The Quest for an Improved Dialog Between Modeler and Experimentalist

Jan Seibert

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

Jeffrey J. McDonnell

Department of Forest Engineering, Oregon State University, Corvallis, Oregon

Multi-criteria calibration of runoff models using additional data, such as groundwater levels or soil moisture, has been proposed as a way to constrain parameter values and to ensure the realistic simulation of internal variables. Nevertheless, in many cases the availability of such 'hard data' is limited. We argue that experimentalists working in a catchment often have much more knowledge of catchment behavior than is currently used for model calibration and testing. While potentially highly useful, this information is difficult to use directly as exact numbers in the calibration process. We present a framework whereby these 'soft' data from the experimentalist are made useful through fuzzy measures of model-simulation and parameter-value acceptability. The use of soft data is an approach to formalize the exchange of information and calibration measures between experimentalist and modeler. This dialog may also greatly augment the traditional and few 'hard' data measures available. We illustrate the value of 'soft data' with the application of a three-box conceptual model for the Maimai catchment in New Zealand. The model was calibrated against hard data (runoff and groundwater-levels) as well as a number of criteria derived from the soft data (e.g., percent new water, reservoir volume). While very good fits were obtained when calibrating against runoff only (model efficiency = 0.93), parameter sets obtained in this way showed, in general, poor internal consistency. Inclusion of soft-data criteria in the model calibration process resulted in lower model-efficiency values (around 0.84 when including all criteria) but led to better overall performance, as interpreted by the experimentalist's view of catchment runoff dynamics.

INTRODUCTION

Many different conceptual models of catchment hydrology have been developed during the last few decades [*Singh*, 1995]. These models have become valuable tools for water management problems (*e.g.*, flood forecasting, water balance studies and computation of design floods). The increasing awareness of environmental problems has given additional impetus to hydrological modeling. Runoff models have to meet new requirements when they are intended to deal with problems such as acidification, soil erosion and land degradation, leaching of pollutants, irrigation, sustainable water-resource management or possible consequences of land-use or climatic changes. Linkages to geochemistry, ecology, meteorology and other sciences must be considered explicitly and realistic simulations of internal processes become essential.

Calibration of Watershed Models
Water Science and Application Volume 6

10/1029/006WS22

Despite much effort [*Hornberger and Boyer*, 1995], hydrological modeling is faced by fundamental problems such as the need for calibration and the equifinality of different model structures and parameter sets (*i.e.*, the phenomenon that equally good model simulations might be obtained in many different ways, *Beven*, 1993). These problems are linked to the limited data availability and the natural heterogeneity of watersheds [*e.g.*, *Beven*, 1993; *O'Connell* and *Todini*, 1996; *Bronstert*, 1999]. Problems also can be related to the procedures used for model testing. Traditional tests such as split-sample tests are often not sufficient to evaluate model validity or to assess the pros and cons of different model approaches. More powerful and rigorous methods for model calibration and testing are clearly required [*Kirchner et al.*, 1996; *Mroczkowski et al.*, 1997, *Kavetski et al.*, this issue].

Multi-criteria Model Calibration

Manual calibration of a model by trial and error is a time-consuming method and results may be subjective. This is particularly true when calibrating against more than one hydrological variable. Therefore, various automatic calibration methods have been developed [*Sorooshian* and *Gupta*, 1995; *Gupta et al.*, this volume, *Duan*, this volume]. In general, these methods allow for a quick and 'objective' calibration. On the other hand there is the danger that model calibration becomes a 'dumb' curve fitting exercise. By this we mean that unlike the manual calibration process where the hydrologist will implicitly make use of his/her process knowledge (e.g. by examining different aspects of the hydrograph or the simulation of internal variables), in the automatic approach, only explicitly stated criteria are considered. Thus, there appears to be a need for methods to infuse hydrological reasoning into the automatic calibration process.

Two 'ways forward' on the equifinality issue include: (1) making more detailed use out of the comparison between simulated and observed runoff series [e.g., *Boyle et al.*, 2000; this volume; *Burges*, this volume, *Freer*, this volume] or (2) incorporating additional data into the model calibration procedure. Boyle *et al.* [2000; this volume], followed the first approach and proposed a method to combine the strengths of manual and automatic calibration methods. Recognizing that one goodness-of-fit measure is not sufficient to judge the fit of observed and simulated runoff series, they examined different parts of the hydrograph separately. Our work, and this chapter, complements the work of Boyle *et al.* [2000; this volume] by exploring the second approach: *i.e.*, the utilization of additional data in the model calibration process.

The need to utilize additional data for model calibration and testing has been emphasized by others in the recent years [*de Grosbois et al.*, 1988; *Ambroise et al.*, 1995; *Refsgaard*, 1997; *Kuczera* and *Mroczkowski*, 1998; *Seibert*, 1999; *Meixner* and *Bastidas*, this volume]. Testing runoff models against variables other than simply catchment-outlet runoff is important for two main reasons: (1) in many hydrological questions, and for other sciences such as ecology, it may be of much more interest to know what happens *within* a catchment than at the outlet, and (2) to have confidence in model predictions, which are often extrapolations beyond the testable conditions, it must be ensured that the model not only works, but also does so for the right reasons.

Most parameters of conceptual runoff models need to be determined by calibration. Some parameters may have a physical basis but they are *effective* parameters on the catchment or subcatchment scale. The typical problem is that the information contained in the rainfall-runoff relationship usually does not allow the identification of one unique parameter set. Reducing the number of parameters is an unattractive option because it might transform the conceptual gray-box representation of the rainfall-runoff process into a pure black-box description. Another more attractive way to reduce parameter uncertainty is the use of additional data. Franks *et al.* [1998] demonstrated that the known percentage of saturated areas in the catchment helped to constrain calibrated parameter values and model predictions in an application of TOPMODEL. Seibert [2000] found for an application of the HBV model, that groundwater-level data helped to constrain the parameters of the groundwater routine. However, the worth of additional data varies depending on the kind of data, but also on the structure of the applied model. For instance, Kuczera and Mroczkowski [1998] found that groundwater levels helped little to reduce the parameter uncertainty in a hydrosalinity model, whereas stream salinity data more substantially reduced the uncertainties. Blazkova *et al.* [2002] mapped saturated areas and found that this information influenced optimized parameter values for TOPMODEL, but also that the additional information had only limited effect on constraining prediction bounds for stream discharge.

The Concept of Soft Data

In many cases the amount of available additional data is limited. However, a hydrologist might have a perceptual model [*Beven*, 1993], which is a highly detailed yet qualitative understanding of dominant runoff processes even in situations with limited field measurements. Thus, there exists in addition to hard data (streamflow hydrograph, well record) 'soft data' about catchment hydrology and its inter-

nal 'behavior'. While some groups have used the perceptual model to guide the construction of the model elements, little has been done to use this kind of data in the model calibration. The few to do this include Franks *et al.* [1998] who used maps of surface saturated area to constrain parameter ranges for TOPMODEL runs and Franks and Beven [1997] who used related fuzzy measures for evapotranspiration. Soft data can be defined as qualitative knowledge from the experimentalist that cannot be used directly as exact numbers but that can be made useful when transformed into quantitative data through fuzzy measures of model-simulation and parameter-value acceptability. Soft data may be based on 'hard' measurements but these measurements require some interpretation or manipulation by a hydrologist before being useful in model testing. While fuzzy, these soft measures can be exceedingly valuable for indicating 'how a catchment works'. Fuzzy measures, which implement the concept of partial truth with values between completely true and completely false, have been found to be useful in hydrological model calibration [*Seibert*, 1997; *Aronica et al.*, 1998; *Franks et al.*, 1998; *Hankin* and *Beven*, 1998]. *Aronica et al.* [1998], for instance, used a fuzzy-rule based calibration for a system containing highly uncertain flood information. A fuzzy measure varies between zero and one and describes the degree to which the statement 'x is a member of Y' or, in our case, 'this parameter set is the best possible set' is true.

Different methods are available for automatic optimization. Evolution-based optimization methods have been found to be suitable tools for the calibration of conceptual runoff models [*Wang*, 1991; *Duan et al.* 1992; *Franchini*, 1996; *Kuczera*, 1997; *Yapo et al.*, 1998, *Duan*, this volume]. Genetic algorithms are one class of these methods. The goal of genetic algorithms, originally suggested by Holland [1975; 1992], is to mimic evolution. Parameter sets are encoded to chromosome-like strings and different recombination operators are used to generate new parameter sets. The optimization starts with a population of randomly generated parameter sets. These are evaluated by running the model; those sets that give a better simulation according to some objective function, are given more chances to generate new sets than those sets that gave poorer results. Seibert [2000] used a genetic algorithm to find the true parameter values for a theoretical, error-free test case with synthetic data. For a real-world case, with calibration against observed runoff, he found that parameter values varied considerably for different calibration trials. However, approximately the same model efficiency was achieved in almost every trial. This possibility for different parameter sets in the case of a flat goodness-of-fit surface allows one to utilize the genetic algorithm to evaluate parameter uncertainty using the variation of calibrated parameter values as a measure of parameter identifiability [*Seibert*, 2000]. The genetic algorithm can, thus, provide an indication of parameter uncertainty and serve as an alternative to Monte Carlo approaches like, the Generalized Likelihood Uncertainty Estimation (GLUE) techniques of Freer *et al.* [1996].

In this chapter we present a method for how to use the additional data that often exists in experimental catchments for the calibration of conceptual runoff models. We present a number of 'soft data' measures as means to improve the dialog between modeler and experimentalist. We describe and use the implementation of a genetic algorithm for calibration, as proposed by Seibert [2000], and illustrate these methods for the Maimai watershed in New Zealand. Our main message in this chapter is that additional soft data may be a useful way to ensure that a model of catchment hydrology not only works (for runoff simulation), but also does so for the right process reasons.

MATERIAL AND METHODS

Soft Data

We define soft data as knowledge from the experimentalist that cannot be used directly for model calibration and testing but that can be made useful through fuzzy measures of model-simulation and parameter-value acceptability. It is important to note that soft data may be based on 'hard' measurements that require some interpretation or manipulation by a hydrologist before being useful in model testing. Model simulations may be judged in more process-based, ways when soft data is used compared to when only the hard data is considered For instance, the experimentalist might have some observations concerning the range in which groundwater levels fluctuate within a given zone of the catchment, or conceptual model box (based on field campaign information or observations made over some irregular time periods) or the contribution of rainfall or 'new' water [*McDonnell et al.*, 1991] to peak flow (from event-based isotope tracing studies). Soft data can be used to constrain the calibration by: (1) evaluating the model with regard to simulations for which there might be no hard data available for comparison, and (2) assessing how reasonable the parameter values are, based on field experience. This range of 'reasonable' parameter values might be wide, especially when the parameter values are effective values at some larger scale.

When comparing model simulations or parameter values with soft data, there may be a relatively wide range of acceptable simulations or values. Furthermore, there might be a range of values that fall between 'fully acceptable' and

'not acceptable', based on the experimentalist's experience in the field and other synoptic measurements. Fuzzy measures of acceptance can be used to consider these ranges [*Franks et al.*, 1998]. For each soft data type, a trapezoidal function (Eq. 1), where the experimentalist is asked to assign values to the variables a_i, is used to compute the degree of acceptance, μ, from the corresponding simulated quantity or parameter value x. This trapezoidal function is a simple way to map experimentalist experience into a quantity, which then can be optimized (Fig. 1). Other functions with different shapes might be used instead of the trapezoidal function.

$$\mu(x)=\begin{cases}0 & if\ x\le a_1\\ \dfrac{x-a_1}{a_2-a_1} & if\ a_1\le x< a_2\\ 1 & if\ a_2\le x< a_3\\ \dfrac{a_4-x}{a_4-a_3} & if\ a_3\le x< a_4\\ 0 & if\ x> a_4\end{cases} \qquad (1)$$

An important point is that that uncertainty exists in the experimentalist's view of the catchment and that data collected in the field have their related uncertainties [*Sherlock et al.*, 2002]. Thus, the trapezoidal function provides a way for the experimentalist to also provide his or her uncertainty bounds on the delivered rules to the modeler.

The general acceptability of a parameter set was defined by three components: (1) the goodness-of-fit measures for the hard data such as the model efficiency [*Nash* and *Sutcliffe*, 1970] for runoff (A_1), (2) the goodness of the model simulations with regard to soft data (*e.g.*, maximum groundwater levels) as quantified using Eq. 1 (A_2), and (3) the acceptability of the parameter values based on the experimentalist's experience (A_3). For all three components, a value of one for A_i corresponds to a perfect fit (or complete acceptability).

The overall acceptability, A, of a parameter set is computed as a weighted geometric mean with the weights n_1, n_2, and n_3 (Eq. 2). A can then be used as optimization criterion.

$$A = A_1^{n1} A_2^{n2} A_3^{n3} \quad with \quad n_1 + n_2 + n_3 = 1 \qquad (2)$$

The selection of the weights in Eq. 2 n_1, n_2, and n_3 determines which solution along the pareto-optimality sub-space will be found. The weights allow placement of more (or less) emphasis on the different types of data. A higher value for n_1, for instance, might be justified if there is much useful and accurate hard data, whereas a smaller value might be appropriate if the hard data consists of only runoff.

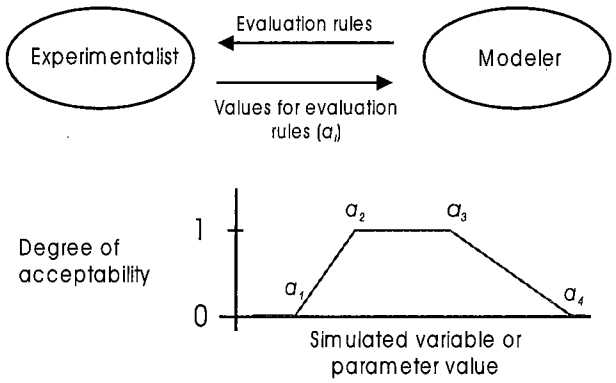

Figure 1. Framework for formalized dialog between experimentalist and modeler using a trapezoidal function as a means of assigning values to the soft data.

Description of the Genetic Algorithm

A genetic algorithm utilizes an evolution of parameter sets with elements of selection and recombination to find optimized parameter sets [*Duan*, this volume]. An initial population of n (set to 50) parameter sets is selected randomly within the parameter space. The 'fitness' of an individual set is quantified as the value of an objective function. A new population (generation) is generated from this population by n times combining two parameter sets, which are chosen randomly but with a higher chance of being picked for sets with a higher 'fitness' (*i.e.*, objective function). From the two parent sets (sets A and B) the new parameter set is generated by applying for each parameter randomly (with some probability, p_i), each of the following four rules: (1) value of set A (p_1=0.41), (2) value of set B (p_2=0.41), (3) random value between the values of set A and set B (alternatively, if both values were equal, a random value close to this value) (p_3=0.16), or, (4) random value within the limits given for the parameter (mutation) (p_4=0.02). The first two rules preserve the values of the preceding generation, whereas the other two rules provide an amount of random search. A balance between these rules is important for the success of the algorithm. However, within reasonable ranges adjustments to the probabilities for the different rules have only minor effects on the performance of the algorithm. Finally the fitness of each set in the new population is evaluated and the new generation replaces the old one. However, the best set is retained if there is no better set in the proceeding generation. This process is repeated for a number of generations.

The results of a genetic algorithm can be improved by combination with a local search method [*Wang*, 1991]. For instance the parameter set found by a genetic algorithm can be used as starting point for a local optimization [*Franchini*,

1996]. In addition to this form of subsequent 'fine-tuning', a local search approach can also be implemented during the 'evolution' process [Seibert, 2000]. At a small probability (p=0.02), the new parameter set is not found by the parameter-by-parameter combinations as described above; instead the new parameter set is the result of a one-dimensional optimization along the line determined by the two parameter sets using Brent's method [*Press et al.*, 1992]. In this chapter we divide the total number of 2500 model runs into 2000 runs for the genetic algorithm and 500 runs for the subsequent local optimization. We use Powell's quadratically convergent method for this multidimensional, local optimization, as described in Press *et al.* [1992].

Our genetic algorithm includes stochastic elements such as the randomly generated initial set of parameter sets and the partly random generation of offsprings during the 'evolution' of parameter sets. Thus, the calibrated parameter values may vary for different calibration trials, when different parameter sets result in similarly good simulations according to the goodness-of-fit measure. This makes this optimization algorithm suitable to address parameter uncertainty using the variation of calibrated parameter values as a measure of parameter identifiability. For the results presented in this study, sixty calibration trials were performed for each goodness-of-fit measure and the best 50 parameter sets were used for further analysis of model performance and parameter identifiability.

The Maimai Watershed

Maimai M8 is a small 3.8 ha headwater catchment located to the east of the Paparoa Mountain Range on the South Island of New Zealand. Slopes are short (<300 m) and steep (average 34°) with local relief of 100-150 m. Stream channels are deeply incised and lower portions of the slope profiles are strongly convex. Areas that could contribute to storm response by saturation overland flow are small and limited to 4-7 % [*Mosley*, 1979; *Pearce et al.*, 1986]. Mean annual precipitation is approximately 2600 mm, producing an estimated 1550 mm of runoff. There were 11 major runoff events during the period of record used for model simulation in this study (August-December, 1987) with a maximum runoff of 6 mm/h. Additional to rainfall and runoff data, groundwater levels extracted from the tensiometer data in McDonnell [1989, 1990], were available for two locations (one in the riparian and one in the hollow zone). Mean monthly values of potential evaporation estimated by L. Rowe [1992, pers.comm.] were distributed using a sine curve for each day [*J. Freer*, 2000, pers. comm.].

The Maimai M8 watershed is a well-studied watershed with ongoing hillslope research by several research teams since the late 1970s. During these studies a very detailed yet qualitative perceptual model of hillslope hydrology evolved (for review see McGlynn *et al.* [2002]).

Conceptual Three-box Model

While this chapter focuses on soft data for multi-criteria calibration, the soft data first helped guide the box-model construction. Our conceptual model is based on the three reservoirs identified from the experimental studies at M8: riparian, hollow and hillslope zones (Fig. 2, Table 1). These zones (or model boxes) display very different groundwater dynamics [*McDonnell*, 1990] and group clearly based on their isotopic characteristics [*McDonnell et al.*, 1991]. Water is simulated to flow from the hillslope zone into the hollow zone and from the hollow zone into the riparian zone. Outflow from the riparian zone forms the flow in the stream. Most importantly, and most novel for this model, is the formulation used to model the unsaturated and saturated storage. Due to the shallow groundwater (groundwater levels 0 – 1.5 m below the ground surface) growth of the (transient) saturated zone occurs at the expense of the unsaturated zone thickness. Thus, a coupled formulation of the saturated and unsaturated storage was used, as proposed by Seibert *et al.* [2002]. In this formulation, the amount of saturated storage determines the maximum space for unsaturated storage. For a more detailed description and equations of the three-box model the reader is referred to [*Seibert* and *McDonnell*, 2002].

Table 1. List of parameters used in the three-box model.

Parameter	Description	Unit
z_{max}	Soil depth [a] [mm]	
c	Parameter corresponding to water content at field capacity divided by porosity	[-]
d	Parameter corresponding to water content at wilting point divided by porosity	[-]
B	Shape coefficient determining groundwater recharge	[-]
$k_{1,riparian}$	Outflow coefficient, riparian box	[h^{-1}]
$k_{1,hollow}$	Outflow coefficient, hollow box, lower outflow	[h^{-1}]
$k_{2,hollow}$	Outflow coefficient, hollow box, upper outflow	[h^{-1}]
$k_{1,hillslope}$	Outflow coefficient, hillslope box	[h^{-1}]
$z_{threshold}$	Threshold storage for contribution from upper outflow in the hollow box	[mm]
p	Porosity [a]	[-]
$f_{riparian}$	Areal fraction of the riparian zone	[-]
f_{hollow}	Areal fraction of the hollow zone	[-]

[a] Different values were allowed for riparian, hollow and hillslope box

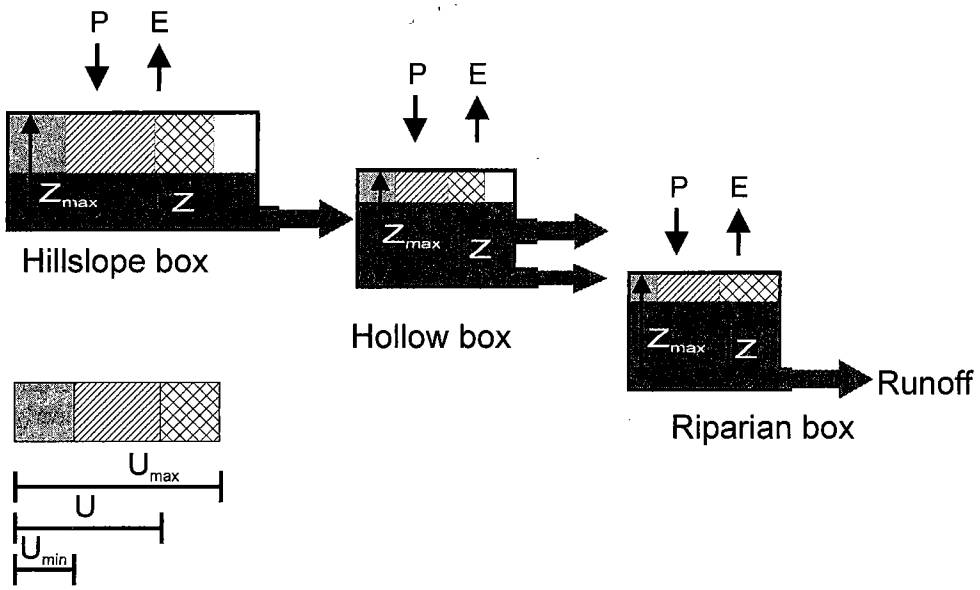

Figure 2. Structure of the three-box model developed for the Maimai M8 watershed including hillslope, hollow and riparian zone reservoirs. (P: precipitation, E: evaporation, z: groundwater level above bedrock, U: unsaturated storage). See also Table 1.

As for any model, several simplifications and assumptions are made to derive this conceptual three-box model [*Seibert* and *McDonnell*, 2002]. The model structure is guided by experimental findings at Maimai. Obviously these simplifications and assumptions are not universally applicable; for other watersheds, a different model structure may be more appropriate (perhaps different box configurations, different number of boxes or different sizes and connections of boxes). The dialogue between experimentalist and modeler using the soft-data framework might guide this construction of conceptual models for particular catchments.

Application of the Soft-Data Framework

For presentation in this chapter we include a subset of the available soft data for demonstration purposes: groundwater levels in the three boxes, the new-water contribution to peak runoff, and some other parameter values. Evaluation rules were developed using Eq. 1 to judge model performance with regard to minimum and maximum groundwater levels as well as the frequency of levels being above a specified level (Table 2). The values for these rules were motivated by field studies reported in McDonnell [1990], McDonnell et al. [1991] and Stewart and McDonnell [1991] for the same August-December 1987 period where groundwater response in the riparian and hollow zones were quantified with recording tensiometers that show distinctly different wetting, filling, draining behavior. Riparian zones were characterized by rapid conversion of tension to pressure potential (*i.e.*, rapid conversion of unsaturated zone to a saturated zone by storage filling and water table rise from below). Water tables were sustained in this zone for 1-2 days following the cessation of rainfall. These data provided the soft data measures for minimum and maximum groundwater levels and frequency of levels above a specified level (listed in Table 2). The hollow zone response was much more sensitive to rainfall inputs: conversion of unsaturated zone to transient saturation occurred within the few hours of the hydrograph rising limb and pore pressure recession rates closely matched stream and subsurface-trench hydrograph recession rates. Soft data for the hillslope positions were gathered from previous throughflow pit analysis by Mosley [1979] including continuously recorded pit outflow from a number of distinct linear hillslope segments. Hillslope sections (unlike hollows and riparian zones) show very infrequent water table development—when water tables were present, they were restricted vis-à-vis the soft data measure trapezoidal function classification (see numbers in Table 2). The soil catena sequences in the Maimai catchment as mapped by McKie [1978] confirm these interpretations.

Hillslope soils show no evidence of any gleying whereas gleying appears in the hollow zone and is most dominant in the riparian zone. We view this as a long-term expression of the spatial delineation of boxes and water table longevity applied in this study. Table 2 includes also a number of soft-data rules including isotope hydrograph separation-derived new-water estimates (at peakflow). Values for these rules were based on results from hydrograph separations reported

Table 2. Evaluation rules based on soft data used for model calibration (the values for a_i define the trapezoidal function used to compute the degree of acceptance, see Eq. 1).

Type of soft information	Specific soft information	a_1	a_2	a_3	a_4	Motivation
New water contribution to peak runoff [-]	870930 18.00	0.03	0.06	0.12	0.15	McDonnell *et al.* [1991]
	871008 3.00	0.05	0.13	0.31	0.40	"
	871010 17.00	-	0	0.03	0.06	"
	871013 11.00	0.17	0.23	0.35	0.41	"
	871113 19.00	-	0	0.03	0.06	"
	871127 8.00	0.04	0.07	0.13	0.16	"
Range of groundwater levels, min./max. fraction of saturated part of the soil [-]	Maximum hillslope	0	0.2	0.5	0.7	Mosley [1979]
	Maximum hollow	0	0.5	0.75	1	McDonnell [1990]
	Minimum hollow	0	0.05	0.1	0.2	"
	Minimum riparian	0.05	0.1	0.3	0.5	"
Frequency of groundwater levels above a certain level (as fraction of soil) [-]	Hillslope, above 0.5 during events	-	0	0.1	0.3	Mosley [1979]
	Hollow above 0.7 during events	-	0	0.1	0.2	McDonnell [1990]
	Hollow above 0.9 during events	-	-	0	0.1	"
	Riparian above 0.2	0.6	0.8	1	1	"
	Riparian above 0.9 during events	0	0.25	0.75	1	"
Parameter values	Fraction of riparian zone [-]	0.01	0.03	0.07	0.10	Mosley [1979]
	Fraction of hollow zone [-]	0.05	0.10	0.15	0.20	McDonnell [1990]
	Porosity in hillslope zone [-]	0.45	0.6	0.7	0.75	McDonnell [1989]
	Porosity in hollow zone [-]	0.45	0.55	0.65	0.75	"
	Porosity in riparian zone [-]	0.45	0.5	0.6	0.75	"
	Soil depth for hillslope zone [m]	0.1	0.3	0.8	1.5	McDonnell *et al.* [1998]
	Soil depth for hollow zone [m]	0.5	1	2	2.5	"
	Soil depth for riparian zone [m]	0.15	0.4	0.75	1	"
	Threshold level in hollow zone, fraction of soil depth [-]	0	0.1	0.4	1	McDonnell [1990] McDonnell *et al.* [1991]

in McDonnell [1989] and McDonnell *et al.* [1991]. These evaluation rules allowed computation of degree of acceptance with respect to the simulated new-water. New water percentage is a very useful integrated measure of the relative contribution of rainfall versus displaced stored water contributions at various times through the storm hydrograph. Unlike the point-based water level measures and rules, the new water percentage subsumes point scale variability into an integrated measure of catchment runoff dynamics. In our dataset, the new-water percentages varied, from event to event, and some storms did not have rain isotopic concentration suitable for application of the two-component mass balance separation technique. The flexibility of the soft data is such that even for isolated measures from field campaigns or experiments (or when hydrograph separation was possible) rules may be developed to guide the model calibration process, even if this information is derived from periods outside the simulated calibration period.

We computed degrees of acceptance for a number of parameters using the soft data evaluation rules. Acceptance in this instance is defined as the degree to which model parameter values agree with the field experience and the perceptual model of the catchment runoff process. These acceptance values varied from one, if the value was within the desirable range and decreased towards zero with increasing deviations from this range (Table 2). For example, we allowed values from 1 to 10 percent for the areal fraction of the riparian zone (*i.e.*, the variable source area in this case), but the degree of acceptance was one, only for values between 3 and 7 percent (based on mapped saturated areas in the M8 catchment reported in Mosley [1979]). Based on the individual parameters the acceptability of a certain parameter set was computed as the geometric mean of the respective degrees of acceptance.

We quantified the acceptability of calibrations using hard data (A_1) using a combination of the efficiency measure, R_{eff}, and the relative volume error, V_E, (=accumulated difference divided by sum of observed runoff) for the runoff simulations as proposed by Lindström [1997] (Eq. 3). Following Lindström [1997], a value of 0.1 was used for the weighing coefficient, ω, which determines the relative emphasis on the volume error. The coefficient of determination, r^2, was used to assess the performance of the simulations for the groundwater levels in the riparian and the hollow zone, and A_1 is computed as average of these different goodness-of-fit measures (Eq. 3).

$$A_1 = \frac{1}{2}\left(R_{eff} - \omega|V_E| + \sqrt{r^2_{gw\ hollow}\ r^2_{gw\ riparian}}\right) \qquad (3)$$

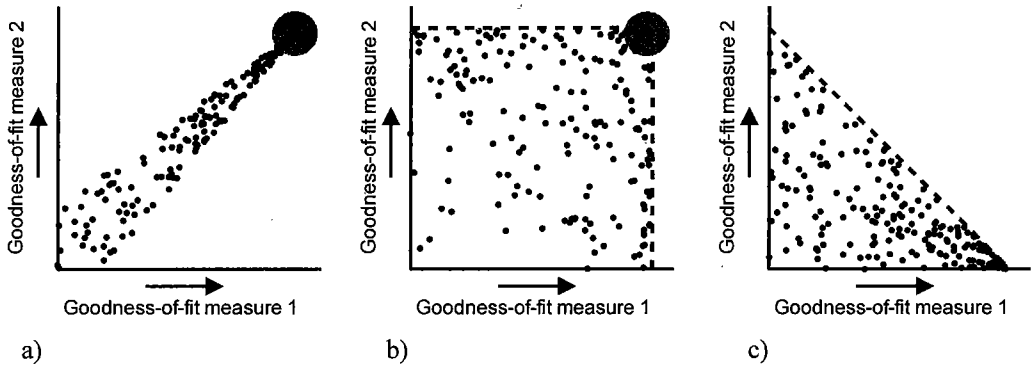

Figure 3. Three different types of relations between goodness-of-fit measures for the best realizations: a) a strong positive correlation, b) no correlation, and c) a negative correlation. Each dot represents one realization (or parameter set), the dashed line represents the pareto-optimality and the gray circle indicates the region in which the 'best' parameter sets are found.

Using the coefficient of determination, r^2, we did not force the model to *exactly fit* the observations, but allowed for an offset and a different amplitude. We argue that it is the dynamics, rather than the exact levels, that should be used from this kind of data where we compare the point observation from the field with a simulated average behavior of an entire zone (i.e., box within the model). By also utilizing soft data, there is no need to 'over fit' the model to the levels obtained from tensiometer observations at a few observation locations – in our case, one point in the hollow zone and another mid-way up the main valley bottom in the riparian zone (see McDonnell [1990] for field details).

Acceptability of the model simulations using soft data (A_2) was computed as the arithmetic mean of 15 evaluation rules of the soft data for groundwater levels and contribution of new water (Table 2). The arithmetic mean was used in this instance since the geometric mean is less suitable when values can become zero. Acceptability of the parameter values based on soft data (A_3) was computed as the geometric mean of nine evaluation rules of the different parameters (Table 2).

When plotting two different goodness-of-fit measures against each other for a number of realizations (parameter sets), the relations for the best realizations can be grouped into three basic cases: (1) a strong positive correlation, (2) no correlation, and (3) a negative correlation (Fig. 3). In case 1 the second criterion does not contribute with additional information and only one of the goodness-of-fit measures needs to be calculated. The situation is different for the case 2, where the both criteria provide different information. However, in both cases it is quite apparent from which region one would choose parameter sets to achieve optimal model performance, *i.e.*, from a region where one can find realizations that are optimal for both criteria (see gray circle in Fig.3). In case 3 the two criteria also provide different information, but here the two criteria are not unrelated and "conflict' one another. In other words, a good solution according to one criterion can only be obtained at the price of a poor performance according to the second criterion. It is therefore not possible to find a solution that is optimal according to the two criteria simultaneously, since the best values for the two criteria are negatively correlated. The best solutions are found along a pareto-optimality line (*i.e.*, 'compromise-solutions'). If the 'compromise-solutions' are too poor compared to the individual best solutions, this might indicate a problem with the model structure [*Seibert*, 2000]. As mentioned above, the selection of the weights n_1, n_2, and n_3 in Eq. 2 determines which solution along the pareto-optimality sub-space (lines in Fig. 3) will be found.

We tested different combinations to examine the relations between the different criteria. We quantified the value of the soft data by testing how the measures helped in ensuring internal model consistency and reducing parameter uncertainty. First we examined how model performance, as judged by the various criteria, varied when the model was calibrated considering different sets of criteria. Second, we compared the magnitude of parameter uncertainty when calibrating against runoff only and when calibrating against different combinations of criteria. For this part of the analysis we used values of 0.4, 0.4 and 0.2 for the weights in Eq. 2 n_1, n_2, and n_3 respectively to place more emphasize on the acceptability with regard to the simulations (both hard and soft data) and less weight on the acceptability of the parameter values.

RESULTS

Model Performance

The model was able to reproduce observed runoff very well. When calibrated with runoff data only, the model was

able to simulate the observed runoff with values of 0.93 for the model efficiency [*Nash* and *Sutcliffe*, 1970]. Notwithstanding, while high model efficiency was obtained with the runoff-only (hard data) calibration, goodness-of-fit statistics for percent new water and soft groundwater measures for example, were very poor (Fig.4, shaded area). If one examines the simulated groundwater levels for each of the three boxes for the runoff-only calibration, several different response patterns are produced—each with a high model efficiency for runoff (Fig. 5a-c). In Fig. 5a, the riparian and hollow box fail to behave like observed reservoir dynamics reported in McDonnell [1990], with too much water remaining in the hollow box, especially between events. Fig. 5b is an example where each of the three boxes filled and drained too quickly during events. Fig. 5c shows an appropriate riparian box response but poor representation of the hollow zone where the zone is drained too quickly. This is a compelling example of how relying only on the traditional single-criterion, hard-data model calibration, can produce 'right answers for the wrong reasons'. It each case, without the insight of soft data, one may have been tempted to assume that the model 'worked well' given the high model efficiency for any of the very similar runoff simulations.

As additional hard and soft data were entered into the model calibration, the model efficiency for runoff decreased (from the 0.93 value to 0.84) but goodness-of-fit for the process description (*i.e.*, soft data on groundwater, percent-new-water and parameter values) increased dramatically (Fig. 4 and 6). The combined objective function A (Eq. 2) increased from 0.46 to 0.79 when adding A_2 and A_3 to the optimization criterion. In general, the variability in the various goodness-of-fit measures decreased when more criteria were included into the calibration. Most importantly perhaps, the groundwater dynamics simulated with a parameter set obtained by this multi-criteria calibration are in keeping with experimental observations on reservoir response. The goodness-of-fit of the groundwater level simulations increased from 0.53 to 0.82 for the hard data and from 0.34 to 0.60 for the soft data, for parameter sets optimized using the combination of all criteria compared to the simulations using parameter sets calibrated to only runoff. Furthermore, the range of objective-function values generally decreased when a criterion was considered during calibration.

The simulation with the best overall performance caused a somewhat reduced model efficiency for runoff but displayed more 'realistic' internal dynamics (Fig. 6). Fig. 6 also shows the decrease of unsaturated storage through the event, indicative of the coupled formulation of saturated and unsaturated storage. We argue that this formulation is an important and new feature of the three-box approach because it is a more realistic conceptualization of the unsaturated-saturated storage interactions given the shallow groundwater. While application of the model to other catchments might involve different arrangements and numbers of boxes, the

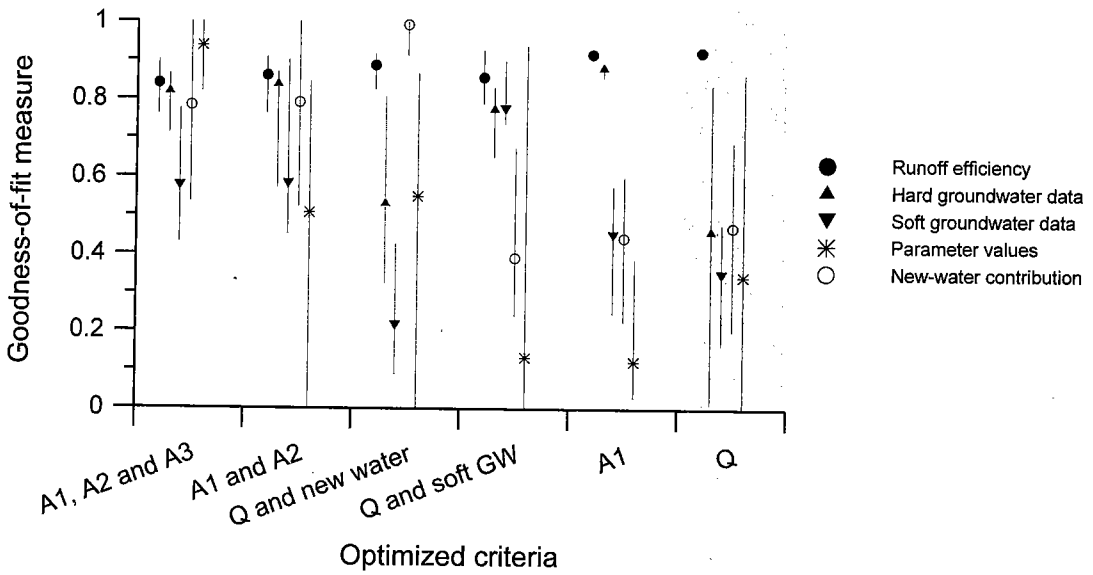

Figure 4. Goodness-of-fit measures for runoff, groundwater levels, new water ratios, soft groundwater measures, and parameter-value acceptability for calibrations against various combinations hard and soft information (see text for definition of the different optimization criteria). The symbol shows the median of 50 calibration trials and the vertical lines indicate the range of these trials. The shaded area relates to the traditional calibration approach using only runoff data and highlights the problem of internal consistency when calibrating against only runoff.

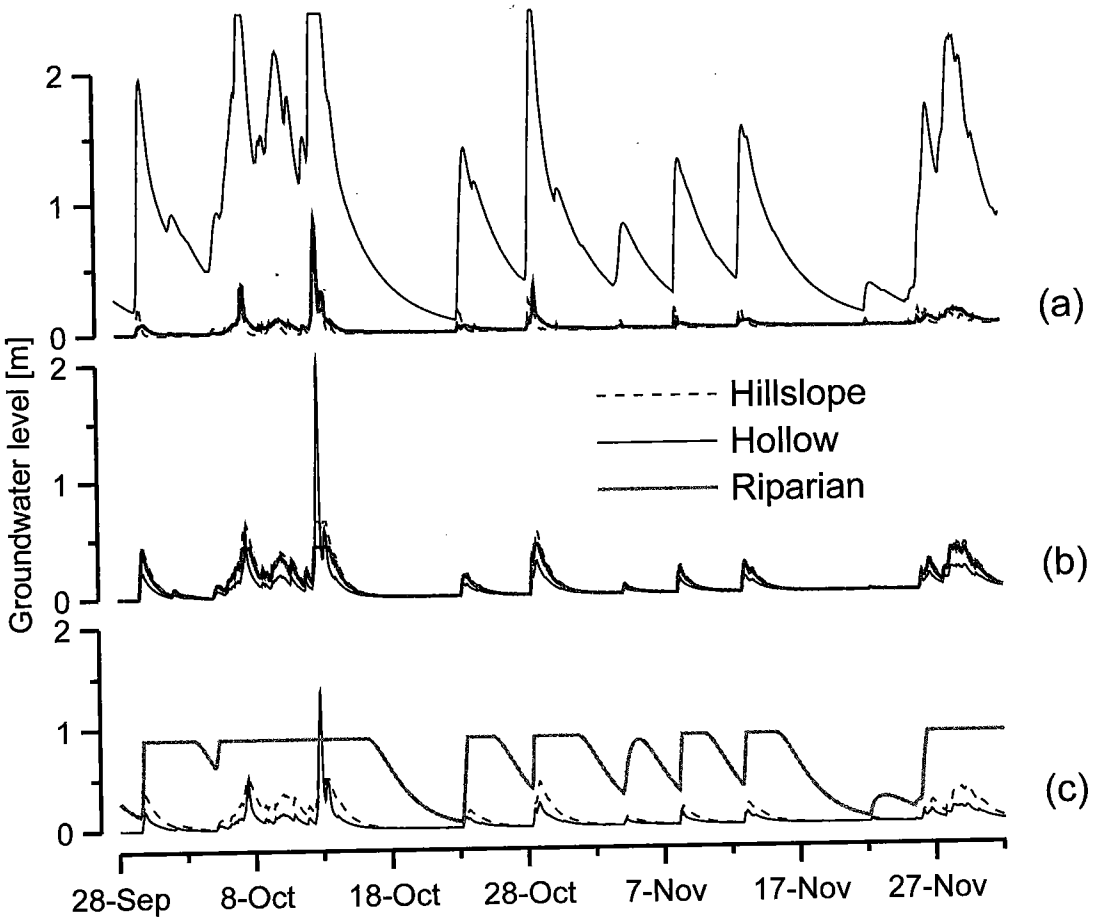

Figure 5. Three model runs with different parameter sets resulting in different groundwater dynamics (levels in [m] above bedrock). All three parameter sets had been calibrated to observed runoff and gave an almost similar goodness-of-fit (model efficiency ~0.93). None of the three sets of groundwater time series agrees with the perceptual model of the watershed.

coupled saturated-unsaturated zone formulation is one that is common to many headwater catchment conditions.

Relation between Optimization Criteria

Different parameter sets will be found through calibration if different weights (n_i) are used for the overall acceptability in Eq. 2. Using different combinations of A_1 and A_2 as well as A_1 and A_3 demonstrated that both soft-data criteria (A_2 and A_3) gave different information than the hard data (A_1) (Fig. 7). There is no conflict between the hard data and the soft data on parameter values (A_3) (Fig. 7b), *i.e.*, the calibrated solutions all follow the 'no-correlation'-pattern (compare Fig. 3b). On the other hand, there is a trade-off between the hard data and the soft data on model simulations (A_2) (Fig. 7a), *i.e.*, it is not possible to find a solution that is optimal according to both criteria simultaneously. The solutions form a curve that lies in between the 'negative-correlation' and the 'no-correlation'-patterns (compare Fig. 3 b,c) indicating that there is some conflict between the criteria, but not total disagreement.

Parameter Uncertainty

For each parameter, 50 different values were obtained by the different calibration trials. The range between the 0.1 and 0.9 percentile divided by the median was computed for each parameter as measure of parameter uncertainty. The ratio between the values obtained from multi-criteria soft data calibrations and those derived from runoff-only hard data calibrations indicated a general reduction of parameter uncertainty (*i.e.*, the variation of calibrated parameter values decreased) when adding different criteria, but results varied from model parameter to model parameter. When optimizing the combination of all criteria (A_1, A_2 and A_3) the ratio varied between 0.03 and 0.65. The median was

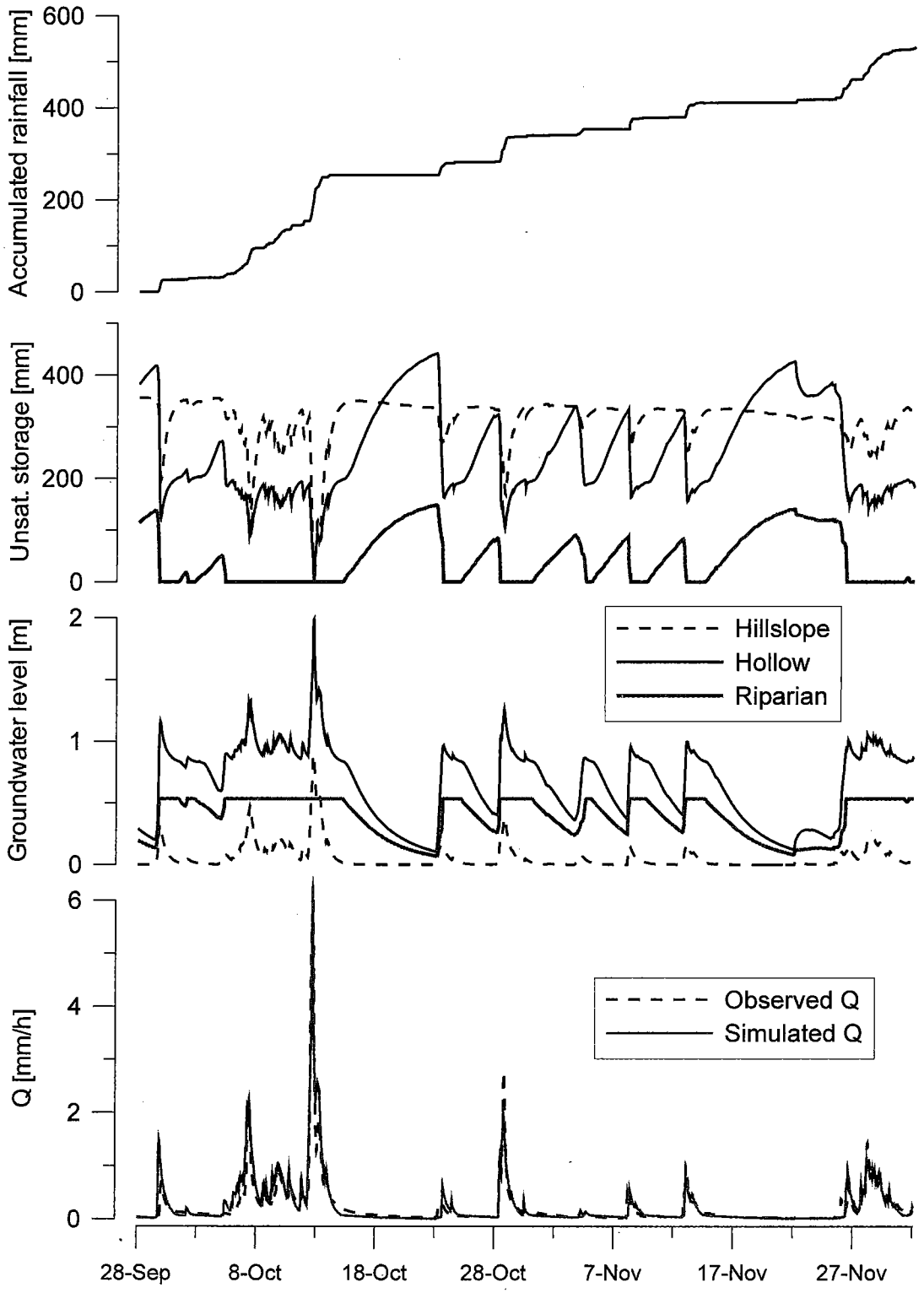

Figure 6. Simulation with best overall performance. Accumulated rainfall, simulated unsaturated storage and simulated groundwater levels (m above bedrock), as well as observed and simulated runoff. The model efficiency for runoff is 0.84 and the simulated groundwater dynamics agree in general with the perceptual model.