Gauging the Ungauged Basin: Relative Value of Soft and Hard Data

J. Seibert[1] and J. J. McDonnell[2]

Abstract: The long-standing issue of hydrological predictions in ungauged basins has received increased attention due to the recent International Association of Hydrological Sciences (IAHS) decade on predictions in ungauged basins (PUB). Since the outset of PUB, many have noted that the best way to confront an ungauged basin is to first make some basic streamflow measurements. This study explores the value of a rudimentary gauging campaign for making predictions in an ungauged basin. The well-studied Maimai watershed in New Zealand was used as a hypothetical ungauged basin, and this study was designed to start with no runoff data and add iteratively different subsets of the available data to constrain the calibration of a simple three-reservoir conceptual catchment model. These subsets included single runoff events or a limited number of point values—in other words, what could be measured with limited, campaign-like field efforts in an ungauged basin. In addition, different types of soft data were explored to constrain the model calibration. Model simulations were validated using the available runoff data from different years. It was found that surprisingly little runoff data were necessary to derive model parameterizations that provided good results for the validation periods, especially when these runoff data were combined with soft data. The relative value of soft data increased with decreasing amount of streamflow data. The findings from the Maimai watershed suggest that, when starting with no flow information, one event or 10 observations during high flow provide almost as much information as three months of continuously measured streamflow for constraining the calibration of a simple catchment model. **DOI: 10.1061/(ASCE)HE.1943-5584.0000861.** © *2014 American Society of Civil Engineers.*

Author keywords: Prediction in ungauged basins; Catchment modelling; Soft data; Maimai; Value of data.

Introduction

Making predictions in ungauged basins is a grand challenge for hydrology (from Sivapalan 2003). The first International Hydrological Decade (IHD, 1965–1974) resulted in the gauging of over 3,500 catchments around the world (Falkenmark and Chapman 1989). For the modeler, these data were a boon to model development, testing and validation, but left a legacy of reliance on data for model calibration. In the past decade, the modeling community has recognized the limitations of model calibration based on streamflow data with its associated uncertainty, issues of parameter identifability and equifinality (Beven 2001). Furthermore, models relying on calibration are usually restricted to catchments where a long series of runoff data already exists. Along the way, some modelers have assumed that more physically based, fully distributed approaches might overcome the need for model calibration. This assumption has been increasingly questioned due to the difficulties of measuring the relevant parameters at the relevant scale. A consensus is emerging that simple models of catchment hydrology can play an important role if the requirement of long data series for calibration can be relaxed (Beven 2002).

For the experimentalist, the IHD was a formative period during which most of the main post-Hortonian runoff formation processes were defined (Beven 2006). While its legacy is that of intense hydrological discovery, the IHD provided little instruction on what to measure, in what order and why. Most gauging was aimed either at rainfall-runoff description or very detailed point-based measurements aimed at process discovery in certain zones within a study watershed. Indeed, these two extremes characterize watershed gauging even today—where rather than a systematic diagnosis of catchment function and behavior, studies often jump from whole-watershed mass balance analysis to hillslope trenching or plot-scale macropore flow monitoring (Weiler and McDonnell 2007).

Indeed, one might characterize current—experimental approaches to catchment gauging as ad hoc (McDonnell et al. 2007). There is little post-IHD consensus as to where to locate a weir or rain gauge, let alone where to take any other supplemental measurements to aid in the development of a local model structure or test information. Only slowly has the dialog between experimentalist and modeler begun (Woods and Rowe 1996).

Here the challenge of how to gauge the ungauged basin by exploring the value of limited streamflow measurements and additional soft data is addressed. Soft data are data that can be obtained in the field during limited field campaigns and represent qualitative knowledge from the experimentalist that cannot be used directly as exact numbers. Soft data can be made useful through fuzzy measures of model-simulation and parameter-value acceptability (Seibert and McDonnell 2002). 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

[1]Dept. of Geography, Univ. of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland; Dept. of Earth Sciences, Uppsala Univ., Sweden; and Dept. of Physical Geography and Quaternary Geology, Stockholm Univ., Sweden (corresponding author). E-mail: jan.seibert@geo.uzh.ch

[2]Global Institute for Water Security, Univ. of Saskatchewan, National Hydrology Research Centre, 11 Innovation Boulevard, Saskatoon, SK, Canada S7N 3H5; and School of Geosciences, Univ. of Aberdeen, Aberdeen, UK.

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 (Aronica et al. 1998; Hankin and Beven 1998; Seibert 1997). Pappenberger et al. (2007), for instance, used a fuzzy-rule based calibration motivated by uncertain remotely sensed flood inundation 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 this case, *this parameter set is the best possible set* is true. Soft data are a rather different type of information than traditional hard data measures. Soft data are often spotty, discontinuous, and numerically approximate.

Since the concept of soft data was introduced (Seibert and McDonnell 2002), it has proved useful in augmenting data in gauged catchments for model structural development (Weiler and McDonnell 2007) and multicriteria model calibration (Vache and McDonnell 2006; Winsemius et al. 2009). In many ways, soft data can be seen as the quintessential PUB data. When encountering an ungauged basin for the first time, the hydrologist's perceptual model of the site is often a highly detailed yet qualitative understanding of dominant runoff processes. Thus, there exists in addition to any limited hard data (e.g., limited streamflow gauging) *soft data* about catchment hydrology. Recently, Seibert and Beven (2009) have shown that a few runoff measurements can contain much of the information content of continuous runoff time series, showing that even limited efforts at stream gauging could yield useful weighted ensemble means for modeling in poorly gauged catchments. In similar studies the value of limited amounts of other types of data, additional to runoff, has been demonstrated for groundwater levels (Juston et al. 2009) and glacial mass balances (Konz and Seibert 2010).

This paper goes beyond these analyses by asking: *What is the additional value of soft data relative to the value of limited stream gauging?* and, *How might this guide gauging of ungauged basins and the dialog between experimentalist and modeler?* The work is based on using the well-studied Maimai watershed in New Zealand as a hypothetical ungauged basin, where the study was designed to start with no runoff data and add different subsets of the available data to constrain the calibration of a simple catchment model. These subsets mimic and test different possible field sampling strategies and include single runoff events or a limited number of point values; in other words these data represent what could be measured with limited efforts in an ungauged basin. Besides these runoff data, different types of soft data were used (also obtainable in a couple of weeks of field work) to constrain the model calibration and to facilitate communication between experimentalist and modeler for new ways to test models and to improve parameter identifiabilty.

## Material and Methods

Study Site

The Maimai research catchments are a set of highly responsive, steep, and wet watersheds on the west coast of the South Island of New Zealand. Maimai has a long history of hillslope hydrological research [for a complete review see McGlynn et al. (2002)]. More importantly, unlike sites of other experimental work, Maimai shows striking simplicity in catchment response. Several recent model studies have used the Maimai dataset as the basis for the development and testing of new model structures (Fenicia et al. 2008; Seibert and McDonnell 2002; Weiler and McDonnell 2007) and multicriteria model calibration techniques (Freer et al. 2004;

Vache and McDonnell 2006; Vache et al. 2004). Similarly steep and wet catchments have been a focal point for experimental studies and conceptual model development since the mid-1970s (Graham et al. 2010; McDonnell et al. 1990, 1991; McGlynn et al. 2003; McGlynn and McDonnell 2003; Mosley 1979, 1982; Sidle et al. 2000; Stewart and McDonnell 1991; Woods and Rowe 1996). The simplicity in catchment response is driven largely by the lack of seasonality and the chronically wet state of the system. Soils rarely drain below 10% of saturation (Mosley 1979) and overlie a effectively impermeable, compacted and cemented pliestocene conglomerate. Quickflow [QF as defined originally by Hewlett and Hibbert (1967) comprises 65% of the mean annual runoff and 39% of annual total rainfall (P) (Pearce et al. 1986)].

Data from the three-month time series of McDonnell (1989) were used—the same data series used in the aforementioned model studies. These data cover the period mid-September to December for the year 1987 and include 11 rainfall events that range in size from 26 to 105 mm. Rainfall and discharge data collected for the months August–December in 1985 and 1986 were used for validation. Discharge data was collected at the outlet of the 3.8 ha catchment with a 90° v-notch weir at 10 min intervals. Rainfall data was collected near the gauge with a tipping bucket gauge (Campbell Scientific) accumulated over 10-min intervals.

Three-Box Model

A simple conceptual model developed by Seibert and McDonnell (2002) was used in this study. This model builds on Hydrologiska Byråns Vattenbalansavdelning (HBV) model concepts (Bergström 1992; Lindström et al. 1997), but represents different hydrological and geomorphological zones through simple boxes or reservoirs: hillslope, hollow, and riparian zones (Fig. 1). A similar modeling approach, with two boxes, has previously been used for an application in Sweden (Seibert et al. 2003). While the present three-box model was originally developed with the Maimai catchment in mind, it is generic enough that one might use such a model in any steep, humid catchment. In this conceptual hydrologic model, the three reservoirs are characterized by different groundwater dynamics (McDonnell 1990) and distinguishable by their water isotopic characteristics (McDonnell et al. 1991). Water flow is then simulated to occur from the hillslope box into the hollow box and from the hollow zone into the riparian zone. Streamflow is simulated as flow from the riparian zone. For each box, a coupled formulation of the saturated and unsaturated storage was used, which implies that the maximum space for unsaturated storage
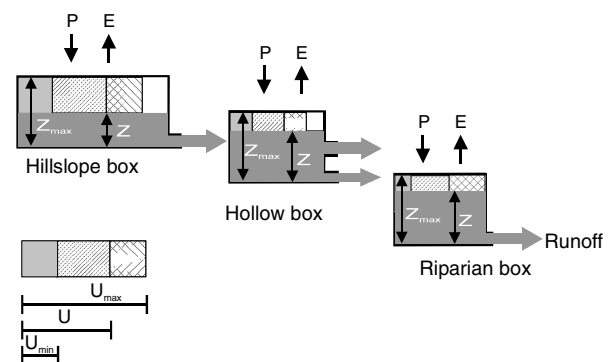
Fig. 1. Structure of the three-box model including hillslope, hollow and riparian zone reservoirs (P: precipitation, E: evaporation, z: groundwater level above bedrock, z_{max}: maximal groundwater level above bedrock, U: unsaturated storage) (Seibert and McDonnell 2002)

is determined by the simulated amount of saturated storage. For a more detailed description and equations of the three-box model, the reader is referred to Seibert and McDonnell (2002) and Seibert et al. (2003).

Soft Data

For model evaluation based on soft data the same evaluation rules as in the previous study (Seibert and McDonnell 2002) were used here. These rules constrain model calibrations in two ways: by evaluating the model with regard to simulations and by assessing the acceptability of certain parameter values. Note that 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. The rules were, thus, defined as trapezoidal functions, where the experimentalist was asked to assign values to four variables a_i [Eq. (1)]. These fuzzy measures of acceptance were used to compute the degree of acceptance, μ, for a certain parameter set from the corresponding simulated quantity or parameter value x. In other words, the trapezoidal function is used to map the experimentalist's knowledge into a quantity, which then can be used for evaluation of model simulations

$$\mu(x) = \begin{cases} 0 & \text{if } x \leq a_1 \\ \frac{x-a_1}{a_2-a_1} & \text{if } a_1 \leq x < a_2 \\ 1 & \text{if } a_2 \leq x < a_3 \\ \frac{a_4-x}{a_4-a_3} & \text{if } a_3 \leq x < a_4 \\ 0 & \text{if } x > a_4 \end{cases} \quad (1)$$

Altogether, 15 evaluation rules were used and for each of these the experimentalist provided the a_i values defining the trapezoidal function to the modeler (Seibert and McDonnell 2002, Table 3). These rules evaluated model performance with regard to soft data describing maximum and minimum groundwater levels for the different zones, frequencies of groundwater levels above and below certain levels and the contribution of new (event precipitation) water to event runoff. The overall acceptability of the model simulations with regard to soft data was evaluated as the arithmetic mean of the 15 evaluation rule values.

As in the previous study the overall acceptability of a parameter set was defined by three components: (1) the model efficiency (R_{eff}) values (Nash and Sutcliffe 1970) for the hard runoff data (calculated based on subsets of the total runoff series, see below) (A_1), (2) the acceptability of the model simulations with regard to soft data (A_2), and (3) the acceptability of the parameter values based on the experimentalist's understanding (A_3). In each case, a value of one for A_i implies a perfect fit (or full acceptability). The overall acceptability, A, of a parameter set, which then was used as the objective function, was computed as a weighted geometric mean with the weights n_1 (set to 0.4), n_2 (0.4), and n_3 (0.2) [Eq. (2)]

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

Model Calibration and Validation

Different subsets of the available runoff data were tested from the period mid-September to mid-December 1987 (pretending that the full time series was not available). Different types of subsets were used, as shown in Fig. 2, meaning different assumed field data gathering campaigns with respect to flow. The first subset mimicked a field party would capturing one of four different runoff events during the three-month period. The assumption was that
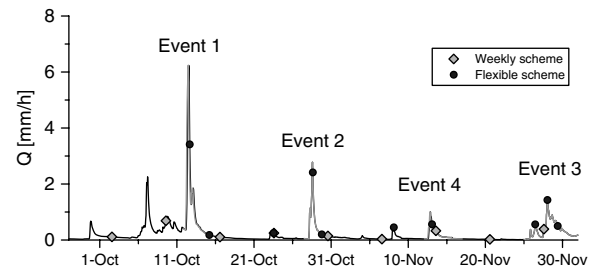
Fig. 2. Hydrograph showing in grey and with circle and diamond symbols the different subsets of the 1987 runoff series used for model calibration; for the flexible scheme with two respective four observations, the two and four observations latest in the series (i.e., to right) were used

runoff was measured continuously for each of these four measured events. Next, a scenario with one gauging outing per week with no conditioning on flow situations (i.e., the uninformed observer scenario) was created. Finally, a flexible flow-based gauging scenario was used that might be typical for real cases with gauging during high-flow conditions but missing the peaks by some hours. This flexible scheme was tested for three different numbers of gauging occasions: ten, four, and two observations (Fig. 2).

For each of the four gauging subsets, the three-box model was calibrated using a generic algorithm (Seibert 2000). The sum of squared errors was used as objective function for the optimization. The generic algorithm has stochastic elements and might, thus, find solutions at different locations in the parameter space when different parameter sets provide similarly good results. Therefore, 50 different calibration trials were performed for each subset of runoff data. The calibrated parameter sets were then evaluated based on their ability to reproduce the entire runoff series from the different years. This was done both for data from 1987 (mid-September–mid–December), which represents the period from which the different subsets of limited streamflow data had been selected, and for two completely independent five-month periods in 1985 and 1986 (August–December). Based on the 50 model runs with different parameter sets median values as well as the 10 and 90% were computed.

The limited runoff data were combined with the soft data used by Seibert and McDonnell (2002) for additional evaluation according to a number of soft data criteria. While the particular soft data used in this study were those used previously (Seibert and McDonnell 2002) and admittedly based on several years' fieldwork in this catchment, in an ungauged basin context similar soft data measures could be formulated based on expert knowledge from similar catchments or that collected in a few weeks of field work. These soft data include information on groundwater dynamics, new-water contributions and parameter values based on field experience (such as soil depth, saturated area extent, etc.). Both using all available soft data and using different subsets of the soft data was tested, i.e., the different types of soft data were also tested individually. For calibration, the combined objective function [Eq. (2)] was used, as suggested previously (Seibert and McDonnell 2002).

For calibrations based only on streamflow data, 50 trials were performed and the 50 parameter sets were evaluated based on their runoff simulations for the periods described above. For model validation the focus was on runoff simulations only, and the model efficiency was used to evaluate model performance over the entire observation period.
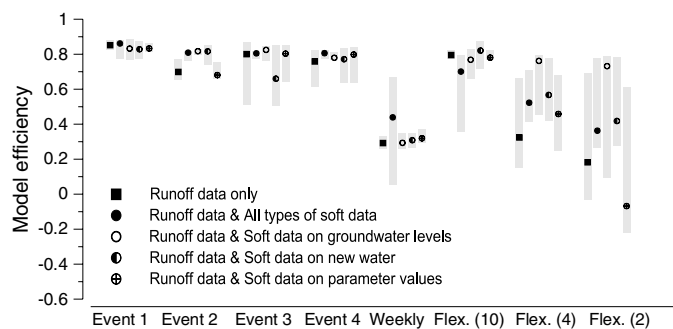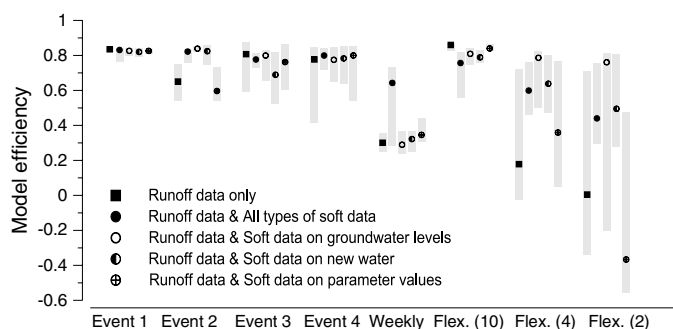
Fig. 3. Performance of runoff simulations using parameter sets calibrated based on different subsets of runoff data from 1987; the symbol indicates the median of 50 calibration trials and the grey shade indicates the 10–90% range of all calibration trials

Results

Significant differences were found in model performances for the complete runoff series compared to the different scenarios. The highest median model efficiency values were between 0.8 and 0.85. These values, while high were lower compared to the highest values obtained when using all runoff data for calibration, which were 0.92 for the scenario where calibration was made against runoff only and 0.84 when also considering the soft data and optimizing different criteria (Seibert and McDonnell 2002).

There was a clear performance difference between the different scenarios of the runoff data for calibration (Fig. 3). When using only subsets of runoff data for calibration, entire events generally resulted in good model performances with a tendency for better performances for the larger events. Using single weekly runoff observations resulted in rather poor performances. The 10 flexibly chosen observations resulted in good performances, whereas performances decreased when using only two or four observation dates (Fig. 3). Similar results were obtained when validating the different model calibrations based on runoff simulation performance for the two periods in 1985 and 1986 (Figs. 4 and 5).

Including soft data resulted generally in better model performance during the different validation periods across all scenarios (Fig. 3). The only exception was the *10 flexibly chosen days* scenario for which a slightly poorer model performance was obtained when including soft data in the model calibration. For most cases there was also less variation in performance between the different parameter sets obtained in the 50 different calibration trials; in

Fig. 4. Performance of runoff simulations using parameter sets calibrated based on different subsets of runoff data from 1985; the symbol indicates the median of 50 calibration trials and the grey shade indicates the 10–90% range of all calibration trials
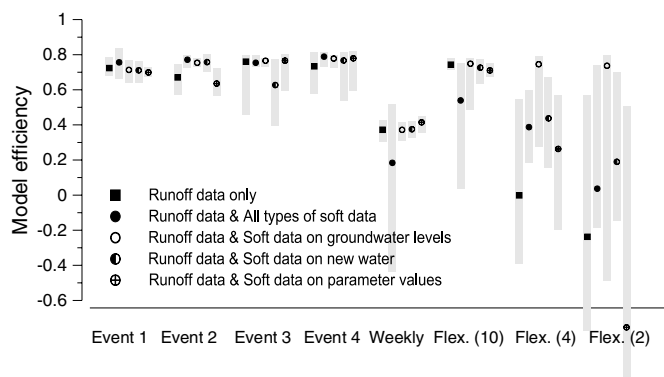
Fig. 5. Performance of runoff simulations using parameter sets calibrated based on different subsets of runoff data from 1986; the symbol indicates the median of 50 calibration trials and the grey shade indicates the 10–90% range of all calibration trials

other words, the calibrations were better constrained. Exceptions included the largest event (Event 1) and the flexible scheme with 10 observation dates. Testing the different calibrations on runoff data from the fully independent years 1985 and 1986 again resulted in similar results (Figs. 4 and 5).

In terms of the different types of soft data it was found that using soft data related exclusively to groundwater dynamics resulted in better model performances overall as opposed to inclusion of all soft data. Use of soft data related to new water contributions or parameter values resulted, at least partly, in poorer model performances (Fig. 3). Especially for the cases with only two or four runoff observations, the soft groundwater data helped greatly to improve model performances (as measured by median model efficiencies), although some calibration trials also resulted in poorer model performances. Again, results were confirmed when evaluating the model performance based on runoff simulation for the two fully independent years, 1985 and 1986 (Figs. 4 and 5).

Discussion

Making hydrological predictions in ungauged basins remains a grand challenge in catchment hydrology. Knowing what to measure, in what order and why is still an open question. The findings in the current study suggest that when starting with no flow information, measurement of one event or 10 observations during different high flow situations provide almost as much information as three months of continuously measured stream flow for constraining the calibration of a simple catchment model. This is surprising as it is normally assumed that much longer time series are needed for good calibration. It is not necessarily that longer time series are not beneficial, but the results of this study show that there are alternative strategies in cases where such data are not available. Within the PUB framework, these results suggest that much can be accomplished with small amount of targeted, clever field work.

These findings build upon recent work of Seibert and Beven (2009) that showed that a relatively small number of streamflow measurements (~6 to 16) helped to constrain model calibration. Unlike Seibert and Beven (2009), additional hard and soft data were also used here in combination with such a limited number of streamflow observations and found that field-experience based soft data on groundwater dynamics, new-water contributions and parameter values could at least partly compensate for a smaller number of stream flow measurements. These findings are consistent also with studies that have used only hard data as additional

data to runoff measurements (Juston et al. 2009; Konz and Seibert 2010). Both of these studies demonstrated that data in addition to streamflow observations can provide useful information for model calibration especially if only few streamflow observations are available. Such soft data can help to obtain good streamflow simulations.

Here, it has been demonstrated that soft data, in addition to being of value to ensure internal model consistency, can also be of similar value as other additional hard data in constraining model calibrations and finding suitable parameter sets for streamflow simulations. These findings suggest that soft data are especially useful if only few hard data observations are available. The value of soft data increased with decreasing availability of hard runoff data available to be used for calibration. Like Seibert and Beven (2009), the current findings suggest that low-flow observations—the subset consisting of assumed weekly gauging in this case consisted mainly of low-flow observations—provided relatively little information content for the model calibration process.

In general, continuous measurements during one event were more valuable for parameter value selection than single streamflow observations distributed over a longer period. This is contrary to the hypothesis going into this work, where it was expected that the less autocorrelated data (in the latter case) would have provided more information. The continuous data provided information for higher flow conditions and also information on the shape of the recession curve, resulting in a powerful combination for calibration.

The current findings suggest that using soft data as additional information in the calibration process can affect model performance in two ways: (1) this information might help to select better performing parameter values, and (2) alternatively, optimizing against different criteria may result in compromise solutions and, thus, lower efficiency values for runoff. Which of these two effects dominates depends on the amount of runoff data and its information content. With much data, especially if high flow conditions are included, the *compromise-effect* predominates, whereas with less hard runoff data available the *information effect* gets more important. These findings go beyond what Seibert and McDonnell (2002) were able to show with soft data, where they used the entire observed streamflow time series and the focus was the potential of soft data for ensuring internal consistency and being *right for the right reasons*. In the present study the focus is on the value of soft data to support the selection of suitable parameter sets, in combination with a limited number of streamflow observations, in order to get as good streamflow simulations as possible.

Finally, the current findings suggest that a new, more thoughtful, top-down approach to gauging ungauged catchments may be warranted. The traditional approach implemented during the IHD—a *gauge it, and wait and see* philosophy—appears to be less useful for model development, calibration and use than targeted field gauging aimed at high flow periods. The results here suggest that a diagnostic approach to ungauged basins may be very fruitful and tractable within a ~2-week reconnaissance field campaign. Similar to a medical diagnosis whereby a doctor would measure a patient's temperature, heart rate, blood pressure, cholesterol, etc., probing of a catchment's flow signal, flow source components, soil depth, bedrock permeability etc., may be a better way to both constrain a conceptualization of flow generation and most efficiently calibrate and validate a catchment model. Much soft data exist even for ungauged catchments. Within the PUB framework, one could envisage other forms of soft data that may be gathered from locals living/working in the area that may have an acute awareness of the hydrology of their basin: how frequently a river floods, how extensive surface saturation becomes during a rainy season, approximate soil depth ranges, when fields are difficult/easy to plough, highest or lowest water levels in a river, length of periods with little or no flow, etc. Such approaches could for instance include historic accounts of flooding (Brázdil et al. 2006; Glaser et al. 2010; Schmocker-Fackel and Naef 2010) for model calibration. Such creative use of soft data going forward, combined with a better awareness of the value of hard streamflow data, will be a powerful tool for work in ungauged basins.

Conclusion

The current study has shown the value of a rudimentary gauging campaign for making predictions in an ungauged basin using the well-studied Maimai watershed as a hypothetical ungauged basin. When starting with no runoff data and adding different subsets of the available data series to constrain the calibration of the simple catchment model, it is found that one event or 10 observations during high flow provided almost as much information as the entire three-month flow time series. Within the PUB framework, these results suggest that much can be accomplished with small amount of targeted clever field work. The value of soft data increases when streamflow measurements are limited, although the relative value of soft versus hard data must be expected to change for other, often more complex, catchments. It was found that soft data were important for internal consistency of the simple conceptual rainfall-runoff model. If few streamflow observations are available, then the use of soft data leads to better streamflow simulations. Consequently, soft data are most helpful when streamflow measurements are most lacking. These findings suggest a new way forward to field diagnosis and model use, guided by an improved dialog between experimentalist and modeler.

References

Aronica, G., Hankin, B., and Beven, K. (1998). "Topographic sensitivity and parameter uncertainty in the predictions of a 2D inundation model." *Hydroinformatics '98*, V. Babovic, and L. C. Larsen, ed., A. A. Balkema, Brookfield, VT, 1083–1088.

Bergström, S. (1992). "The HBV model: Its structure and applications." *Rep. No. 4*, Swedish Meteorological and Hydrological Institute (SMHI), Hydrology, Norrköping, 35.

Beven, K. (2001). *Rainfall-runoff modelling, the primer*, Wiley, New York, 360.

Beven, K. (2002). "Towards a coherent philosophy for modelling the environment." *Proc., R. Soc. A: Math., Phys. Eng. Sci.*, 458(2026), 2465–2484.

Beven, K. (2006). "A manifesto for the equifinality thesis." *J. Hydrol.*, 320(1–2), 18–36.

Brázdil, R., Kundzewicz, Z. W., and Benito, G. (2006). "Historical hydrology for studying flood risk in Europe." *Hydrol. Sci. J.*, 51(5), 739–764.

Falkenmark, M., and Chapman, T. (1989). "Comparative hydrology: An ecological approach to land and water resources." UNESCO, Paris, 309.

Fenicia, F., McDonnell, J. J., and Savenije, H. H. G. (2008). "Learning from model improvement: On the contribution of complementary data to process understanding." *Water Resour. Res.*, 44(6), W06419.

Freer, J. E., McMillan, H., McDonnell, J. J., and Beven, K. J. (2004). "Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures." *J. Hydrol.*, 291(3–4), 254–277.

Glaser, R., et al. (2010). "The variability of European floods since AD 1500." *Clim. Change*, 101(1), 235–256.

Graham, C. B., Woods, R. A., and McDonnell, J. J. (2010). "Hillslope threshold response to rainfall:(1) A field based forensic approach." *J. Hydrol.*, 393(1), 65–76.

Hankin, B. C., and Beven, K. J. (1998). "Modelling dispersion in complex open channel flows: Equifinality of model structure (1)." *Stochastic Hydrol. Hydraul.*, 12(6), 377–396.

Hewlett, J. D., and Hibbert, A. R. (1967). "Factors affecting the response of small watersheds to precipitation in humid areas." *Forest hydrology*, W. E. Sopper and H. W. Lull, eds., Pergamon Press, New York, 275–291.

Juston, J., Seibert, J., and Johansson, P. O. (2009). "Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment." *Hydrol. Process.*, 23(21), 3093–3109.

Konz, M., and Seibert, J. (2010). "On the value of glacier mass balances for hydrological model calibration." *J. Hydrol.*, 385(1–4), 238–246.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S. (1997). "Development and test of the distributed HBV-96 hydrological model." *J. Hydrol.*, 201(1–4), 272–288.

McDonnell, J. J. (1989). "The age, origin and pathway of subsurface storm-flow in a steep humid catchment." Univ. of Canterbury, Christchurch, 270.

McDonnell, J. J. (1990). "A rationale for old water discharge through macropores in a steep, humid catchment." *Water Resour. Res.*, 26(11), 2821–2832.

McDonnell, J. J., Bonell, M., Stewart, M. K., and Pearce, A. J. (1990). "Deuterium variations in storm rainfall - Implications for stream hydrograph separation." *Water Resour. Res.*, 26(3), 455–458.

McDonnell, J. J., et al. (2007). "Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology." *Water Resour. Res.*, 43(7), W07301.

McDonnell, J. J., Stewart, M. K., and Owens, I. F. (1991). "Effect of catchment-scale subsurface mixing on stream isotopic response." *Water Resour. Res.*, 27(12), 3065–3073.

McGlynn, B., McDonnell, J., Stewart, M., and Seibert, J. (2003). "On the relationships between catchment scale and streamwater mean residence time." *Hydrol. Process.*, 17(1), 175–181.

McGlynn, B. L., and McDonnell, J. J. (2003). "Quantifying the relative contributions of riparian and hillslope zones to catchment runoff." *Water Resour. Res.*, 39(11), SWC21–SWC220.

McGlynn, B. L., McDonnell, J. J., and Brammer, D. D. (2002). "A review of the evolving perceptual model of hillslope flowpaths at the Maimai catchments, New Zealand." *J. Hydrol.*, 257(1), 1–26.

Mosley, M. P. (1979). "Streamflow generation in a forested watershed, New-Zealand." *Water Resour. Res.*, 15(4), 795–806.

Mosley, M. P. (1982). "Subsurface flow velocities through selected forest soils, South Island, New-Zealand." *J. Hydrol.*, 55(1–4), 65–92.

Nash, J. E., and Sutcliffe, J. (1970). "River flow forecasting through conceptual models part I—A discussion of principles." *J. Hydrol.*, 10(3), 282–290.

Pappenberger, F., Frodsham, K., Beven, K., Romanowicz, R., and Matgen, P. (2007). "Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations." *Hydrol. Earth Syst. Sci.*, 11(2), 739–752.

Pearce, A. J., Stewart, M. K., and Sklash, M. G. (1986). "Storm runoff generation in humid headwater catchments: 1. Where does the water come from?" *Water Resour. Res.*, 22, 1263–1272.

Schmocker-Fackel, P., and Naef, F. (2010). "Changes in flood frequencies in Switzerland since 1500." *Hydrol. Earth Syst. Sci.*, 14(8), 1581–1594.

Seibert, J. (1997). "Estimation of parameter uncertainty in the HBV model." *Nordic Hydrol.*, 28(4), 247–262.

Seibert, J. (2000). "Multi-criteria calibration of a conceptual runoff model using a genetic algorithm." *Hydrol. Earth Syst. Sci.*, 4(2), 215–224.

Seibert, J., and Beven, K. J. (2009). "Gauging the ungauged basin: How many discharge measurements are needed?" *Hydrol. Earth Syst. Sci.*, 13(6), 883–892.

Seibert, J., and McDonnell, J. J. (2002). "On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration." *Water Resour. Res.*, 38(11), 23.1–23.14.

Seibert, J., Rodhe, A., and Bishop, K. (2003). "Simulating interactions between saturated and unsaturated storage in a conceptual runoff model." *Hydrol. Process.*, 17(2), 379–390.

Sidle, R. C., et al. (2000). "Stormflow generation in steep forested headwaters: A linked hydrogeomorphic paradigm." *Hydrol. Process.*, 14(3), 369–385.

Sivapalan, M. (2003). "Prediction in ungauged basins: A grand challenge for theoretical hydrology." *Hydrol. Process.*, 17(15), 3163–3170.

Stewart, M. K., and McDonnell, J. J. (1991). "Modeling base-flow soil-water residence times from deuterium concentrations." *Water Resour. Res.*, 27(10), 2681–2693.

Vache, K. B., and McDonnell, J. J. (2006). "A process-based rejectionist framework for evaluating catchment runoff model structure." *Water Resour. Res.*, 42(2), W02409.

Vache, K. B., McDonnell, J. J., and Bolte, J. (2004). "On the use of multiple criteria for a posteriori model rejection: Soft data to characterize model performance." *Geophys. Res. Lett.*, 31(21), L21504 1.

Weiler, M., and McDonnell, J. J. (2007). "Conceptualizing lateral preferential flow and flow networks and simulating the effects on gauged and ungauged hillslopes." *Water Resour. Res.*, 43(3), W03403.

Winsemius, H., Schaefli, B., Montanari, A., and Savenije, H. (2009). "On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information." *Water Resour. Res.*, 45(12), W12422.

Woods, R., and Rowe, L. (1996). "The changing spatial variability of subsurface flow across a hillside." *J. Hydrol.*, 35(1), 51–86.