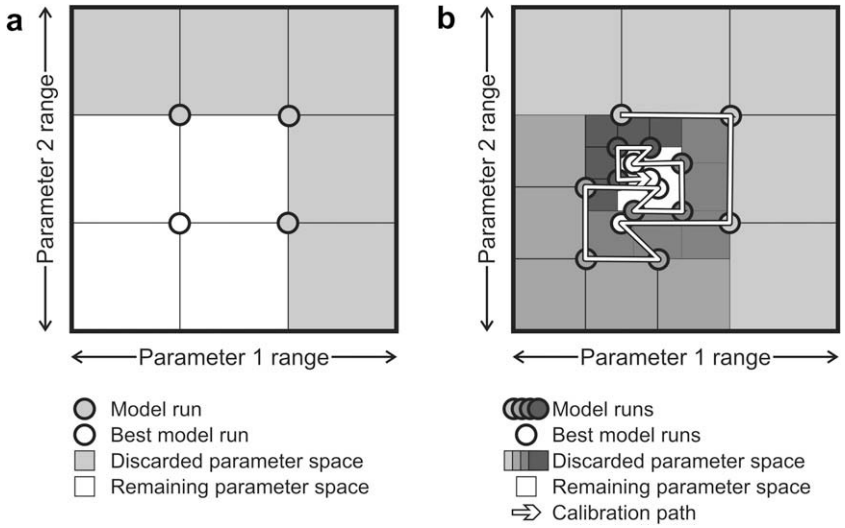


Fig. 5. Iterative search algorithm: (a) shows a single iteration, (b) the result of four successive iterations.

Table 2
Grouping of parameters for pair-wise calibration, with initial values and calibration range.

Pair	Parameters	Parameter description and unit	Initial value	Calibration range
1	PKOR	Precipitation adjustment (%)	0	−30–30
	SNOKOR	Snow adjustment (%)	0	−50–50
2	T0	Threshold temperature snowmelt (°C)	0	−1.00–1.00
	TMFSNOW	Temperature melt factor for snow (mm d^{-1} K^{-1})	1.5	0.10–3.00
3	T0	Threshold temperature snowmelt (°C)	0	−1.00–1.00
	RMFSNOW	Radiation melt factor for snow (mm h^{-1} K^{-1} W^{-1} m^2)	1×10^{-4}	5×10^{-5}–3×10^{-4}
4	SGR	Threshold storage for surface runoff (mm)	30	10–50
	K0H	Storage time for surface runoff (h)	10	10–30
5	K1H	Storage time for interflow (h)	75	50–150
	PERC	Percolation rate (mm h^{-1})	0.1	0.04–0.20
6	CG1H	Storage time for quick baseflow (h)	750	200–1000
	SLZ1MAX	Maximum content of the quick baseflow storage (mm)	150	25–250
7	K2H	Storage time for slow baseflow (h)	2500	1000–4000
	PERC	Percolation rate (mm h^{-1})	0.1	0.04–0.20
8[a]	ICETMF	Temperature melt factor for ice (mm d^{-1} K^{-1})	2	0.50–3.00
	ICERMF	Radiation melt factor for ice (mm h^{-1} K^{-1} W^{-1} m^2)	2×10^{-5}	1×10^{-5}–3×10^{-4}

[a] For glaciated catchments only.

The 1994–1997 period (with 1993 as the model warm-up year) was chosen for calibration since it offers the best conditions concerning meteorological network density. Furthermore, it is the time period for which the largest number of the selected 140 catchments has gauged runoff data, which allows the calibration period to be kept highly constant. Further details concerning the suitability of this calibration period will be provided in the Discussion Section. For catchments where no runoff data are available for this period, the alternative calibration periods 1984–1987, 1989–1993 or 1999–2003 are used, each with a preceding warm-up year. The years not used for calibration serve for model validation; typically, these are 15 years (1984–1993 and 1998–2003, warm-up year 1983).

The calibration procedure that has been introduced in this section will be referred to as 'standard procedure' below.

Flood calibration

With a view to flood estimation, the standard calibration procedure introduced above had to be extended with peak-flow-sensitive efficiency scores in order to improve results in peak-flow simulation. This entails compromises in standard model quality since it is not feasible to reproduce both standard and flood conditions with the same parameter set due to a combination of inadequacies in model structure and inaccuracies in rainfall and discharge data (Lamb, 1999; see also Cullmann, 2007). This is especially true if high temporal resolutions and long simulation periods are considered.

Extending the scores introduced by Lamb (1999), an attempt was made to construct a combination of peak-flow-sensitive measures which features maximum objectivity and allows calibration which requires – contrary to Lamb's study – no user intervention.

Peak-flow-sensitive efficiency scores

A total of four additional peak-flow-sensitive scores are used. Two of them are computed on the basis of a peak-over-threshold (POT) series (Naden, 1992) which encompasses, for a total period of k years, the $3 \times k$ largest records of the time series in question. To extract POT series, the independency of two subsequent peaks P_{t1} and P_{t2} needs to be ensured. Therefore, runoff is required to fall below a threshold value of $\overline{Q} + \{[\min(P_{t1}, P_{t2}) - \overline{Q}]/2\}$, i.e. below half the height of the smaller of the two peaks, relative to mean annual runoff $\overline{Q}$ (see Maniak, 2005). If this condition is not met, only the larger of the two peaks is extracted.

The first score is sensitive to the extent and timing of the peaks in the POT record and was proposed by Lamb (1999):

$$QT_{POT} = \sum_{i=1}^{m} \sqrt{[(P_i - p_i)/P^*]^2 + [(T_i - t_i)/T^*]^2}, \quad QT_{POT} \in [0, \infty[\qquad (9)$$

For a POT series with m records, P_i and p_i are the ith-largest observed and simulated peak-flow values, respectively, P^* is the difference between the largest and the smallest observed peak-flow values, T_i and t_i are the occurrence times of the ith-largest observed and simulated peak-flow values, respectively, and T^* is the time difference between the latest and the earliest observed peak flows. A QT_{POT} value of 0 indicates perfect correspondence of observed and simulated peaks; with growing deviation, the function value increases. QT_{POT} has no upper limit and is a relative value, i.e. it can be used only to assess the improvement between two different model runs.

Also introduced by Lamb (1999), the second score is the sum of absolute errors in the POT series:

$$SAE_{POT} = \sum_{i=1}^{m} |P_i - p_i|, \quad SAE_{POT} \in [0, \infty[\qquad (10)$$

Again, this is a relative function; a SAE_{POT} value of 0 means perfect correspondence, its increase points at growing deviations between observed and simulated peaks.

The third score is a sum of weighted absolute errors which considers the entire flow record. By multiplying the absolute errors with observed runoff Q to the power of a, the sensitivity for high flow values is increased:

$$SWAE = \sum_{t=1}^{n} (Q_t^a |Q_t - q_t|), \quad SWAE \in [0, \infty[\qquad (11)$$

A value of $a = 1.5$ was used, as proposed by Lamb (1999), for evaluation of peak-flow conditions.

Finally, the particular conditions of high-Alpine catchments are accounted for with a fourth score, which assesses the k annual peak-flow (APF) values. Particularly in glaciated catchments, the highest annual runoff value can be very similar from year to year. It therefore occurs quite frequently that the ith-largest observed peak value P corresponds very well to the ith-largest simulated peak value p, although P_i and p_i originate from different years (comparable issues were reported by Cameron et al., 1999). Although QT_{POT} considers this effect to some extent, it was necessary to introduce a specific score based on Spearman's rank correlation coefficient:

$$SRC_{APF} = 1 - \frac{6 \cdot \sum_{i=1}^{k} (P_i - p_i)^2}{k(k^2 - 1)}, \quad SRC_{APF} \in [-1, 1] \qquad (12)$$

Combination with standard calibration

The aim of deterministic long-term simulation requires a hydrologically plausible model parameterisation, even for the flood-adjusted version (see e.g. Beven, 2001). Therefore, the flood calibration starts from the parameter set established using the above standard procedure; it consists in running the iterative search procedure one more time, now incorporating the additional flood-sensitive scores. The weight w_{HQ} defines the influence of the flood-sensitive scores, with $w_{HQ} = 1$ meaning a pure calibration to flood scores and $w_{HQ} = 0$ a pure calibration to standard scores.

To determine the value of w_{HQ}, six representative catchments were examined with values of $w_{HQ} = 0$, 0.25, 0.5, 0.75 and 1. Fig. 6 shows that as to the flood-sensitive scores (left), no improvement is achieved for values of w_{HQ} larger than 0.5. It is also evident that only the QT_{POT} and SAE_{POT} scores can be improved markedly in all six representative catchments. In contrast, $SWAE$ and SRC_{APF} show individually different results for the six catchments. Therefore, the average changes observed in Fig. 6 are small and indicate no clear tendency. The behaviour of standard scores (Fig. 6, right) shows a slight increase for $w_{HQ} = 0.25$ and then a steady decrease with increasing w_{HQ} – obviously, compromises have to be made in order to improve reproduction of flood events. With a w_{HQ} value of 0.5, a maximum increase in flood scores is achieved while the decrease in standard scores remains at an acceptable level.

Results

The calibration and validation results from 49 representative catchments are summarised in Fig. 7. We will focus on Nash–Sutcliffe efficiency (NSE) and annual average volumetric deviation (SDV_a) since these scores are well known and most easily interpreted.

For the standard calibrated version, the loss in NSE when moving from the calibration period (median NSE: 0.75) to the validation period (median NSE: 0.72) is small, and the characteristics of interquartile range and spread do not change significantly. It can

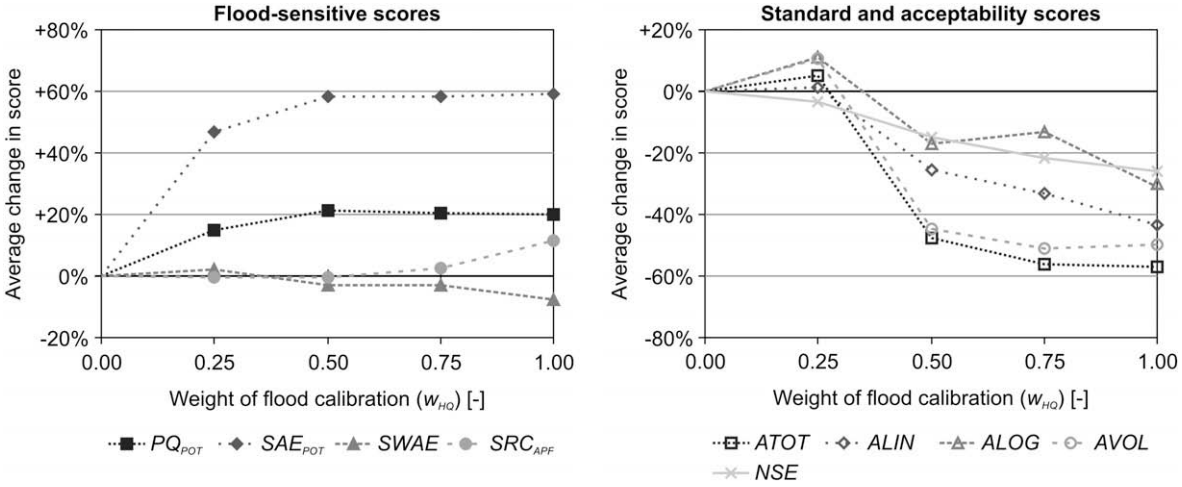


Fig. 6. Average increase or decrease in flood (left) as well as standard and acceptability (right) scores for six representative test catchments, relative to weight of flood calibration w_{HQ} (indexed to $w_{HQ} = 0$).

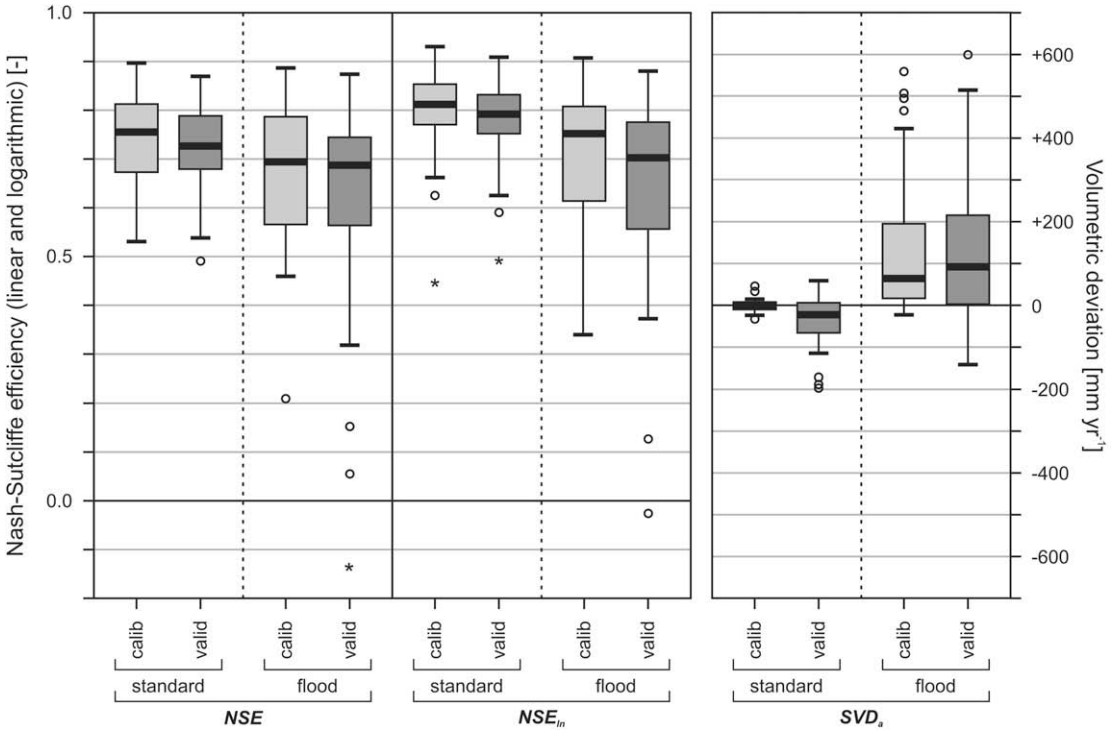


Fig. 7. Box-plot for *NSE*, NSE_{ln} and SVD_a model scores from 49 representative catchments in calibration (calib) and validation (valid) periods for standard and flood-calibrated parameter sets. Circles denote outliers (distance from upper or lower quartile is between 1.5 and 3 times the quartile range), stars extreme values (distance from upper or lower quartile is greater than 3 times the quartile range).

therefore be assumed that the parameter sets are representative of the catchment's dynamics. The same is true for the NSE_{ln} score, which even achieves a slightly higher median value of 0.81 (calibration) and 0.79 (validation). When comparing these results to those from other studies, it should be borne in mind that the assessment of hourly flows with the Nash–Sutcliffe efficiency is more demanding than the assessment of daily flows. Furthermore, the present study considers catchments with a wide range of different properties in a complex mountainous environment subject to large regional and seasonal differences in meteorological conditions.

The excellent median calibrated SDV_a of +0.9 mm yr^{-1} slightly decreases to –21.0 mm yr^{-1} in validation, and at the same time, the quartile spread rises from 14.9 mm yr^{-1} to 72.3 mm yr^{-1}. Against the backdrop of average runoff of Switzerland (991 mm yr^{-1}; see Weingartner et al., 2007), these values are interpreted as an annual runoff overestimation of 0.1% (quartile spread: 1.5%) for calibration and an underestimation of 2.1% (quartile spread: 7.3%) for validation. These figures are very reasonable particularly when the large uncertainties in interpolated precipitation are taken into consideration (see e.g. Chaubey et al., 1999; Ahrens and Jaun, 2007). Therefore, the representativity of our parameter sets is confirmed also with regard to water balance.

Looking at the flood-calibrated parameter sets, it is apparent that they perform worse than the standard calibrated variant with respect to the normal flow scores assessed here. This highlights the necessity to make compromises in overall model quality in order to achieve an improvement of the range of peak-flow values. The compromises are apparent particularly in the rather marked increase in volumetric deviation (SVD_a). Nevertheless, the *NSE* and NSE_{ln} scores indicate that the simulation is still hydrologically plausible. Furthermore, the decrease from calibration to validation is as small as that for standard calibration, although a few outliers to the bottom are observed. Still, stability and representativity of the flood-calibrated parameter sets can be assumed.

The improvements achieved for the two particularly relevant peak-over-threshold scores QT_{POT} and SAE_{POT} through flood calibration (*calHQ*) are shown in Fig. 8. Since both scores are relative only, the improvements are given relative to standard calibration (*calMQ*):

$$QT_{POT}^{relMQ} = [QT_{POT}(calMQ) - QT_{POT}(calHQ)]/QT_{POT}(calMQ) \quad (13)$$

$$SAE_{POT}^{relMQ} = [SAE_{POT}(calMQ) - SAE_{POT}(calHQ)]/SAE_{POT}(calMQ) \quad (14)$$

For QT_{POT}, a median improvement of 32% is observed in calibration. Validation falls noticeably short of this because the assessment of flood peaks concerning both their size and their timing is very demanding. Nevertheless, with a median improvement of 13%, a visible improvement is achieved in the majority of cases. Results for SAE_{POT}, on the other hand, are very stable, with a median improvement of 54% both for the calibration period and for the validation period. Detailed results concerning flood estimation are discussed in the companion paper by Viviroli et al. (2009a).

A similar evaluation for all 140 catchments is not feasible since for some catchments, validation data (i.e. hourly observed runoff) are available only for short periods; this means that an assessment would be too heterogeneous. The median scores achieved in calibration, however, are very similar to those of the representative sample above: for standard calibration, median values are 0.74 for *NSE*, 0.77 for NSE_{ln} and 2.2 mm yr^{-1} for SDV_a; for flood calibration, they are 0.67 for *NSE*, 0.63 for NSE_{ln} and 53.0 mm yr^{-1} for SDV_a (for details see Viviroli, 2007).

On average, standard calibration required 309 model runs (range: 204–408 runs), while the additional flood calibration took another 139 model runs (range: 120–144 runs). Compared to other parameter identification methods such as grid search (see "Search algorithm") or Monte-Carlo (see "Discussion"), this is highly efficient. High efficiency is relevant considering that PREVAH is a semi-distributed HBV-type model, where a single model run for

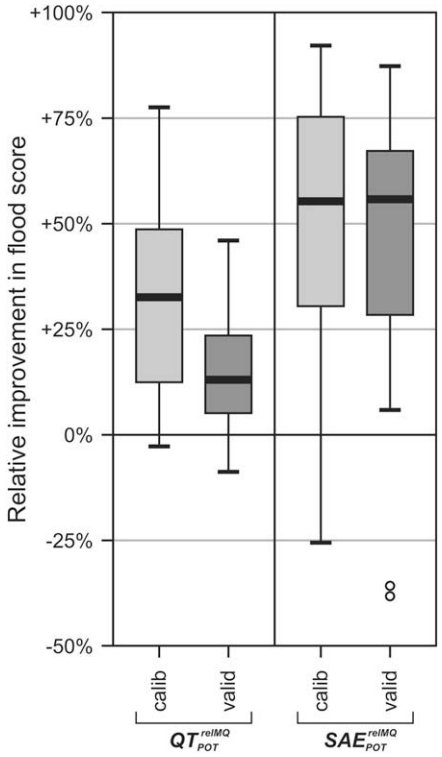


Fig. 8. Box-plot for improvement of flood-sensitive scores QT_{POT} and SAE_{POT} from standard to flood calibration, relative to standard calibration. Data from 49 representative catchments in calibration (calib) and validation (valid) periods.

calibration at hourly resolution typically takes several minutes using modern computer resources. Efficiency regarding results will be commented on in the Discussion.

Fig. 9 examines *NSE* relative to various catchment properties:

As to catchment size, best results are obtained in the scale range of 100–500 km^2; for catchments smaller than about 25 km^2, the maximum model efficiency achieved decreases noticeably. A general tendency is observed in that smaller catchments show a higher number of low *NSE* values, which is consistent with expected effects of error averaging for intermediate scales (see Blöschl, 2001).

As to the other properties, connections are observed insofar as the number of catchments showing low *NSE* values decreases with increasing value of the examined property. From the HRU size pattern it can be deduced that catchments with more homogeneous properties (i.e. bigger HRUs) are slightly easier to model. The patterns for mean catchment altitude and particularly for average simulated snowmelt show that the reliability of snowmelt and glacier melt simulation is high (similar findings were obtained for Austria by Parajka et al., 2005). It is assumed, however, that these patterns are partly caused by the particular sensitivity of the *NSE* score to the discharge regime type (see "Elementary scores used" and Schaefli and Gupta, 2007). Precipitation and runoff characteristics suggest that drier catchments are more difficult to simulate with the present model set-up. Results from catchments which were discarded because of insufficient efficiencies have shown that below 1250 mm yr^{-1} of areal precipitation, there are a greater number of catchments with low *NSE* values (Viviroli, 2007). A similar threshold appears to be constituted by 1000 mm yr^{-1} of observed runoff. Neither of these properties, however, seem to limit model efficiency, i.e. high *NSE* values are observed across the entire range of precipitation and runoff characteristics.

The above six catchment characteristics were also evaluated with regard to SDV_a and the improvement of QT_{POT} from standard to flood calibration (QT_{POT}^{relMQ}). No clear patterns, however, were found in these scores except a larger improvement of QT_{POT} in basins with smaller HRU size (Viviroli, 2007). This means that neither the adjustment of volumetric error nor the flood calibration routine are particularly dependent on the properties examined.

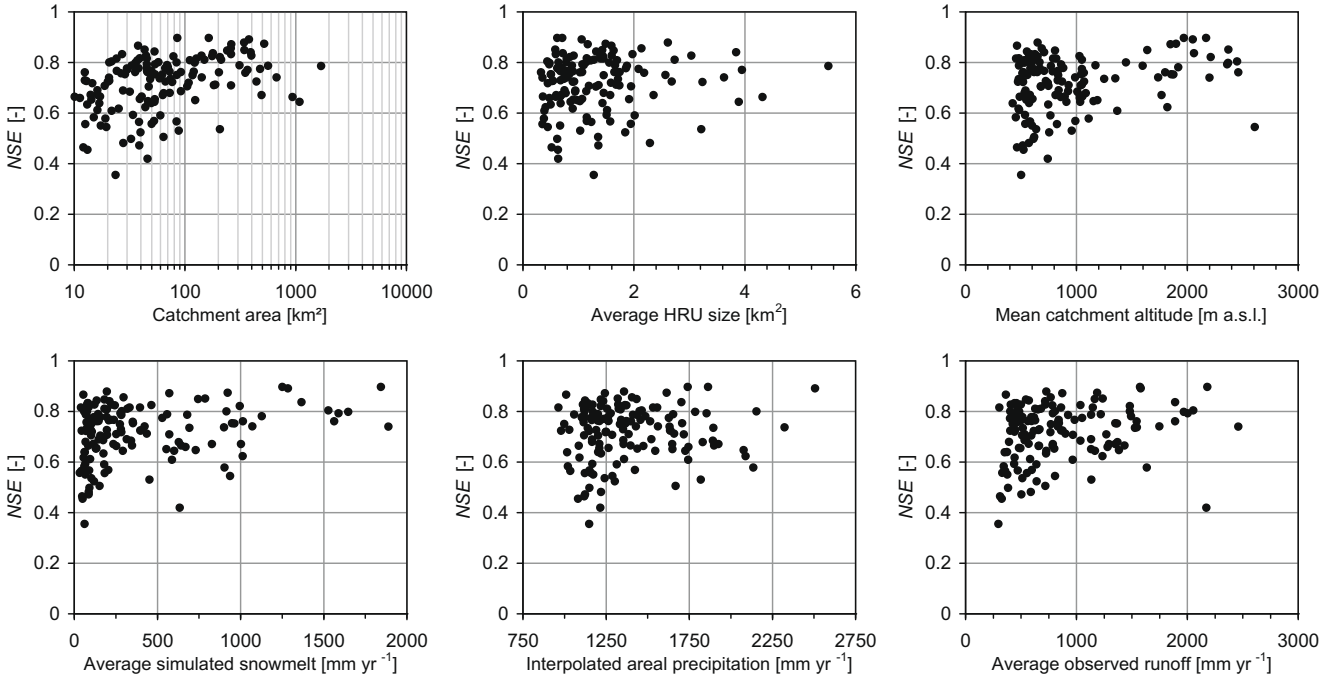


Fig. 9. Nash–Sutcliffe efficiency (*NSE*) from standard calibration compared to catchment area, average HRU size, mean catchment altitude, average annual simulated snowmelt, average annual areal precipitation and average observed runoff. Data from 140 catchments, calibration period.

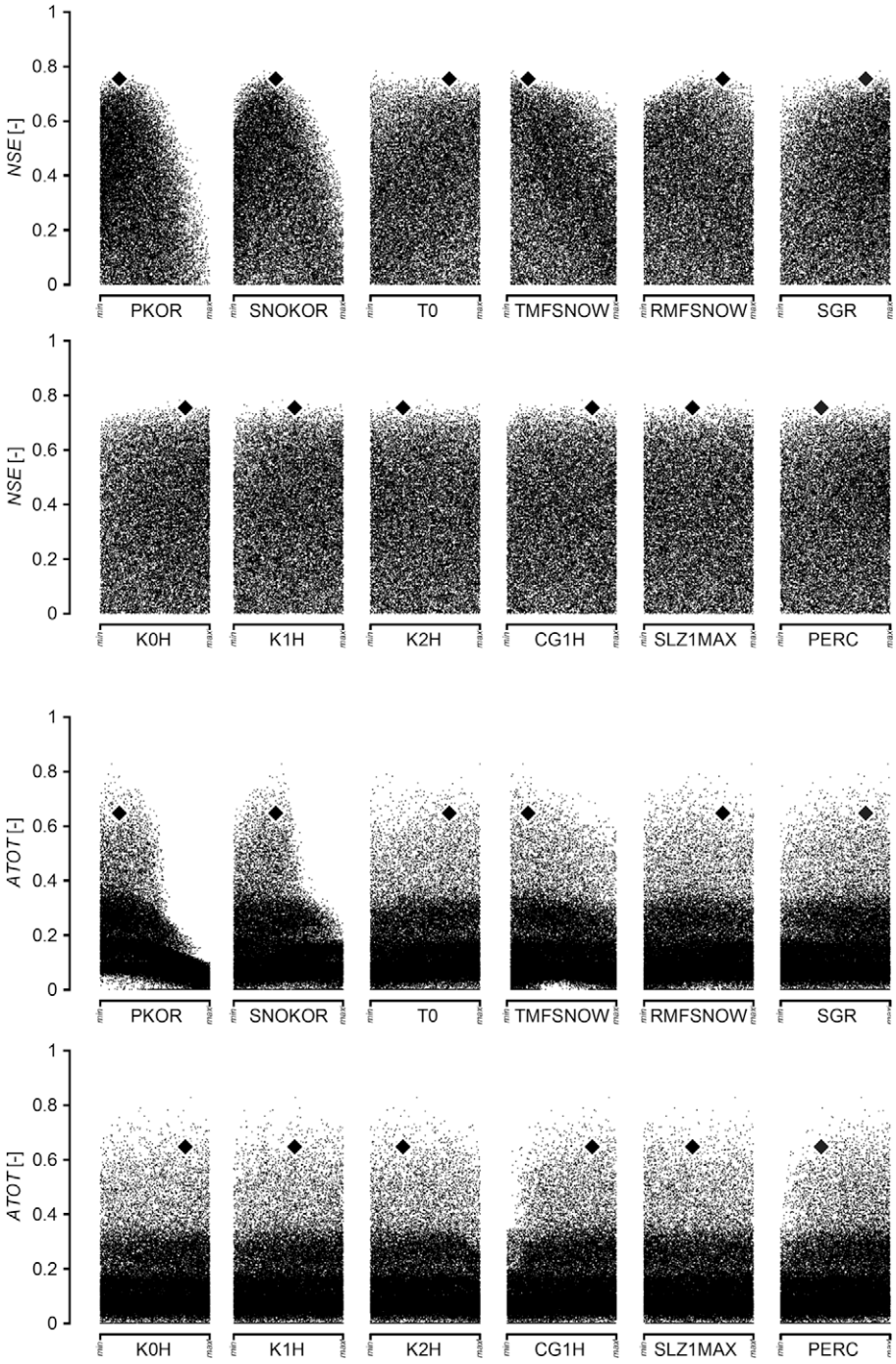


Fig. 10. Nash–Sutcliffe efficiencies (*NSEs*, upper half) and PREVAH summary scores (*ATOT*, lower half) for 50,000 randomly generated Monte-Carlo sets of the 12 tuneable model parameters; test catchment of Allenbach@Adelboden, 1994–1997 simulation period. The diamond signature denotes the value obtained using standard calibration, 1994–1997. For parameter limits and units see Table 2.

Discussion

Effectiveness

The effectiveness of our calibration procedure and the plausibility of the parameters determined were assessed by comparing them to the results of a Monte-Carlo analysis. For this purpose, 50,000 parameter sets were generated randomly for a few selected catchments, with parameter limits identical with those used in calibration (see Table 2). The large number of random parameter combinations is necessary for gaining sufficient resolution of the 12-dimensional parameter space. The comparison is based on the standard calibrated parameter sets (as opposed to the additional flood calibration variant), which are expected to have higher overall hydrological plausibility. On a current x86 machine, such a Monte-Carlo experiment with PREVAH running in hourly time steps typically takes several weeks per catchment.

Fig. 10 shows the results of the Monte-Carlo analysis for the Alpine, non-glaciated catchment of Allenbach@Adelboden (area: 28.8 km^2). Concerning *NSE* in the 1994–1997 calibration period (upper half of Fig. 10), the parameter set identified with our pragmatic calibration routine (black diamonds) scores as high as the

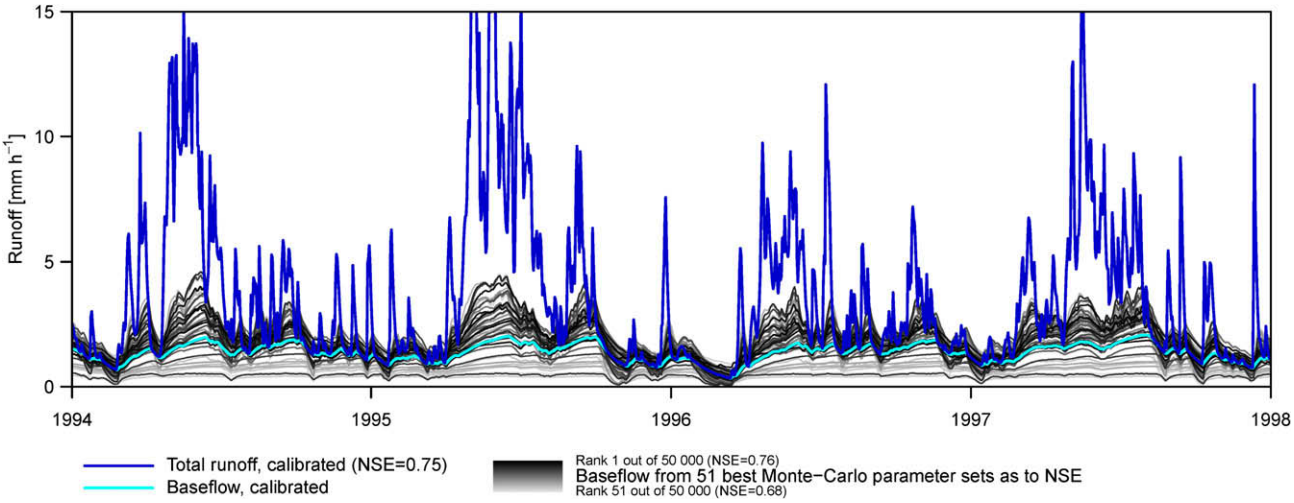


Fig. 11. Baseflow component from the 51 (out of a total of 50,000) Monte-Carlo experiments with best performance according to Nash–Sutcliffe efficiency (*NSE*), compared to baseflow and total runoff from calibrated model. Total runoff from Monte-Carlo experiments is not shown since the differences from calibrated total runoff are relatively small.

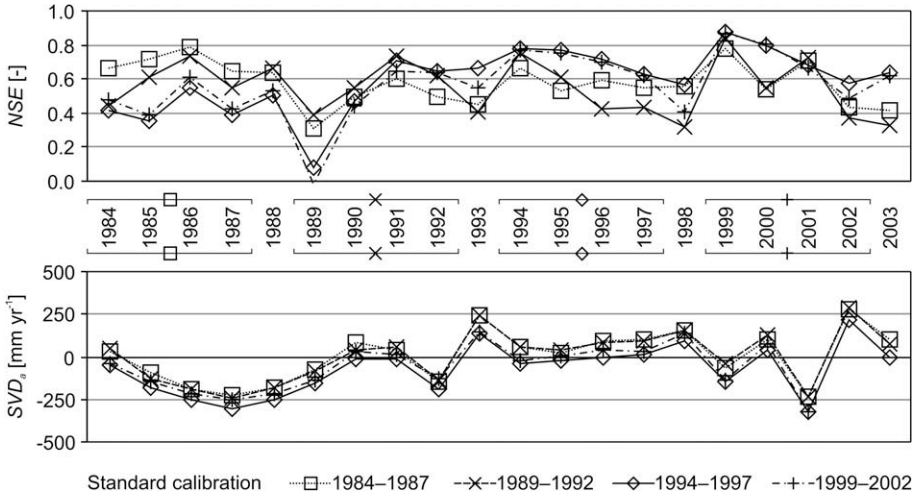


Fig. 12. Nash–Sutcliffe efficiencies (*NSEs*) and mean annual volumetric deviations (SVD_a) for Allenbach@Adelboden catchment, 1984–2003, using four standard calibrated parameter sets from calibration periods 1984–1987, 1989–1992, 1994–1997 and 1999–2002.

best parameter sets from the Monte-Carlo sample: the standard calibrated set ranks 13th of the 50,000 random sets and is therefore within the best 0.03%. The time taken to obtain the figures for that set, however, is shorter by a factor of roughly 100.

The lower part of Fig. 10 reveals that the PREVAH summary score *ATOT* is indeed able to further differentiate the simulation results. In the case of GG1H or PERC, for instance, the relatively indistinct *NSE* pattern is complemented with a more discernible range of best results regarding *ATOT*. Fig. 10 reveals that in the case of the Allenbach River, the *ATOT* scores for the parameter sets identified using our calibration routine are lower than those from the Monte-Carlo analysis – although their rank of 145th is still within the best 0.29% of the 50,000 random realisations. It should, however, be noted that these results for *ATOT* are not entirely comparable. In our calibration, *ATOT* is used to guide the individual pair-wise calibration steps with a view to obtaining plausible parameter interactions and, ultimately, a representative parameter set. With this goal, *ATOT* combines the *NSE*, NSE_{log} and *VD* efficiency scores with three time periods (see "Efficiency score system") and introduces valuable knowledge of parameter sensitivity into the calibration process. In Monte-Carlo sampling, however, *ATOT* is not able to function as intended because it is directly applied to the entire parameter set.

A visual analysis of the baseflow modelled (Fig. 11) confirms that our calibration leads to highly plausible behaviour of the groundwater module for the Allenbach River: the modelled baseflow component is neither too inert (which would mean that it is substituted by interflow) nor too active (which would mean that it substitutes for interflow). The 51 best Monte-Carlo realisations, in contrast, show a wide range of less plausible groundwater responses. On the one hand, baseflow is frequently misused as a partial substitute for interflow. On the other hand, baseflow is in some cases nearly constant over all seasons, meaning that the groundwater module merely produces a more or less uniform offset runoff value. There is no apparent connection between model performance and plausibility of baseflow simulation in the Monte-Carlo experiments, except that baseflow responses that are clearly too inert achieve rather low scores in the majority of cases (not always, though). Whereas the results from the Monte-Carlo experiments seem to show only low sensitivity towards baseflow and may therefore be implausible in that respect, more highly sensitive model components such as snowmelt show a clear response in

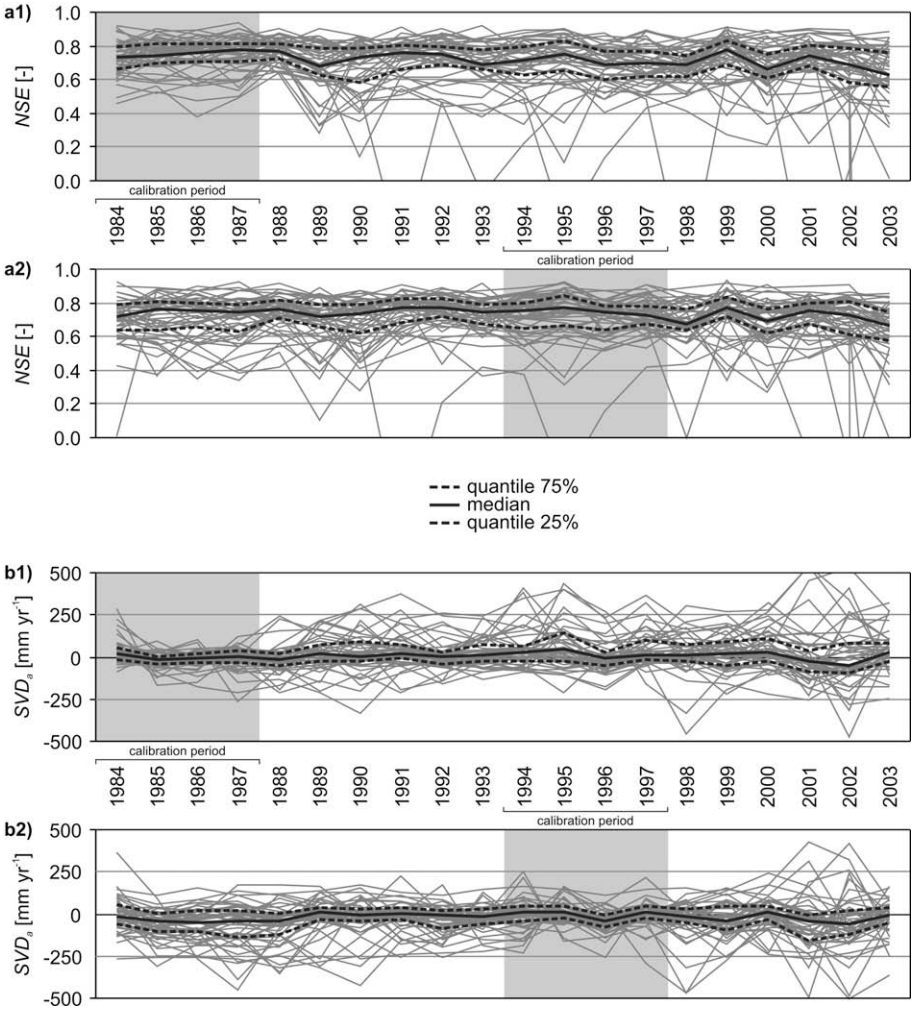


Fig. 13. Annual Nash–Sutcliffe efficiencies (*NSEs*, a...) and annual volumetric deviations (SVD_a, b...) for 49 representative test catchments using standard calibration for periods 1984–1987 (...1) and 1994–1997 (...2). Each grey line represents the results from one of the 49 catchments, black lines indicate corresponding quartile and median values. The occurrence of negative *NSE* values is due to failure of meteorological gauging stations (see main text).

model efficiency and are therefore plausible even if selected from randomly generated parameter sets.

It has already been shown in the "Results" chapter, and will be demonstrated in more detail in "Calibration period", that our procedure leads to representative parameter sets which show only small decreases in model performance when moving from calibration to validation period. Such representativity and robustness is, however, not guaranteed at all for a random procedure that runs the risk of overfitting the model and therefore yielding parameter sets with poor physical plausibility (Schoups et al., 2008) and inferior suitability for parameter regionalisation.

The above analysis also made it apparent that our model is particularly sensitive to the two water balance correction factors PKOR and SNOKOR (see Fig. 10). It was hoped that an additional Monte-Carlo experiment with these two values fixed at 0, would lead to an increase in sensitivity for the other parameters. However, the resultant changes were rather small; the number of parameter sets with reasonable results was still high for 10 parameters (Viviroli, 2007).

Calibration period

In the following section, we will examine whether our calibration procedure yields parameter sets that show similar performance irrespective of the calibration period in use.

Assessment of an exemplary test catchment

In a first detail analysis for an exemplary catchment, standard calibration was carried out for Allenbach@Adelboden for the four independent periods 1984–1987, 1989–1992, 1994–1997 and 1999–2002, each with a preceding warm-up year, which was discarded. Fig. 12 shows Nash–Sutcliffe efficiencies (*NSEs*) and summarised annual volumetric deviations (SVD_a) for these four calibration periods and suggests that there is no dominant influence of calibration period on simulation results. The pattern of annual *NSE* and SVD_a scores is comparable for each of the variants, leading to the conclusion that model performance is mainly governed by the hydrological processes occurring and the representativity of meteorological input; this will be further discussed in the subsequent section. Concerning average *NSE*, no parameter sets are found which perform better than those determined in the 1993–1997 standard calibration period. The unsatisfactory results in *NSE* for 1989 are due to problems concerning the meteorological data.

Assessment of 49 representative test catchments

In a second analysis, the 49 representative catchments were calibrated on two independent calibration periods: 1984–1987 and 1994–1997. A detailed analysis of model performance in these 49 representative catchments is given in Fig. 13, which shows SVD_a and *NSE* results for all simulation years (1984–2003).

Comparing the results from calibrations on 1984–1987 and 1994–1997, respectively, differences emerge as to the behaviour of *NSE* in the individual catchments (Fig. 13a). However, the patterns of both median and interquartile range are remarkably similar, which emphasises the robustness of our calibration scheme. The occurrence of negative *NSE* values is due to failure of meteorological stations in regions with a sparse network of such stations, which cannot be amended for either calibration period.

A slightly more stable picture is observed in the average annual summarised volumetric deviations (SVD_a) when using the 1994–1997 calibration period. It might be relevant that this calibration period is closer to the middle of the entire 1984–2003 simulation period and is therefore more representative. The 1984–1987 calibration period, by contrast, is at the edge of the overall simulation period, and the respective parameter sets are therefore applied for years more distant from the actual calibration. Consequently, long-term oscillations in the water balance of the catchments may show through. Another factor, as mentioned above, is that the network of meteorological stations is denser in the later of the two calibration periods; therefore, more information is available to the model, which, in turn, allows more effective calibration.

Model warm-up

As mentioned in "Calibration period", each simulation is preceded by a warm-up year, which is a common practice in hydrological modelling (see e.g. Seibert, 1997 and Zappa et al., 2003 for use in HBV-type models). McIntyre et al. (2005) are most specific in this respect and recommend neglecting the first 20% of the data time series to reduce sensitivity to initial conditions. This is in accordance with our experience from using PREVAH, which led us to use 1 year for warm-up and the subsequent 4 years for actual calibration (i.e. exactly the relation of 20–80%). This model spin-up particularly serves to fill the low-frequency storages for baseflow and snow. Especially for these more inert model components, it seems important that the warm-up period is of sufficient length to achieve reasonable initial conditions for calibration. Further significance of this is derived from the slight drift in the long-term water balance of the catchments which was revealed by analysis of SVD_a in the preceding "Assessment of 49 representative test catchments".

For 49 representative catchments of our sample, we examined the deviation of a freshly started model (M93, begin: 1.1.1993) as compared to a model which had already been running for 10 years (M83, begin: 1.1.1983). The freshly started model refers directly to the calibration situation in this paper, as described in "Calibration period": simulation starts in 1993, the first 12 months are discarded and the calibration refers to the 1994–1997 period. Not surprisingly, the most noticeable deviations are observed for the lower zone (saturated) storage SLZ and the corresponding slow runoff component *R*2, both of which refer to baseflow. The median deviation of the 49 representative catchments after 12 months is, however, still relatively low (SLZ: +7%, *R*2: +3%) and therefore representative for calibration. The initialisation of snow storage (SSNO) poses no problem since all catchments are clear of snow in summer (with the exception of snow in the accumulation zone of glaciers; this is treated separately, however). Consequently, the snow storages of M93 and M83 are replenished with similar water equivalents after the first summer in M93. For the fast and delayed runoff components (*R*0 and *R*1, respectively) as well as the corresponding upper zone storage (SUZ), temporary deviations are observed at the very beginning of M93 when SUZ is empty yet, and in the first snowmelt season (April/May, i.e. months 4/5 of M93) when fast runoff from snowmelt is too small. The relative deviations of further higher-frequency storages (e.g. soil moisture storage SSM, interception storage SI) are smaller than 1% after 30 days at the latest and remain negligible afterwards.

The above analyses lead us to the conclusion that the one-year model warm-up chosen in this study is sufficient. The deviations in SLZ and *R*2 at the beginning of the actual calibration period seem acceptable, particularly when it is borne in mind that the focus of the present paper is on peak flows rather than low flows. To achieve even more accurate initial conditions in future, various procedures may be adopted. One possibility would be to use a second warm-up year. Although such a procedure could further improve the accuracy of initial conditions, it seems unfavourable since valuable gauge data are scarified. Another possibility would be to assign initial snow storage values, which is difficult, though, since snow conditions may vary strongly from year to year (Auer et al., 2007). Furthermore, these storage values would even need to be specified for horizontal layers of 100 m height, which is difficult to determine but crucial for snowmelt behaviour in spring. Finally, regionally differentiated values for an a priori initialisation of SLZ could be sought (see e.g. Schwarze et al., 1999).

Complementary information content of flood efficiency scores

Being of particular relevance to the flood calibration variant, the complementary information content of the peak-flow-sensitive scores is assessed using the Monte-Carlo experiment introduced above. The Nash–Sutcliffe efficiency (*NSE*), a typical standard efficiency score, serves as a reference value. Complementarity can be assumed if high values in a peak-flow-sensitive score are not connected with high *NSE* scores and vice versa.

The patterns resulting for our exemplary catchment of Allenbach@Adelboden (Fig. 14) can be interpreted as follows:

- For QT_{POT} and *NSE*, no common optimum is found: for best results concerning QT_{POT}, (upper front of dot cloud), some compromises have to be made concerning *NSE* (right front of dot cloud) and vice versa. Therefore, it can be assumed that QT_{POT} introduces information which is not contained in *NSE*. However, finding a parameter combination which is plausible with regard to standard flow (*NSE*) and gives good results for peak flows (QT_{POT}) seems possible (upper right front of dot cloud).

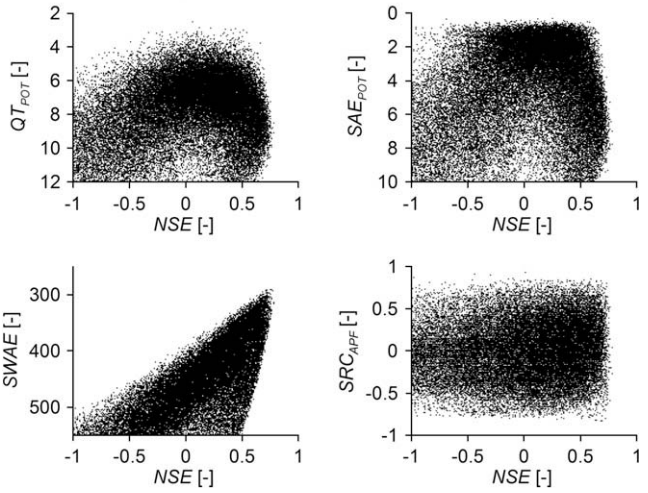


Fig. 14. Comparison of flood-sensitive scores QT_{POT}, SAE_{POT}, *SWAE* and SRC_{APF} with standard Nash–Sutcliffe efficiency (*NSE*); the common optimum is always located to the upper right. Data from 50,000 randomly created (Monte-Carlo) parameter sets for Allenbach@Adelboden catchment (28.8 km^2), 1994–1997 simulation period.

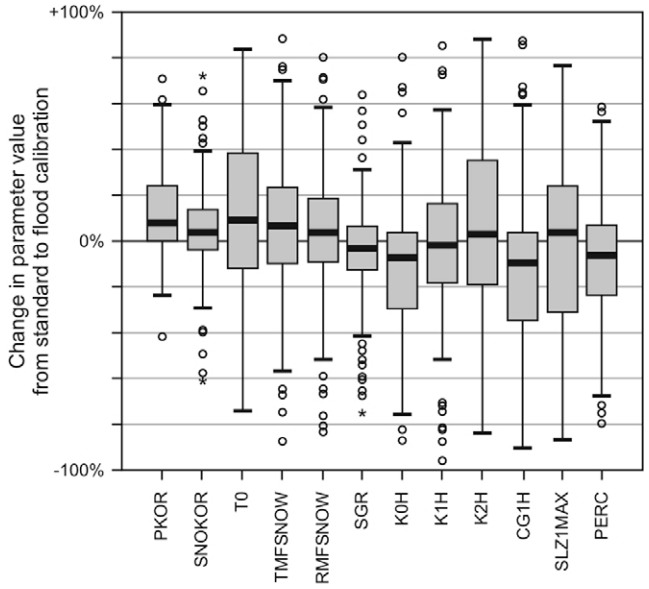


Fig. 15. Change in parameter value from standard calibration to flood calibration; values are relative to parameter range observed in standard calibration and refer to 140 successfully calibrated catchments.

- The interpretation of SAE_{POT} is similar to that of QT_{POT}, although minor compromises are necessary to obtain good results for both SAE_{POT} and *NSE*. Since good results for SAE_{POT} are also found when using parameter sets with low *NSE*, the overall hydrological plausibility of SAE_{POT} must be rated lower.
- The *SWAE* and *NSE* scores clearly share a common area of best results (upper right tip of dot cloud). When *NSE* is high, only little additional information can be expected from *SWAE*; with lower *NSE* values, however, *SWAE* can be used to discern parameter sets with better reproduction of high flow conditions.
- For SRC_{APF}, there seems to be only a small connection with *NSE*, which strongly suggests that independent information is provided by SRC_{APF}. At the same time, the very limited agreement of high SRC_{APF} values with high *NSE* values indicates that the overall hydrological plausibility of SRC_{APF} is low – not surprisingly, since it only assesses the annual peak-flow record. Therefore, SRC_{APF} should always be used in combination with other efficiency scores such as *NSE*.

In conclusion, it can be said that the QT_{POT} and SAE_{POT} scores are most valuable for adding peak-flow information to the calibration process. The *SWAE* score, on the other hand, contains less extra information on peak flows and rather serves to complement flood calibration. Finally, SRC_{APF} may be valuable for further differentiating results from the above scores, although it is of an auxiliary nature only due to its low plausibility for overall simulation.

Change in parameters from standard to flood calibration

This final section of the "Discussion" chapter examines how the standard calibration parameters changed after additional flood calibration (Fig. 15). With a focus on fast and delayed runoff (see model schematic, Fig. 1), the picture can be interpreted as follows:

With a noticeable increase in the water balance correction factor for rain (PKOR) and a slight decrease in the percolation rate (PERC), more water is available in the upper zone runoff storage (SUZ) for formation of surface runoff (*R*0) and interflow (*R*1). At the same time, the storage time for surface runoff (K0H) is reduced, which leads to accelerated surface runoff (*R*0). As to the storage time for interflow (K1H), no clear tendency is observed, and the threshold value for formation of surface runoff (SGR) is reduced only slightly.

The noticeably larger quartile distances for the parameters governing baseflow (K2H, CG1H and SLZ1MAX) suggest that here, sensitivity for flood calibration is smaller. The same is true for the threshold temperature for snowmelt (T0), while the temperature and radiation melt factors for snowmelt (TMFSNOW, RMFSNOW) show a slight increase.

Conclusions

Our pragmatic iterative calibration procedure does not claim to find with absolute certainty a global optimum in parameter space. Rather, it was demonstrated that it identifies robust parameter sets which maintain high performance scores in model validation. Besides reliability and objectivity, the procedure's major advantage is its cost-effectiveness, which makes it particularly useful for models demanding a great deal of CPU time, such as PREVAH. Therefore, many basins can be calibrated simultaneously by a single user, which also makes this method particularly valuable for national hydrometeorological surveys that use hydrological modelling for water resources management or flood estimation tasks at an extensive number of gauge sites. Furthermore, it is possible to re-calibrate a model for many sites within a reasonable period of time when additional or improved data become available (e.g. improved weather radar products; see Wüest et al., submitted) or when a specific event should be analysed for many sites (e.g. analysis of the 2005 flood events by the Swiss Federal Office for the Environment; see Bezzola et al., 2008 and Schwanbeck et al., 2008).

While there is no need for user intervention, the modeller can still adjust details of the procedure, e.g. by adapting the weights of the individual scores to the purposes of the specific study (e.g. focus on low flows or water balance). Furthermore, specific field experience about processes can be included by adapting the acceptable parameter range (see "Search algorithm"; also see Seibert and McDonnell, 2002). The additional flood calibration mode provides a well-balanced procedure to improve the standard parameter sets with respect to modelling high flows. Although compromises have to be made regarding the range of standard flow conditions, overall hydrological plausibility is retained.

With the methods presented here, it was possible to calibrate and validate a process-oriented, semi-distributed hydrological model in hourly resolution for an extensive set of 140 mesoscale catchments in Switzerland. This is an essential step in tackling the regionalisation of the hydrological model PREVAH for Switzerland. Adding the flood calibration mode provides the basis for flood estimation in ungauged Swiss basins through continuous process-based modelling (Viviroli et al., 2009a).

Acknowledgements

This work was mainly funded through a PhD grant for Daniel Viviroli from the Swiss Federal Office for the Environment (FOEN). We gratefully acknowledge Manfred Spreafico's support in accompanying the project. We also thank Massimo Corti, Daniel Gasser and Christof Sonderegger for their help with model calibration. The constructive comments of an anonymous reviewer were very helpful in improving the manuscript.

References

Ahrens, B., Jaun, S., 2007. On evaluation of ensemble precipitation forecasts with observation-based ensembles. Advances in Geosciences 10, 139–144.

Auer, M., Meister, R., Stoffel, A., 2007. Mean Snow Depths, 1983–2002. Hydrological Atlas of Switzerland, Plate 3.11. Federal Office for the Environment, Bern, CH.
Bergström, S., 1995. The HBV model. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, CO, pp. 443–476.
Beven, K.J., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water Resources 16, 41–51.
Beven, K.J., 2001. Rainfall–Runoff Modelling: The Primer. Wiley, Chichester, UK.
Beven, K.J., 2006a. On undermining the science? Hydrological Processes 20, 3141–3146.
Beven, K.J., 2006b. A manifesto for the equifinality thesis. Journal of Hydrology 320, 18–36.
Bezzola, G.-R., Hegg, C., (Eds.), 2008. Ereignisanalyse Hochwasser 2005. Teil 2 – Analyse von Prozessen, Massnahmen und Gefahrengrundlagen. Umwelt-Wissen 08-25. Federal Office for the Environment, Bern. <http://www.bafu.admin.ch/publikationen/publikation/00100/index.html?lang=de>.
Blöschl, G., 2001. Scaling in hydrology. Hydrological Processes 15, 709–711.
Botterweg, P., 1995. The user's influence on model calibration results: an example of the model SOIL, independently calibrated by two users. Ecological Modelling 81 (1–3), 71–80.
Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Towards improved calibration in hydrologic models: combining the strengths of manual and automatic methods. Water Resources Research 36 (12), 3663–3674.
Bronstert, A., Bárdossy, A., Bismuth, C., Buiteveld, H., Busch, N., Disse, M., Engel, H., Fritsch, U., Hundecha, Y., Lammersen, R., Niehoff, D., Ritter, N., 2003. LAHoR – Quantifizierung des Einflusses der Landoberfläche und der Ausbaumassnahmen auf die Hochwasserbedingungen im Rheingebiet. CHR Report II-18. International Commission for the Hydrology of the Rhine Basin (CHR), Lelystad.
Cameron, D., Beven, K.J., Tawn, J., Blazkova, S., Naden, P., 1999. Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty). Journal of Hydrology 219, 169–187.
Chaubey, I., Haan, C.T., Grunwald, S., Salisbury, J.M., 1999. Uncertainty in the model parameters due to spatial variability of rainfall. Journal of Hydrology 220, 48–61.
Cullmann, J., 2007. Enhancing flood forecasting with the help of transient model parameters. In: Ubertini L., Manciola P., Casadei S. (Eds.) Earth: Our Changing Planet. Proceedings of IUGG XXIV General Assembly in Perugia, Italy, p. 4661.
Duan, Q., Sorooshian, S., Gupta, V.K., 1992. Effective and efficient global optimisation for conceptual rainfall–runoff models. Water Resources Research 28 (4), 1015–1031.
Duan, Q., Gupta, H.V., Sorooshian, S., Rousseau, A.N., Turcotte, R., 2003. Calibration of watershed models. Water Science and Applications, vol. 6. American Geophysical Union (AGU), Washington.
Eckhardt, K., 2002. Vergleich zweier Verfahren zur automatischen Modellkalibrierung. Hydrologie und Wasserbewirtschaftung 46 (2), 69–73.
FOEN (Federal Office for the Environment), 2008. Observed Discharge Time Series of Swiss Rivers. Bern.
Frei, C., Schär, C., 1998. A precipitation climatology of the Alps from high-resolution raingauge observations. International Journal of Climatology 18, 873–900.
Garen, D.C., Marks, D., 2001. Spatial fields of meteorological input data including forest canopy corrections for an energy budget snow simulation model. In: Dolman, A.J., Hall, A.J., Kavvas, M.L., Oki, T., Pomeroy, J.W. (Eds.), Soil–Vegetation–Atmosphere Transfer Schemes and Large Scale Hydrological Models. IAHS publication 270. International Association of Hydrological Sciences, pp. 349–353.
Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resources Research 34 (4), 751–763.
Gupta, H.V., Beven, K.J., Wagener, T., 2005. Model calibration and uncertainty estimation. In: Anderson, M.G. (Ed.), Encyclopedia of Hydrological Sciences. Wiley, Chichester, UK.
Gurtz, J., Baltensweiler, A., Lang, H., 1999. Spatially distributed hydrotope-based modeling of evapotranspiration and runoff in mountainous basins. Hydrological Processes 13, 2751–2768.
Gurtz, J., Zappa, M., Jasper, K., Lang, H., Verbunt, M., Badoux, A., Vitvar, T., 2003. A comparative study in modeling runoff and its components in two mountainous catchments. Hydrological Processes 17 (2), 297–311.
Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York.
Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall–runoff-model? Water Resources Research 29 (8), 2637–2649.
Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research 42, W03S04.
Kron, W., 2006. Summer 2005 in Central Europe: Many Alpine valleys under water. Topics Geo 2005, Munich Reinsurance Company, Munich.
Kuczera, G., Mroczkowski, M., 1998. Assessment of hydrological parameter uncertainty and the worth of multiresponse data. Water Resources Research 34 (6), 1481–1489.
Lamb, R., 1999. Calibration of a conceptual rainfall–runoff model for flood frequency estimation by continuous simulation. Water Resources Research 35 (10), 3103–3114.
Legates, D.R., McCabe Jr., G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research 35, 233–241.
Madsen, H., 2000. Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. Journal of Hydrology 235, 276–288.
Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Advances in Water Resources 26, 205–216.
Madsen, H., Wilson, G., Ammentorp, H.C., 2002. Comparison of different automated strategies for calibration of rainfall–runoff models. Journal of Hydrology 261, 48–59.
Maniak, U., 2005. Hydrologie und wasserwirtschaft. Eine Einführung für Ingenieure. Springer, Berlin.
Mantovan, P., Todini, E., 2006. Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. Journal of Hydrology 330, 368–381.
Margot, A., Schädler, B., Sigg, R., Weingartner, R., 1992. Influence on rivers by water power stations (>300 kW) and the lake control. Hydrological Atlas of Switzerland, Plate 5.3. Federal Office for the Environment, Bern, CH.
McIntyre, N., Lee, H., Wheater, H.S., Young, A.R., Wagener, T., 2005. Ensemble predictions of runoff in ungauged catchments. Water Resources Research 41, W12434.
Merz, R., 2002. Understanding and Estimating Flood Probabilities at the Regional Scale. PhD Thesis, Technical University of Vienna. Wiener Mitteilungen Wasser Abwasser Gewässer 181, Vienna.
Merz, R., Blöschl, G., 2004. Regionalisation of catchment model parameters. Journal of Hydrology 287, 95–123.
MeteoSwiss (Federal Office for Meteorology and Climatology), 2008. Time Series of Meteorological Variables. Zürich.
Naden, P.S., 1992. Analysis and use of peaks-over-threshold data in flood estimation. In: Saul, A.J. (Ed.), Floods and Flood Management. Kluwer, Dordrecht, pp. 131–143.
Pappenberger, F., Beven, K.J., 2006. Ignorance is bliss: or seven reasons not to use uncertainty analysis. Water Resources Research 42, W05302.
Parajka, J., Merz, R., Blöschl, G., 2005. A comparison of regionalisation methods for catchment model parameters. Hydrology and Earth System Sciences 9, 157–171.
Perrin, C., Michel, C., Andréassian, V., 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. Journal of Hydrology 242, 275–301.
Ross, B.B., Contractor, D.N., Shanholtz, V.O., 1979. A finite element model of overland and channel flow for assessing the hydrologic impact of landuse change. Journal of Hydrology 41, 1–30.
Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrological Processes 21, 2075–2080.
Schoups, G., van de Giesen, N.C., Savenije, H.H.G., 2008. Model complexity control for hydrologic prediction. Water Resources Research 44, W00B03.
Schwanbeck, J., Viviroli, D., Weingartner, R., Frei, C., Schumann, A., 2008. Modellbasierte Sensitivitätsanalysen für das Berner Oberland. In: Bezzola, G.-R., Hegg, C., (Eds.), Ereignisanalyse Hochwasser 2005. Teil 2 – Analyse von Prozessen, Massnahmen und Gefahrengrundlagen. Umwelt-Wissen 08-25. Federal Office for the Environment, Bern, pp. 48–58. <http://www.bafu.admin.ch/publikationen/publikation/00100/index.html?lang=de>.
Schwarze, R., Droege, W., Opherden, K., 1999. Regional analysis and modelling of groundwater runoff components from catchments in hard rock areas. In: Diekkrüger, B., Kirkby, M.J., Schröder, U. (Eds.), Regionalisation in Hydrology. Proceedings of a conference held at Braunschweig. IAHS publication 254. International Association of Hydrological Sciences IAHS, Wallingford, UK, pp. 221–232.
Seibert, J., 1997. Estimation of parameter uncertainty in the HBV model. Nordic Hydrology 28, 247–262.
Seibert, J., McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in the catchment hydrology: use of soft data for multi-criteria model calibration. Water Resources Research 38 (11), 1241.
SFSO (Swiss Federal Statistical Office), 2003. GEOSTAT Database Products. Licence No. G158000315, ©SFSO. Neuchâtel, CH.
Sorooshian, S., Gupta, V.K., 1995. Model calibration. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, Colorado, pp. 23–68.
Viviroli, D., 2007. Ein prozessorientiertes Modellsystem zur Ermittlung seltener Hochwasserabflüsse für ungemessene Einzugsgebiete der Schweiz. PhD Thesis, Faculty of Science, University of Bern. Geographica Bernensia G77. Institute of Geography, University of Bern. ISBN: 978-3-905835-00-7.
Viviroli, D., Gurtz, J., Zappa, M., 2007. The Hydrological Modelling System PREVAH. Geographica Bernensia P40. Institute of Geography, University of Bern. ISBN: 978-3905835-01-0.
Viviroli, D., Mittelbach, H., Gurtz, J., Weingartner, R., 2009a. Continuous Simulation for Flood Estimation in Ungauged Mesoscale Catchments of Switzerland – Part II: Parameter Regionalisation and Results. Journal of Hydrology 377 (1-2), 208–225.
Viviroli, D., Zappa, M., Gurtz, J., Weingartner, R., 2009b. An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools. Environmental Modelling & Software 24 (10), 1209–1222.
Vrugt, J.A., Diks, C.G.H., Gupta, H.V., Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. Water Resources Research 41, W01017.
Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrology and Earth System Sciences 5 (1), 13–26.

Wagener, T., McIntyre, N., Lees, M.J., Wheater, H.S., Gupta, H.V., 2003. Towards reduced uncertainty in conceptual rainfall–runoff modelling: dynamic identifiability analysis. Hydrological Processes 17, 455–476.
Weingartner, R., Barben, M., Spreafico, M., 2003. Floods in mountain areas – an overview based on examples from Switzerland. Journal of Hydrology 282, 10–24.
Weingartner, R., Viviroli, D., Schädler, B., 2007. Water resources in mountain regions: A methodological approach to assess the water balance in a highland-lowland-system. Hydrological Process 21, 578–585.
Wilby, R.L., 1996. Contemporary Hydrology. Wiley, Chichester, UK.
Wüest, M., Frei, C., Altenhoff, A., Hagen, M., Litschi M., Schär, C. A gridded hourly precipitation dataset for Switzerland using rain-gauge analysis and radar-based disaggregation. International Journal of Climatology, submitted.
Zappa, M., Gurtz, J., 2003. Simulation of soil moisture and evapotranspiration in a soil profile during the 1999 MAP-Riviera Campaign. Hydrology and Earth System Sciences 7, 903–919.
Zappa, M., Kan, C., 2007. Extreme heat and runoff extremes in the Swiss Alps. Natural Hazards and Earth System Sciences 7, 375–389.
Zappa, M., Pos, F., Strasser, U., Warmerdam, P., Gurtz, J., 2003. Seasonal water balance of an Alpine catchment as evaluated by different methods for spatially distributed snowmelt modelling. Nordic Hydrology 34, 179–202.