Advances in Geosciences, 5, 83–87, 2005
SRef-ID: 1680-7359/adgeo/2005-5-83
European Geosciences Union

Advances in Geosciences

Investigation of the transferability of hydrological models and a method to improve model calibration

G. Hartmann and A. Bárdossy

IWS, Universität Stuttgart, Pfaffenwaldring 61, 70550 Stuttgart, Germany

Received: 7 January 2005 – Revised: 1 August 2005 – Accepted: 1 September 2005 – Published: 16 December 2005

Abstract. In order to find a model parameterization such that the hydrological model performs well even under different conditions, appropriate model performance measures have to be determined. A common performance measure is the Nash Sutcliffe efficiency. Usually it is calculated comparing observed and modelled daily values. In this paper a modified version is suggested in order to calibrate a model on different time scales simultaneously (days up to years). A spatially distributed hydrological model based on HBV concept was used. The modelling was applied on the Upper Neckar catchment, a mesoscale river in south western Germany with a basin size of about 4000 km^2. The observation period 1961–1990 was divided into four different climatic periods, referred to as "warm", "cold", "wet" and "dry". These sub periods were used to assess the transferability of the model calibration and of the measure of performance. In a first step, the hydrological model was calibrated on a certain period and afterwards applied on the same period. Then, a validation was performed on the climatologically opposite period than the calibration, e.g. the model calibrated on the cold period was applied on the warm period. Optimal parameter sets were identified by an automatic calibration procedure based on Simulated Annealing. The results show, that calibrating a hydrological model that is supposed to handle short as well as long term signals becomes an important task. Especially the objective function has to be chosen very carefully.

1 Introduction

The antagonism between what is available for hydrological modelling and what is really needed includes not only the spatial and temporal resolution of input variables, but also a need for statistically correct relations between these variables. Uncertainties within hydrological models can increase the variance of the output substantially. However, these kind of uncertainties are not the only consequence: uncertainties can even lead to biases, which are oftentimes not detected. Such models might work well for the situation they were calibrated for, with more or less stationary conditions. Yet, little is known about their reaction to changed circumstances, e.g. changes in climate or in land use.

If a model is to be used under non-stationary conditions, its parameters and process descriptions should be transferable. This means, the parameters should be identified in a way, that they give good results not only for the situation for which they were calibrated, but also for as many other situations as possible. This is illustrated in Fig. 1 where different model performances are given. Some of those models perform well for situation 1, but fail for situation 2, or vice versa, whereas transferable models and model parameterizations show consistent model performance for both situations.

Figure 2 gives an example for a theoretical case, where two different models with good performance are transferred to an unknown situation. For example the models might be used to calculate a land use change scenario or a climate change scenario. Although the parsimonious model B has a smaller range of possible output, this whole range might lie far from reality for the changed situation. On the other hand, there may be a model A with a broad range of results, but the observations are included within this range. Therefore, it is not the width of the uncertainty bounds for the changed situation that we should be concerned about but instead, the bias of a model.

The goal of this paper was to find a model calibration method and a corresponding measure that enables us to avoid biases and gives good results for different situations with different time scales.

In order to assess this transferability, a hydrological model was calibrated on different climatic periods and then validated on other climatic periods. Thus, different 10-year periods with different climatic conditions were compiled as follows. Mean annual temperature and total annual precipitation were calculated for the observation period 1961–1990.

Correspondence to: G. Hartmann
(hartmann@iws.uni-stuttgart.de)

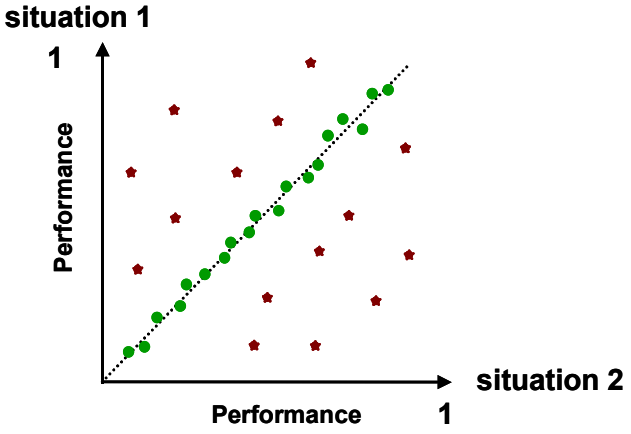

Fig. 1. Different model performances: some models give good results for situation 1 but bad results for situation 2 or vice versa (dark stars). Transferable models give similar results for both conditions (light dots).

Then, this period was subdivided into three sub-periods, first representing 10 warm, 10 normal, and 10 cold years, and, second, 10 wet, 10 normal, and 10 dry years.

Figure 3 explains the choice of the sub periods. The hydrological model was calibrated for one sub period in turn and validated on the others. The first step was to adapt the model to the same period it was calibrated to. Then the model was applied to other 10 years, e.g. the model calibrated on the cold years was examined for the warm years. Although the calibration was done only on the chosen years, the modeling itself was always performed for the entire observation period.

2 The objective function

A typical performance measure for a model is the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970) (NS) between observed and modelled daily values. However, if calibration is only performed on the daily scale, small systematic under or over-estimations will not be detected. Therefore, model performance was considered not only on daily values but also on aggregations of different time scales: In a first step, the mean value for aggregations for weeks, then for the aggregations for months, for all four seasons and for the entire year was calculated. For the aggregations up to one season (90 days), the performance increased steadily (see Fig. 4), which was expected, since averaging over a certain time means that small scale details are not considered anymore. However, all the aggregations smaller than the annual aggregation receive their quality partly from the annual cycle, which is not related to the quality of the model itself. The performance of the annual mean, however, cannot be improved by the annual cycle. Therefore, the performance of the annual aggregation is – although smaller than the previous performances – very important, because this performance is only due to the model quality.

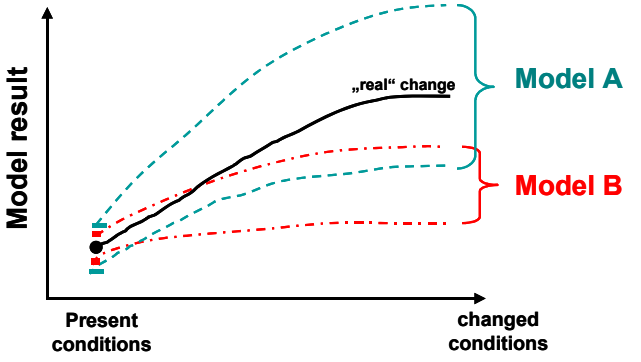

Fig. 2. Theoretical description of different model types showing similar results for the present situation but producing different results for the changed situation.

Finally, not only the NS between observed and modelled daily values, but also a weighted NS emphasizing extreme values and the NS between observed and modelled annual values were used. The different aggregation times are calculated as follows. Suppose $Q_O(t_i)$ is the observed discharge series and $Q_M(\theta, t_i)$ is the modeled series with model parameter θ for the time t_i. According to the selected time period P and whether extremes are considered or not, the weight for time t_i is defined as $w(t_i, P, x)$. Suppose the time step of the model is $t_i - t_{i-1} = \Delta t$, I is the total number of time steps and l is the summation index. Then, NS can be defined for time steps $j\Delta t$ as

$$NS(j, P, \theta, x) = 1 - \frac{\sum_{l=1}^{L} (Q_O^{(J)}(\tau_l) - Q_M^{(J)}(\theta, \tau_l))^2}{\sum_{l=1}^{L} (Q_O^{(J)}(\tau_l) - \overline{Q_O^{(J)}})^2} \quad (1)$$

with

$$Q_O^{(J)}(\tau_l) = \sum_{j=1}^{J} Q_O(\tau_l + j\Delta t) \cdot w \quad (2)$$

$$Q_M^{(J)}(\tau_l, \theta) = \sum_{j=1}^{J} Q_M(\theta, \tau_l + j\Delta t) \cdot w \quad (3)$$

where, in case extremes are not emphasized (x=1), only the chosen period is considered, or, in case extremes are emphasized (x=2), the extremes are multiplied with their square root. Thus, the extreme values compared to the other daily values become higher and are therefore more stressed:

$$w = \begin{cases} 0 & \text{if } i \notin P \\ 1 & \text{if } i \in P \text{ and } x = 1 \\ \sqrt{Q_O(t_i)} & \text{if } i \in P \text{ and } x = 2 \end{cases} \quad (4)$$

A linear combination of the NS-values on different time scales is used to measure the performance of the model and forms the overall objective function S for automatic calibration:

$$S(P, \theta) = \alpha_1 NS(1, P, \theta, 1) + \alpha_2 NS(1, P, \theta, 2) + \alpha_3 NS(365, P, \theta, 1) \quad (5)$$

This objective function reflects the fact that the model should perform reasonably well for a set of different time scales and

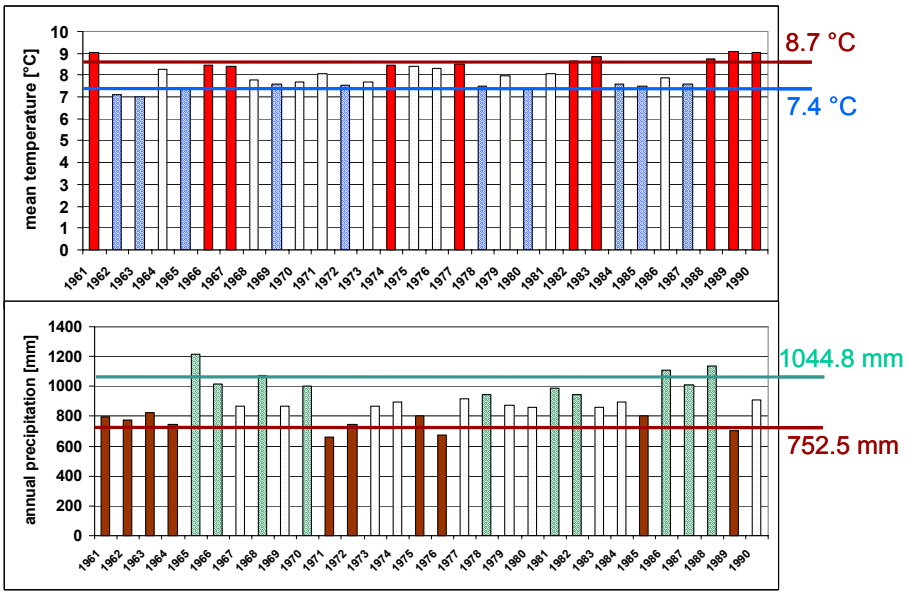

Fig. 3. Division of the observation period 1961 to 1990 into three sub-periods of, first, in terms of mean annual temperature 10 warm (solid bars), 10 normal, and 10 cold (dashed bars) years, and, second, in terms of annual precipitation 10 wet (dashed bars), 10 normal, and 10 dry (solid bars) years.

not only for the computational time step. The first part of the objective function considers the overall performance, the second part ensures the representation of the extremes, and the third part considers the interannual variability. Different optimization methods were set up, where the different parts of the objective function were weighted differently. The combination of time scales used for each optimization method is determined by different weights α, shown in Table 1. The calibration of the model was performed for different time intervals P – warm, cold, dry and wet years as specified above.

Table 1. Weights α used for different optimization methods.

Method-No.	1	2	3	4	5	6	7	8	9	10
Day (α_1)	1	1	1	1	1	1	1	1	1	1
"Extremes" (α_2)	1	1	1	1	1	0	0	1	1	0
Year (α_3)	1	1	1	1.5	2	0	1	1	0	1

3 The optimization algorithm

Since different model parameters θ can lead to similar performance (problem of equifinality, see Beven and Binley, 1992), the same objective function was used for multiple runs. A logical procedure had to be introduced to find the parameter values that optimize the numerical value of the objective function. For each optimization method, multiple simulations with the model are executed, each searching for an optimal parameter set. A parameter set that produces good results but is totally unrealistic has to be avoided. Therefore, certain constraints are necessary.

In this study, the optimal parameter sets were identified by an automatic calibration procedure based on Simulated Annealing (Aarts and Korst, 1989). With this procedure it is possible to include all kinds of known pre-conditions on model parameters. Here for example, close constraints on soil properties were applied according to the soil maps (e.g. the conceptual parameter "field capacity" was always kept higher than the wilting point). A certain range of possible values for each parameter was determined. The parameters were forced to stay within these ranges during the calibration.

4 The hydrological model and catchment details

A distributed hydrological model based on HBV (Bergström and Forsman, 1973) concept was used. It was applied on the Upper Neckar catchment, a mesoscale river in south western Germany with a basin size of about 4000 km^2. The catchment was divided into 13 subcatchments representing different land use and topographical conditions. Then each of the subcatchments was further divided into up to 6 zones, which represent different soil characteristics. The sizes of these zones range from 4 km^2 to 240 km^2. Runoff concentration was calculated on the subcatchment scale, the calculation of runoff formation was performed on the zones and was thus spatially more detailed. Daily discharge data from 13 gauging stations, as well as daily temperature data from 44 stations and precipitation data from 288 stations within and around the study area were obtained for the period 1961–1990. With such a dense observation network deficits in the model results should not be due to measuring errors.

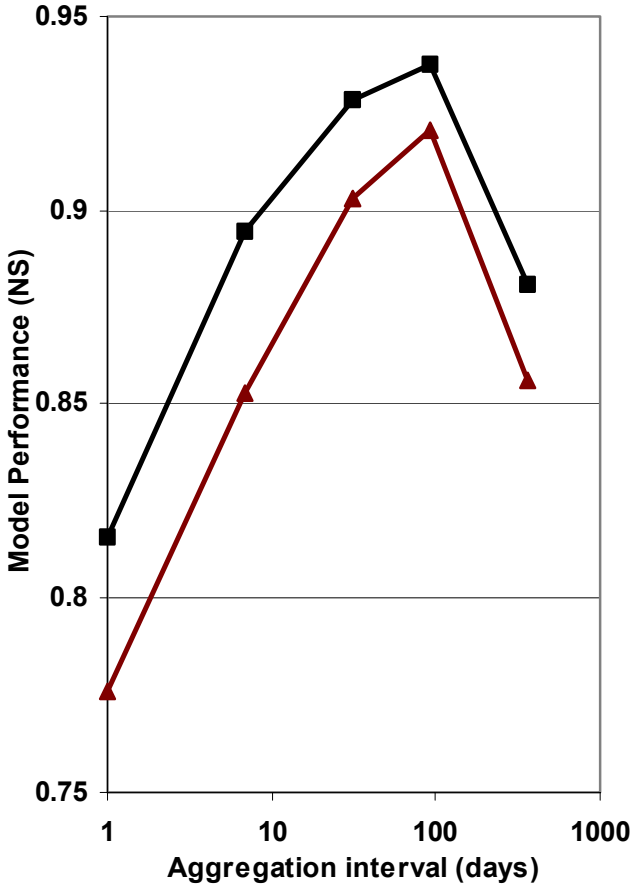

Fig. 4. Two examples for the increase and decrease of the model performance with increasing aggregation intervals.

5 Evaluation

For the evaluation of the results different approaches were used: the performance of all optimization methods on the annual cycles of the hydrographs was investigated on daily as well as on annual scales. The increases and decreases of water storage in all subcatchments were inspected to see if certain limits were exceeded. If this was the case the reasons were checked. Runoff duration curves were established and analysed for all subcatchments. During the course of the investigation it was found that comparing the results only on a daily scale is not sufficient and the use of the NS-efficiency on different time scales was included in the evaluation – like for the calibration of the model. Besides daily discharge, also aggregations of daily discharge for the weekly, monthly, seasonal and annual means were investigated. The NS-values for different calibration and validation periods are compared in Fig. 5. The difference between NS calibration and NS validation shows the loss (or gain) in model performance, when a model calibrated on an opposing climatological situation is used.

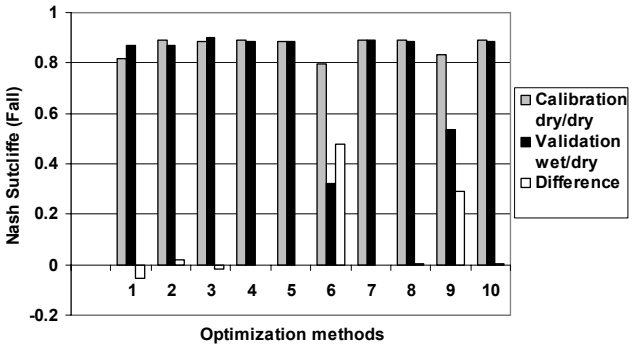

Fig. 5. Example for the different transferability of the different optimization methods. "wet/dry" = calibrated on wet periods, but applied on dry periods. The differences between calibration and validation for methods 6 and 9, which only use daily values for calibration (α_3=0 in Eq. 5) are much higher than for the other methods. Therefore, these methods are not as transferable as the others.

Table 2. Mean difference in NS between calibration and validation in terms of different calibration time scales. "warm/cold" = calibrated on warm periods, but validated on cold periods. Bold values indicate problematic cases.

	warm/cold	cold/warm	wet/dry	dry/wet
Day	<0.1	<0.1	<0.1	<0.1
Week	<0.1	<0.1	<0.1	<0.1
Month	<0.1	<0.1	<0.1	<0.1
Spring	**<=0.25**	<=0.12	<=0.15	<=0.12
Summer	<0.1	**<=0.54**	<=0.14	<0.1
Fall	**<=0.34**	<0.1	**<=0.27**	<=0.17
Winter	<=0.14	<=0.16	<0.1	**<=0.31**
Year	<0.1	**<0.3**	**<0.5**	**<=0.77**

6 Results

Optimization methods considering only daily values for calibration show severe problems. This was found for different evaluation approaches. For example subcatchments in karstic areas with problems due to ungauged transfer of water to areas outside the catchment counterbalance this by an increase in their water storage during the modeling. The highest increases were found using optimization methods in which the model is calibrated only on daily values.

Comparisons of the model performance on different time scales show that problems cannot be detected for short time periods (days, weeks and months) (see Table 2). Especially with problematic optimization methods such as those using only daily values for calibration, the mean differences in NS between calibration and validation for these short time periods are negligible.

For aggregated longer time periods, problems become obvious (bold values in Table 2). Figure 5 gives an example for the transferability of different optimization methods to different humid conditions. All optimization methods were

first calibrated as well as validated on the dry periods (so-called "calibration"). In the second step, they were calibrated on the wet periods, then validated on the dry periods (so-called "validation"). The example shows the calculated performances for calibration and validation as well as their differences. The optimization methods which were calibrated only on daily values (methods 6 and 9, see Table 1, α_3=0 in Eq. 5), clearly failed to follow the change in humidity. However, those methods which were not only calibrated on daily values, but also on annual values (α_3=1 in Eq. 5) still perform well.

The example shows the evaluation of the NS for the aggregation period "Fall". Although none of the 10 optimization methods uses the aggregation for "Fall" in their calibration, those which use the annual aggregation perform much better on the time period "Fall" than those, which only use daily values during their calibration.

The investigation of the annual cycle shows that if only daily values are compared, almost no differences can be detected between the different optimization methods. However, if annual values are compared, the results of different methods can be clearly distinguished as shown in Fig. 6. The optimization methods 6 and 9, which consider only daily values (α_3=0 in Eq. 5) have difficulties to handle a changed signal (here for example the results for dry years with the model calibrated on wet years), whereas those methods (e.g. method 8), which calibrate not only on daily values but also on annual values (α_3=1 in Eq. 5) still perform well.

7 Conclusions

The introduced calibration method allows a good performance on different time scales simultaneously. This is an important aspect, since one result of this study is, that calibrating only on short time aggregations (days, weeks, months) does not reveal all problems, such as small biases. Only an aggregation on longer periods allows the detection of such problems. If a hydrological model shall be transferable, the calibration should thus include aggregations for longer time periods than only daily values.

In general, automatic calibration methods are "blind" and might therefore lead to unreasonable parameter values and model performances. With the calibration method presented here it is possible to force the procedure to give reasonable results on different time scales at the same time already during the calibration process.

If models are calibrated manually, longterm balances (in general for more than one year) are used as a control – with the automated method described in this paper, for example one of several focuses can be set on the representation of the interannual variability by fixing one of the objective functions on the good representation of the annual values. This focus can be flexibly set to other time spans according to one's needs.

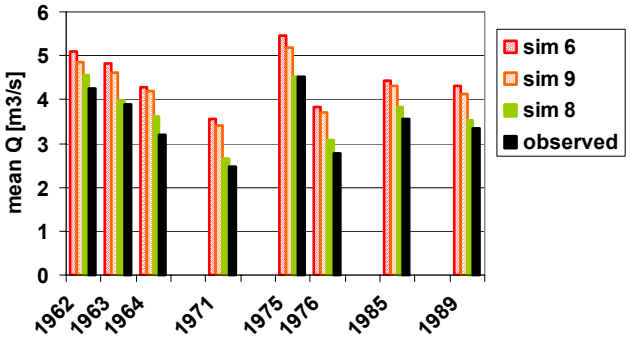

Fig. 6. Mean runoff for each year for the subcatchment Rottweil. The model was calibrated on the wet years. The performance of optimization method 8 (sim 8) for the dry years 1962–1964, 1971, 1975, 1976, 1985 and 1989 is better than with the methods 6 and 9 (sim 6 and sim 9). This means the model optimized with method 8 has a better transferability.

Edited by: P. Krause, K. Bongartz, and W.-A. Flügel
Reviewed by: anonymous referees

References

Aarts, E. and Korst, J.: Simulated Annealing and Boltzmann machines: a stochastic approach to combinatorical optimization and neural computing, John Wiley and Sons, Chichester, 284 pp., 1989.

Beven, K. J. and Binley, A. M.: The Future Of Distributed Models: Model Calibration And uncertainty prediction, Hydrol. Processes, 6, 279–298, 1992.

Bergström, S. and Forsman, A.: Development of a conceptual deterministic rainfall-runoff model, Nordic Hydrol., 4, 147–170, 1973.

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