Hydrol. Earth Syst. Sci., 13, 883–892, 2009
www.hydrol-earth-syst-sci.net/13/883/2009/

Hydrology and Earth System Sciences

Gauging the ungauged basin: how many discharge measurements are needed?

J. Seibert[1,2] **and K. J. Beven**[3,4]

[1]University of Zurich, Zurich, Switzerland
[2]Stockholm University, Stockholm, Sweden
[3]Lancaster University, Lancaster, LA1 4YQ, UK
[4]Uppsala University, Uppsala, Sweden

Received: 26 February 2009 – Published in Hydrol. Earth Syst. Sci. Discuss.: 12 March 2009
Revised: 26 May 2009 – Accepted: 8 June 2009 – Published: 22 June 2009

Abstract. Runoff estimation in ungauged catchments is probably one of the most basic and oldest tasks of hydrologists. This long-standing issue has received increased attention recently due to the PUB (Prediction in Ungauged Basins) initiative. Given the challenges of predicting runoff for ungauged catchments one might argue that the best course of action is to take a few runoff measurements. In this study we explored how implementing such a procedure might support predictions in an ungauged basin. We used a number of monitored Swedish catchments as hypothetical ungauged basins where we pretended to start with no runoff data and then added different sub-sets of the available data to constrain a simple catchment model. These sub-sets consisted of a limited number of single runoff measurements; in other words these data represent what could be measured with limited efforts in an ungauged basin. We used a Monte Carlo approach and predicted runoff as a weighted ensemble mean of simulations using acceptable parameter sets. We found that the ensemble prediction clearly outperformed the predictions using single parameter sets and that surprisingly little runoff data was necessary to identify model parameterizations that provided good results for the "ungauged" test periods. These results indicated that a few runoff measurements can contain much of the information content of continuous runoff time series. However, the study also indicated that results may differ significantly between catchments and also depend on the days chosen for taking the measurements.

Correspondence to: J. Seibert
(jan.seibert@geo.uzh.ch)

1 Introduction

1.1 Prediction of Ungauged Basins (PUB)

The PUB initiative of the International Association of Hydrological Scientists is a 10 year project seeking to improve the prediction of catchment responses in ungauged basins by improving the scientific basis of hydrology (Sivapalan et al., 2003b). As recognised in the PUB Science Plan (Sivapalan et al., 2003a), this is essentially an exercise in the constraint of uncertainty since any approach to extrapolate process or parameter information from gauged sites to ungauged sites must inevitably result in uncertainty in the representation of the responses of the ungauged site of interest. Raising the "awareness for the value of data, especially the gauging of hydrologic variables" has therefore been listed as one of the objectives of PUB (Sivapalan et al., 2003a).

Beven (2007) has suggested that this can be treated as a problem of learning about places in the context of "models of everywhere". In the near future it will be possible to have hydrology and water quality models for large catchments (or all the catchments in a country) that can be used for a variety of decision making processes. In the first instance, nearly all the places being represented in such models will be "ungauged" and subject to (perhaps considerable) uncertainty. Since such uncertainty can be a serious issue for risk-assessment decision making, it will be important to limit uncertainty by taking additional measurements. In this way, the appropriate ways of representing different places will gradually improve, with scope for using different model structures in different places, and different (uncertain) parameter sets in apparently similar places, as part of the learning strategy.

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

This raises the question of how many measurements might be necessary to achieve a desired and cost-effective reduction in uncertainty. Unfortunately, there is little guidance in the hydrological modelling literature about the worth of measurements in model identification, except for some suggestions about how long a discharge record is necessary to obtain an optimal model calibration (of the order of several years of data) (Sorooshian et al., 1983; Yapo et al., 1996). If these suggestions are correct then it is unlikely that funding for long term data collection in an ungauged catchment would be made available, except perhaps for high capital expenditure projects such as dams. An alternative approach is to investigate how far one comes with a limited number of observations. Obviously, this will not provide as good fits as a "full" calibration using several years of data, but it might be a pratical approach when one has to make predictions for an ungauged catchment.

Perrin et al. (2007) investigated how many streamflow observations are needed to obtain calibrations similar to those of a full calibration. They found that calibrating a simple runoff model using about 100–350 observation days spread randomly over a longer time period provided robust parameter estimates and that results hardly improved when including more days for calibration. Using a similar approach it was also shown that good parameter estimates could be obtained by combining the information of a few observation days with using prior knowledge in the form of regionalised parameter estimates (Rojas-Serna et al., 2006).

The value of a limited number of observations might further be increased if the observations are scheduled in a clever way. McIntyre and Wheater (2004) tested the effect of using different subsets of data for the calibration of an in-river phosphorus model. They found that a relatively small amount of all data was necessary to obtain good results as long as these data were taken in an event-based way rather than at fixed intervals. Juston et al. (2009) demonstrate that, if selected in an intelligent way, a small fraction of data points in a longer time series might contain almost all information of the entire data series. Eng and Milly (2007) found that "a single pair of strategically timed streamflow measurements" considerably improved the estimation of base flow recession coefficients compared to estimates based on catchment area. Rode et al. (2007) applied a water quality model to the Elbe River and found that a subset of the entire calibration data provided good results. Binley and Beven (2003) also showed that a single set of geophysical measurements of a deep soil water profile contained most of the information from 18 months of weekly measurements in conditioning a model of groundwater recharge. This is an indication of the potential value of limited observation data for constraining model prediction uncertainties even for ungauged basins. More generally, during such a learning process, we would seek to constrain prediction uncertainties of models with prior estimates of model parameters based on much more limited or more readily available data sets, such as national databases of model parameter values (e.g., Yadav et al., 2007; Zhang et al., 2008).

1.2 Gauging the ungauged basin

One option for making predictions at an ungauged site is to take discharge measurements there. The costs of such measurements will decrease as the new generation of floating acoustic doppler velocity measurement devices becomes available. Even a small number of discharge estimates using current metering methods may be feasible for some applications if it can be shown that the measurements have value in reducing uncertainties for decision making. In this paper we explore the effects of different numbers of site specific discharge measurements taken within a one year period on the quality of the predictions using the HBV runoff model for 11 catchments in one region of Sweden treated as if they were ungauged. We used a Monte Carlo approach and predicted runoff as a weighted ensemble mean of simulations using acceptable parameter sets. This approach is similar to that suggested by McIntyre et al. (2005), who computed the runoff for ungauged basins as an ensemble mean of acceptable parameter sets derived for a number of different gauged catchments, for which each 10 000 different parameter sets were evaluated. For our study, however, we used a small portion of the data from the "ungauged" catchments to select acceptable parameter sets instead of using information from other catchments.

2 Methods

2.1 The test catchments

This study was based on the eleven catchments from Seibert (1999). The catchments are located in central Sweden north of Uppsala. Elevation differences are generally small (about 100 m) and coniferous forest is the prevailing land use (Table 1). Seibert (1999) found relationships between calibrated values of HBV model parameters and three catchment characteristics: catchment area and the percentages of lake and forest. It should be noted that in this region the land cover can be used as a proxy for the distribution of different soil types. In general the areas with till soils are forested while clay soils underlie agricultural lands. Based on data from a total of 17 stations, the areal, corrected precipitation for each catchment was calculated by Seibert (1994) using the Thiessen polygon method with correction factors given by Eriksson (1983). Temperature data were interpolated from four measurement stations. The monthly long-term mean potential evaporation was taken from Eriksson (1981). The simulation period of September 1981 to August 1990 was preceded by a warming-up period of eight months. For all 11 catchments good simulations can be achieved (Seibert, 1999); calibration to the entire 10-year series resulted in model efficiency (Nash and Sutcliffe, 1970) values of about

Table 1. Characteristics of the study catchments.

River	Station	Abbrev.	Area (km^2)	Forest (%)	Field or meadow (%)	Lake (%)
Lillån	Gränvad	GR	168	41.0	59.0	0
Örsundaån	Härnevi	HA	305	55.0	44.0	1.0
Hågaån	Lurbo	LU	124	77.7	27.0	0.3
Sävaån	Ransta	RA	198	66.1	33.0	0.9
Sävjaån	Sävja	SA	727	64.0	34.0	2.0
Sagån	Sörsätra	SO	612	61.0	37.9	1.1
Stabbybäckenn	Stabby	ST	6.6	87.0	13.0	0
Stalbobäcken	Tärnsjö	TA	14	84.5	14.0	1.5
Fyrisån	Ulva Kvarn	UL	950	61.0	36.0	3.0
Vatthomaån	Vattholma 2	VA	284	71.0	24.2	4.8
Svartån	Åkesta Kvarn	AK	730	69.0	27.0	4.0

0.8 for most catchments, with the exception of three catchments with values above 0.85 (AK, SA and VA) and the two smallest catchments (ST, TA) with values around 0.7.

2.2 The HBV model

The HBV model (Bergström, 1992; Lindström et al., 1997) is a conceptual model of catchment hydrology that simulates daily discharge based on time series of daily rainfall and air temperature as well as monthly estimates of potential evaporation. The long-term mean evaporation rates are modified based on deviations of the actual daily temperature from the long-term mean temperature for the respective month. Different routines are used to represent the major components of catchment hydrology: a snow routine where snow accumulation and snow melt is computed by a degree-day method, a soil routine where groundwater recharge and actual evaporation are simulated as functions of actual water storage, a response routine with three linear reservoir equations, and a routing routine using a triangular weighting function. More detailed descriptions of the model can be found elsewhere (Bergström, 1992, 1995; Harlin and Kung, 1992; Seibert, 1999).

In this study the "HBV light" version (Seibert, 1997, 1999) was used that corresponds to the HBV-6 version described by Bergström (1992) with two slight changes. Instead of starting the simulation with some user-defined initial state values, HBV light uses a warming-up period during which state variables evolve from standard initial values to their appropriate values according to meteorological conditions and parameter values. Furthermore, the restriction that only integer values are allowed for the routing parameter MAXBAS had been removed. In this study the HBV model was applied using only one land use and one elevation zone. Furthermore, to reduce the number of parameters in the response routine, the upper outflow from the upper groundwater box was excluded (i.e., the parameters UZL and K_0 were not used and the response routine, thus, consisted of two linear boxes).

2.3 Modelling approach

In order to test the value of a limited number of stream gaugings taken during one year we posed three questions:

1. Can a limited number of stream gaugings help to distinguish between "good" and "poor" parameter sets?
2. How does the performance of runoff simulated by a weighted mean of the 100 "best" parameter sets depend on the number of streamflow gaugings?
3. Can we identify good strategies to select a certain number of measurement dates within one year?

These questions were addressed within a GLUE-type model conditioning framework (e.g., Beven and Binley, 1992; Beven et al., 2008; Beven, 2009) assuming that a time series of model input data were available to drive the model but that only a small number of discharge observations were available to evaluate model performance. We started by generating 10 000 random parameter sets sampled from uniform distributions, for which upper and lower limits were specified. We used the same limits as in previous studies (Seibert, 1997, 1999), which are defined based on model applications in different parts of Sweden (Bergström, 1990). By using a uniform distribution it is assumed that there is no prior information and that all values within the possible prior ranges are equally possible. Different subsets of the entire 10-year series of runoff data were used to evaluate these model simulations. With these subsets it was assumed that runoff had been observed only at a certain number of days during a one year period.

For each hydrological year during the 10-year period, subsets were generated by randomly selecting 1, 2, 4, 8, 16, 32, 64, 128, or 256 observation days. These random selections of a certain number of observation days were each repeated 100 times. For each of these subsets the following steps were taken to produce constrained runoff simulation:

1. The model performance was evaluated based on the subset using the sum of squared errors (SSE) as objective function. According to their performance the 10 000 parameter sets were ranked and the best 100 sets were selected for further analysis.

2. The performance of each of the 100 parameter sets was evaluated based on the entire 10-year period using the model efficiency, R_{eff}, as objective function and the median model efficiency was computed.

3. Based on the ensemble of 100 runoff simulations one series of simulated runoff was computed as weighted ensemble mean, $Q_{\text{ensemblemean}}(t) = \sum_{i=1}^{100} w_i \; Q_i(t)$. The weights w_i were taken from a linear decreasing function so that the "best" parameter set received a weight of 0.02 and the 100th parameter set received a weight of zero.

4. The time series computed in this way were evaluated based on the entire 10-year period using the model efficiency, R_{eff}, as objective function.

Please note that the same ranking would have been obtained for using the model efficiency rather than the SSE in step 1 with the exception that the model efficiency can not be calculated for cases where there is only one observation. For comparison we also performed this analysis for the case of zero observations days; in this case 100 of the 10 000 parameter sets were randomly selected and ranked. We assumed a random selection of gauging dates to address the first two questions. Additionally we compared a number of strategies to select 6 gauging dates within one year. The number of 6 dates was chosen based on initial information indicating that good results might be obtained with this number of observations. With fewer observations there would be too little information regardless of strategy and with an increasing number of observations the strategy becomes less important. In this study we tested the following strategies:

- 6 days with the bimonthly maxima (MAX6)
- 6 days with bimonthly minima (MIN6)
- 6 days with bimonthly mean flows (here the day with the flow least different from the mean flow was selected; in case of several days with a zero difference the first one was used) (MEAN6)
- The day with the annual maximum flow and 5 days along the recession (10, 20, 30, 40, and 50 days after peak flow) (MAX1REC5)
- The 2 days with maximum flow in spring and fall and for both cases 2 days during the recession (10 and 20 days after peak flow) (MAX2REC4)

As two benchmark strategies we also selected 6 days evenly distributed over the year (the first and 15th days of each two-month period) (BENCH1 and BENCH15).

Similarly to the randomly selected dates, we tested each strategy for the 10 different hydrological years for each catchment, i.e., the simulations were evaluated based on runoff at 6 selected days during one year and an ensemble of good parameter sets was compiled. The weighted ensemble mean was then evaluated based on its fit with the observed runoff during the entire 10-year period.

The issue of formal statistical and informal methods for model calibration has been receiving considerable attention in the hydrological literature (e.g., Beven, 2006; Mantovan and Todini, 2006; Beven et al., 2008; Smith et al., 2008). Here, using daily data from small catchments there is an expectation that rainfall input errors and timing, and model structural errors may make it difficult to formulate a formal statistical model of the residual errors. We have therefore chosen to examine the issue of the value of data within a more traditional, efficiency based, calibration framework in this initial study, but intend to explore some of these issues in future work.

3 Results

First we analyzed the ability of a limited number of runoff observations to select parameter sets that performed better than others for the entire 10-year period. The distribution of model efficiency values for individual parameter sets selected as the top one percent of parameter sets clearly moved towards better model performances with an increasing number of runoff observations. The median model performance increase was highest when the number of runoff observations increased by 2 to 16 observations while a plateau was reached at about 32 observations when additional observations did not greatly improve the average model performance (Fig. 1).

For the further analyses, weighted ensemble means were used to predict runoff for the entire 10-year period instead of simulations using individual parameter sets. The results clearly showed that the ensemble mean outperformed the individual simulations. Most convincing was that the ensemble mean was better than the prediction using only the one best parameter set in almost all cases (Fig. 2).

As could be expected from the shift of the distribution of model performance of individual parameter sets (Fig. 1), the performance of the ensemble mean also increased with an

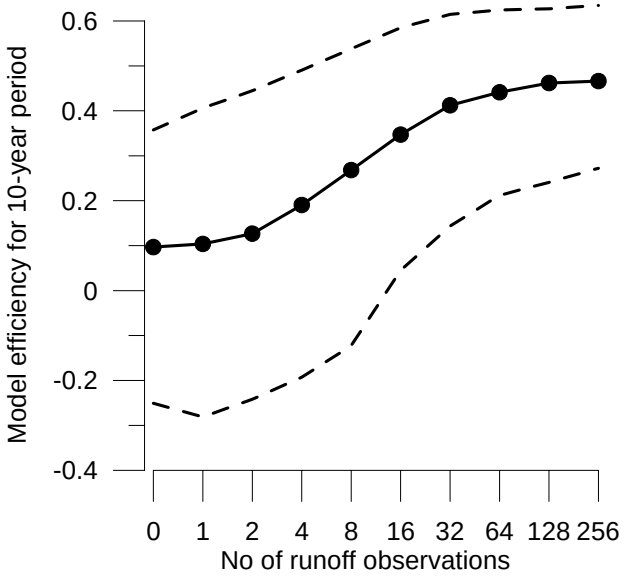

Fig. 1. Median (solid line) and percentiles (10 and 90%, dashed line) for the performance of the 100 best parameter sets according to nQ measurements for simulating the entire time series (median values for the different catchments, different years and different selection of the n days).

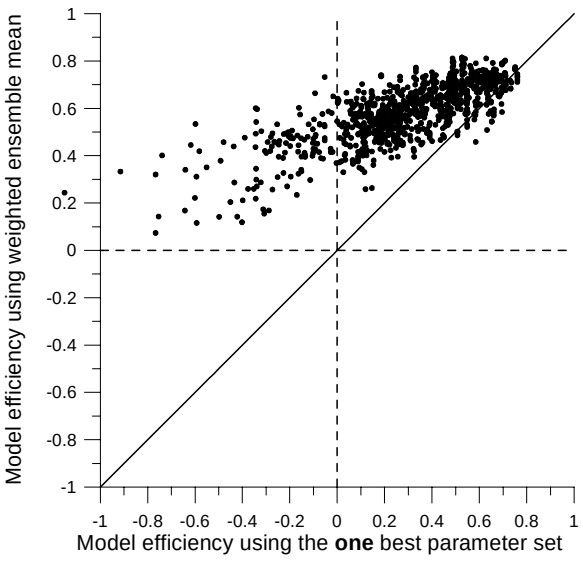

Fig. 2. Performance of the ensemble mean compared to the performance of the single one best parameter value. Each dot represents the average of the 100 realizations of n observation days during one year and in one catchment.

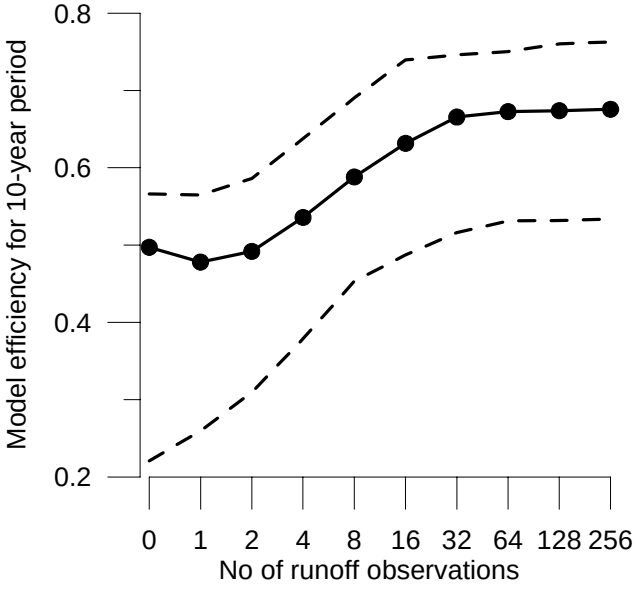

Fig. 3. Model efficiencies for the entire 10-year period of the weighted ensemble mean where the ensemble has been selected based on n measurements during one year. The solid line and the circles represent the median over all years, catchments and random realisations of the selection of the n days. The dashed lines show the percentiles (10 and 90%) for the different catchments and years.

increasing number of runoff observations. The agreement of the runoff series computed as ensemble mean with the observed 10-year series was low when less than 4 runoff observations were used to evaluate the parameter sets (Fig. 3). On the other hand, the performance was significantly better when 8 or more observations were used to select the 100 best parameter sets. There was a considerable variation among the catchments as to how good the ensemble mean was when not being constrained by any measurement and the catchments also varied in how much the additional measurements helped to increase the efficiency (Fig. 4). In general the increase in performance was largest for those catchments for which the difference between the "non-informed" (i.e. using only prior parameter estimates with no conditioning observations) performance of the ensemble mean and the efficiency that could be achieved by calibration to the entire 10-year period was largest. However, for each catchment there was a large variation of the ensemble mean performance depending on the year in which observation days were selected (spread of lines in Fig. 4).

In some cases constraining the model by a small number of runoff observations actually caused a decrease in model performance. This is especially evident when we look at the variation of the different realisations to different selections of n observations (Fig. 5). Depending on which days were selected to constrain the model, the agreement of the ensemble mean with the observed runoff varied significantly (percentiles in Fig. 5). It is interesting to note that while constraining the model on average helps to increase model performance, a poor (random) selection of observation days can actually result in worse model predictions. This suggests that the observations can be in conflict with the model representation of a catchment. This could be due to model

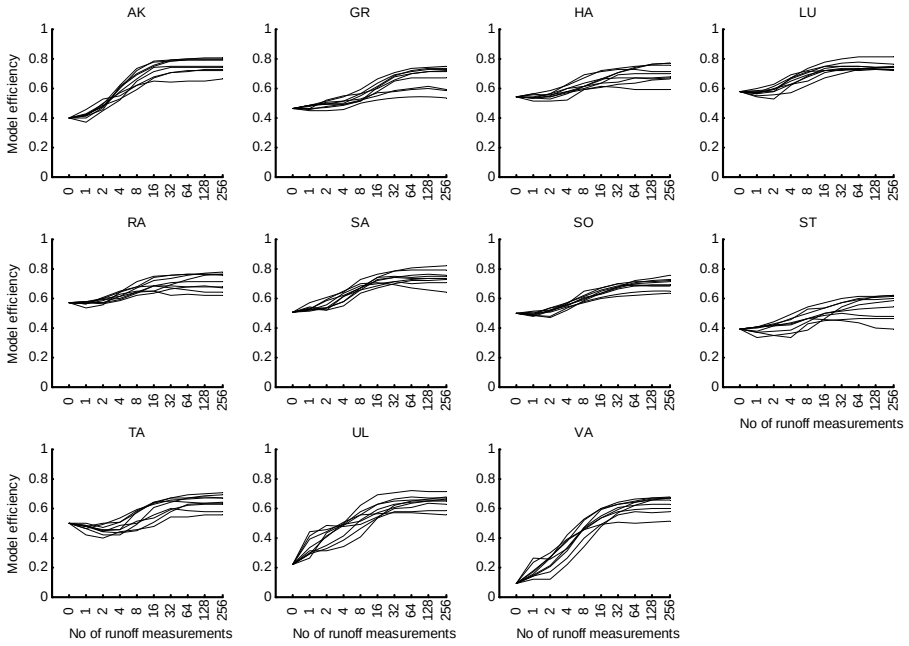

Fig. 4. Model efficiencies for the entire 10-year period of the weighted ensemble mean where the ensemble has been selected based on n measurements during one year. Each line represents one year used for this selection. Calibration to the entire 10-year series results in model efficiency (Nash and Sutcliffe, 1970) values of about 0.8 for most catchments, with the exception of three catchments with values above 0.85 for three catchments (AK, SA and VA) and around 0.7 for the two smallest catchments (ST, TA) (Seibert, 1999).

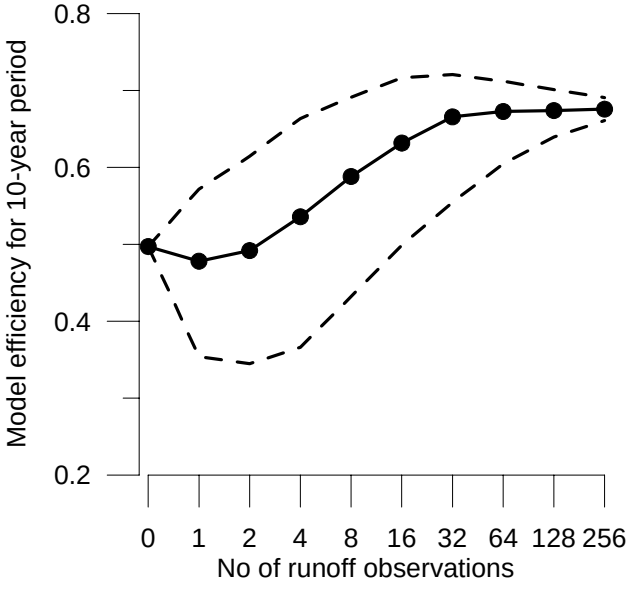

Fig. 5. Model efficiencies for the entire 10-year period of the weighted ensemble mean where the ensemble has been selected based on n measurements during one year. The solid line and the circles represent the median over all years, catchments and random realisations of the selection of the n days. The dashed lines show the medians of the percentiles (10% and 90%) for the different realisations of the selection of the n days.

structural error, but may also occur if the observations (by chance) are not representative of the longer term catchment response, such as when either the particular discharge observations or the associated rainfall event data are in error or when a particular selection is only representative of the low flow response. When only a small number of discharge observations are assumed available, such errors will take on a greater significance than when a full time series is available. This has been seen, for example, in the identification of rainfall multipliers for individual events in full calibration (Kavetski et al., 2006; Kuczera et al., 2006; Vrugt et al., 2008). These multipliers will reflect the influence of model structural errors and discharge measurement errors but in some cases take on values far from unity. Selection of an observation point for discharge in one of those events might then not be that informative (and might indeed be disinformative) in model conditioning (Beven et al., 2008).

The different tested strategies resulted in different model fits. There was a considerable scatter in the achieved model performances for the different catchments (evaluated as an average over the 10 different years for each catchment) (Fig. 6). On average the strategies that included sampling of maximum flows (MAX6, MAX1REC5, MAX2REC4) performed better than the benchmark strategies for the efficiency measure, whereas the strategies involving minimum or mean flows resulted in poorer model simulations on this measure. These results of course depend on the chosen objective function. With the two benchmark strategies efficiency values of

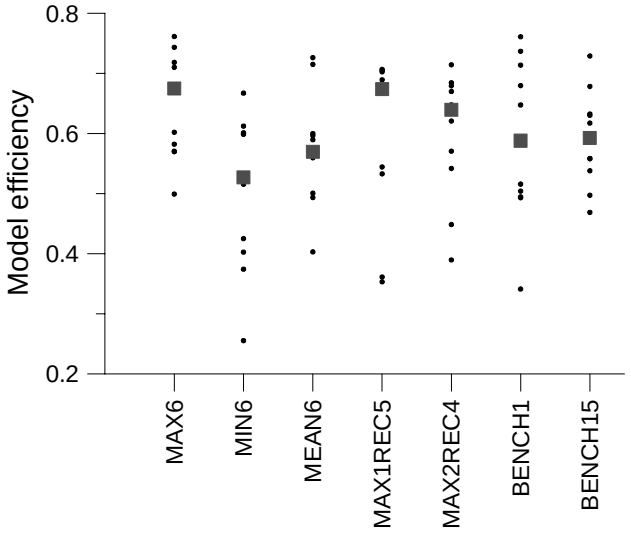

Fig. 6. Performance of the different strategies to select 6 days during one year. Each black dot represents the median of 10 years for one catchment, the square symbol represents the median of all catchments.

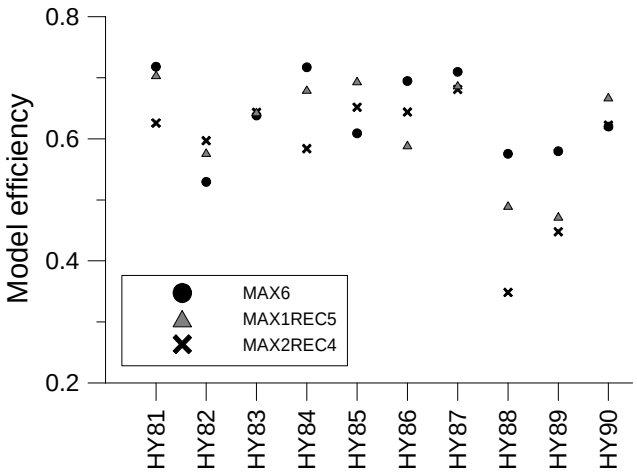

Fig. 7. Performance of the three best strategies applied in different years. Each point represents the median of all catchments when dates from one year were used. The three years 1982, 1988 and 1989 were the years with particularly little snow accumulation.

0.59 (median) were obtained, which can be compared to the average median efficiency for the different realisations selecting 4 or 8 random days of 0.54. The averages computed from the different catchments for different years showed that the same strategy provided varying results depending on the year used for the 6 gaugings. It can be noted that the three years for which the 6 measurements resulted in the poorest results were the three years with the least snow accumulation during winter (Fig. 7). There also was a tendency toward poorer results for the two smallest catchments and for the catchments with highest percentage of lakes.

While this study has demonstrated that a limited number of streamflow observations can be sufficient to constrain a model it remains an open question how these measurements should be distributed in time to maximise the information content. Certainly, the guided, hydrologically intelligent sampling that covers both high and low flows performs as well as the random sampling of a larger number of observations.

4 Discussion

The results in this paper suggest that for the type of daily runoff modelling studied, a limit to the information content in a series of observations is reached after a relatively small number of measurements are used in constraining model predictions, although some parts of the time series appear to be more informative in conditioning the model than others. With a very small number of samples, however, it has been shown that the weighted ensemble model performance might possibly decrease in both calibration and, more so, in prediction, relative to prior parameter estimates. While these results are based on a few catchments in Sweden and results might be different in other regions, our results are encouraging and motivate further studies on the question of how to gauge the ungauged catchment.

This type of assessment of data information content needs to be explored in other catchments. A number of issues might lead to different conclusions. The first is the representativity of a discharge measurement in relation to the time scale of the response of a catchment. Here we have taken the already available daily discharge observations from gauged sites (treated as ungauged for the purposes of the analysis), as if they were "point" measurements of flow. Recognising that any "point" measurement extends over a period of time that depends on the measurement technique and size of the river, this is more representative of catchments where flow is changing relatively slowly than of small flashy catchments. For this particular set of mesoscale Swedish catchments diurnal variability of stream flow is small because of the flat topography and the dampening effect of lakes along the stream network.

Second, the results will be dependent on the interaction of input errors (primarily precipitation and temperature in this case) and model structural error. As in any hydrological modelling study this is one reason why we expect results in model evaluation to be more uncertain than in model calibration where the effective values of the parameters will in part reflect the particular sequence of estimated inputs rather than the true inputs to the system. It is also perhaps one reason why the performance of a single "best" calibrated parameter set is generally worse in evaluation than the ensemble weighted mean of the top 100 parameter sets. The ensemble of models might be more robust to any particular sequence

of input errors than a single good model fit. Here it can be mentioned that most previous studies looking on the necessary amount of data for model calibration relied on using one single best parameter set rather than allowing for several suitable parameter sets (Sorooshian et al., 1983; Yapo et al., 1996; Perrin et al., 2007).

It is encouraging, although it perhaps should not be surprising, that a hydrologically intelligent choice of a small number of observations performs well relative to either regularly or randomly chosen measurement times. Choosing days with the highest discharge followed by days during the recession provided similar results to using the 6 days with highest discharge during a two-month period. One could expect that the latter should result in better calibrations since the different observations provide more independent information. However, sampling days during the recession allows constraining the model not only based on absolute discharge but also on rates of change. It is worth noting that in this study we have assumed that one can pick the day with the highest discharge to make a measurement. Our results should therefore be seen as a best-case scenario; in reality the observation days will be distributed less optimally for logistic reasons. This problem poses interesting questions which could be addressed by adding some prediction uncertainty in making decisions about which days to select for gauging. This could also be studied as an iterative problem. Once we have a few gaugings to constrain our model, our ability to predict suitable days for further measurements might improve.

The intelligent choice should reflect what we expect the model to do. There may also be other hydrologically intelligent sampling strategies in addressing the ungauged basin problem. It has, for example, become much less expensive to install a networked transducer that can provide continuous level measurements in real time, and that can easily be moved from site to site (Hughes et al., 2006). A small number of observations could then be used to produce a rating curve and hence a continuous estimate of discharge. The rating curve derived in this way would be expected to have significant uncertainty, but more complete time coverage and temporal resolution might more than compensate for allowing for rating curve uncertainty in model calibration. Another type of information for ungauged catchments are parameter estimates based on regionalisation approaches. A few streamflow observations may then be used to further constrain these parameter estimates. The value of the streamflow measurements obviously will depend on how well the parameter values had been estimated initially.

5 Concluding remarks

In some situations, a solution to the ungauged basin problem will be to take some (a small number) discharge measurements to help constrain model predictions. However, very little guidance is to be found in the literature about the value of different measurement strategies in this context. This paper provides a framework for looking at different choices of numbers and selection of measurements in model conditioning. We have shown, by applying the HBV model to a number of small Swedish catchments, that only a few measurements can be effective in constraining prediction uncertainties. There is, however, always the possibility that as the number of measurements is reduced the real information content becomes more sensitive to particularities of the time the measurement is taken, especially the possibility of rainfall input errors and model structural errors even if the discharge measurement is itself accurate.

Thus, many more studies similar to the one presented here are needed to develop guidelines on what to measure and when to measure in ungauged basins. Well-instrumented catchments are needed for this type of approach where the basic idea is to pretend that there is only a subset of these data available. This situation has been mentioned as the PUB-paradox (Bonell et al., 2006): data-rich catchments are needed to test methods for data-poor environments. The results in this study differed for different catchments and different years, demonstrating that misleading results might be obtained if one would test the value of data based on only one or few cases. In a real application for one particular case it is impossible to know exactly whether good simulations could be obtained based on a certain number of observations, but using many test cases, as in this study, allows some probability distribution of errors to be assessed as an index of the uncertainty that might be expected in a real application. This suggests that the ungauged basin problem must be treated as a learning process, with more data being added if the application requires it, either because it appears that predictions are wrong, or because more constraint of uncertainty is required for decision making (Beven, 2007).

An interesting outcome of this study is that a hydrologically intelligent choice of when the measurements are made might help to maximise the information content of the observations, although much more work is required on identifying the most useful observations. It was also shown that mean ensemble predictions generally produced better results than any single model prediction after conditioning on a small number of observations.

We intend to explore additional sampling strategies in future. For practical applications the cost of the measurements must be considered. This includes posing questions such as: Are more but less accurate measurements or fewer but more accurate measurements more useful? For more remote catchments where travel to the site is the major cost of a single discharge measurement the optimal strategy might be different from an easily accessible catchment, where the cost is more determined by the actual measurement. It is also possible that the accuracy or frequency of discharge measurements in the model calibration process might be dominated by poor knowledge of catchment inputs. Here, as is common in rainfall-runoff modelling, we have assumed that the

estimated inputs to the catchment are sufficiently accurate that model performance over the 10 year evaluation period is not overly constrained. Similar issues arise in regionalisation approaches to the ungauged basin problem but we contend that the eventual solution to constraining predictions of ungauged basins will be to have rapid ways of feeding in more observations to the modelling and learning process.

Acknowledgements. The data used in this study has been collected by SMHI. We thank Bettina Schaefli, Steve Lyon, Vazken Andréassian, Charles Perrin, Ludovic Oudin and Thibault Mathevet as well as two anonymous reviewers for their valuable comments on an earlier draft of this paper.

Edited by: L. Pfister

References

Bergström, S.: Parametervärden för HBV-modellen i Sverige: erfarenheter från modellkalibreringar under perioden 1975–1989, Swedish Meteorological and Hydrological Institute, Norrköping, 35 pp., 1990.

Bergström, S.: The HBV Model: Its Structure and Applications, Swedish vMeteorological and Hydrological Institute (SMHI), Hydrology, Norrköping, 35 pp., 1992.

Bergström, S.: The HBV model (Chapter 13), in: Computer Models of Watershed Hydrology, edited by: Singh, V. P., Water Resources Publications, Highlands Ranch, Colorado, USA, 443–476, 1995.

Beven, K.: On undermining the science?, Hydrol. Process., 20, 3141–3146, 2006.

Beven, K.: Towards integrated environmental models of everywhere: uncertainty, data and modelling as a learning process, Hydrol. Earth Syst. Sci., 11, 460–467, 2007, http://www.hydrol-earth-syst-sci.net/11/460/2007/.

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

Beven, K. J.: Environmental Modelling: An Uncertain Future?, Routledge London, 310 pp., 2009.

Beven, K. J., Smith, P. J., and Freer, J. E.: So just why would a modeller choose to be incoherent?, J. Hydrol., 354, 15–32, 2008.

Binley, A. and Beven, K.: Vadose Zone Flow Model Uncertainty as Conditioned on Geophysical Data, Ground Water, 41, 119–127, 2003.

Bonell, M., McDonnell, J. J., Scatena, F. N., Seibert, J., Uhlenbrook, S., and Van Lanen, H. A. J.: HELPing FRIENDs in PUBs: charting a course for synergies within international water research programmes in gauged and ungauged basins, Hydrol. Process., 20, 1867–1874, 2006.

Eng, K. and Milly, P. C. D.: Relating low-flow characteristics to the base flow recession time constant at partial record stream gauges, Water Resour. Res., 43, W01201, doi:10.1029/2006WR005293, 2007.

Eriksson, B.: Den potentiella evaporationen i Sverige, Swedish Meteorological and Hydrological Institute, SMHI, 40 pp., 1981 (in Swedish, The potential evaporation in Sweden).

Eriksson, B.: Data rörande Sveriges Nederbördsklimat-Normalvärden för perioden 1951–80 (Data concerning the precipitation climate of Sweden-Normal values for the period 1951–80), SMHI Rapport, Norrköping, 92 pp., 1983 (in Swedish).

Harlin, J. and Kung, C. S.: Parameter uncertainty and simulation of design floods in Sweden, J. Hydrol. (Amsterdam), 137, 209–230, 1992.

Hughes, D., Greenwood, P., Coulson, G., Blair, G., Pappenberger, F., Smith, P., and Beven, K.: GridStix: Supporting Flood Prediction using Embedded Hardware and Next Generation Grid Middleware, International Workshop on Wireless Mobile Multimedia, 621–626, 2006.

Juston, J., Seibert, J., and Johansson, P. O.: Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment, Hydrol. Process., in press, 2009.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resour. Res., 42, W03407, doi:10.1029/2005WR004368, 2006.

Kuczera, G., Kavetski, D., Franks, S., and Thyer, M.: Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, J. Hydrol., 331, 161–177, 2006.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272–288, 1997.

Mantovan, P. and Todini, E.: Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, J. Hydrol., 330, 368–381, 2006.

McIntyre, N., Lee, H., Wheater, H., Young, A., and Wagener, T.: Ensemble predictions of runoff in ungauged catchments, Water Resour. Res., 41, 3307–3323, 2005.

McIntyre, N. R. and Wheater, H. S.: Calibration of an in-river phosphorus model: prior evaluation of data needs and model uncertainty, J. Hydrol., 290, 100–116, 2004.

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

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., and Mathevet, T.: Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models, Hydrolog. Sci. J., 52, 131–151, 2007.

Rode, M., Suhr, U., and Wriedt, G.: Multi-objective calibration of a river water quality model—Information content of calibration data, Ecol. Modell., 204, 129–142, 2007.

Rojas-Serna, C., Michel, C., Perrin, C., and Andreassian, V.: Ungauged catchments: how to make the most of a few streamflow measurements?, Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment – MOPEX, IAHS publication, 307, 230–236, 2006.

Seibert, J.: Estimation of parameter uncertainty in the HBV model, Nord. Hydrol., 28, 247–262, 1997.

Seibert, J.: Regionalisation of parameters for a conceptual rainfall-runoff model, Agr. Forest Meteorol., 98, 279–293, 1999.

Seibert, P.: Hydrological characteristics of the NOPEX research area, Undergraduate thesis, Institute of Earth Sciences/Hydrology, Uppsala University, Uppsala, Sweden, 51 pp., 1994.

Sivapalan, M., Schaake, J., and Sapporo, J.: PUB Science and Implementation Plan, V5., online available at: http://pub.iwmi.org/

UI/Images/PUB_Science_Plan_V_5.pdf, IAHS Decade on Predictions in Ungauged Basins (PUB), 2003a.

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J., Mendiondo, E., O'Connell, E. P., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., and Zehe, E.: IAHS decade on predictions in ungauged basins (PUB), 2003-2012: Shaping an exciting future for the hydrologic sciences, Hydrolog. Sci. J., 48, 857–880, 2003b.

Smith, P., Beven, K. J., and Tawn, J. A.: Informal likelihood measures in model assessment: Theoretic development and investigation, Adv. Water Res., 31, 1087–1100, 2008.

Sorooshian, S., Gupta, V. K., and Fulton, J. L.: Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility, Water Resour. Res., 19, 251–259, 1983.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., and Robinson, B. A.: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, Water Resour. Res., 44, W00B09, doi:10.1029/2007WR006720, 2008.

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Res., 30, 1756–1774, 2007.

Yapo, P. O., Gupta, H. V., and Sorooshian, S.: Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, J. Hydrol., 181, 23–48, 1996.

Zhang, Z., Wagener, T., Reed, P., and Bhushan, R.: Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization, Water Resour. Res, 44, W00B04, doi:10.1029/2008WR006833, 2008.