On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multi-criteria model calibration

Jan Seibert*
Swedish University of Agricultural Sciences, Department of Environmental Assessment, S-750 07 Uppsala, Sweden

Jeffrey J. McDonnell
Oregon State University, Department of Forest Engineering, Corvallis OR 97331, U.S.A.

Revised version, submitted to Water Resources Research

March, 2002

Short Title: Soft data

* Corresponding author:
Jan Seibert
Swedish University of Agricultural Sciences
Department of Environmental Assessment
Box 7050
S-750 07 Uppsala
Sweden
e-mail: jan.seibert@ma.slu.se
Tel.: + 46-18-67 3145
Fax : + 46-18-67 3156

Abstract

The dialog between experimentalist and modeler in catchment hydrology has been minimal to date. The experimentalist often has a highly detailed yet highly qualitative understanding of dominant runoff processes—thus there is often much more information content on the catchment than we use for calibration of a model. While modelers often appreciate the need for 'hard data' for the model calibration process, there has been little thought given to how modelers might access this 'soft' or process knowledge. We present a new method where soft data (*i.e.*, qualitative knowledge from the experimentalist that cannot be used directly as exact numbers) are made useful through fuzzy measures of model-simulation and parameter-value acceptability. We developed a three-box lumped conceptual model for the Maimai catchment in New Zealand, a particularly well-studied process-hydrological research catchment. The boxes represent the key hydrological reservoirs that are known to have distinct groundwater dynamics, isotopic composition and solute chemistry. 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, etc). We achieved very good fits for the three-box model when optimizing the parameter values with only runoff (R_{eff}=0.93). However, parameter sets obtained in this way showed in general a poor goodness-of-fit for other criteria such as the simulated new-water contributions to peak runoff. Inclusion of soft-data criteria in the model calibration process resulted in lower R_{eff}-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. The model performance with respect to soft data (like, for instance, the new water ratio) increased significantly and

parameter uncertainty was reduced by 60% on average with the introduction of the soft data multi-criteria calibration. We argue that accepting lower model efficiencies for runoff is 'worth it' if one can develop a more 'real' model of catchment behavior. The use of soft data is an approach to formalize this exchange between experimentalist and modeler and to more fully utilize the information content from experimental catchments.

Introduction

Catchment hydrology is at a cross-roads. While complex descriptions of the age, origin and pathway of surface and subsurface stormflow abound in the literature (reviewed recently by Bonell [1998]), most catchment modeling studies have not been able to fully use this information for model development, calibration and testing. As a result, process hydrological studies of dominant runoff producing processes and model studies of runoff generation are often poorly linked. Recently there has been a tendency away from fully-distributed, physically-based models back to conceptual models due to concerns that the small-scale physics may not be appropriate at the scale of model (grid) applications and the inability to determine physical parameters a priori. These issues give rise to problems like of overparameterisation, parameter uncertainty and model output uncertainty [Beven, 1993; 2001]. While conceptual models may be much more simplified and lumped, they offer the potential for development based on process understanding of key zones or reservoirs of catchment response. A problem in conceptual modeling of catchment hydrology is that parsimonious models, which may allow identification of parameter values through calibration against runoff, in general are too simple to allow a realistic

representation of the main hydrological processes and, thus, provide only limited possibilities for internal model testing. As interest in the geochemical dimensions of streamflow modeling increases, reservoir (or box model) conceptual approaches that explicitly treat volume-based mixing and water (and ultimately tracer) mass balance become increasing useful [Harris *et al.*, 1995; Hooper *et al.*, 1998; Seibert *et al.*, 2002a]. Spatial distinction into different zones is motivated for many catchments based on, for instance, different hydrochemical functioning [Cirmo and McDonnell, 1997] or differences in groundwater dynamics [Seibert *et al.*, 2002b]. As such, box models with explicit reservoir volumes have indeed a physical basis, because water and tracer mass balances can be accounted for explicitly during each model step. However, there are important issues yet to be solved concerning the use of box models for representation of solute transport such as the consideration of preferential flow or incomplete mixing.

A major obstacle in moving forward with conceptual modeling approaches is how to fully utilize experimental data for internal calibration and validation. Currently, the use of this field data for model calibration is often limited beyond simple streamflow information despite the general acceptance that internal state information is necessary for ensuring model consistency. The usefulness of having various criteria for assessment of model performance is widely accepted [Kuczera and Mroczkowski, 1998]. When multiple criteria are used for calibration or validation, this has often meant only the use of two or three so-called hard data criteria (*e.g.*, runoff and groundwater levels) as compared to only one criterion (*i.e.*, runoff) [*e.g.*, Kuczera, 1983; Hooper *et al.*, 1988; Refsgaard, 1997; Kuczera and Mroczkowski, 1998; Seibert, 2000]. The willingness to use only hard data is a hindrance to moving forward. The dilemma is clear: modelers recognize that

more criteria are desirable but in most cases there are no suitable hard data available. Others have commented on the dilemma that we have; on one hand, a knowledge of catchment behavior by the experimentalist that is highly complex and highly qualitative, but on the other hand the need for simplification when developing model structures caused by data and computational limitations [Beven, 1993]. While some groups have used this perceptual model [Beven, 1993] 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 fuzzy measures for evapotranspiration.

The hydrologist's perceptual model is often a highly detailed yet qualitative understanding of dominant runoff processes. Thus, there exist in addition to hard data (streamflow hydrograph, well record) 'soft data' about catchment hydrology. Soft data are a rather different type of information than traditional hard data measures. Soft data are often 'spotty', discontinuous and numerically approximate. 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 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;

Hankin and Beven, 1998]. Aronica *et al.* [1998], for instance, used a fuzzy-rule based calibration motivated by 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.

We argue that soft data represent a new dimension to the model calibration process that might: (1) enable a dialog between experimentalist and modeler, (2) be a formal check on the 'reasonableness' and consistency of internal model structures and simulations, and (3) specify realistic parameter ranges often ignored in today's automatic calibration routines. When calibrating a conceptual rainfall-runoff model manually, some of this qualitative understanding might implicitly influence the calibration. The search for optimal parameters is thus restricted to certain parameter values and the modeler might visually inspect simulated internal variables such as groundwater levels and consider how reasonable these simulations are. Model parameters in conceptual models are not directly measurable. Parameters may be related to measurable quantities but they are effective values for a much larger scale than the measurement scale. However, for some parameters, the field hydrologist experimentalist might reject or prefer values within certain ranges based on his/her knowledge of the catchment and its behavior during and between events. Usually the search of parameter values by calibration is constrained only by the specification of feasible ranges. The concept of soft data enables one to specify a narrower desirable range for a number of parameters. During calibration, the model is 'punished' for values outside these desirable ranges, but such values may still be chosen by the calibration if they lead to better fits.

The explicit use of soft data has two advantages: (1) the goodness-of-fit criteria are stated a priori, while still being subjective, and (2) the method can be used in automatic calibration routines. In other words, the procedure injects some experimentalist common sense into the automatic calibration process. Similarly, the use of soft data can be seen as a proactive way to reduce parameter uncertainty where the modeler and experimentalist together specify additional criteria to judge model simulations. As such the soft data approach developed in this paper is a way to specify model performance criteria 'up front'. This complements methods used to quantify parameter uncertainty and its effects such as the generalized uncertainty estimation (GLUE) approach [*e.g.* Freer *et al.*, 1996] or the other philosophical approach using the Pareto optimal set methodology for defining parameter sets that are in some way optimal [*e.g.* Gupta *et al.*, 1999].

This paper explores the new philosophy and approach for development of more realistic models of catchment behavior using soft data where multiple criteria are used to constrain the model calibration. We argue that this method is the necessary dialog that should occur between the modeler and the experimentalist to enable a better process representation of catchment hydrology in conceptual runoff models. We use the well-characterized Maimai watershed (recently reviewed by McGlynn *et al.* [2002]) as the testing ground for these new ideas. This paper: (1) presents a new three-box model of headwater catchment response based on an extension of ideas developed in Seibert *et al.* [2002a], (2) incorporates a number of soft-data measures from experimental studies at the catchment for model calibration, and (3) assesses the value of soft data relative to traditional hard information measures for model calibration. While the paper does not advocate the use of soft data over hard data, we make the case that soft data may be an

important augmentation to hard data for model calibration and should be actively sought out where available.

Material and methods

The Maimai watershed

Maimai M8 is a small 3.8 ha study 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^{o}) 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. The summer months are the driest; monthly rainfall from December to February averages 165 mm and for the rest of the year between 190 to 270 mm. On average, there are 156 rain days per year and only about 2 snow days per year [Rowe *et al.*, 1994]. In addition to being wet, the catchments are highly responsive to storm rainfall. Quickflow comprises 65% of the mean annual runoff and 39% of annual total rainfall [Pearce *et al.*, 1986]. The period of record used for model simulation in this study was August-December, 1987. There were 11 major runoff events during this period 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 the quintessential headwater research catchment: it is underlain by a firmly compacted poorly impermeable conglomerate and seepage losses to deep groundwater are negligible (estimated at 100 mm/yr based on 25 years of water balance data). The wet and humid climatic environment, in conjunction with topographic and soil characteristics, results in the soils normally remaining within 10% of saturation [Mosley, 1979]. As such, the catchment shows clear and unambiguous catchment-wide response to storm rainfall. The thin nature of the soils promotes the lateral development of root networks and channels. Soil profiles reveal extensive macropores and preferential flow pathways at vertical pit faces which form along cracks and holes in the soil and along live and dead root channels [Mosley 1979]. Lateral root channel networks are evident in the numerous tree throws that exist throughout the catchments. Preferential flow also occurs along soil horizon planes and the soil-bedrock interface.

Perceptual model of the Maimai watershed

M8 has been the site of ongoing hillslope research by several research teams since the late 1970s. These studies have facilitated the development of a very detailed yet qualitative perceptual model of hillslope hydrology, reviewed recently by McGlynn *et al.* [2002]. While dye tracer studies by Mosley [1979] showed that storm rainfall follows preferential flow pathways at the hillslope scale, subsequent water isotopic tracing studies in the catchment by Pearce *et al.* [1986] and Sklash *et al.* [1986] showed (paradoxically) that there was little if any event water in the stream during stormflow periods. Thus, stored soil water and groundwater comprise the majority of channel stormflow. McDonnell [1990] developed a perceptual model to explain the mechanism of stormflow generation by constraining the dominant processes using recording tensiometer

observations, isotope tracing and various other chemical and hydrometric approaches. For small events of less than about 15 mm rainfall, McDonnell *et al.* [1991] found that the riparian zone (*i.e.*, the near-stream valley bottom) could account for the volume of old water in the channel hydrograph. During larger events, McDonnell [1990] found that hillslope hollows (*i.e.*, topographic convergent zones on the slopes) were the dominant runoff producing zones where new water moved to depth and created a perched water table at the soil-bedrock interface. Lateral pipeflow then formed along the soil bedrock interface [McDonnell *et al.*, 1998], conveying quantities of old water laterally downslope sufficient in quantity and quality to explain measured old water volumes. Topographic convergence of flowpaths from planar hillslopes to the hollows enabled hollows to be well-primed for rapid conversion of matrix to pressure potentials. Soil water isotopic composition [McDonnell *et al.*, 1991] and chemical composition [Grady and Elsenbeer, 2000, pers. comm.] all followed a similar pattern of distinct and unambiguous response zones and inter-storm reservoirs: hillslopes, hollows and riparian zones. These zones display very different groundwater dynamics [McDonnell, 1990] and group clearly, based on their isotopic characteristics. Data in McDonnell *et al.* [1991], although not fully appreciated at the time of publication, revealed, using a cluster analysis, the three distinct isotopic groupings from suction lysimeter data extracted from 11 devices across the catchment. Finally, according to the perceptual model of McDonnell *et al.* [1991] flow occurs from the hillslope zone to the hollow zone and from there to the riparian zone before contributing to runoff. The soil catena sequences observed in the catchment by McKie [1978] confirm this perception based on soil characteristics.

Conceptual three-box model

The conceptual model is based on the three reservoirs identified from the experimental studies at M8: riparian, hollow and hillslope zones (Fig. 1). 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.* [2002a]. In this formulation, the amount of saturated storage determines the maximum space for unsaturated storage.

The maximum amount of saturated storage (with the groundwater table reaching the ground surface), S_{max}, is computed as product of porosity, p, and soil depth, z_{max}, (see Table 1 for a list of all model parameters). Based on the calculated actual value of the saturated storage, S, the maximum unsaturated storage at drainage equilibrium ('field capacity'), U_{max}, is computed (Eq. 1). Similarly the amount of water stored in the unsaturated zone below wilting point, U_{min}, is computed (Eq. 2).

$$U_{\max} = c \left(S_{\max} - S\right) \quad (1)$$

$$U_{\min} = d \left(S_{\max} - S\right) \quad (2)$$

S, U and U_{min} represent volumes of water per unit ground area, whereas the model parameters c and d are dimensionless. From equations 1 and 2, it follows that c corresponds to field capacity divided by porosity and d corresponds to wilting point divided by porosity.

For the unsaturated zone, an approach similar to that used in the HBV model [Bergström, 1995] is used. The amount of rainfall, P, is divided into recharge to groundwater, R, and addition to the storage in the unsaturated zone using a non-linear function (Eq. 3, β [-] is a shape factor). Evaporation from the soil, E_{act}, is estimated based on the actual storage in the unsaturated zone, U, and the potential evaporation, E_{pot} (Eq. 4).

$$R = P\left(\frac{U - U_{\min}}{U_{\max} - U_{\min}}\right)^{\beta} \tag{3}$$

$$E_{act} = E_{pot}\,\frac{U - U_{\min}}{U_{\max} - U_{\min}} \tag{4}$$

Outflow from the hillslope and riparian reservoir is computed as a simple linear function of the groundwater level in each box, $z_{riparian}$, z_{hollow} and $z_{hillslope}$ [m above bedrock]. The groundwater levels are computed from the saturated storage using a porosity parameter for each box. The hollow reservoir is given an additional threshold-based linear function based on the McDonnell [1990; pp. 2830 Fig 10] perceptual model (Eq. 5-7):

$$Q_{hillslope} = k_{1,hillslope}\; z_{hillslope} \tag{5}$$

$$\begin{aligned} Q_{hollow} &= k_{1,hollow}\; z_{hollow} && if\ z_{hollow} \le z_{threshold} \\ Q_{hollow} &= k_{1,hollow}\; z_{threshold} + k_{2,hollow}\left(z_{hollow} - z_{threshold}\right) && if\ z_{hollow} > z_{threshold} \end{aligned} \tag{6}$$

$$Q_{riparian} = k_{1,riparian}\; z_{riparian} \tag{7}$$

The use of a threshold in the hollow reservoir is also motivated by field observations reported by McDonnell *et al.* [1998] that indicate large fluxes through macropores along the bedrock-soil interface. The threshold level, $z_{threshold}$, corresponds to the level at which these fluxes become significant, whereby transient water table develops in the hollow during events and initiates a lateral pipeflow at in the lower soil profile.

Based on rainfall, simulated runoff (Q_i) and simulated evaporation (E_{act}), the amount of unsaturated and saturated storage in each box is updated for each time step. In the case of falling groundwater levels, a certain amount of saturated storage changes its status to 'unsaturated'. The change of storage in the saturated zone (ΔS) equals the difference between recharge (for the riparian and hollow box including lateral inflow) and runoff plus a portion of the change, which is the amount of water changing its status from saturated to unsaturated (Eq. 8). Eq. 8 can be rearranged to allow direct calculation of ΔS (Eq. 9) and computation of the corresponding change in unsaturated storage due to groundwater level change, ΔU_{gc} (Eq. 10).

$$\Delta S = R - Q - E_{act} + c\Delta S \tag{8}$$

$$\Delta S = \frac{R - Q - E_{act}}{1 - c} \tag{9}$$

$$\Delta U_{gc} = -c\Delta S \tag{10}$$

When the groundwater level rises, an amount of unsaturated storage in a similar way alters its status to 'saturated' (Eq.s 11-13).

$$\Delta S = R - Q - E_{act} + c\frac{U}{U_{\max}}\Delta S \tag{11}$$

$$\Delta S = \frac{R - Q - E_{act}}{1 - c\dfrac{U}{U_{\max}}} \tag{12}$$

$$\Delta U_{gc} = -c\frac{U}{U_{\max}}\Delta S \tag{13}$$

The fraction U/U_{max} appears in these equations since drainage equilibrium ($U=U_{max}$) cannot be assumed when the water table rises. When the water table falls, on

the other hand, drainage equilibrium can be expected in the drained soil layer and U/U_{max}=1 (Eqs. 8-10).

Important assumptions and simplifications of the three-box model include the following: (1) no lateral flow is assumed to take place in the unsaturated zone of individual reservoir boxes (based on matric potential data from previous experimental work [McDonnell, 1990] that shows downward hydraulic gradients in the unsaturated zone between and during events), (2) no bypass flow from hillslope reservoir to the stream is allowed (again based on experimental work of Mosley [1979] that examined topographic convergence in the colluvial filled hollows), (3) no substream or hyporheic exchange is considered between the channel and the riparian zone. This last assumption may be a gross simplification based on recent comments by Bencala [2000], and our group is actively researching the issue of hyporheic exchange at Maimai [McGlynn, 2002, pers.comm.]. A preliminarily guess, however, is that the amount of in the Maimai M8 catchment may be quite limited, due to the tight nature of the underlying substrate and the fact that the stream flows on bedrock for much of its length. Finally, whilst these assumptions and simplifications are supposed to be appropriate for the Maimai M8 catchment, we do not advocate that they are universally applicable. The model structure is guided by experimental findings and application of any model and articulation of box/reservoir numbers, configurations and characteristics would be framed on a catchment-by-catchment basis. It is also important to recognize that the important step of relating the model boxes to actual landscape units, which is a prerequisite for comparing internal simulations with observations, is not trivial and may in many catchments be more difficult than at Maimai where landforms are fairly simple.

Utilization of soft data

Given the relatively large number of parameters (16) in the three-box model, the information contained in the hard data (runoff and two groundwater-level series) is insufficient for the identification of parameter values through calibration. Consequently, parameter uncertainty would be expected to be large. Soft data enable additional judgment of model simulations in more ways and more process-based ways than using only the available hard data. For instance, the experimentalist might have some observations concerning the range in which groundwater levels fluctuate within a given box (based on field campaign information or observations made over some irregular time periods) or the contribution of new water to peak flow (from event-based isotope tracing studies). Soft data can be used in two ways to constrain the calibration: (1) to evaluate aspects of the model simulations for which there is no hard data available and (2) to assess how reasonable the parameter values are based on field experience (Table 2).

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, we defined a trapezoidal function (Equation 14) to compute the degree of acceptance from the corresponding simulated quantity or parameter value. This trapezoidal function is a simple way to map experimentalist experience into a quantity, which then can be optimized (Fig. 2). This approach recognizes that there is uncertainty in even the experimental data [Sherlock *et al.*, 2000]—using a fuzzy membership

function like this trapezoidal form enables the modeler-experimentalist dialog to explicitly recognize this.

$$\mu(x) = \begin{cases} 0 & if\ x \le a_1 \\ \dfrac{x - a_1}{a_2 - a_1} & if\ a_1 \le x < a_2 \\ 1 & if\ a_2 \le x < a_3 \\ \dfrac{a_4 - x}{a_4 - a_3} & if\ a_3 \le x < a_4 \\ 0 & if\ x > a_4 \end{cases} \tag{14}$$

In this study, we used soft data measures for a number of groundwater level measures in the three boxes. Evaluation rules were developed using Equation 14 to judge model performance with regard to minimum and maximum groundwater levels as well as the frequency of levels being above specified level (Table 3). The values for these rules were motivated by field studies reported in McDonnell [1990] for the same August-December 1987 period where groundwater responses 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 enabled the soft data measures of minimum and maximum groundwater levels and frequency of levels above a specified level (listed in Table 3) to be defined. 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] where he 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 classification (see numbers in Table 3). Here again, the soil catena sequences as mapped by McKie [1978] confirm these interpretations. Hillslope soils show no evidence of any gleying whereas gley appears in the hollow zone and is most dominant in the riparian zone.

Table 3 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 in McDonnell [1989] and McDonnell *et al.* [1991]. These evaluation rules allowed computing of a degree of acceptance with respect to the simulated new-water. The new-water percentages varied, of course, from event to event and some storms did not have rain isotopic concentration suitable for using 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, rules may be developed to guide the model calibration process. The isotope hydrograph separation soft data are particularly useful since the M8 catchment has such large (and repeatable) old water contributions to peakflow. This measure is a tremendous perceptual constraint on how a conceptual box model may allow flow of new water into the channel during events. We view this use of isotopic data as one of the first formal attempts to include isotope-based hydrograph separation results into a model exercise. While a few studies in the past have used stream isotope concentrations through time for model calibration and

testing [Hooper *et al.*, 1988; Seibert *et al.*, 2002a], there are no studies we are aware of that make use of computed new water ratios. O-18 time series could be used directly (*i.e.*, as hard data), but often the observed signal is weak and observed time series are short and discontinuous (for review see Buttle [1994]). In many cases it might, thus, be more suitable to use information derived from the observations such as the new-water contribution to peak flow. This is an example for soft data which is based on hard data, but where interpretation by a hydrologist is needed to transform the actual measurements into data, which might be imprecise as it is in the case of the new-water contributions to peak flow. Given the fact that much such information exists for experimental catchments around the world, we see much potential in moving forward with soft data calibration in the future.

For a number of the parameters a degree of acceptance was computed. Acceptance in this instance is defined as the degree to which parameter values agree with the field experience and the perceptual model of the catchment. These acceptance values varied again from one, if the value was within the desirable ranges and decreased towards zero with increasing deviations from this range (Table 3). For example, we allowed values from 1 to 10 percent for the spatial 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.

Quantifying the acceptability and value of hard and soft data

We quantified the acceptability of calibrations using hard data (A_1) using the Nash and Sutcliffe [1970] efficiency measure, R_{eff}, and the relative volume error, V_E, for the runoff simulations. Following Lindström [1997], a value of 0.1 is chosen 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 hard-data 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 (Equation 15).

$$A_1 = \frac{1}{2}\left(R_{eff} - \omega|V_E| + \sqrt{r^2_{gw\,hollow}\; r^2_{gw\,riparian}}\right) \qquad (15)$$

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 observed level at one location with a simulated average behavior of an entire zone. By utilizing also 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.

Acceptability of the model simulations using soft data (A_2) was computed as arithmetic mean of the 15 evaluation rules of the soft data regarding groundwater levels and contribution of new water (Table 3). 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 the nine evaluation rules of the different parameters (Table 3).

The overall acceptability, A, of a parameter set was computed as a weighted geometric mean (Equation 16). Values of 0.4, 0.4 and 0.2 were chosen for n_1, n_2, and n_3 respectively to place more emphasize on the acceptability with regard to the simulations.

$$A = A_1^{n_1} A_2^{n_2} A_3^{n_3} \quad with \quad n_1 + n_2 + n_3 = 1 \tag{16}$$

The selection of the weights n_1, n_2, and n_3 determines which solution along the pareto-optimality sub-space will be found. The trade-offs between the various criteria can be studied using different weights [Seibert and McDonnell, 2002].

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 a varying set of criteria. Second, we compared the magnitude of parameter uncertainty when calibrating against only runoff and when calibrating against different combinations of criteria. We used a genetic algorithm, as described by Seibert [2000], for model calibration. This algorithm, which mimics evolution, 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 similar 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. Sixty calibration trials, each using 2500 model runs, were performed for each goodness-of-fit measure and the best 50 (of 60) parameter sets were used for further analysis of model performance and parameter identifiability.

Results

Model output

The model was able to reproduce observed runoff during the Aug-Dec period (Fig. 3). Model simulations calibrated with only hard data runoff values led to very good fits, with a model efficiency of 0.93. 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). This is not a new finding—hydrologists have known for years that getting a model to reproduce a hydrograph is not necessarily a robust test of how accurate or 'real' the model structure might be. Parameter ranges were poorly constrained when hard data only were used for calibration and the agreement of the calibrated parameter values with the experimentalist's knowledge was less than 0.4 (Fig. 4).

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 Figure 5a, the riparian and hollow box fail to behave like observed reservoir dynamics reported in McDonnell [1990], with too much water in the hollow box, especially between events. Figure 5b is an example where each of the three boxes filled and drained too quickly. Figure 5c shows an appropriate riparian box response but poor representation of the hollow zone, which is drained to quickly too. This is a compelling example of how relying only on the traditional single-criteria, 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 the runoff.

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 (based on soft groundwater, percent-new-water and parameter-value data) increased dramatically (Fig. 4 and 6). The combined objective function A (Eq. 16) increased from 0.46 to 0.79 when adding A_2 and A_3 to the calibration. 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.

The three-box model captured the water level dynamics extracted from the tensiometer data for both the riparian and the hollow box (Fig. 7) as also indicated by high r^2-values. It should be noted that using the coefficient of determination, r^2, we emphasized the dynamics and did not force the model to exactly fit the point observations (we allowed for an offset and a different amplitude). We also tested an alternative goodness-of-fit measure, which corresponded to the coefficient of determination, but with the constraint that the slope of the regression line was fixed to a value of one, *i.e.*, we still allowed for an offset but not for differences in amplitude. In that way the model was forced to better reproduce the amplitudes, which were observed at the two points in the catchment (see dotted lines in Fig. 7). Other results such as overall model performance

did not change significantly when using the alternative goodness-of-fit measure for the hard groundwater data.

The simulation with the best overall performance caused a somewhat reduced model efficiency for runoff but displayed more 'realistic' internal dynamics (Figure 6). Figure 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.

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 a measure of parameter uncertainty. The ratio between these values obtained from multi-criteria calibrations and those derived from runoff-only 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 (Fig. 8). 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 0.4, implying that using all criteria helped to reduce parameter uncertainty on average by 60% relative to the single criterion calibration against only runoff. The reduction of parameter uncertainty was most obvious for the coefficients of the linear outflow equations, despite the fact that no 'desirable' parameter ranges were specified for these parameters. Including hard groundwater data or soft data for new-water contribution to peak runoff

also reduced parameter uncertainty, but not as significantly as for the combination of all criteria.

Discussion

On the experimentalist's contribution to model development and calibration

The compilation of evaluation rules for model performance such as shown in Table 3 force the experimentalist to put numbers to his or her qualitative knowledge. This has been lacking in catchment hydrology for years. Dunne [1983], Klemeš [1986] and many others have called for experimentalists and modelers to unite—but this has been very slow in happening. We argue that the soft data discussions are a formal attempt at addressing Klemeš's and Dunne's challenge. The soft data numbers may themselves reflect some considerable uncertainty (as shown recently in experimental work by Sherlock *et al.* [2000]). The soft data approach also requires a number of subjective decisions, such as the specification of the evaluation rules and the weighing of the different objective functions. Nevertheless, we argue that the use of these data is still better than the alternative of neglecting this knowledge and using only hard data in the calibration process! While automatic model calibration has many advantages compared to the time-consuming manual trial-and-error method, others have argued that the automatic calibration reduces the modeling to simply a curve fitting exercise. Boyle *et al.* [2000] 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. The use of soft data is another step in the direction of infusing hydrological reasoning in automatic calibration. Our work complements the work of Boyle *et al.* [2000] by offering other forms of data to embrace in the calibration

process—information that often languishes in the data banks of experimental watersheds around the world, that hitherto have not yet been brought into the formal modeling process. We think that the reasons for this are due to the fact that modelers perceive this information to be too qualitative and not robust enough to be useful in any quantitative way. While we would agree that these soft data measures are often fuzzy, they **are** the type of data needed to move to more realistic simulations of catchment behavior. Furthermore, while not superior to hard data, soft data represent an untapped source of information available for calibration.

Types of soft data

The soft data measures used in this paper vary from static measures (*e.g.*, the spatial extent of the riparian zone) to data on groundwater level variations and highly integrated measures like the percent of new water at peakflow. We expect that if this soft data approach were attempted in other experimental catchments, choice of soft data measures could, and would, be different. In fact, the point is that 'one should use what one has on hand' for their catchment. Admittedly there is a plethora of ad hoc decisions to be made when using soft data measures to evaluate model performance. However, this should not discourage the modeler to heed these decisions by the experimentalist. Not using any of these data may be the poorest subjective decision of all.

The results of isotopic hydrograph separations have the advantage that the new-water contribution is an integrated measure of catchment response and offers much constrain on the perceptual model of runoff generation. Few studies to date have used isotope data in model calibration—despite the now common use of this in watershed analysis [Kendall and McDonnell, 1998]. Hooper *et al.* [1988] used continuous stream O-

18 to calibrate the Birkenes model—another simple conceptual box model of runoff response. Similarly, Seibert *et al.* [2002a] have used continues stream O-18 for model testing. In the present study, we use the new water ratio for discrete events rather than a continuous time series of O-18. Unlike Scandinavia where previous attempts have been made, the Maimai catchment shows several periods of rainfall 'cross-over' with stream baseflow and groundwater because of the lower amplitude of the seasonal O-18 variations—making continuous time series modeling less valuable. Nevertheless, the new-water soft-data measure is an example of making the most of data available for a given situation. In many catchment studies, additional (soft) data may be available that could be used to constrain model simulations. In snow-dominated environments, for instance, snow cover information may be used. In cases where the expansion and contraction of surface-saturated areas is important (and considered in the model), knowledge of the maximal portion of the catchment that might become saturated can be used. Franks *et al.* [1998] derived information on the extent of saturated areas from remote sensing and this information helped to constrain parameter values of TOPMODEL. In most cases measurements on the extent of saturated areas are not available, but hydrological reasoning and field experience might allow specifying a range of reasonable values (*e.g.* based on topography). The extent and spatial distribution of saturated areas might also be derived from vegetation and soil information [Güntner *et al.*, 1999].

Overall performance and internal consistency improvements with soft data

The model performance based on the various criteria varied between the parameter sets, which had been calibrated using different combinations of these criteria (Fig. 4).

Calibration against only one or two criteria resulted in poor simulations according to the other criteria, which were not used for calibration. For example, the best parameter sets according to runoff (with a median efficiency 0.92) were poor in their ability to correctly reproduce hard and soft groundwater levels (median r^2 =0.41 and median μ_{gw} =0.29). While the calibration against all criteria did not provide the best fits according to single criteria, the best overall performance was obtained in this way, as judged by the hard and soft data. Thus, while the runoff model efficiency dropped from 0.92 to 0.84 (median values) moving from 'no soft data' to 'all soft data', important process descriptors like the contribution of new water to peak runoff were much better reproduced (median $\mu_{new\ water}$=0.8 compared to 0.67), compared to the calibration using only hard data (A_1, runoff and groundwater).

Even in catchments where there is some groundwater-level data available as hard data for comparison, this data often only represents a limited number of locations. In our case we had data on groundwater levels from only two locations. We assumed that the dynamics were representative for each zone, but not necessarily for the mean depth to groundwater and the amplitude of the level variations. By using the coefficient of determination as goodness-of-fit measure for the hard groundwater data we did force the model to reproduce the dynamics, but used soft data to constrain the groundwater level simulations with regard to their absolute values and amplitudes. Results did not change significantly when using an alternative goodness-of-fit measure, which required the model to also reproduce the amplitude of the hard groundwater data. Probably results would have changed if we had used a goodness-of-fit measure that evaluated the simulated levels also with regard to their absolute values, but given the fact that the hard

groundwater data only represented one location in each zone, such a strong constraint would not have been warranted.

On the value of soft information

Runoff simulation for the Maimai watershed is relatively easy by comparison to many other catchments since there is only minimal seasonality and soils are highly transmissive and underlain by impermeable substrate. Previous TOPMODEL simulations at the site [Beven and Freer, 2001] and the present study have all achieved good fits for streamflow. However, simply modeling runoff with a high efficiency is of course not a robust test of model performance. Our work shows that sometimes lower R_{eff}-values are 'the price we have to pay' to obtain a better overall model performance and better adherence to the perceptual model of runoff generation. The question then becomes: Is this runoff efficiency reduction worth accepting in order to achieve a better conceptualization with respect to the soft data available? We argue from data presented in this paper that it is indeed worth accepting lower runoff-efficiency values if one can develop a more 'real' model of the catchment. The parameter set determined by using several criteria for calibration (based on hard and soft data) will in most cases lead to a poorer fit of simulated and observed catchment, but move the model to one that better captures the key processes that the experimentalist feels is important in controlling catchment response.

While this paper deals mostly with soft data and multi-criteria calibration, it should be stated that soft data first helped guide the box model construction. The three boxes chosen represent the experimentalist's objective definition of the key runoff producing reservoirs in the catchment based on observed water table dynamics, well

chemistry and soil/groundwater isotopic composition. The unsaturated-saturated zone coupling in the model was implemented because in catchments like Maimai, with shallow soils (~1.5 m) and impermeable substrate, tensiometric observations often highlight the importance of unsaturated zone conversion to transient groundwater during events. Future applications of the three-box model and soft data strategy to other catchments though, need to make these decisions according to their conceptualization of runoff generation processes at their locale. There are certainly other box configurations that could be envisioned. Again, the point is here that the first step, before using soft data for calibration, is the construction of a model that is appropriate for the catchment of concern. We believe the explicit volume-based box structure is a way forward and we are actively re-working this structure for other watersheds.

Lastly, we should acknowledge that this exercise has changed the way that we might conduct the next field campaign at Maimai. Future experiments need to move away from detailed hillslope transect and point-scale studies and more towards capturing the first order controls on different landscape units. Groundwater-streamflow relationships may provide further guidance on how to discretize key catchment reservoirs and how to parameterize their response function. We would envision new experiments comparing the well response in different landscape units to streamflow—with the direction and magnitude of hysteresis as the objective measure of unit response [Seibert *et al.*, 2002b]. The model work in this paper has also emphasized the importance of the soil's drainable porosity or specific yield on water table responses in the different landscape units. New experimental work at Maimai should put more weight on the estimation of this variable and its variation with location and depth.

Concluding remarks

The study used multi-criteria soft data for model development and for internal calibration and validation. We show that conceptual modeling of catchment hydrology can include identification of parameter values through calibration against hard and soft data. We believe that this approach is the way forward for development of more realistic models of catchment behavior using soft data where multiple criteria can now be used to constrain the model in various ways. These soft data are a representation of qualitative knowledge from the experimentalist, which cannot be used directly as exact numbers but is made useful through fuzzy measures of model-simulation and parameter-value acceptability. We argue that the necessary dialog that must occur between the modeler and the experimentalist can be made explicit in this way. We propose that this approach is also useful for comparing the value of different field measurements that experimentalists might make in support of modeling. We are currently exploring other types of soft data (*e.g.* mean residence time data) as we move to larger watershed scales and begin to incorporate conservative mixing between reservoirs. Our main message in this work is that rather than being 'right for the wrong reasons', a better process representation of catchment hydrology in conceptual runoff modeling should be 'less right, for the right reasons'.

Acknowledgments

We thank Keith Beven and the two anonymous reviewers for valuable comments and Jim Freer for compiling the data. This research was partly funded by the Swedish Research Council (grant 620-20001065/2001) and NSF grant EAR-9805475.

References

Aronica, G., B.Hankin and K.Beven, 1998. Uncertainty and equifinality in calibrating distributed roughness coefficients in a flood propagation model with limited data. *Advances in Water Resources*, 22: 349-365.

Bencala, 2000. Hyporheic zone hydrological processes. *Hydrological Processes,* 14: 2797-2798.

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

Beven, K., 1993. Prophecy, reality and uncertainty in distributed hydrological modeling, *Advances in Water Resources*, 16: 41-51.

Beven, K., 2001. How far can we go in distributed modeling? *Hydrology and Earth System Science*, 5: 1-12.

Beven, K. and Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology, *Journal of Hydrology*, 249: 11-29.

Bonell, M., 1998. Selected challenges in runoff generation research in forests from the hillslope to headwater drainage scale. *Journal of the American Water Resources Association*, 34: 765-785.

Boyle, DP, H.V. Gupta and S. Sorooshian, 2000. Towards improved calibration of hydrological models: combining the strengths of manual and automatic methods. *Water Resources Research*, 36: 3663-3674.

Buttle, J.M., 1994. Isotope hydrograph separations and rapid delivery of pre-event water from drainage basins. *Progress in Physical Geography*, 18: 16-41.

Cirmo, C.P. and McDonnell, J.J., 1997. Linking the hydrologic and biochemical controls of nitrogen transport in near-stream zones of temperate-forested catchments: a review. *Journal of Hydrology* 199: 88-120.

Dunne, 1983, Relation of field studies and modeling in the prediction of storm runoff. *Journal of Hydrology*, 65: 25-48.

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

Franks, S. and K. Beven, 1997. Estimation of evapotranspiration at the landscape scale: A fuzzy disaggregation approach. *Water Resources Research*, 33: 2929-2938.

Freer, J., Beven, K.J. and Ambroise, B., 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research* 32: 2161-2173.

Güntner, A., Uhlenbrook, S., Seibert, J. and Leibundgut, Ch., 1999. Multi-criterial validation of TOPMODEL in a mountainous catchment. *Hydrological Processes* 13: 1603-1620

Gupta, H., S. Sorooshian and P. Yapo, 1999. Towards improved calibration of hydrologic models: Multiple and noncommensureable measures of information. *Water Resources Research*, 34: 751-763.

Hankin, B.G., and K.Beven, 1998. Modelling dispersion in complex open channel flows: Fuzzy calibration (2), *Stochastic Hydrology and Hydraulics*, 12: 397-412.

Harris, D.M., J.J. McDonnell and A. Rodhe, 1995. Hydrograph separation using continuous open-system isotopic mixing. *Water Resources Research*, 31: 157-171.

Hooper, R.P., Stone, A., Christophersen, N., de Grosbois, E., and Seip, H.M., 1988. Assessing the Birkenes model of stream acidification using a multisignal calibration methodology. *Water Resources Research* 24: 1308-1316.

Hooper, R., B. Aulenbach, D. Burns, J.J. McDonnell, J. Freer, C. Kendall and K. Beven, 1998. Riparian control of streamwater chemistry: Implications for hydrochemical basin models. *International Association of Hydrological Sciences*, Publication 248: 451-458.

Klemeš, 1986. Dilettantism in hydrology: transition or destiny. *Water Resources Research* 22: 177S-188S.

Kuczera, G., 1983. Improved parameter inference in catchment models, 2. Combining different kinds of hydrologic data and testing their compatibility. *Water Resources Research*, 19: 1163-1172.

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

Lindström, G., 1997. A simple automatic calibration routine for the HBV model. *Nordic Hydrology*, 28: 153-168.

Kendall, C. and J.J. McDonnell (eds.), 1998. Isotope tracers in catchment Hydrology, Elsevier Science Publishers, 816 pp.

McDonnell, J.J., 1989. The age, origin and pathway of subsurface stormflow. PhD Thesis, *University of Canterbury, Christchurch*, 270 pp.

McDonnell, J.J., 1990. A rationale for old water discharge through macropores in a steep, humid catchment. *Water Resources Research*, 26: 2821-2832.

McDonnell, J.J., M.K. Stewart and I.F. Owens, 1991. Effects of catchment-scale subsurface watershed mixing on stream isotopic response. *Water Resources Research*, 26: 3065-3073.

McDonnell, J.J., D. Brammer, C. Kendall, N. Hjerdt, L. Rowe, M. Stewart and R. Woods, 1998. Flow pathways on steep forested hillslopes: The tracer, tensiometer and trough approach, In Tani *et al.* (eds). *Environmental Forest Science*, Kluwer Academic Publishers, pp. 463-474

McGlynn, B, J.J. McDonnell and D. Brammer, 2002. An evolving perceptual model of hillslope flow in a steep forested humid catchment: A review of the Maimai catchment. *Journal of Hydrology*, 257: 1-26.

McKie, D.A., 1978. A study of soil variability within the Blackball Hill Soils, Reefton, NewZealand. M.Ag.Sc. Thesis, University of Canterbury, 180 pp.

Mosley, M., 1979. Streamflow generation in a forested watershed, New Zealand. *Water Resources Research*, 15: 795-806.

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

Pearce, A., M Stewart and M. Sklash, 1986. Storm runoff generation in humid headwater catchments, 1: Where does the water come from? *Water Resources Research*, 22: 1263-1272.

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

Rowe, L., A. Pearce and C. O'Loughlin, 1994. Hydrology and related changes after harvesting native forest catchments and establishing Pinus radiata plantations. Part 1: Introduction to the study. *Hydrological Processes*, 8: 263-279.

Seibert, J., 1997. Estimation of parameter uncertainty in the HBV model, *Nordic Hydrology*, 28: 247-262

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

Seibert, J. and McDonnell, J., 2002. Multicriteria calibration of conceptual runoff models – the quest for an improved dialog between modeler and experimentalist. *Submitted to AGU Book "Advances in Calibration of Watershed Models",* in review

Seibert, J., A.Rodhe, K.Bishop, 2002a. Simulating interactions between saturated and unsaturated storage in a conceptual runoff model, *Hydrological Processes*, in press.

Seibert, J, K. Bishop, A. Rodhe and J. McDonnell, 2002b. Groundwater dynamics along a hillslope: A test of the steady-state hypothesis. *Water Resources Research*, in review.

Sklash, M., M. Stewart and A. Pearce, 1986. Storm runoff generation in humid headwater catchments, 2, A case study of hillslope and low-order stream response, *Water Resources Research*, 22: 1273-1282.

Sherlock, M.D., Chappell, N.A. and McDonnell, J.J., 2000. Effects of experimental uncertainty on the calculation of hillslope flow paths, *Hydrological Processes*, 14: 2457-2471.

Tables

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	[-]
β	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

Table 2. The three different ways of evaluating model acceptability based on hard data (A_1) and soft (A_2 and A_3) data.

	Acceptability according to …	**Example**	**Measure**
$\boldsymbol{A_1}$	Fit between simulated and observed data	Runoff	Efficiency
$\boldsymbol{A_2}$	Agreement with perceptual (qualitative) knowledge	New water contribution	Percentage of peak flow for certain events
$\boldsymbol{A_3}$	Reasonability of parameter values	Spatial extension of riparian zone	Fraction of catchment area

Table 3. 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. 14)

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]

Figure captions

1. 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, z_{max}: maximal groundwater level above bedrock, U: unsaturated storage).

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

3. Accumulated rainfall, runoff model error, and observed and simulated runoff for the period September-December 1987. Measured data is shown as dashed line. The simulation of runoff (solid line) is based on the calibration using only runoff data.

4. Goodness-of-fit measures for runoff, groundwater levels (as derived from the tensiometer data), 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.

5. Three model runs with different parameter sets resulting in different groundwater dynamics. All three parameter sets had been calibrated to observed runoff and gave a

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.

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.

7. Comparison of groundwater level simulations (dashed line) for the riparian and the hollow box and levels extracted from tensiometer observations (solid line) (levels are here given in m below ground surface). The dotted line shows the simulated groundwater levels when using the alternative goodness-of-fit criteria.

8. Reduction of parameter uncertainty by using additional calibration criteria compared to a single-criterion calibration. The ratio (v_{multi} / v_{single}, where v is the range between the 0.1 and 0.9 percentile divided by the median) is shown for all 16 model parameters and the median ratio is shown to the right. A ratio below one (dashed line) indicates a reduction of parameter uncertainty. The vertical bars show the ratio when using the combination of all criteria, and the symbols show the ratio for different combinations of criteria based on hard and soft data.

Figures

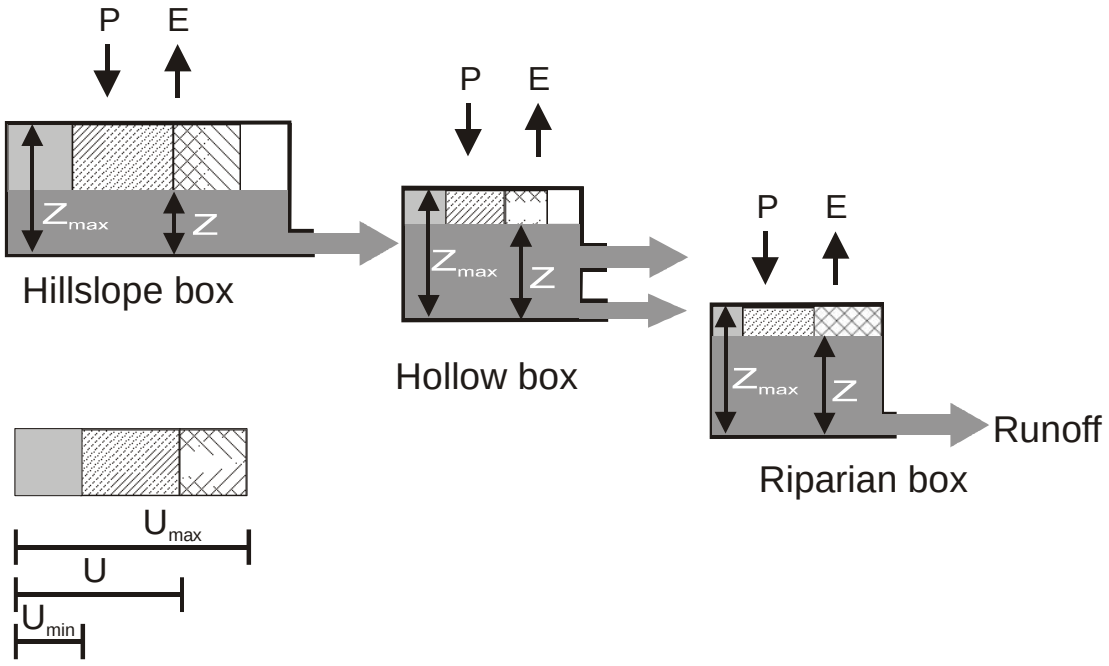

Figure 1

Fig. 1.

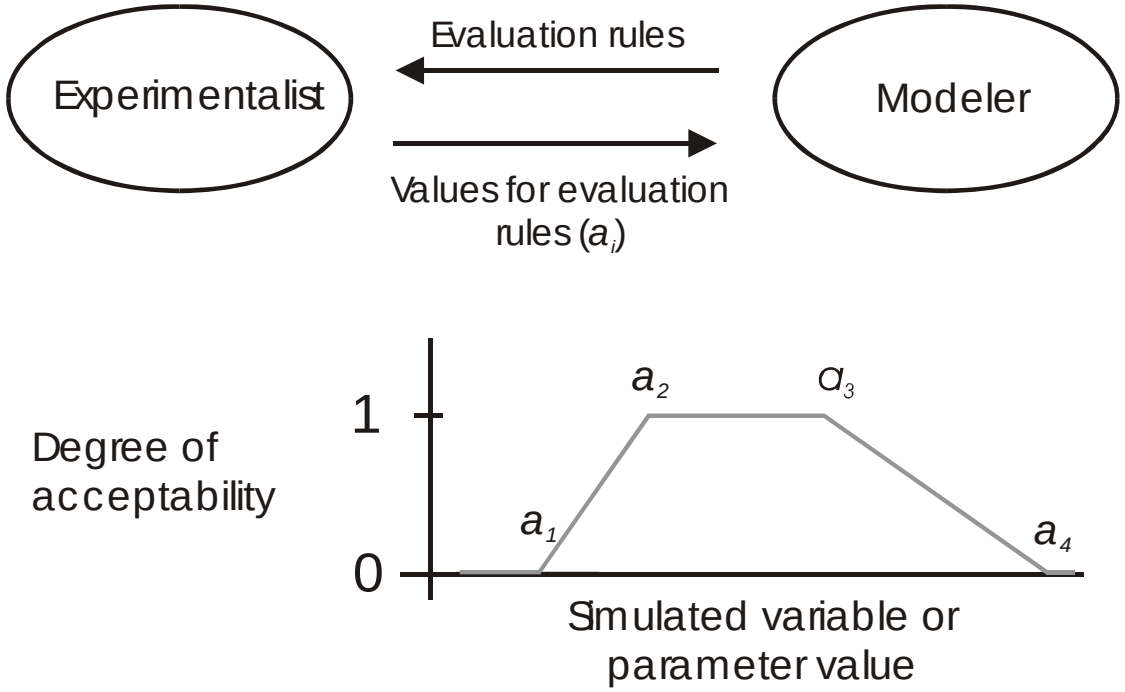
Experimentalist
Evaluation rules
Values for evaluation rules (a_i)
Modeler
Degree of acceptability
1
0
a_1
a_2
a_3
a_4
Simulated variable or parameter value

Fig. 2

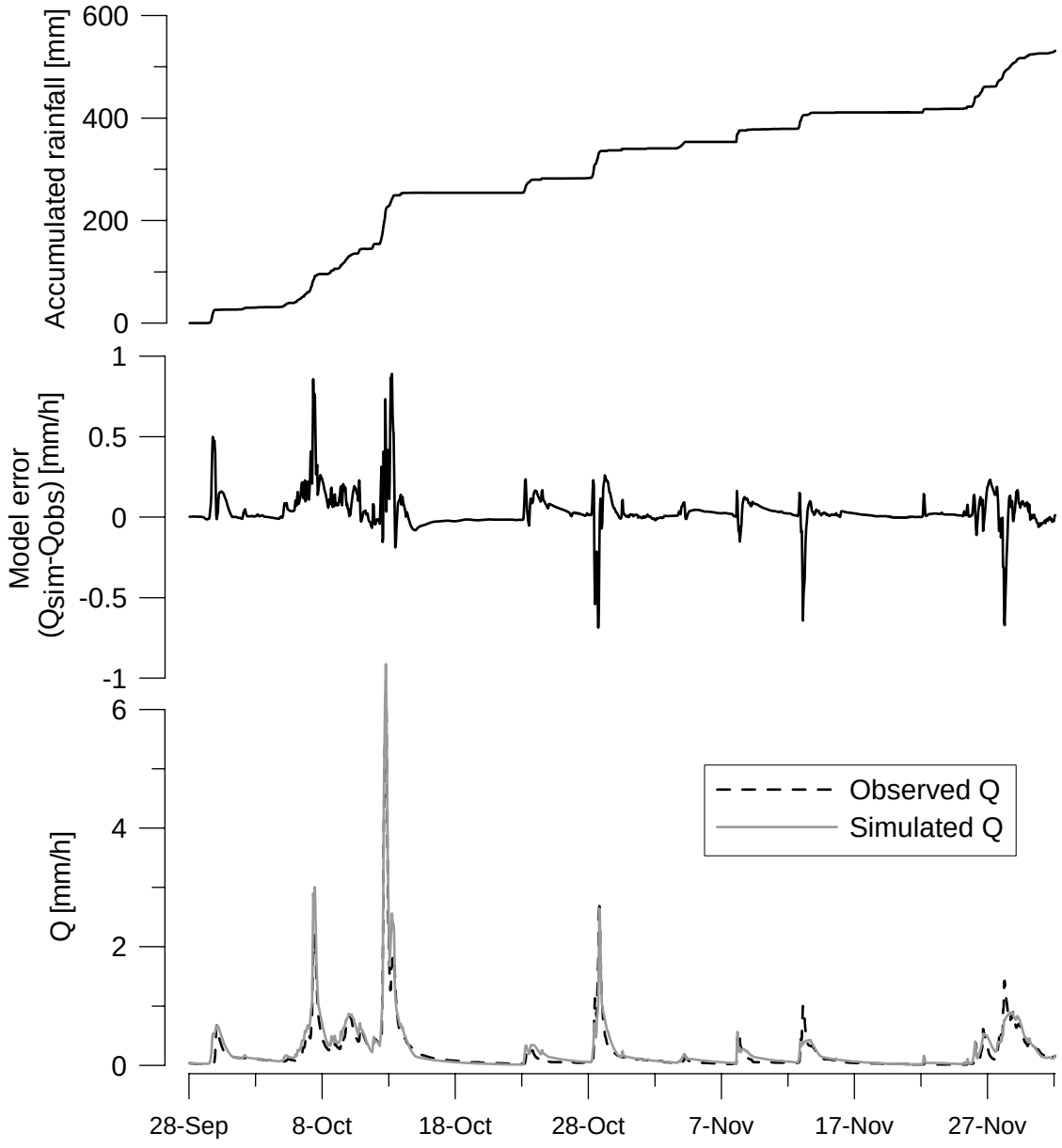
Accumulated rainfall [mm]
600
400
200
0
Model error
(Qsim-Qobs) [mm/h]
1
0.5
0
-0.5
-1
Q [mm/h]
6
4
2
0
Observed Q
Simulated Q
28-Sep
8-Oct
18-Oct
28-Oct
7-Nov
17-Nov
27-Nov

Fig. 3

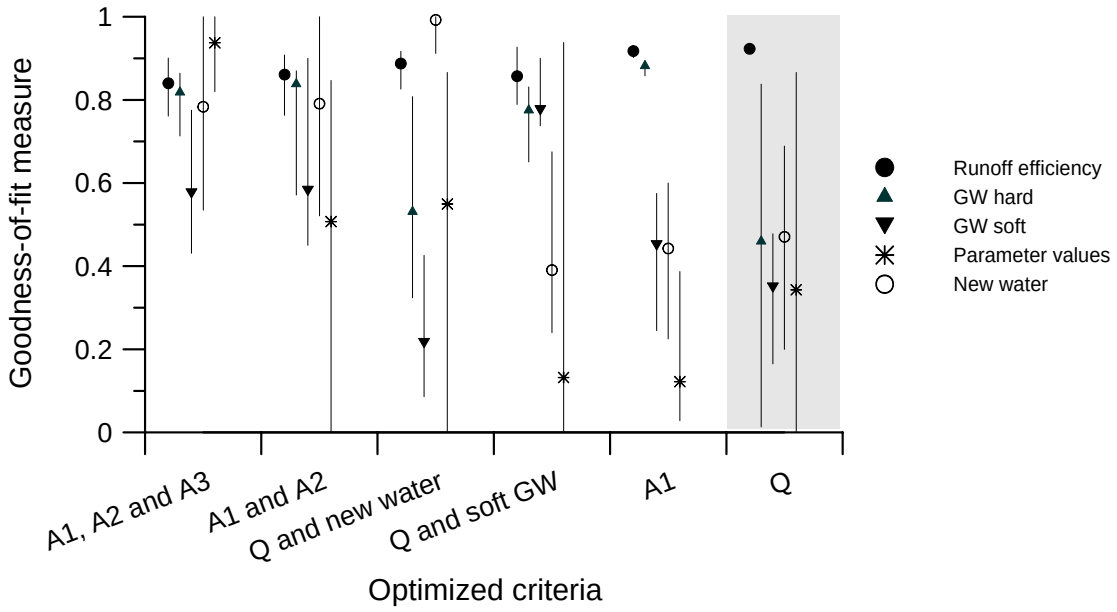
Goodness-of-fit measure
1
0.8
0.6
0.4
0.2
0
A1, A2 and A3
A1 and A2
Q and new water
Q and soft GW
A1
Q
Optimized criteria
Runoff efficiency
GW hard
GW soft
Parameter values
New water

Fig. 4

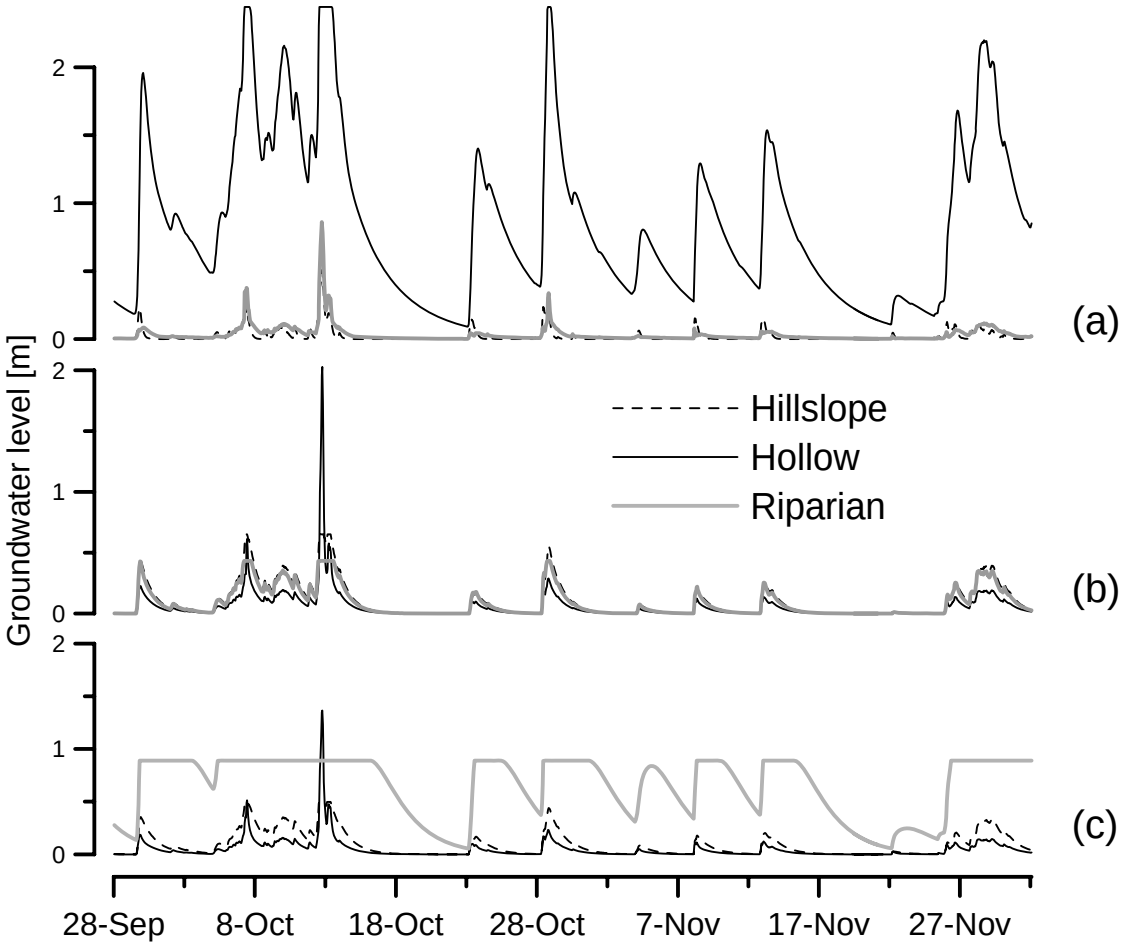
Groundwater level [m]
2
1
0
(a)
(b)
(c)
Hillslope
Hollow
Riparian
28-Sep
8-Oct
18-Oct
28-Oct
7-Nov
17-Nov
27-Nov

Fig 5.

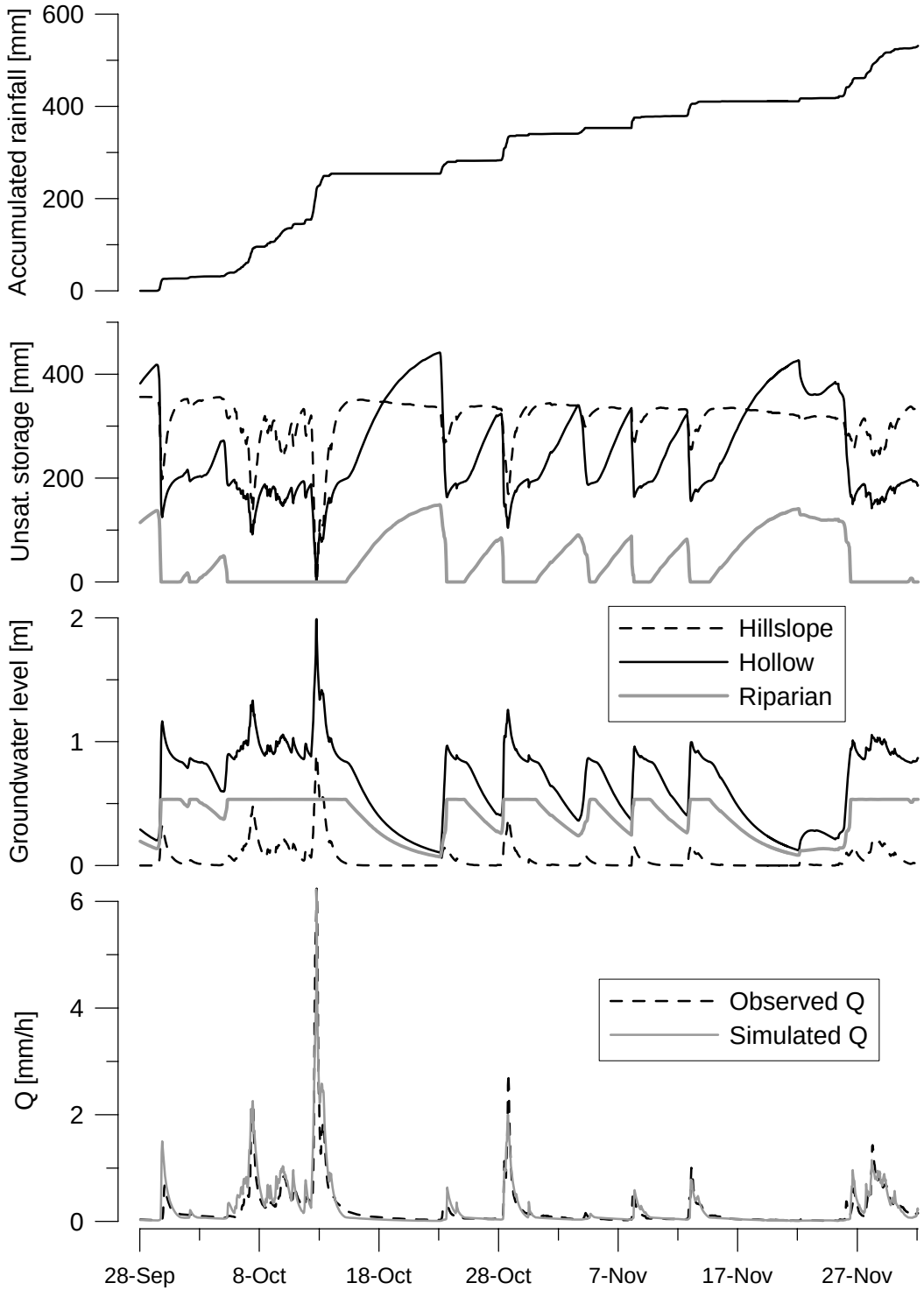
Accumulated rainfall [mm]
600
400
200
0
Unsat. storage [mm]
400
200
0
Groundwater level [m]
2
1
0
Hillslope
Hollow
Riparian
Q [mm/h]
6
4
2
0
Observed Q
Simulated Q
28-Sep
8-Oct
18-Oct
28-Oct
7-Nov
17-Nov
27-Nov

Fig. 6

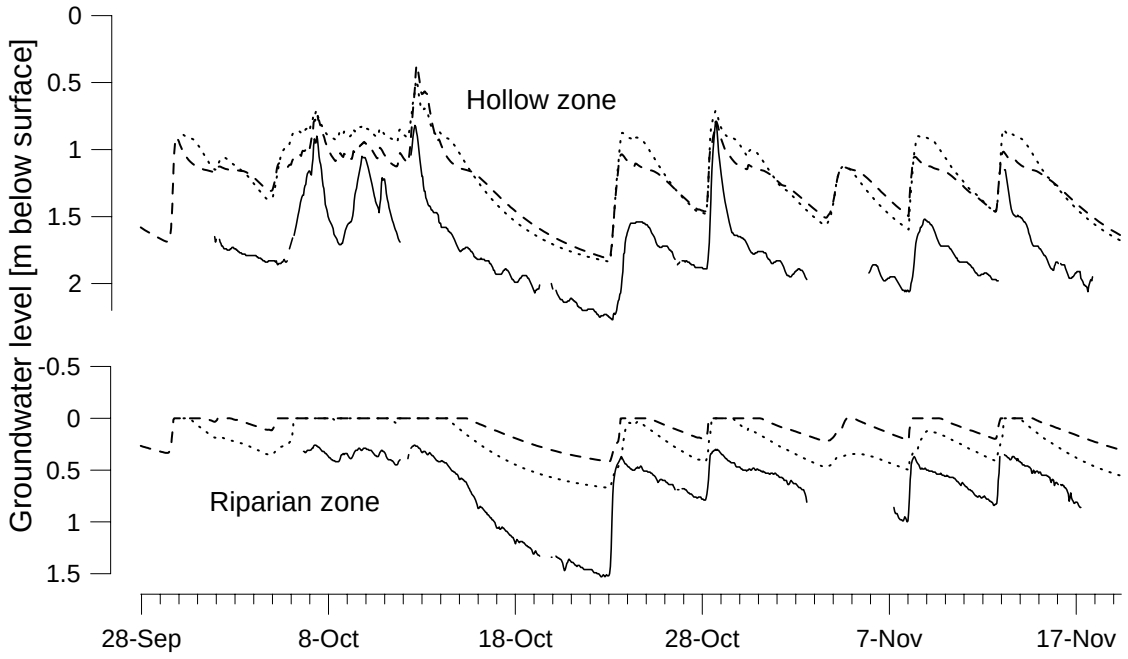
Groundwater level [m below surface]
0
0.5
1
1.5
2
Hollow zone
-0.5
0
0.5
1
1.5
Riparian zone
28-Sep
8-Oct
18-Oct
28-Oct
7-Nov
17-Nov

Fig. 7

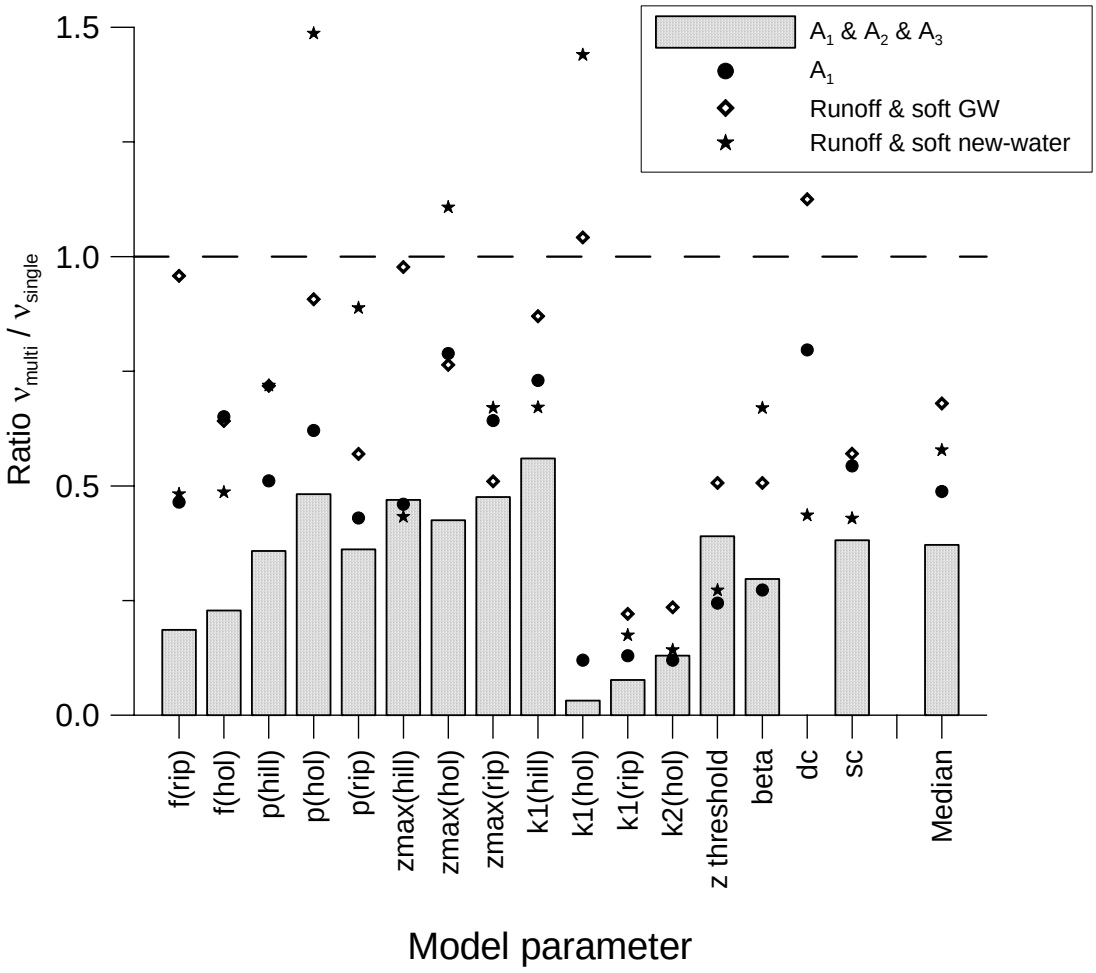
1.5
1.0
0.5
0.0
Ratio v_{multi} / v_{single}
A_1 & A_2 & A_3
A_1
Runoff & soft GW
Runoff & soft new-water
f(rip)
f(hol)
p(hill)
p(hol)
p(rip)
zmax(hill)
zmax(hol)
zmax(rip)
k1(hill)
k1(hol)
k1(rip)
k2(hol)
z threshold
beta
dc
sc
Median
Model parameter

Fig. 8