

Scale effects in conceptual hydrological modeling

R. Merz,¹ J. Parajka,¹ and G. Blöschl¹

Received 17 February 2009; revised 15 June 2009; accepted 24 June 2009; published 9 September 2009.

[1] We simulate the water balance dynamics of 269 catchments in Austria ranging in size from 10 to 130,000 km² using a semidistributed conceptual model with 11 parameters based on a daily time step. The simulation results suggest that the Nash-Sutcliffe model efficiencies increase over the scale range of 10 and 10,000 km². The scatter of the model performances decreases with catchment scale, particularly the volume errors. This implies that the model simulates the long-term water balance more reliably as one goes up in scale. Most calibrated parameters do not change with catchment scale, but there is a trend with catchment area of the upper and lower envelope curves of some parameters. We also examine time scale effects. Calibration efficiencies decrease and verification efficiencies increase with the number of years available for calibration. The change in efficiencies is largest between 1 and 5 years used for calibration. This suggests that a calibration period of 5 years captures most of the temporal hydrological variability, so this would be the minimum for achieving a reasonable predictive model performance. The correlation of model parameters between different calibration periods, as a measure of the degree to which parameters can be identified, increases with increasing length of the calibration period. For some parameters, the correlation increases beyond 5 years of calibration. This suggest that although runoff may be simulated well using 5 years of calibration, some parameters may not be well constrained and hence internal state variables and fluxes may still be associated with larger uncertainties than with a larger calibration period.

Citation: Merz, R., J. Parajka, and G. Blöschl (2009), Scale effects in conceptual hydrological modeling, *Water Resour. Res.*, 45, W09405, doi:10.1029/2009WR007872.

1. Introduction

[2] Understanding and modeling the water balance dynamics of catchments is important from both engineering and scientific perspectives. Water balance models play an important role in managing the water resources of river basins. They can be used in the context of assessing anthropogenic effects on water quantity and quality, for estimating design values, and for streamflow forecasting [e.g., Beven, 2001a]. In the past decades a host of water balance models have been developed that range in structure from simple black box models, to conceptual models and complex physically based models [Singh and Frevert, 2001a, 2001b]. In conceptual models, the basic processes such as interception, infiltration, evaporation, surface and subsurface runoff, etc., are separated to some extent, but the algorithms that are used to describe the processes are essentially calibrated inputoutput relationships, formulated to mimic the functional behavior of the process in question [e.g., Beven, 2001a].

[3] It has been argued that the functional behavior of catchments may differ vastly with catchment scale, so it may not be appropriate to use the same model structure and the same model parameters for small and large catchments alike [*Beven*, 1989, 1991]. However, while scale issues in

catchment modeling have attracted a lot of attention in the past years [Blöschl and Sivapalan, 1995], there exist very few studies that have systematically examined whether the model structure should change with catchment scale and if so, in which way. For small catchments, often very complex, spatially distributed models are used while for large catchments, often simple lumped models suffice [Gravson and Blöschl, 2000], the rationale being that the spatial variability averages out with scale, so less detail is needed as one moves up in scale [Blöschl et al., 1995]. However, there also exists the counterexample that streamflow forecasts are usually based on subcatchment models, the rationale being that as one moves up in scale, more diverse hydrologic conditions are encountered, so more subcatchments are needed to represent these differences explicitly [*Sivapalan*, 2003].

[4] One possibility to test the research question of whether the model structure should change with catchment scale is to reverse the question and apply the same structure to both small and large catchments and examine whether the model performance changes with scale. If it does change with scale, one can conclude that the model structure used is more appropriate for one scale than for the other, so at the scale where performance is poor a different structure should be used. On the other hand, if the model performance does not change with catchment scale one can conclude that the same model structure is equally suitable for all scales. While numerous studies have examined model performance in general [e.g., *Jakeman and Hornberger*, 1993; *Refsgaard*

¹Institute of Hydraulic Engineering and Water Resources Management, Vienna University of Technology, Vienna, Austria.

Copyright 2009 by the American Geophysical Union. 0043-1397/09/2009WR007872\$09.00

and Knudsen, 1996], very few of them have explicitly considered model performance as a function of catchment scale. One notable example is that of *Perrin et al.* [2001], who tested the model performance of several lumped conceptual models on 429 catchments of different scales. They found that most models performed similarly for large catchments as they did for small catchments. They also noted that the data quality strongly influenced model performance, particularly in their case where catchments came from different parts of the world.

[5] In a similar fashion one would expect model parameters to change with catchment scale, as different processes are likely to dominate in small and large catchments [Beven, 1991, 2001b; Gottschalk et al., 2001; Fenicia et al., 2008]. In small catchments, hillslope processes including macropore effects tend to control the streamflow response. In large catchments, channel processes and flood inundation, as well as regional aquifers, may significantly modulate hillslope response, and the response is often much slower than in small catchments. Typically in Austria, catchments of 1 km² have a time of concentration of about 1 h, catchments of 100,000 km² have a time of concentration of about 3 days, and the larger lag is mainly due to the channel travel times. One can test the scale dependence of model parameters, in a similar fashion as that of model structure, by calibrating the same model to a range of catchment scales and then examining the calibrated parameters as a function of scale. One of the few examples of testing the scale dependence of model parameters in this fashion is given by Bergström and Graham [1998], who applied the HBV model to 25 subcatchments of the Baltic Sea catchment. They found that the runoff generation parameters were relatively stable over a wide range of scales and concluded that there may be no scale problem with conceptual water balance modeling, as "the large basin is just a sum of many small ones" [Bergström and Graham, 1998, p. 261].

[6] There is another important scale dependence of catchment models, which is the time scale dependence of model performance. In the calibration process one usually attempts to use a sufficiently long period of runoff observations to ascertain that the calibration is not just a fit to the data set but genuinely represents the population of streamflow variability [Bergström, 1991]. In practice, however, the choice of record length is often determined by data availability, so in many cases only a few years of runoff data are used for calibration. The generic scale question here is whether the model efficiency changes with the period of runoff data used for calibration and what is a sufficiently long period to acquire sufficient confidence in the performance of the model for future applications. In general, one would expect that the minimum calibration period needed is one that samples all different types of hydrological conditions, including extreme events. This is usually checked by comparing model efficiencies for the calibration and verification periods [Refsgaard, 2000] rather than by varying the length of the calibration period. If the verification efficiency is not much poorer than the calibration efficiency, one concludes that the model genuinely represents the population of streamflow variability.

[7] Both the estimation of model parameters and the issues of calibration and verification efficiencies are con-

founded by problems of parameter uncertainty [Montanari, 2005, 2007; Refsgaard et al., 2006; Götzinger and Bárdossy, 2008; Freer et al., 1996]. An analysis of calibrated model parameters as a function of catchment scale is only useful if the parameter uncertainty is smaller than possible scale effects. Similarly, in the analysis of model performance as a function of spatial and temporal scales, robust parameter estimation is an advantage. Most of the analyses of parameter uncertainty in the literature are based on Monte Carlo simulations for the same catchment [see, e.g., Gupta et al., 1998]. Uhlenbrook et al. [1999], for example, analyzed the parameter uncertainty of a conceptual water balance model (the HBV model) for a small mountainous catchment using Monte Carlo simulations. They found some of the parameters such as the maximum soil moisture storage and the lower zone recession coefficient to be poorly defined while other parameters such as the degree-day factor were much better constrained. A similar study was performed by Seibert [1997] for a number of Swedish catchments, but the uncertain parameters were not the same as those in the study of Uhlenbrook et al. [1999]. This implies that parameter uncertainty significantly depends on the catchments studied and data aspects in addition to the model structure. An alternative to Monte Carlo studies is calibrating the model on different subperiods and comparing the calibrated parameters for the respective subperiods. This is in fact a more stringent test of parameter robustness than Monte Carlo analyses, as it tests both the identifiability of parameters and the stationarity of the data and their quality. If the calibrated model parameters for the subperiods are similar, then the uncertainty can be assumed to be small. However, relatively long data series are needed for this type of test to be meaningful.

[8] The aim of this paper is to investigate scale effects in modeling the water balance dynamics of catchments as discussed above. Specifically, we examine four research questions: (1) Does model performance change with catchment scale? (2) Do model parameters change with catchment scale? (3) Does model performance change with the length of the calibration period? (4) Do model parameters change with the length of the calibration period?

[9] We used a semidistributed conceptual water balance model, based on the structure of HBV, to address these questions and simulate daily runoff for 269 Austrian catchments ranging in scale from 10 km² to 130,000 km². We examine the parameter uncertainty by comparing model parameters calibrated for different periods. In section 2 we present the data, followed by a description of the model. We then address each of the four research questions in sequence.

2. Data

[10] This study was carried out in Austria using data from the period 1976–2005. Austria is flat or undulating in the east and north, and Alpine in the west and south. Elevations range from 115 m above sea level (asl) to 3797 m asl. Mean annual precipitation is less than 400 mm/a in the east and almost 3000 mm/a in the west. Land use is mainly agricultural in the lowlands and forest in the medium elevation ranges. Alpine vegetation and rocks prevail in the highest catchments. The data set used in this study includes measurements of daily precipitation at 1091 stations and daily air temperature at 212 climatic stations. Daily runoff data

	Model Parameter j	Model Component	p_l	p_u	α	β	p_{max}
SCF	snow correction factor ()	snow	1.0	1.5	1.2	4.0	1.03
DDF	degree-day factor (mm/°C d)	snow	0.0	5.0	2.0	4.0	1.25
T_M	melt temperature (°C)	snow	-1.0	3.0	2	4	0.0
FC	maximum soil moisture storage (mm)	soil	0.0	600	1.1	1.5	100
LP/FC	ratio of limit for potential evaporation and FC ()	soil	0.0	1.0	4.0	1.2	0.94
В	nonlinearity parameter of runoff generation ()	soil	0.0	20	1.1	1.5	3.4
K_0	storage coefficient of additional outlet (d)	runoff	0.0	2.0	2.0	4.0	0.5
K_1	fast storage coefficient (d)	runoff	2.0	30	2.0	4.0	9.0
K_2	slow storage coefficient (d)	runoff	30	250	1.05	1.05	105
LS_{UZ}	storage capacity threshold (mm)	runoff	1.0	100	3.0	3.0	50
C_P	percolation rate (mm/d)	runoff	0.0	8.0	2.0	4.0	2.0

Table 1. A Priori Distribution of Parameter Values^a

^aHere p_l and p_u are the lower and upper bounds of the parameter space used in all iterations; u and v are the initial parameters of the a priori distribution (equation (3)); and p_{max} is the initial parameter value at which the a priori distribution is at a maximum.

from 269 gauged catchments were used with areas ranging from 10 km² to 130,000 km² and a median of 243 km².

[11] The inputs to the hydrologic catchment model were prepared in two steps. First, the daily values of precipitation, snow depth, and air temperature were spatially interpolated by methods that use elevation as auxiliary information. External drift kriging was used for precipitation and snow depth, and the least squares trend prediction method was used for air temperature. The spatial distribution of potential evapotranspiration was estimated by a modified Blaney-Criddle method [Parajka et al., 2005] using daily air temperature and potential sunshine duration calculated by the Solei-32 model [Mészároš et al., 2002] that incorporates shading by surrounding terrain. In a second step, a digital elevation model of 1×1 km grid resolution was used for deriving 200-m elevation zones in each catchment. Time series of daily precipitation, air temperature, potential evaporation, and snow depth were then extracted for each of the elevation zones to be used in the water balance simulations.

3. Hydrological Model

3.1. Model Structure

[12] The model used in this paper is a semidistributed conceptual rainfall-runoff model, following the structure of the HBV model [Bergström, 1992]. The model equations are given in the appendix of Parajka et al. [2007a]. Each catchment is subdivided into elevation zones of 200-m vertical range. The model runs on a daily time step and consists of a snow routine, a soil moisture routine, and a flow routing routine. The snow routine represents snow accumulation and melt by a simple degree-day concept, involving the degree-day factor DDF and melt temperature T_{M} . Catch deficit of the precipitation gauges during snowfall (i.e., systematic measurement errors due the wind effects) is corrected by a snow correction factor, SCF. If air temperature is above a threshold temperature T_R , precipitation is considered to occur as rainfall; below a threshold temperature T_S , it is considered to occur as snowfall; and a mix of rain and snow is in between. The soil moisture routine represents runoff generation and changes in the soil moisture state of the catchment and involves three parameters: the maximum soil moisture storage FC, a parameter representing the soil moisture state above which evaporation is at its potential rate, termed the limit for potential evaporation LP, and a parameter in the nonlinear function

relating runoff generation to the soil moisture state, termed the nonlinearity parameter *B*. Runoff routing on the hillslopes is represented by an upper and a lower soil reservoir. Excess rainfall enters the upper zone reservoir and leaves this reservoir through three paths: outflow from the reservoir based on a fast storage coefficient K_1 ; percolation to the lower zone with a constant percolation rate C_P ; and if a threshold of the storage state LS_{UZ} is exceeded, through an additional outlet based on a very fast storage coefficient K_0 . Water leaves the lower zone based on a slow storage coefficient K_2 . The outflow from both reservoirs is then routed by a triangular transfer function representing runoff routing in the streams. The base of the transfer function decreases linearly with discharge, and the factor of proportionality is C_R , which is a calibration parameter.

[13] The model was run for all 269 gauged catchments in Austria. Daily inputs (precipitation, air temperature, and potential evapotranspiration) were allowed to vary with elevation within a catchment, and the soil moisture accounting and snow accounting were performed independently in each elevation zone. However, the same model parameters were assumed to apply to all elevation zones of a catchment. In order to reduce the number of calibrated model parameters, Parajka et al. [2007b] performed a sensitivity analysis for Austrian catchments (including most of the catchments of this study). Ranking the sensitivity of each model parameter in Austrian catchments revealed that many of the model parameters are sensitive in some catchments but insensitive in others. Three parameters that were among those that generally showed the least sensitivity were preset $(T_R = 2^{\circ}\text{C}, T_S = 0^{\circ}\text{C}, C_R = 25 \text{ d}^2/\text{mm})$, and 11 parameters (Table 1) were estimated by calibration.

3.2. Model Calibration and Verification

[14] We calibrated the model parameters to observed runoff making use of an automated procedure that involves an objective function. Selection of the objective function is not straightforward because different objective functions tend to test different aspects of the similarity between model output and observations. However, as *Weglarczyk* [1998] points out, the objective functions commonly used in hydrology are often not independent and based on a small number of fundamental measures that are related to bias or random errors, or combinations thereof.

[15] We used the *Nash and Sutcliffe* [1970] efficiency, *ME*, of flows and the Nash and Sutcliffe efficiencies of the

logarithmic flows, ME_{ln} , for comparing simulated and observed runoff in this paper:

$$ME = 1 - \frac{\sum_{i=1}^{n} (Q_{obs,i} - Q_{sim,i})^2}{\sum_{i=1}^{n} (Q_{obs,i} - \overline{Q}_{obs})^2}$$
(1a)

$$ME_{\ln} = 1 - \frac{\sum\limits_{i=1}^{n} \left(\ln(Q_{obs,i}) - \ln(Q_{sim,i}) \right)^2}{\sum\limits_{i=1}^{n} \left(\ln(Q_{obs,i}) - \overline{\ln(Q_{obs})} \right)^2},$$
(1b)

where $Q_{obs,i}$ and $Q_{sim,i}$ are observed and simulated runoff on day *i*, respectively, and $\overline{Q_{obs}}$ is the mean of observed runoff over the calibration period of *n* days. A perfect match between simulated and observed runoff implies ME = 1 and for a less than perfect match ME < 1.

[16] In calibration procedures the parameter values are usually bounded between two limits [*Duan et al.*, 1992], and otherwise no a priori assumptions are made about the parameters. This implies that the a priori distribution of the parameters is a uniform distribution. We believe that it is possible to make a more informed guess about the shape of the a priori distribution and introduced a penalty function ε_p based on a Beta distribution for each parameter:

$$\varepsilon_p = \sum_{j=1}^{k} \frac{f_{\max,j} - f_j \left(\frac{p_j - p_{l,j}}{p_{u,j} - p_{l,j}}\right)}{f_{\max,j}}$$
(2a)

$$f_{\max,j} = f j \left(\frac{p_{\max,j} - p_{l,j}}{p_{u,j} - p_{l,j}} \right), \tag{2b}$$

where p_j is the model parameter *j* to be calibrated, p_1 and p_u are the lower and upper bounds of the parameter space, p_{max} is the parameter value at which the Beta distribution is at a maximum, and *k* is the number of parameters to be calibrated. Here *f* is the probability density function of the Beta function:

$$f(x|\alpha,\beta) = \frac{1}{Beta(\alpha,\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

for $0 < x < 1, \alpha > 0, \beta > 0$ (3)

with

$$Beta(\alpha,\beta) = \int_{0}^{1} x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

[17] We assumed values of α and β for each parameter k based on our own assessment of the hydrologic characteristics of the study region and our prior experience with hydrological modeling in Austria (Table 1). The α and β vary between the parameters but do not vary between the catchments. We chose the lower and upper bounds of the parameters based on literature values [*Bergström*, 1992;

Seibert, 1997] and on our own assessment. The bounds were the same for all catchments (Table 1).

[18] The entire objective function now consists of the following parts:

$$Z = w_1(1 - ME) + w_2(1 - ME_{\ln}) + w_3\varepsilon_p,$$
(4)

where the weights w_i were set to $w_1 = 0.4$, $w_2 = 0.4$ and $w_3 = 0.2$ based on our prior experience with hydrological modeling in Austria, giving a relative importance of 80% to a good fit of observed and simulated runoff measured by Nash and Sutcliffe model efficiency and 20% to the a priori distribution of the model parameters on average of the 269 catchments. This objective function was minimized using the shuffled complex evolution (SCE-UA) method [*Duan et al.*, 1992]. Model parameters were calibrated to runoff from 1 November to 31 October for each calibration period.

[19] Warm-up periods from January to October were used in all calibrations, in order to set the initial conditions for the simulations in the calibration and verification periods. We judged the model performance by a split sample test in the terminology of *Klemeš* [1986]. We compared simulated and observed runoff in terms of model efficiencies *ME* for verification periods that were not used for calibration. We also compared these efficiencies and errors with those obtained for the calibration period. Only if the model performs similarly well for both periods can the model be used with confidence in a predictive mode. A second measure of model performance used here is the volume error *VE*, which is a measure of bias and is defined as

$$VE = \frac{\sum_{i=1}^{n} Q_{sim,i} - \sum_{i=1}^{n} Q_{obs,i}}{\sum_{i=1}^{n} Q_{obs,i}}.$$
 (5)

VE = 0 implies no bias, and values larger and smaller than 0 imply an overestimation and an underestimation of the total runoff volume, respectively.

[20] In Figure 1 the Nash Sutcliffe efficiencies ME and volume errors VE of the calibration and verification periods have been plotted against each other. Figures 1a and 1c show the results for parameters calibrated to 1976-1990 and verified by 1991-2005, and Figures 1b and 1d show the results for the swapped periods. Figures 1a and 1b show Nash Sutcliffe efficiencies ME, and Figures 1c and 1d show volume errors VE. The mean model efficiencies for the periods are given in Table 2.

[21] The Nash Sutcliffe efficiencies *ME* for calibration and verification are correlated with a correlation coefficient of r = 0.79 and r = 0.68 in Figures 1a and 1c and Figures 1b and 1d, respectively. This means that a good calibration efficiency tends to imply a good efficiency for the verification period. For a number of catchments, the verification efficiency is higher than the calibration efficiency. This is not surprising, as the objective function used for calibration (equation (4)) consists of a number of components, one of which is the Nash-Sutcliffe efficiency. Also, the two periods may have different hydrologic characteristics. The median of the calibration and verification efficiencies in Figure 1a is 0.75 and 0.71, respectively. This means that on average over



Figure 1. Model performance of verification versus calibration periods. (a and c) Model calibrated on the period from 1976 to 1990. (b and d) Model calibrated on the period from 1991 to 2005. *ME* is the Nash-Sutcliffe model efficiency (equations (1a) and (1b)), and *VE* is the volume error (equation (5)). Each point relates to one catchment.

all the catchments, one would expect a loss in model efficiency of 0.04 when using the parameters calibrated to 1976-1990 for predicting runoff for the period 1991-2005. For the swapped periods, the change in the median model efficiency is 0.74 to 0.72; that is, one may expect a loss in model efficiency of only 0.02 when moving from the calibration to the verification period. If we compare the median model efficiencies for the same periods, calibration efficiencies are higher than verification efficiencies and the difference is 0.03 for both periods. On the other hand, if we compare the calibration efficiencies for the different periods, efficiencies are higher for 1976-1990 than they are for 1991-2005 and the difference is 0.01. The difference for

the verification efficiencies is also 0.01. This implies that the change in model efficiency, when moving from calibration to verification, consists of two components. The first component is a loss in model efficiency of 0.03 (in both cases), due to a general tendency of models to better represent calibration data than verification data. The model is calibrated to represent the hydrologic conditions for the calibration period, and as they are never exactly the same in the verification period there is a loss of accuracy. The second component is a change in model efficiency of 0.01 (in both cases), due to the model performing poorer in the more recent period than in the earlier period. This difference is likely related to differences in the data quality. It is also possible that part of the difference is related to nonstationarities in the hydrologic conditions that result in a slightly poorer model performance in the more recent period.

[22] The model efficiencies for the calibration and verification periods found in this paper are similar or better than those studies in the literature that are based on an analysis of a similar number of catchments [see, e.g., Oudin et al., 2008; Perrin et al., 2001, 2008]. The model efficiencies may be slightly lower than those of studies where only a small number of catchments were analyzed [e.g., World Meteorological Organization (WMO), 1986]. There may be two reasons for this. The first is that unlike many studies in the literature analyzing only one or a few catchments, we have not handpicked the model for each catchment. For some of the catchments the model structure may be less than perfect. The second is that there may also be some data problems that we have not detected, but in an individual case study for a small set of catchments, as commonly reported in the literature, one would remove outliers and focus on the data for which the model gives consistent results. The overall ratios of model efficiencies of the verification and calibration periods are close to unity. This means that there is no over-calibration and the model can be used with confidence for the analyses in this paper.

[23] Before analyzing the scale effects of model parameters we examined to what degree the parameters represent real hydrologic conditions in the catchments rather than model calibration artifacts related to parameter uncertainty. We analyzed the parameter uncertainty by comparing the

Table 2. Mean Nash-Sutcliffe Model Efficiencies *ME*, Mean Volume Errors *VE*, and Mean Model Parameters for Two Calibration and Verification Periods, as Well as Correlation Between the Two Calibration Periods

	Calibration 1976-1990/	Calibration 1991-2005/	Coefficient of Correlation	
	Verification 1991–2005	Verification 1976–1990		
Number of catchments	269	269		
Mean ME (calibration)	0.75	0.74	0.79	
Mean ME (verification)	0.71	0.72	0.68	
Mean VE (calibration)	-0.03	-0.01	0.75	
Mean VE (verification)	0.03	-0.09	0.18	
Mean SCF	1.08	1.07	0.77	
Mean DDF	1.74	1.60	0.70	
Mean T_M	-0.1	0.0	0.73	
Mean LP/FC	0.93	0.92	0.75	
Mean FC	148	183	0.82	
Mean B	2.8	4.0	0.88	
Mean K_0	0.41	0.39	0.74	
Mean K_1	10.5	10.4	0.85	
Mean K ₂	115	106	0.78	
Mean LS_{UZ}	49.3	49.7	0.81	
Mean C_P	2.06	2.12	0.91	



Figure 2. Model parameters calibrated on the period from 1976 to 1990 versus model parameters calibrated on the period from 1991 to 2005.

parameter sets of two independent calibration periods. We believe that the kind of uncertainty analysis is a more meaningful test of parameter uncertainty than the Monte Carlo simulations usually performed in the literature [e.g., Beven and Binley, 1992]. Monte Carlo simulations only assess the parameter uncertainty due to model structure, while the differences in the two parameter sets examined here represent both uncertainties due to model structure and data errors. Also, we are using here a large number of catchments rather than a single catchment with a large number of realizations, as is usually the case with Monte Carlo studies. Our analyses give a measure of the uncertainty of each parameter relative to the range of parameter values encountered in different catchments of a given region. In Figure 2 selected model parameters of the snow model, the soil moisture accounting scheme, and the response function calibrated to the period 1976–1990 are plotted against those of the period 1991–2005. The average parameter values for the two calibration periods are given in Table 2. The coefficients of correlation between the parameter values of the two periods are always greater than 0.7. Because of the good correlation, we believe that the parameter uncertainty will not significantly affect the interpretation of the parameter values. To add credence to the analysis, we analyzed the periods separately and checked for consistency of the results.

4. Results

4.1. Does Model Performance Change With Catchment Scale?

[24] In Figure 3 Nash-Sutcliffe model efficiencies *ME* have been plotted against the catchment area. The medians and standard deviations within scale classes are plotted as lines in Figure 3. Model performance tends to increase with catchment scale for catchments between 10 and 10,000 km² (Figure 3a). For the calibration period 1976–1990, the median of the model performance increase from 0.72 to

0.82. For the 1991-2005 calibration period the median increases from 0.7 to 0.8. This indicates that the model tends to better simulate the water balance dynamics in larger catchments. However, for catchments larger than 10,000 km² the median model efficiencies for both calibration periods drop to 0.75 and 0.71, respectively. These catchments consist of one Inn catchment of about 26,000 km² and the Danube catchments with catchment areas larger than 90,000 km². These are only a few catchments so the drop in efficiency may not be very significant. Also, part of the catchment area of the Danube catchments is outside Austria and the input data have been extrapolated from stations within Austria. The drop in efficiency may therefore partly be related to a decrease in data quality. A similar trend can be found for the verification periods with increasing model efficiencies for catchments between 10 and 10,000 km² and a decrease in model efficiencies for catchments larger than 10,000 km^2 (Figure 3b). The ratio of the efficiencies, on average, is close to one (Figure 3c). The standard deviations of the Nash-Sutcliffe model efficiencies for different catchment scales are given in Figure 3d. The standard deviation decreases with scale, which suggests that larger catchments can be modeled most consistently, i.e., the model performance is never very poor. Only for the class with catchment areas larger than 10,000 km² is the standard deviation of the model efficiencies for the verification period 1991-2005 larger than the trend. In order to assess the scale trends of the standard deviation of model efficiencies more qualitatively, a Breusch-Pagan test [Breusch and Pagan, 1979] was performed in which the squared residuals of a regression with scale are plotted against scale. The null hypothesis was that the slope of a regression to those residuals is zero. In case of the model efficiencies for the calibration period 1976-1990 and the verification period 1991-2005 the null hypothesis was rejected at a significance level of 0.9 or higher. In the case of the model efficiencies for the calibration period 1991-2005 the decrease was not significant at the 90% level, and in the case of the verification period



Figure 3. Nash Sutcliffe model efficiencies *ME* plotted versus catchment area. (a) Calibration, (b) verification, (c) ratio of verification and calibration efficiencies, and (d) standard deviation of model efficiencies within catchments scale classes. Solid lines in Figures 3a-3c show mean efficiencies for a scale range of $3-30 \text{ km}^2$, $30-300 \text{ km}^2$, etc. Black and gray dots represent the model performance of catchments calibrated to 1976-1990 and 1991-2005, respectively.

1991–2005 there was even a slight increase. However, the statistical test should be treated with care as the few very small and very large catchments dominate the test statistics.

[25] For interpretation, we have plotted the Nash-Sutcliffe model efficiencies ME of the verification periods against the ratio of liquid precipitation and total precipitation (Figure 4). A ratio of 0 means that all the precipitation falls as snow, while a ratio of 1 means that all the precipitation falls as rain. There is a clear decrease in model performance with increasing proportion of rain. For the mountainous catchments about 30% of precipitation falls as rain (70% as snow) and the model performance is of the order of 0.8. For the lowland catchments about 90% of precipitation falls as rain (10% as snow) and the model performance varies around 0.6. Additionally, the variability of the model performance between catchments is smaller for snowdominated mountainous catchments. Clearly the hydrological regime has a stronger effect on model performance than catchment scale per se. The hydrologic regime and catchment scale are not well correlated in Austria.

[26] As a second measure of model performance, the volume errors *VE* have been plotted in Figure 5 against catchment scale. It is clear from Figure 5 that the median volume errors are not related to the catchment area, neither for the calibration period nor for the verification period, and the same applies to the differences of the two. However, the scatter of the volume errors does change with scale. For the verification period 1976–1990, the standard deviation of the volume errors decreases from 0.14 to 0.04 as one moves up in scale from 10 to 30 km² to larger than 10,000 km², and there is a similar decrease for the other verification period and the calibration periods. The Breusch-Pagan test



Figure 4. Nash-Sutcliffe model efficiencies ME plotted versus the long-term ratio of liquid rainfall to precipitation. Black and gray dots represent the model performance of catchments calibrated to 1976–1990 and 1991–2005, respectively.

indicates that for all the cases the decrease of the standard deviations of the volume error with scale is significant at the 90% level or higher. This implies that the model simulates the long-term water balance more consistently as one goes up in scale. It is likely that this scale effect is related to the larger number of stations for input data available for each catchment as one moves up in catchment scale.

4.2. Do Model Parameters Change With Catchment Scale?

[27] In Figure 6 some of the calibrated parameters of the snow model and the soil moisture accounting scheme have been plotted against catchment area. Figure 6a shows that the degree-day factors DDF vary between 1 and 2 mm/(d °C) for most of the catchments. The upper envelope decreases with catchment area. The largest DDF values of 2.5 occur around 500 km², while for the large Danube catchments the DDF values are about 1.5. Spatial patterns of the DDF are given by Parajka et al. [2007b, Figure 9]. The largest DDF values are found for catchments in southwestern Austria and for some catchments in northern Austria. Southwestern Austria is a high mountain region where large snow densities and hence high melt rate can be expected. In northern Austria, rain-on-snow events are frequent [Merz and Blöschl, 2003]. These events, for a given air temperature, tend to produce larger melt rates than radiation-dominated snowmelt events, which explains the larger DDF values. Catchments in these two regions are usually smaller than a few hundred square kilometers. Larger catchments in Austria extend from the Alpine region to the lower parts of Austria. Because of the spatiotemporal variability of the snowpacks, smaller melt rates can be expected and rain-of-snow events are less important, which results in smaller DDF values. This suggests that the changes in the DDF with catchment scale, while significant, are not primarily a scale effect per se but a result of the temporal and spatial variability in the dominant processes in different parts of the catchments. The snow correction factor SCF (not shown here) similarly shows little scale dependence, but some of the small catchments in the high alpine area give larger than average *SCF* because of the wind catch deficit of snowfall.

[28] The parameters of the soil moisture accounting scheme, *FC*, *LP*, and *B*, do not show much scale dependence. A decrease in the scatter when moving up in scale from 1000 to 100,000 km² is found for maximum soil moisture storage *FC*, but this may be due to the small number of catchments for the larger scales (Figure 6b). The nonlinearity parameter *B* (Figure 6c) does not show any scale dependence but shows an increase in the calibrated values for the larger catchments when moving from the 1976–1990 calibration period to the 1991–2005 period. The increase is likely related to nonstationarities in the hydrologic conditions.

[29] The snow and soil moisture accounting processes can be considered local-scale processes because at every point in a catchment, snow accumulation, snowmelt, soil moisture replenishment through precipitation, and soil moisture depletion by evapotranspiration occur rather independently. Of course, there are atmospheric redistribution processes of moisture across the landscape, but these do not depend on catchment size. One would therefore assume the entire catchment to simply represent the lumped response of the local response, so the scale effects may be small. In contrast, runoff routing on the hillslopes and in the streams is a lateral rather than a local processes, so one would expect that the associated parameters of the response and transfer functions may be more strongly dependent on scale. In Figure 7 selected parameters of the response and transfer function have been plotted against catchment area in a similar fashion as in Figure 6. There exists an upper envelope on K_0 that decreases with catchment area and a lower envelope of the parameter values on K_1 that increases with catchment area. The scale effect of the storage coefficient K_1 is not surprising because large catchments usually do have a much slower response than small catchments and are never very flashy. This is a reflection of the space-time variability of runoff generation and routing processes in different subcatchments that tend to smooth runoff response. It is interesting that K_0 decreases with scale. K_0 is the storage coefficient of an outlet from the upper zone. It is possible that the smaller K_0 values in the larger catchments reflect the contribution of subcatchments, which produce flashy response once the storage threshold LS_{uz} is exceeded. Also, the decrease in K_0 may reflect the contribution of nearstream floodplains to runoff, which may be larger in large catchments. This, however, only occurs rarely as suggested by the large LS_{uz} values in the large catchments (not shown here). LS_{uz} varies from 30 to 60 mm for 100 km², while LS_{uz} is always larger than 55 mm for catchments larger than 100,000 km². There is no scale effect on K_2 , which controls base flow. This is not surprising, as one would expect K_2 to depend mainly on geology.

4.3. Does Model Performance Change With the Length of the Calibration Period?

[30] To analyze the effect of the length of the calibration period on model performance, the model was calibrated for periods of 1, 3, 5, 10, 15, and 30 years. All of the periods were nonoverlapping starting in 1976; that is, for the 30-year calibration period the model was calibrated to the period 1976–2005; for the 15-year periods the model was calibrated to 1976–1990 and 1991–2005; for the 10-year



Figure 5. Volume errors *VE* plotted versus catchment area. (a) Calibration, (b) verification, (c) differences of verification and calibration efficiencies, and (d) standard deviation of model efficiencies for catchments scale classes. Solid lines in Figures 5a-5c show mean volume error for a scale range of $3-30 \text{ km}^2$, $30-300 \text{ km}^2$ etc. Black and gray dots represent the model performance of catchments calibrated to 1976-1990 and 1991-2005, respectively.

periods the model was calibrated to 1976–1985, 1986–1995, and 1996–2005; and so on. The records were split into periods of the same length for verification.

[31] To analyze the dependence of model performance on the length of the calibration period, we calculated the mean model efficiency of the different nonoverlapping periods of the same length for each catchment and calculated the spatial mean of all catchments in the study region, termed spatial mean of model efficiencies $Mean_{sp}(ME)$ (Figures 8a and 8c). As a second measure, we calculated the standard deviations of model efficiency of different calibration periods and plotted the spatial mean of all catchments, termed spatial mean of temporal standard deviation $Mean_{sp}(Sdev_t)$, against the length of the calibration period (Figures 8b and 8d). For the calibration case, the spatial mean of the model efficiencies $Mean_{sp}(ME)$ decreases with increasing length of the calibration period, as would be expected (Figure 8a). Clearly, longer hydrological time series are likely to contain more diverse hydrological situations and it is more difficult to represent them equally well with the same parameter set. In contrast, the mean verification efficiencies increase with increasing length of the calibration period. Apparently, with the more general parameter sets obtained from the longer calibration periods, more diverse hydrological conditions can be simulated well, perhaps even those that have not been observed during the calibration period. Similarly, longer calibration periods also result in a smaller variability of model performance, when calibrating the model to different periods of the same length. The variability of model efficiencies between differ-



Figure 6. Model parameters (degree-day factor (*DDF*), maximum soil moisture storage (*FC*), and nonlinearity parameter of runoff generation (*B*)), plotted versus catchment area. Black and gray dots represent the model parameters of catchments calibrated to 1976-1990 and 1991-2005, respectively.

ent calibration periods, as expressed by $Mean_{sp}(Sdev_t(ME))$, is larger for the verification than for calibration case and decreases strongly with increasing calibration length (Figure 8b). This means that when using short calibration periods one may expect slightly larger model efficiencies than when using longer calibration periods, but it is more likely that the model performance is much lower in the verification case.

[32] Figures 8c and 8d show the statistics of the volume errors as a function of the length of the calibration period. There is a similar trend of spatial mean errors to get closer to zero with increasing years of calibration, for both the calibration and the verification periods, although the trend for the calibration period is not so strong (Figure 8c). The spatial mean of the temporal standard deviation of the volume errors of using different calibration periods decreases with increasing calibration period. The decrease is much stronger for the verification period (Figure 8c). The mean standard deviation is much higher for verification than for calibration. This illustrates the ability of calibration of reducing bias.

[33] The spatial variability of the temporal standard deviation between different calibration periods is analyzed in more detail in Figure 9. For each catchment the temporal standard deviation of model performance and volume errors of using different periods for calibration are calculated and the cumulative frequency counts is plotted for the calibration case (Figures 9a and 9c) and the verification case (Figures 9b

and 9d). The temporal standard deviation of the model efficiencies between the different calibration periods increases with decreasing length of the calibration period (Figures 9a and 9b). For example, for about 200 catchments out of the 269 catchments, the temporal standard deviation of the calibration model efficiencies between the two 15-year periods is about 0.04, while the corresponding one for the thirty 1-year calibration periods is about 0.08. The increase in the scatter is much larger for the verification period (Figure 9b). A similar trend can be found for the temporal standard deviations of the volume errors VE (Figures 9c and 9d).

[34] It is interesting that the difference in the frequency counts of the temporal standard deviations between the 1-year calibration periods and the 3-year calibration periods is larger than those between the 3-year and the 30-year calibration periods. This means that when using only 1 year of calibration, one will calibrate the model parameters to rather specific hydrological situations, e.g., the rather dry year 2003 in Europe, and the likelihood that one particular catchment performs poorly in the predictive mode using these calibrated parameters is rather high. Using more than 3 years of calibration, one tends to sample more diverse hydrological conditions and the likelihood that one particular catchment performs poorly in the predictive mode is much lower.

[35] While most time scale effects are continuous, the 15-year period in the case of *VE* seems to be an exception



Figure 7. Model parameters (storage coefficients K_0 , K_1 , K_2), plotted versus catchment area. Black and gray dots represent the model parameters of catchments calibrated to 1976–1990 and 1991–2005, respectively. Dashed lines show envelopes.

(Figures 8d and 9d). For most catchments, the temporal standard deviation of the verification volume errors for the 15-year periods is slightly larger than those of the 10- and 5-year periods. This is likely due to nonstationarities in the hydrologic conditions. Air temperatures have significantly increased in the alpine region over the past years [*Auer et al.*, 2007], resulting in more evaporation and lower runoff volumes. The model calibrated to the first 15-year period tends to overestimate runoff volumes of the more recent period and vice versa. Note that minimizing volume errors are not an explicit part of the objective function used to calibrate the model.

4.4. Do Model Parameters Change With the Length of the Calibration Period?

[36] Model parameters of conceptual models, such as the model used here, are designed to represent the hydrological catchment characteristics, so the parameter values should not depend on the modeling period. On the other hand, it is clear that model parameters are associated with uncertainty and may change if different periods are used for calibration. This ambiguity has serious impacts on parameter and predictive uncertainty [e.g., *Beven and Binley*, 1992] and limits the applicability of conceptual models, e.g., for the simulation of land use or climate change scenarios, or for regionalization studies [*Wagener*, 2007]. Here we address this problem by analyzing if the correlation of model parameters between different calibration periods changes with using calibration periods of varying length.

[37] In Figure 10 the spatial mean of the correlation coefficients of model parameters between different calibration periods are plotted against the length of the calibration period. The spatially averaged correlation coefficients differ between model parameters. *B* and K_1 are better correlated than *DDF* and K_2 (Figure 10), which implies that *DDF* and K_2 are associated with higher uncertainties. This is consistent with analyses of *Parajka et al.* [2007a, 2007b], who also associated higher uncertainty to these parameters.

[38] The mean correlation coefficient increases with increasing length of the calibration period for all parameters. Apparently, the longer the calibration period the better the parameters can be identified, because more diverse situations have been sampled. This suggests that all model parameters of Figure 10 reflect, to some degree, the particular hydrological conditions of the calibration period. For calibration periods between 3 and 5 years, the increase in the average correlation coefficients tends to flatten out for most parameters, which suggest that when using 5 years for calibration diverse hydrological conditions are sampled and the calibrated parameters values are likely to be robust enough to be used in a predictive mode. This is line with the findings of the dependence of model performance on the length of the calibration period in section 4.3. However, for the parameters DDF and K_2 , which are associated with higher uncertainties, the mean correlation coefficient increases monotonically beyond 5 years of calibration and the increase is stronger than for the parameters B and K_1 ,



Figure 8. (a and c) Spatial mean calibration (black line) and verification (gray line) efficiencies *ME* and volume errors *VE*, and (b and d) spatial mean of the temporal standard deviation of *ME* and *VE* versus length of the calibration period (1, 3, 5, 10, 15, 30 years).



Figure 9. Cumulative frequency counts of the temporal standard deviation between different calibration periods of (a) calibration model efficiencies, (b) verification model efficiencies, (c) calibration volume errors, and (d) verification volume errors.



Figure 10. Spatial mean of the correlation coefficients of model parameters between different calibration periods plotted against the length of the calibration period (1, 3, 5, 10, 15 years).

which are associated with lower uncertainties. This implies that the use of 3-5 years for calibration may result in acceptable model performance, with respect to simulating runoff, for calibration and verification, but internal state variables and fluxes, such as soil moisture and fast and slow runoff components, may still be associated with higher uncertainties, due to uncertainty of the related model parameters, than when using longer calibration periods.

5. Discussion and Conclusions

5.1. Does Model Performance Change With Catchment Scale?

[39] The simulation results suggest that the model efficiencies increase over the scale range of 10 and $10,000 \text{ km}^2$. There are a number of factors that may contribute to this scale effect. First, and perhaps most important, the average number of rain gauges per catchment increases with catchment scale. Only half of the catchments smaller than 30 km² contain a rain gauge, while the catchments larger than 1000 km² contain 10 of more stations, on average. Clearly, the larger the sampling density the better rainfall can be estimated [Skøien et al., 2003; Skøien and Blöschl, 2006], so one would expect more reliable runoff simulation. Second, the scale effect may also be related to a simpler rainfall-runoff relationship in larger catchments. As Sivapalan [2003] pointed out, much of the small-scale complexity will integrate, so the model structure used may be better suited at the larger scales. However, the difference in model performance across scales is not very large. For very large catchment scales beyond 10,000 km² the efficiencies drop, but this may not be significant because of the small number of catchments. Also, part of the catchment area of these very large catchments is outside Austria and the input data have been extrapolated from stations within Austria, so the drop in efficiency may also be related to a decrease in data quality.

[40] It is interesting that the hydrological regime appears to have a stronger effect on model performance than catchment scale per se. In the snow-dominated mountainous catchments the model performance is significantly higher than in the rain-dominated lowland catchments. Since small catchments exist in both the mountains and the lowlands, there is no net effect on catchment scale. In the most snow dominated catchments, one can argue that the hydrology is actually relatively simple: The snow just accumulates in the winter, and melts in the spring, so the soil component of the model does not produce a lot of dynamics other than the seasonal pattern. In contrast, in the rain-dominated catchments the soil component may play a more important role in controlling the runoff patterns, depending on the soil moisture status that may vary immensely. The analysis of the model performance with respect to the hydrological regime suggests that the variability of model performance between catchments is much smaller for snow-dominated regimes than for rainfall-dominated regimes. This is in line with the analysis of *Haddeland et al.* [2002] on the influence of the spatial resolution on simulated streamflow in a macroscale hydrologic model. They found that snow-dominated catchments are less sensitive to the aggregation of the inputs and model parameters than are rainfall-dominated catchments.

[41] The scatter of the model performances decreases with catchment scale, particularly the volume errors. This implies that the model simulates the long-term water balance more consistently as one goes up in scale. This scale effect is clearly related to the larger number of rain gauges available for each catchment as one moves up in catchment scale. This effect is quite apparent and points up the importance of getting rainfall inputs right in hydrological modeling. An additional factor may be that in large catchments, the water balance will be more often closed, as deep percolation and/or other losses are relatively less important.

[42] The results seem to confirm that modeling large catchments is easier (or, at least, it is easier to get good results) than for small ones. In the conceptual modeling world, one often forces models with mean areal quantities (most notably, mean areal precipitation). This is the approach often used in operational flood forecasting. If, as this implies, getting the spatial mean right is more important than the spatial variations, which are ignored, then the larger the catchment, the more stations, and the smaller the error in mean areal precipitation. This would lead directly to an inference that the errors should decrease as catchment size increases. If this averaging effect in the forcings dominates, then the result that the errors in runoff simulations are smaller for larger catchments does not necessarily imply simpler rainfall-runoff relationship. It does imply that those effects are smaller than the averaging effect on the inputs.

5.2. Do Model Parameters Change With Catchment Scale?

[43] The calibrated model parameters do not change much with catchment scale, but scale trends exist. For

example, the upper envelope of the degree-day factor DDF decreases with catchment area. It is believed that the decrease is not a scale effect per se but a result of temporal and spatial variability of the snowpack in larger catchments and a change in the dominant processes in different parts of the catchments. The lower envelope of the storage parameter of the upper zone reservoir K_1 increases with catchment area. This is not surprising, as large catchments usually do have a much slower response than small catchments and are never very flashy. In contrast, the response times of small catchments can be small or large, depending on geology and soils [Uchida et al., 2001].

[44] *FC*, *B*, and K_2 do not show any catchment scale effect that is in line with the results of *Bergström and Graham* [1998] for the Baltic Sea catchments. However, it is possible that the effects of encountering increasingly more variability within a catchment and increasingly more smoothing with increasing catchment scale may result in a relatively scale invariant functional behavior of catchments, so the parameters remain relatively stable over a wide range of catchment scales.

[45] It is interesting to put the apparent scale effects that are related to regional differences rather than to scale per se into the context of *Beven*'s [2000] idea of uniqueness of place, which emphasizes the importance of the characteristics and responses of a location to rather an ensemble of parameters that varies randomly across the landscape. It is clear that regional hydrologic patterns are a major control on general scale effects.

5.3. Does Model Performance Change With the Length of the Calibration Period?

[46] The results indicate that calibration efficiencies decrease and verification efficiencies increase with the number of years available for calibration. The effect of decreasing model performance with decreasing number of years available for calibration is probably an effect of less diversity in runoff conditions (i.e., fewer years of high, medium, and low runoff). The decrease in calibration efficiencies is largest between 1 and 5 years, and similarly, the increase in verification efficiencies is largest between 1 and 5 years. This suggests that a calibration period of 5 years captures most of the temporal hydrological variability, so this would be the minimum for achieving a reasonable predictive model performance. Longer periods will only moderately improve the results. For the 15-year calibration period the average verification efficiency is very similar to the average calibration efficiency. This means that for 15 years, the different situations have been sampled exhaustively. The results are consistent with the results of Perrin et al. [2008]. For their set of 900 catchments from Australia, France, and United States, the increase of verification model efficiencies with increasing length of the calibration period is similar to the results of the Austrian case study. These results point to the need for longer calibration periods than are commonly used, certainly longer than the 1-year calibration period recommended by Gan et al. [1997].

5.4. Do Model Parameters Change With the Length of the Calibration Period?

[47] The correlation of model parameters between different calibration periods, as a measure of the degree to which parameters can be identified, increases with increasing length of the calibration period. This suggests that the longer the calibration period the better parameters can be identified. For some parameters the increase in the correlation flattens out at a calibration length of about 5 years. This is in line with the results of the temporal scale effects on model performance. However, for other parameters, the correlation increases beyond 5 years of calibration. This suggest that although runoff may be simulated acceptably, some parameters may not be well defined and hence internal state variables and fluxes may still be associated with larger uncertainties. Longer periods for calibration are needed. However, longer periods for calibration are not always available. An alternative way of reducing the uncertainty of calculated internal state variables and fluxes is to use additional information for calibration [Seibert and McDonnell, 2002], such as snow data [Parajka et al., 2007a], soil moisture data from remote sensing [Parajka et al., 2006], and spatial correlations of model parameters [Parajka et al., 2007b].

5.5. Outlook

[48] This paper has focused on space and time scale trends of model performance and model parameters using averaged statistics of a large number of catchments in Austria. Some of the findings indicate that spatial-scale effects are related to regional differences rather than to scale per se. Similarly, some of the temporal-scale effects are probably related to transient hydrological conditions. In future studies it may hence be worthwhile to analyze the regional patterns of model performance and model parameters and link regional differences in the dominant processes to apparent scale effects. Similarly, it will be of interest to analyze the sensitivity of model parameters to changing catchment and climate conditions [Wagener et al., 2003; Wagener, 2007; Blöschl et al., 2007]. Finding parameter sets that represent future catchment and climate conditions is certainly a challenge hydrologists need to tackle in the years to come.

[49] **Acknowledgments.** We would like to thank FWF project P18993-N10 for financial support. We would also like to thank the Austrian Hydrographic Services for providing the hydrographic data. We want to thank two anonymous reviewers and Dennis Lettenmaier for very useful comments on the manuscript.

References

- Auer, I., et al. (2007), HISTALP—Historical instrumental climatological surface time series of the greater Alpine region 1760–2003, *Int. J. Climatol.*, 27, 17–46, doi:10.1002/joc.1377.
- Bergström, S. (1991), Principles and confidence in hydrological modelling, Nord. Hydrol., 22, 123–136.
- Bergström, S. (1992), The HBV model—Its structure and applications, *Rep.* 4, 32 pp., Swed. Meteorol. and Hydrol. Inst., Norrköping, Sweden.
- Bergström, S., and L. P. Graham (1998), On the scale problem in hydrological modelling, J. Hydrol., 211, 253–265, doi:10.1016/S0022-1694(98)00248-0.
- Beven, K. (1989), Changing ideas in hydrology—The case of physically based models, *J. Hydrol.*, *105*, 157–172, doi:10.1016/0022-1694(89)90101-7.
- Beven, K. (1991), Scale considerations, in *Recent Advances in the Modeling of Hydrologic Systems*, edited by D. S. Bowles and P. E. O'Connell pp. 357–371, Kluwer, Dordrecht, Netherlands.
- Beven, K. J. (2000), Uniqueness of place and process representations in hydrological modelling, *Hydrol. Earth Syst. Sci.*, *4*, 203–214.
- Beven, K. J. (2001a), *Rainfall-Runoff Modelling—The Primer*, 360 pp., John Wiley, Hoboken, N. J.
- Beven, K. J. (2001b), How far can we go with distributed hydrological modelling?, *Hydrol. Earth Syst. Sci.*, *5*, 1–12.

- Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, 6, 279–298, doi:10.1002/hyp.3360060305.
- Blöschl, G., and M. Sivapalan (1995), Scale issues in hydrological modeling— A review, *Hydrol. Processes*, *9*, 251–290, doi:10.1002/hyp.3360090305.
- Blöschl, G., R. B. Grayson, and M. Sivapalan (1995), On the representative elementary area (REA) concept and its utility for distributed rainfallrunoff modelling, *Hydrol. Processes*, 9, 313–330, doi:10.1002/hyp. 3360090307.
- Blöschl, G., S. Ardoin-Bardin, M. Bonell, M. Dorninger, D. Goodrich, D. Gutknecht, D. Matamoros, B. Merz, P. Shand, and J. Szolgay (2007), At what scales do climate variability and land cover change impact on flooding and low flows?, *Hydrol. Processes*, 21, 1241–1247, doi:10.1002/ hyp.6669.
- Breusch, T., and A. Pagan (1979), A simple test of heteroskedasticity and random coefficient variation, *Econometrica*, 47, 1287–1294, doi:10.2307/1911963.
- Duan, Q., S. Sorooshian, and V. K. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour*: *Res.*, 28, 1015–1031, doi:10.1029/91WR02985.
- Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2008), Understanding catchment behavior through stepwise model concept improvement, *Water Resour. Res.*, 44, W01402, doi:10.1029/2006WR005563.
- Freer, J., K. J. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resour. Res.*, 32, 2161–2173.
- Gan, T. Y., E. M. Dlamini, and G. F. Biftu (1997), Effects of model complexity and structure, data quality and objective function on hydrologic modelling, J. Hydrol., 192, 81–103, doi:10.1016/S0022-1694(96)03114-9.
- Gottschalk, L., S. Beldring, K. Engeland, L. Tallaksen, N. R. Sælthun, S. Kolberg, and Y. Motovilov (2001), Regional/macroscale hydrological modelling: A Scandinavian experience, *Hydrol. Sci. J.*, 46, 963–982.
- Götzinger, J., and A. Bárdossy (2008), Generic error model for calibration and uncertainty estimation of hydrological models, *Water Resour. Res.*, 44, W00B07, doi:10.1029/2007WR006691.
- Grayson, R., and G. Blöschl (2000), Spatial modelling of catchment dynamics, in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, edited by R. Grayson and G. Blöschl, chap. 3 pp. 51–81, Cambridge Univ. Press, Cambridge, U. K.
- Gupta, H., S. Sorooshian, and P. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, 34(4), 751–763, doi:10.1029/97WR03495.
- Haddeland, I., B. V. Matheussen, and D. P. Lettenmaier (2002), Influence of spatial resolution on simulated streamflow in a macroscale hydrologic model, *Water Resour. Res.*, 38(7), 1124, doi:10.1029/2001WR000854.
- Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model?, *Water Resour: Res.*, 29, 2637–2649, doi:10.1029/93WR00877.
- Klemeš, V. (1986), Operational testing of hydrological simulation models, *Hydrol. Sci, J.*, 31(1), 13–24.
- Merz, R., and G. Blöschl (2003), A process typology of regional floods, Water Resour. Res., 39(12), 1340, doi:10.1029/2002WR001952.
- Mészároš, I., P. Miklánek, and J. Parajka (2002), Solar energy income modelling in mountainous areas, in *Interdisciplinary Approaches in Small Catchment Hydrology: Monitoring and Research—Proceedings* of the 9th Conference of the European Network of Experimental and *Representative Basins*, edited by L. Holko and P. Miklanek, pp. 127– 135, United Nations Educ., Sci., and Cult. Organ, Paris.
- Montanari, A. (2005), Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfallrunoff simulations, *Water Resour. Res.*, 41, W08406, doi:10.1029/ 2004WR003826.
- Montanari, A. (2007), What do we mean by "uncertainty"? The need for a consistent wording about uncertainty assessment in hydrology, *Hydrol. Processes*, 21(6), 841–845, doi:10.1002/hyp.6623.
- Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models: Part I. A discussion of principles, *J. Hydrol.*, 10, 282–290, doi:10.1016/0022-1694(70)90255-6.
- Oudin, L., V. Andreassian, C. Perrin, C. Michel, and N. Le Moine (2008), Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, *Water Resour. Res.*, 44, W03413, doi:10.1029/2007WR006240.
- Parajka, J., R. Merz, and G. Blöschl (2005), A comparison of regionalisation methods for catchment model parameters, *Hydrol. Earth Syst. Sci.*, 9, 157–171.

- Parajka, J., V. Naeimi, G. Blöschl, W. Wagner, R. Merz, and K. Scipal (2006), Assimilating scatterometer soil moisture data into conceptual hydrologic models at the regional scale, *Hydrol. Earth Syst. Sci.*, 10, 353–368.
- Parajka, J., R. Merz, and G. Blöschl (2007a), Uncertainty and multiple objective calibration in regional water balance modelling, *Hydrol. Pro*cesses, 21, 435–446, doi:10.1002/hyp.6253.
- Parajka, J., G. Blöschl, and R. Merz (2007b), Regional calibration of catchment models: Potential for ungauged catchments, *Water Resour. Res.*, 43, W06406, doi:10.1029/2006WR005271.
- Perrin, C., C. Michel, and V. Andréassian (2001), Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, *J. Hydrol.*, 242, 275–301, doi:10.1016/S0022-1694(00)00393-0.
- Perrin, C., V. Andréassian, C. Rojas Serna, T. Mathevet, and N. Le Moine (2008), Discrete parameterization of hydrological models: Evaluating the use of parameter sets libraries over 900 catchments, *Water Resour. Res.*, 44, W08447, doi:10.1029/2007WR006579.
- Refsgaard, J. C. (2000), Towards a formal approach to calibration and validation of models using spatial data, in *Spatial Patterns in Catchment Hydrology: Observations and Modelling*, edited by R. Grayson and G. Blöschl, chap. 13 pp. 329–354, Cambridge Univ. Press, Cambridge, U. K.
- Refsgaard, J. C., and J. Knudsen (1996), Operational validation and intercomparison of different types of hydrological models, *Water Resour*. *Res.*, 32(7), 2189–2202.
- Refsgaard, J. C., J. P. van der Sluijs, J. Brown, and P. van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Adv. Water Resour.*, 29, 1586–1597, doi:10.1016/j.advwatres.2005.11.013.
- Seibert, J. (1997), Estimation of parameter uncertainty in the HBV model, Nord. Hydrol., 28(4–5), 247–262.
- Seibert, J., and J. J. McDonnell (2002), On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Resour. Res.*, 38(11), 1241, doi:10.1029/ 2001WR000978.
- Singh, V. P. and D. K. Frevert (Eds.) (2001a), *Mathematical Models of Small Watershed Hydrology*, 972 pp., Water Resour. Publ., Highlands Ranch, Colo.
- Singh, V. P. and D. K. Frevert (Eds.) (2001b), Mathematical Models of Large Watershed Hydrology, 914 pp., Water Resour. Publ., Highlands Ranch, Colo.
- Sivapalan, M. (2003), Process complexity at hillslope scale, process simplicity at the watershed scale: Is there a connection?, *Hydrol. Processes*, *17*, 1037–1041, doi:10.1002/hyp.5109.
- Skøien, J. O., and G. Blöschl (2006), Sampling scale effects in random fields and implications for environmental monitoring, *Environ. Monit.* Assess., 114(1-3), 521-552, doi:10.1007/s10661-006-4939-z.
- Skøien, J. O., G. Blöschl, and A. W. Western (2003), Characteristic space scales and timescales in hydrology, *Water Resour. Res.*, 39(10), 1304, doi:10.1029/2002WR001736.
- Uchida, T., K. Kosugi, and T. Mizuyama (2001), Effects of pipeflow on hydrological process and its relation to landslide: A review of pipeflow studies in forested headwater catchments, *Hydrol. Processes*, 15, 2151– 2174, doi:10.1002/hyp.281.
- Uhlenbrook, S., J. Seibert, C. Leibundgut, and A. Rodhe (1999), Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure, *Hydrol. Sci. J.*, 44(5), 779–797.
- Wagener, T. (2007), Can we model the hydrological impacts of environmental change?, *Hydrol. Processes*, 21, 3233–3236, doi:10.1002/ hyp.6873.
- Wagener, T., N. McIntyre, M. J. S. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, *Hydrol. Processes*, 17(2), 455–476, doi:10.1002/hyp.1135.
- Weglarczyk, S. (1998), The interdependence and applicability of some statistical quality measures for hydrological models, *J. Hydrol.*, 206, 98–103, doi:10.1016/S0022-1694(98)00094-8.
- World Meteorological Organization (1986), Intercomparison of models of snowmelt runoff, Oper. Hydrol. Rep. 23, Geneva, Switzerland.

G. Blöschl, R. Merz, and J. Parajka, Institute of Hydraulic Engineering and Water Resources Management, Vienna University of Technology, Karlsplatz 13/223, Vienna A-1040, Austria. (merz@hydro.tuwien.ac.at)