Nordic Hydrology 34(5), 2003, 477-492

Reliability of Model Predictions Outside Calibration Conditions

Jan Seibert

Department of Environmental Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

Short title: Reliability of model predictions

Corresponding author:

Jan Seibert
Swedish University of Agricultural Sciences
Department of Environmental Assessment
Box 7050
S-750 07 Uppsala
Sweden
jan.seibert@ma.slu.se

Abstract

Predictions of probabilities and magnitudes of extreme events are essential for water management. One approach for flood estimation is the use of conceptual runoff models. This approach, however, can be questioned for the same reason as the approach of extreme-value statistics: the model has to be used for conditions far beyond those used for model development and calibration. In this study the HBV model, a conceptual runoff model, was applied to four different catchments and differential split-sample testing (calibration on years with lower runoff peaks and testing it on years with higher peaks) was used to evaluate model performance for the situation when the model has to be used to simulate runoff during conditions different from those observed during calibration. To assess the value of improved calibration different goodness-of-fit measures were used, which allowed to explicitly consider the ability of the model to simulate groundwater-levels and peak flows. The results indicated that applying a model to conditions different from those during the calibration period might not give accurate results and that improved calibration procedures might not automatically provide more accurate flood estimations.

Introduction

Predictions of probabilities and magnitudes of extreme events are essential for water management. The traditional approach of fitting distribution functions to the observed extreme values and extrapolating these functions can be criticised for different reasons (Linsley, 1986; Klemeš, 1986a, 2000a,b). The main criticism is that a distribution function of unknown shape has to be extrapolated far beyond the probabilities that can be justified from the available observations.

Alternatives to the distribution fitting are modelling approaches (*e.g.*, Bergström *et al.*, 1992; Calver and Lamb, 1995). The basic idea is to use a runoff model, which has been calibrated against existing streamflow data, to simulate the streamflow caused by extreme meteorological conditions. Obviously the need for data on extreme circumstances is only shifted from the streamflow to the meteorological data, but it might be more likely to have suitable data on extreme conditions for meteorological data than for streamflow. Firstly, more observations are often available, both in time and space, for meteorological data, such as precipitation or temperature, than for streamflow. Secondly the modelling approach allows combining extreme conditions (*e.g.,* a winter with much snow accumulation, a rapid rise of temperatures in spring and a large rainfall event) (Bergström *et al.*, 1992). The use of a model to estimate extreme runoff events has certainly advantages, but this approach can be criticised in exactly the same way as the fitting of some distribution function: for computation of extreme floods the model has to be applied for conditions far beyond the conditions used for development and calibration. The only reason why we should rely more on the model than on distribution functions is that we have confidence in the validity of the model and, thus, assume that extrapolation of the model calculations are more reliable. In order to have more confidence in a calibrated model than in a fitted distribution function we have to ensure that the model does not only correctly simulate runoff but also does so for the right reasons.

A usual test of a model is a simple split-sample test, where the model is calibrated on data from one period and tested for another, independent, period. This kind of test gives an indication how the model might perform for an independent period with similar conditions. Examples where the result of such a test is called ‘not successful’ are seldom found in literature. This may be mainly because this kind of validation is a simple task (Kirchner *et al.*, 1996; Mroczkowski *et al.*, 1997). Furthermore, this kind of test is not suited to test the models’ ability to give reasonable simulations for conditions that differ from those of the calibration period (Xu, 1999). However, models are most important for problems where we have to apply the model beyond the conditions observed before. The need to apply a model is, for instance, much larger for predicting a 1000-year flood than for predicting a 10-year flood. In the latter case enough data may be available to compute the flood from time series without any model.

In order to test how accurate model predictions might be when applying the model to different conditions a differential split-sample test is more suitable than the simple split-sample test (Klemeš, 1986b; Xu, 1999). The basic idea is to calibrate and to test the model on time periods with dissimilar hydrological conditions such as, for instance, a period with mainly small runoff events and a period with large events. Results of such a test may provide an indication of model performance when we have to extrapolate into unknown conditions. Although this is a more powerful test of a model, the use of this kind of test is by no means widespread. There are only a few noteworthy exceptions where models are tested using a differential split-sample test (*e.g.,* Refsgaard and Knudsen, 1996; Donelly-Makowecki and Moore, 1999).

The issue of parameter uncertainty (*e.g.,* Beven, 1993) has become widely recognized during recent years. Often parameter sets, which perform equally well (according to some criterion) for a calibration period, can be found at very different locations in the parameter space. It may be argued that the problem of identifying a unique parameter set is not an issue for practical model applications, *i.e.*, if different parameter sets were equally suitable to simulate runoff during a calibration period, any one of these sets may be applied. However as shown by, for instance, Seibert (1997) and Uhlenbrook *et al.* (1999), these ‘equally good’ parameter sets may give very different predictions for individual events. Uhlenbrook *et al.* (1999) computed design floods using ‘equally good’ parameter sets and found that the predicted peak discharge of a one-hundred-year flood varied from 40 to almost 60 mm d^{-1}.

The need for improved model calibration and testing has been emphasized in the recent years (de Grosbois *et al.*, 1988; Ambroise *et al.*, 1995; Refsgaard, 1997; Kuczera and Mroczkowski, 1998). Recent studies aiming at improving the calibration of runoff models can be classified into two groups: (1) making more use of the information contained in runoff series (*e.g.*, Boyle *et al.*, 2001) and (2) using additional data (*e.g.*, Franks *et al.*, 1998; Lamb *et al.*, 1998; Seibert, 2000). An implicit assumption is that the improved model calibration not only will reduce parameter uncertainty but also strengthen internal model consistency. It seems reasonable that this improved internal consistency could be associated with more reliable predictions outside the calibration domain. The idea is that, for a model that agrees with the real system in different respects (*e.g.*, with observed internal variables), extrapolation beyond the testable conditions is more reasonable than for a model that just matched runoff during some period. While this assertion might be reasonable, the effects of improved model calibration on prediction errors still have to be investigated.

In this study the HBV model (Bergström, 1995), a conceptual runoff model, which is used, among other purposes, to compute design floods for dam safety in Sweden (Bergström *et al.*, 1992; Lindström and Harlin, 1992), was applied to four different catchments where, in addition to precipitation, temperature and runoff data, also groundwater-level data were available. Differential split-sample testing was used to evaluate model performance for the situation when the model has to be used to simulate runoff during conditions different from those observed during calibration. The model was calibrated on years with lower runoff peaks and tested on years with higher peak flows. To assess the value of improved calibration the model performance was compared for simulations derived from including groundwater-level observations as well as an additional peak-flow criterion into the calibration.

Materials and Methods

HBV model

The HBV model (Bergström, 1976; 1992) is a conceptual model that simulates daily discharge using daily rainfall and temperature, and monthly estimates of potential evaporation as input. The model consists of different routines, where snowmelt is computed by a degree-day method, groundwater recharge and actual evaporation are functions of actual water storage in a soil box, runoff formation is represented by three linear reservoir equations and channel routing is simulated by a triangular weighting function. For both the snow and the soil routine, calculations are performed for each different elevation zone, but the response routine is a lumped representation of the catchment. Further descriptions of the model can be found elsewhere (*e.g.*, Bergström, 1992; 1995; Lindström *et al.*, 1997; Seibert, 1997; 1999).

Study Catchments

Four catchments in Sweden were chosen for this study (Fig. 1, Table 1). The catchments were all mainly forested and ranged from 6 to 18 km^2. On average the annual precipitation was 600-700 mm, and the annual runoff 250-300 mm. Runoff was measured using v-notch or rectangular weirs. Precipitation measurements were available for each catchment from stations within, or less than 5 km outside, the catchment whereas for temperature data stations up to about 30 km away from the respective catchment had to be used. Data series with about twice-monthly observations of groundwater levels were available from 4 to 10 wells for each catchment. The catchments were all mainly covered by till soil with the exception of the Tärnsjö catchment. At Tärnsjön a large esker (ridge of glaciofluvial deposits), rising up to 50 m above the surrounding land, runs through a part of the catchment. The remaining part of the catchment is covered by till soil. Previous studies indicated that the response function of the traditional HBV model might not be appropriate for the Tärnsjö catchment, and that an alternative response function may give better results (Bergström and Sandberg, 1983; Seibert, 2000). The recharge simulated by the soil routine is divided into two parts. A portion C_{PART} [-], related to the portion of the till soil area, is added directly to an linear storage whereas the remaining recharge generated on one day is added evenly distributed over a subsequent period of C_{DELAY} [d] days to another linear storage. The latter storage is thought to represent the esker in which recharge is delayed because of the large unsaturated zone (Seibert, 2000).

Model Application and Differential Split-Sample Test

For all four catchments, calibration and test periods were chosen so that the floods were significantly larger during the test period (Fig. 2). The maximum peak flows during the test period were 50 to 70 percent larger than the largest peak flow during the calibration period (Table 1). The differential split-sample test used in this study consisted of the following steps: (1) Monte Carlo model runs with randomly generated parameter sets for the calibration period, (2) selection of assemblies of the 50 best parameter sets according to three different goodness-of-fit measures (defined below), and (3) simulation of the test period with all parameter sets of the assembly. For the Monte Carlo runs, ranges of possible values were specified for each of the 12 free model parameters based on the range of calibrated values found in previous model applications (Bergström 1990; Seibert, 1999); these ranges were similar to those used by Seibert (1997). For each catchment 3 million parameter sets were drawn randomly using uniform distributions within these ranges. The model was run for each parameter set and the values of three different statistics were computed to evaluate model performance.

The general agreement between observed (Q_{obs}) and simulated (Q_{sim}) catchment runoff was evaluated by the model efficiency, R_{eff} (Eq. 1; Nash and Sutcliffe, 1970).

$$R_{eff} = 1 - \frac{\sum \left(Q_{obs} - Q_{sim} \right)^2}{\sum \left(Q_{obs} - \overline{Q_{obs}} \right)^2} \tag{1}$$

The model performance was also evaluated with regard to the ability to reproduce observed groundwater level variations. The HBV model simulates the groundwater lumped over the catchment and, thus, the local observations could not be compared to the simulations directly. Instead the groundwater observations were spatially averaged, *i.e.*, the arithmetic mean was computed from the observations at the different tubes. To allow comparison with the observed mean groundwater level the storage in the upper (S_{UZ}) and lower (S_{LZ}) groundwater box had to be transformed into a groundwater level, *z* [m a.s.l.]. A linear equation (Eq. 2) with a slope *m*, which corresponded to the inverse of the storage coefficient, and an offset *c* was used. The coefficients were determined by linear regression between the simulated storage and groundwater levels.

$$z = m \left(S_{UZ} + S_{LZ} \right) + c \tag{2}$$

The performance of the groundwater level simulation was evaluated using the coefficient of determination, r^2, as objective function. For the Tärnsjö catchment, where the alternative model structure was used, the wells were grouped according to whether they were located on the esker or not and mean time series were computed for both groups. The geometric mean of the r^2 values of the fit for the two series was computed as objective function.

The third statistic used to evaluate model performance focused on the simulation of peak flows. The model efficiency for runoff simulations, R_{eff}, tends to depend largely on the model fit for periods with high flow conditions. However, since the aim was to simulate extreme floods, another goodness-of-fit measure, which focused even more on high flow conditions, was used additionally. This measure, R_{peak} (Eq. 3), addressed the ability of the model to reproduce peak flows directly by using the absolute differences between observed and simulated peak streamflows ($Q_{peak,\ obs}$ and $Q_{peak,\ sim}$) for all *n* peaks during the simulation period. The set of peaks was determined from the observed runoff series, to be included a peak had to exceed the long-term mean runoff by three times. Furthermore, only the largest peak within any one-month window was used. The corresponding simulated peaks were taken as the largest runoff during a one-week window centred on the date of the observed peak. While a shorter window might have been sufficient for rain events in the relatively small catchments used in this study, the length of the window was chosen to allow a somewhat longer time shift between observed and simulated peaks, which might occur in the case of snowmelt events.

$$R_{peak} = 1 - \frac{\sum_{i=1}^{n} \left| Q_{peak,obs,i} - Q_{peak,sim,i} \right|}{\sum_{i=1}^{n} Q_{peak,obs,i}}$$

Three different assemblies of ‘best’ parameters sets were compiled, each consisting of the 50 parameter sets which performed best according to one goodness-of-fit measure. The three measures used were: model efficiency (R_{eff}), the mean of efficiency and goodness of groundwater level simulations (R_{eff} and r^2), as well as the mean of efficiency and goodness of peak flow simulations (R_{eff} and R_{peak}). Finally, these assemblies were used to simulate runoff for the test periods. In particular the ability to predict floods was tested based on the peak flows during the test period. Both median and the range of 80 percent of the predictions from the 50 parameter sets were computed.

Results

Different objective functions judge the model performance with regard to different aspects. If two criteria are highly correlated no new information is provided by the additional objective function. However, in the case of the three objective functions used in this study there was a 'trade-off' between the objective functions (Fig. 3). This means that the criteria provided different information, but also that it is not possible to find a solution that was optimal according to all criteria simultaneously.

In general good fits could be found for all catchments for the calibration period with efficiency values between 0.76 and 0.82 (Table 2). As in previous studies equally good calibration results could be obtained with very different parameter values. The model efficiencies were considerably lower when using the 'best' parameter sets to simulate the test period (Table 2). The drop of the efficiency values was most pronounced in the two catchments where the conditions were most different between calibration and test periods (Lilla Tivsjön and Tärnsjö).

For the smaller events during the test period, simulated peak flows were simulated more or less acceptably, whereas the peak flow predictions were significantly erroneous for several of the larger events, in particular for the largest events (Fig. 4). Peak flows that were larger than those observed during the calibration periods were systematically underestimated by the model for all four catchments. These results did not vary significantly for the different parameter-set assemblies, although the systematic underestimation was somewhat smaller when the peak-flow criterion was considered (Fig. 4, right column).

Discussion

The simulated peak flows deviated significantly from the observations. Parameter uncertainty caused considerably different predictions for the different peak flows, despite the fact that only the very best parameter sets were included in each assembly (50 best of 3 million runs). Even more important, also the median of the 50 predictions was erroneous in many cases, especially for the largest floods. Contrary to the results of Harlin (1992), who did not find a systematic underestimation of extreme floods, the largest floods were almost all underestimated for all four catchments. There was hardly any improvement when using additional criteria for optimisation; neither the groundwater data nor the extra peak-criterion had the effect one might have hoped for (Fig. 4). At least there was a small reduction of the bias of the peak-flow predictions when using the peak-flow criterion. Considering the groundwater-level simulations for the selection of the best parameter sets provided the highest efficiency values for the test period in the Lilla Tivsjön catchment. This was not the case for the other catchments, where also the drop in model efficiency was of similar size for the different calibration criteria (Table 2).

If calibrated on the test periods, runoff efficiency values were considerably higher than those obtained with the parameter sets determined based on the calibration period (Table 2). Also for R_{peak} significantly higher values could be obtained with calibration on the test period (median for all catchments 0.83 instead of 0.73). The fact that it was possible to obtain much better fits for the test periods indicates that the failure to predict the higher peak flows was not a problem of the model structure. The model was in principle capable to better reproduce also higher peak flows, but other parameter sets than those determined based on the calibration periods were needed.

Although not shown in this paper, it can be noted that the results were similar when other combinations of the objective functions, or using the volume error and the efficiency of the log-runoff values as additional criteria, were used to select the assembly of best parameter sets. Even using just R_{peak} as objective function only slightly reduced the systematic underestimation of large peak flows during the test period.

The idea behind carrying out a differential split-sample test was that the errors made by extrapolation from small and medium sized events to the largest events on record correspond to the errors when using all existing data and extrapolating to events larger than any event on record. It is difficult, if not impossible, to examine this assertion. The errors could be expected to become larger because a catchment might behave more differently for the most extreme events. On the other hand the errors could be supposed to become smaller because runoff during extreme events will approach some limit given by the climatic input data.

The catchments used in this study were all relatively small and, thus, the response times of the catchments might be smaller than the daily time step used for the simulations and the model evaluation. This might cause a degradation of model performance, but it is not obvious that this should affect larger events more than smaller events.

The poor predictions of the larger peak flows might also partly be explained by errors in the observed data for the larger events. The precipitation input can be very uncertain for extreme rainfall events because of spatial variations. Also the observed peak flows can be erroneous. One problem is that rating curves are usually derived from data that does not include the highest observed water stages. For the discharge stations used in this study, the highest runoff values during the test periods were about twice as high as the highest measured runoff used to calculate from the rating curve (SMHI, pers.com.). While the standard error (95% confidence interval) was below 7 percent for all gauging stations the rating curves up to the highest measured discharge, extrapolation of the stage-discharge relation might of course introduce additional errors (Jónsson *et al.*, 2002). At least for the stations used in this study extrapolation could be considered to be more reliable because the gauging stations were all weirs.

While errors in the observed data of the larger events might excuse the poor model predictions, the recognition of these potential errors does not mean that model predictions of, for instance, a one-hundred-year flood based on all existing data are more reliable than indicated by the results of this study. Contrary, if the largest observed events are affected by measurement errors, extrapolation of a model, which has been calibrated based on these data, will become even more uncertain.

Concluding Remarks

There was a significant range of predictions obtained using parameter sets that behaved equally well, according to some goodness-of-fit measure, during the calibration period. This prediction uncertainty caused by parameter uncertainty has been demonstrated before (*e.g.*, Seibert, 1997; Uhlenbrook *et al.*, 1999). Again the results strongly suggest to consider these uncertainties and to present model predictions rather as ranges than as single values.

The results of this study indicate that extrapolations of a model, *i.e.*, the simulation of conditions not observed during the calibration period, should be interpreted with care. This is of special concern when models are to be used to predict extreme events such as in the case of design-flood estimation. Furthermore, the results suggested that improved calibration procedures might not automatically provide more accurate flood estimations. The results presented in this paper are based on four small catchments and relatively short calibration and test periods. Results obviously might be different in other cases, but more research is motivated on the extrapolation of models. The systematic underestimation of the largest peak flows is of special concern since this would imply the possibility of a general underestimation of design floods.

It can be argued that more physically based models might have a greater potential to obtain predictions beyond the range of conditions during calibration. However, Refsgaard and Knudsen (1996) did not find any significant differences between a fully-distributed, physical model and a lumped, conceptual model with regard to model performance in a differential split-sample test. The assertion of the superiority of more physical models, thus, remains to be demonstrated. A differential split-sample test as used in this study provides a more powerful test on model capabilities than the usual split-sample test, because it allows testing the ‘risky’ predictions of a model rather than the ‘safe’ ones.

Acknowledgement

Runoff, precipitation and temperature data were measured by SMHI (Swedish Meteorological and Hydrological Institute). The groundwater level data were provided by Bo Thunholm, SGU (Swedish Geological Survey). Howard Wheater and an anonymous reviewer gave valuable comments on the manuscript. This research was funded by the Swedish Research Council (grant 620-20001065/2001).

References

Ambroise, B., Perrin, J.L., and Reutenauer, D. (1995): Multicriterion validation of a semidistributed conceptual model of the water cycle in the Fecht Catchment (Vosges Massif, France), Water Resources Research, 31: 1467-1481

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

Bergström, S. (1990): Parametervärden för HBV-modellen i Sverige, Erfarenheter från modelkalibreringar under perioden 1975-1989 (Parametervalues for the HBV model in Sweden, in Swedish), SMHI Hydrologi, No.28, Norrköping, 35 pp.

Bergström, S. (1992): The HBV model - its structure and applications. Report RH No.4, Swedish Meteorological and Hydrological Institute (SMHI), Hydrology, Norrköping, Sweden, 35 pp.

Bergström, S. (1995): The HBV model (Chapter 13, pp. 443-476), in: Singh, V.P. (ed.) Computer models of watershed hydrology, Water Resources Publications, Highlands Ranch, Colorado, U.S.A., 1130 pp.

Bergström, S., and Sandberg, G. (1983): Simulation of groundwater response by conceptual models - three case studies, Nordic Hydrology, 14: 85-92

Bergström, S., Harlin, J. and Lindström, G. (1992): Spillway design floods in Sweden: I. New guidelines. Hydrological Sciences Journal 37: 505-519

Beven, K.J. (1993): Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water Resources 16: 41-51

Boyle, D.P.; Gupta, H.V.; Sorooshian, S. ; Koren, V.; Zhang, Z. ; Smith, M. (2001) Toward improved streamflow forecasts: Value of semidistributed modeling, Water Resources Research, 37: 2749-2759

Calver, A. and Lamb R. (1995): Flood frequency estimation using continuous rainfall-runoff modeling, Physics and Chemistry of the Earth, 20(5-6): 479-483

de Groisbois, E., Hooper, R.P., and Christophersen, N. (1988): A multisignal automatic calibration methodology for hydrochemical models: a case study of the Birkenes model, Water Resources Research, 24: 1299-1307

Donelly-Makowecki, L.M. and Moore, R.D. (1999): Hierarchical testing of three rainfall-runoff models in small forested catchments, Journal of Hydrology, 219: 136-152.

Franks, S., Gineste, Ph., Beven, K.J. and Merot, Ph. (1998): On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas into the calibration process. Water Resources Research 34: 787-797.

Harlin, J. (1992): Modelling the hydrological response of extreme floods in Sweden. Nordic Hydrology, 23: 227-244.

Jónsson, P., Petersen-Øverleir, A., Nilsson, E., Edström, M., Iversen, H.L., Sirvio, H. (2002): Methodological and personal uncertainties in the establishment of rating curves, In: Killingtveit, Å. (Editor) *Nordic Hydrological Conference 2002. Volume 1.* Nordic Hydrological Programme, NHP-Report No. 47. XXII Nordic hydrological Conference, Røros, Nordic Association for Hydrology. pp. 35-44

Kirchner, J.W., Hooper, R.P., Kendall, C., Neal, C. and Leavesley, G. (1996): Testing and validating environmental models. The Science of the Total Environment 183: 33-47

Klemeš, V. (1986a): Dilettantism in hydrology: transition or destiny. Water Resources Research 22(9): 177S-188S

Klemeš, V. (1986b): Operational testing of hydrological simulation models. Hydrological Sciences Journal 31: 13-24

Klemeš, V. (2000a): Tall Tales about Tails of Hydrological Distributions. I, Journal of Hydrologic Engineering, 5(3): 227-231

Klemeš, V. (2000b): Tall Tales about Tails of Hydrological Distributions. II, Journal of Hydrologic Engineering, 5(3): 232-239

Kuczera, G. and Mroczkowski, M. (1998): Assessment of hydrological parameter uncertainty and the worth of multiresponse data. Water Resources Research, 34: 1481-1489.

Lamb, R., Beven, K.J. and Myrabø, S. (1998): Use of spatially distributed water table observations to constrain uncertainties in a rainfall-runoff model. Advances in Water Resources, 22(4): 305-317

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S. (1997): Development and test of the distributed HBV-96 hydrological model. Journal of Hydrology, 201: 272-288

Lindström, G., and Harlin, J. (1992): Spillway design floods in Sweden. II: Application and sensitivity analysis. Hydrological Sciences Journal, 37: 521 - 539.

Linsley, R.K. (1986): Flood estimates: how good are they?, Water Resources Research 22(9):159S-164S

Mroczkowski, M., Raper, G.P., and Kuczera, G. (1997): The quest for more powerful validation of conceptual catchment models. Water Resources Research, 33(10): 2325-2335

Nash, J.E., and Sutcliffe, J.V. (1970): River flow forecasting through conceptual models, part 1 - a discussion of principles, Journal of Hydrology, 10: 282-290

Refsgaard, J.C. (1997): Parameterisation, calibration and validation of distributed hydrological models. Journal of Hydrology, 198: 69-97.

Refsgaard, J.C., and Knudsen, J. (1996): Operational validation and intercomparison of different types of hydrological models, Water Resources Research, 32: 2189-2202.

Seibert, J. (1997): Estimation of parameter uncertainty in the HBV model. Nordic Hydrology, 28(4/5), 247-262

Seibert, J. (1999):, Regionalisation of parameters for a conceptual rainfall-runoff model, Agricultural and Forest Meteorology 98-99: 279-293

Seibert, J. (2000): Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. Hydrology and Earth System Sciences, 4: 215-224.

Uhlenbrook, S., Seibert, J., Leibundgut, Ch. and Rodhe, A. (1999): Prediction uncertainty of conceptual rainfall-runoff models caused by problems to identify model parameters and structure, Hydrolgical Sciences - Journal des Sciences Hydrologiques 44(5): 779-798

Xu, C-Y (1999): Operational testing of a water balance model for predicting climate change impacts, Agricultural and Forest Meteorology, 98-99 (1-4), 295-304.

Figures

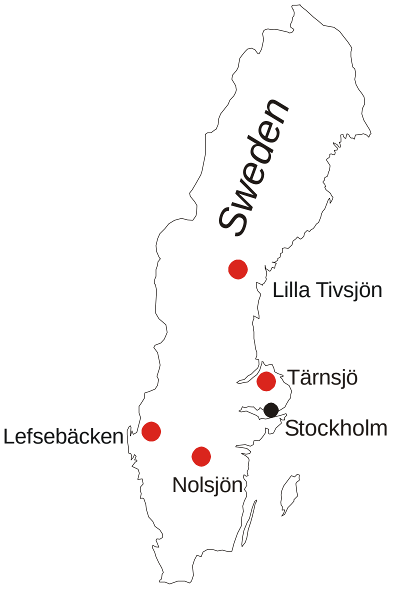
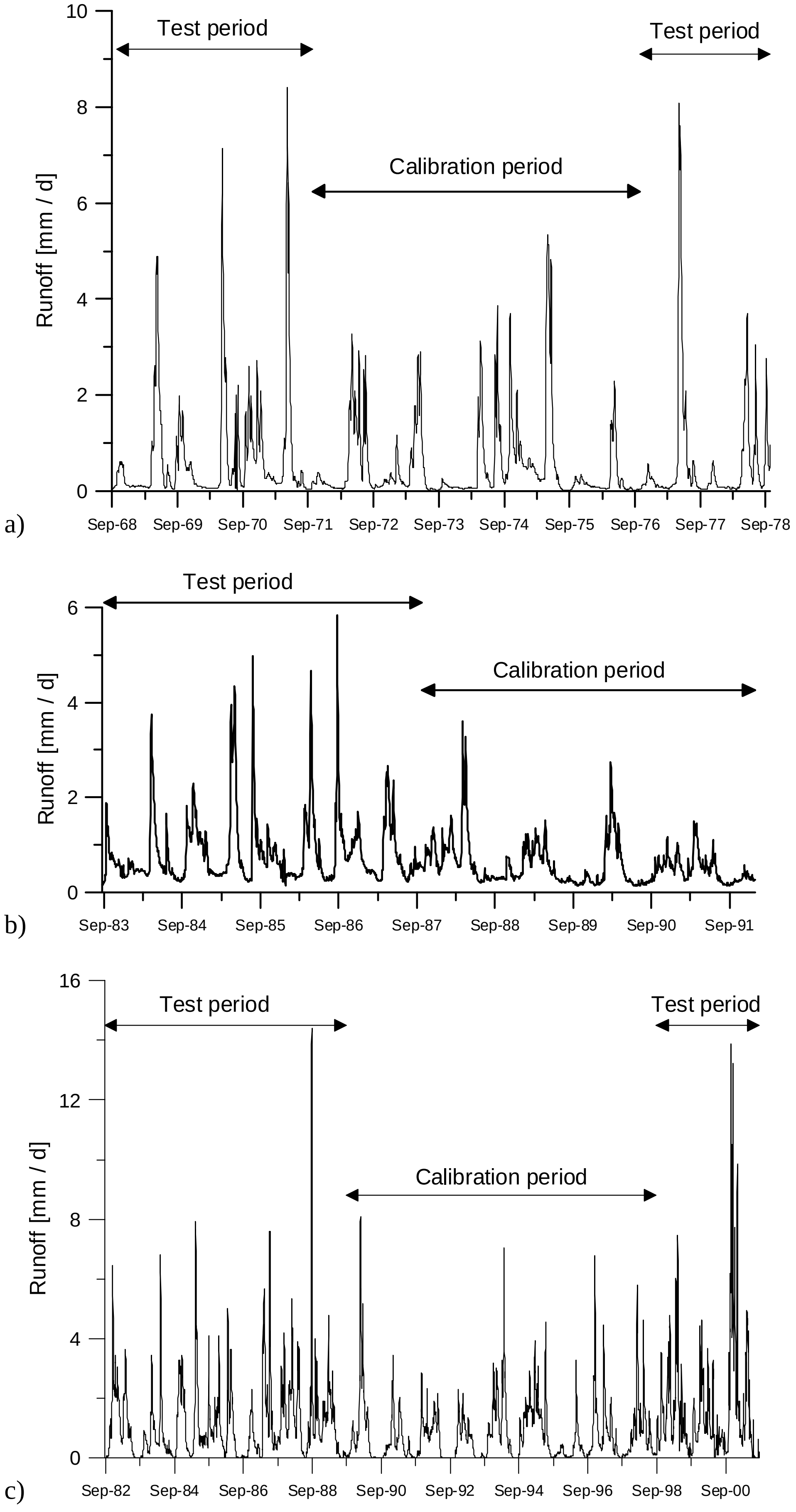


Figure 1. Location of study catchments


Test period
Test period
Calibration period
Runoff [mm / d]
10
8
6
4
2
0
a)
Sep-68
Sep-69
Sep-70
Sep-71
Sep-72
Sep-73
Sep-74
Sep-75
Sep-76
Sep-77
Sep-78
Test period
Calibration period
Runoff [mm / d]
6
4
2
0
b)
Sep-83
Sep-84
Sep-85
Sep-86
Sep-87
Sep-88
Sep-89
Sep-90
Sep-91
Test period
Test period
Calibration period
Runoff [mm / d]
16
12
8
4
0
c)
Sep-82
Sep-84
Sep-86
Sep-88
Sep-90
Sep-92
Sep-94
Sep-96
Sep-98
Sep-00

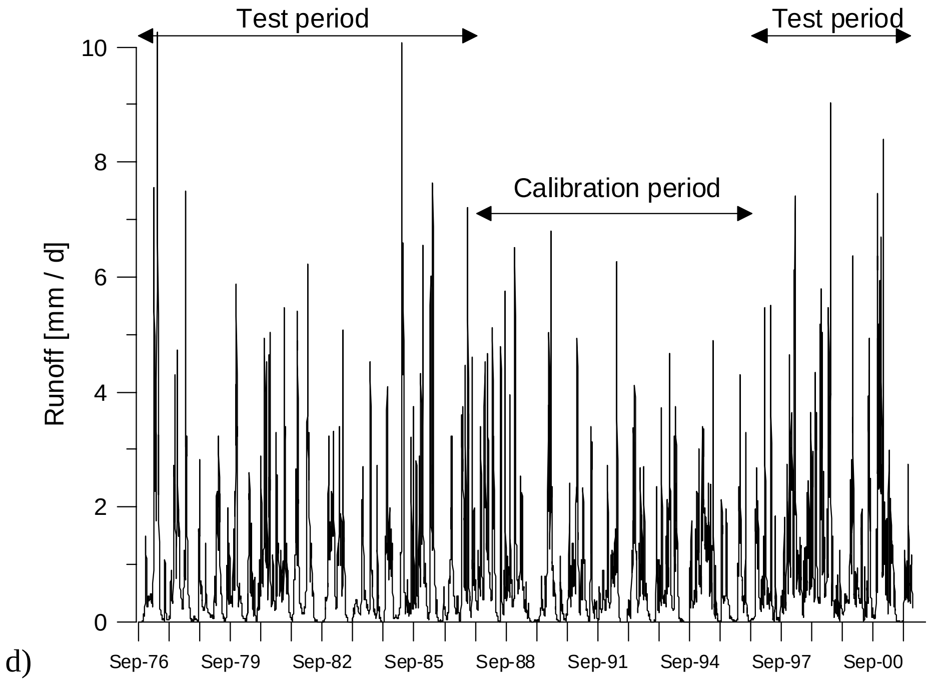


Figure 2. Runoff series for calibration and test periods in the catchments Lilla Tivsjön (a), Tärnsjö (b), Lefsebäcken (c) and Nolsjön (d)

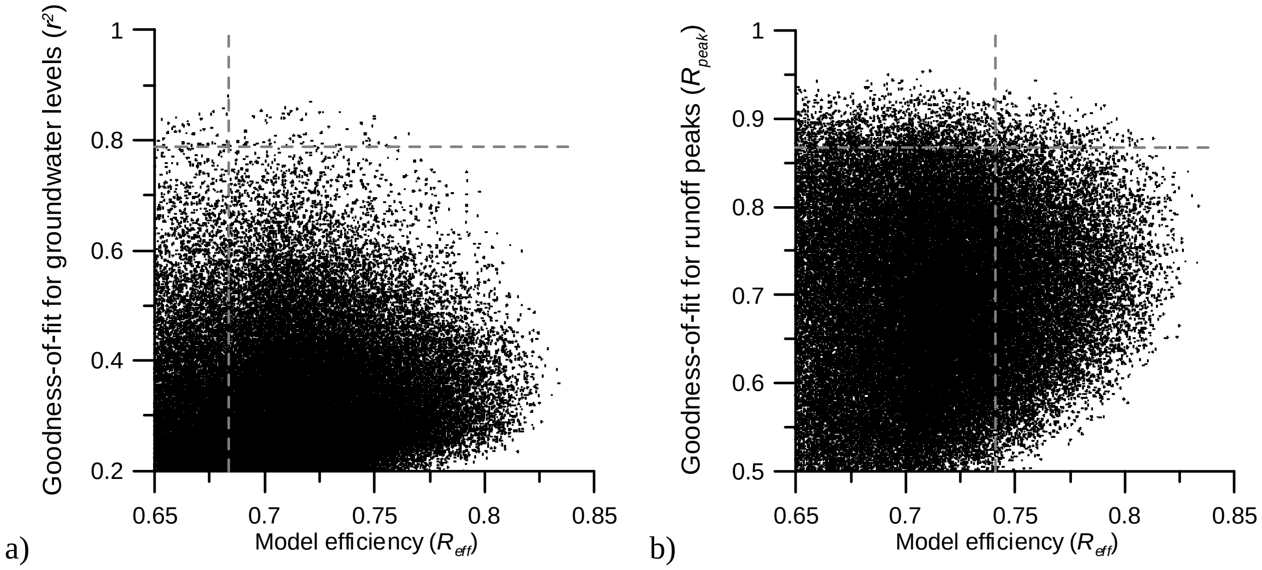


Figure 3. Model performance for the Monte Carlo runs according to the different criteria for the Lilla Tivsjön catchment. Each dot represents one model run with a randomly generated parameter set. The dashed lines indicate the thresholds for the 50 best parameter sets according to the combination of (a) runoff efficiency and goodness of groundwater simulations and (R_{eff} and r^2), as well as (b) efficiency and goodness of peak flow simulations (R_{eff} and R_{peak}).

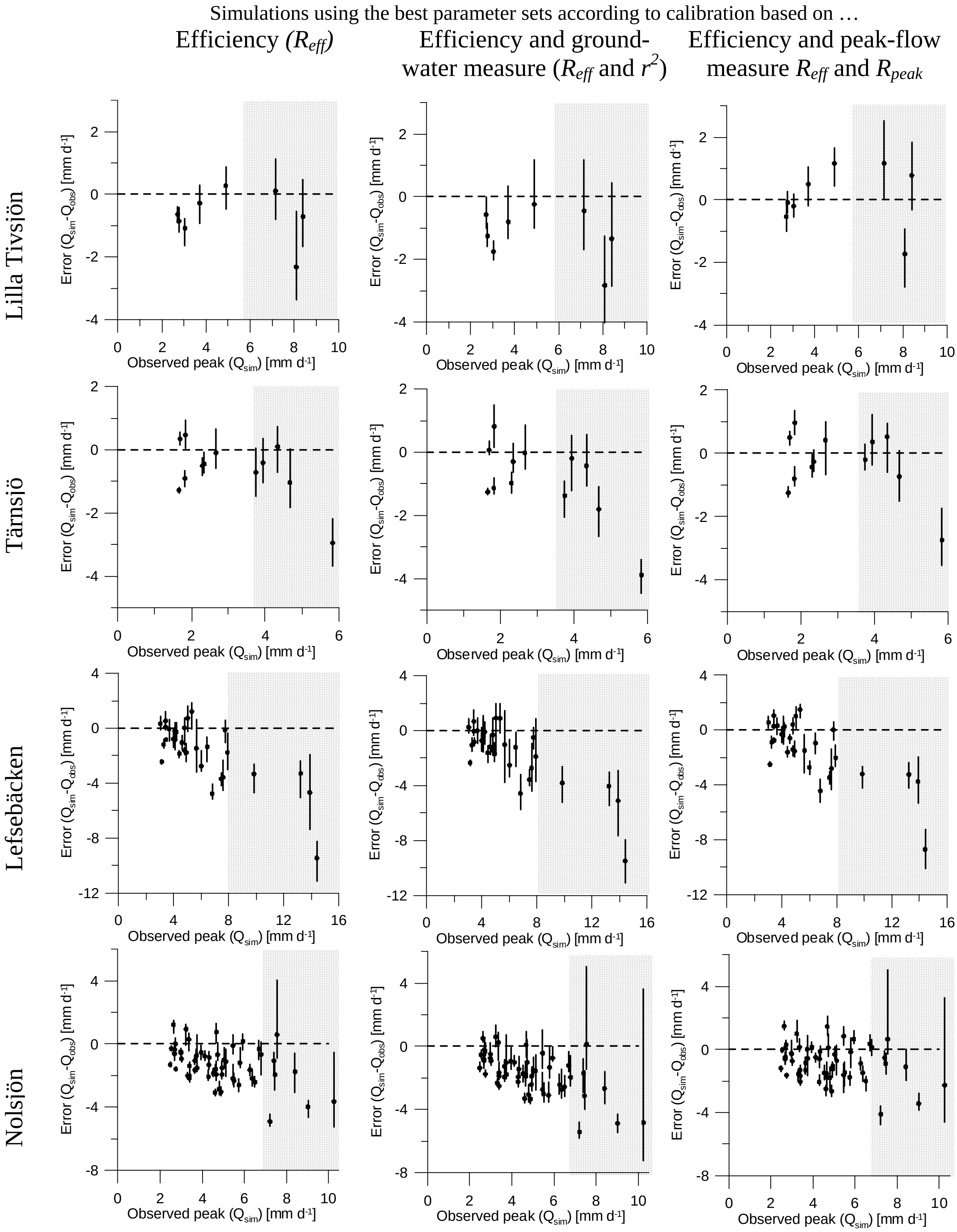


Figure 4. Errors of the simulated peak flows for peaks of different magnitudes during the test period for the four catchments. Both the median and the 80-percent range of the prediction errors obtained using the 50 best parameter sets are shown. The assemblies of best parameter set were determined with regard to the model efficiency (*R_{eff}*, left column), the mean of efficiency and goodness of groundwater level simulations (*R_{eff}* and *r^2*, middle), as well as the mean of efficiency and goodness of peak flow simulations (*R_{eff}* and *R_{peak}*, right). The shaded area indicates where the model was extrapolated, *i.e.*, events that were larger than those observed during the calibration period.

Tables

Table 1. Catchments characteristics

Characteristic	Lilla Tivsjön	Tärnsjö	Lefsebäcken	Nolsjön
SMHI[1] station number	42-1920	54-2299	108-1815	67-1912
Area [km^2]	12.8	14	5.2	18.2
Lake percentage [%]	2.7	1.8	5.4	1.5
Maximum flow during calibration period [mm d^{-1}]	5.3	3.6	8.1	6.8
Maximum flow during test period [mm d^{-1}]	8.4	5.3	14.4	10.3

[1] Swedish Meteorological and Hydrological Institute

Table 2. Model efficiency (R_{eff}) for the calibration and test period using assemblies of the parameter sets which performed best during the calibration period according to the three different objective functions (medians of 50 simulations, CP=calibration period, TP=test period). The efficiency values that could be achieved with calibration on the test period are given for comparison.

Best parameter sets	Lilla Tivsjön		Tärnsjö		Lefsebäcken		Nolsjön	
according to….	CP	TP	CP	TP	CP	TP	CP	TP
R_{eff}	0.82	0.36	0.79	0.65	0.77	0.75	0.76	0.72
$(R_{eff}+r^2)/2$	0.73	0.40	0.69	0.54	0.74	0.73	0.67	0.61
$(R_{eff}+R_{peak})/2$	0.78	0.21	0.75	0.64	0.76	0.75	0.73	0.71
Calibration on test period		0.92		0.78		0.81		0.85