HYDROLOGICAL PROCESSES
Hydrol. Process. **28**, 2451–2467 (2014)
Published online 18 April 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.9726

Catchment properties, function, and conceptual model representation: is there a correspondence?

Fabrizio Fenicia,[1]* Dmitri Kavetski,[2] Hubert H. G. Savenije,[3] Martyn P. Clark,[4] Gerrit Schoups,[3] Laurent Pfister[1] and Jim Freer[5]

[1] *Department of Environment and Agro-Biotechnologies, Centre de Recherche Public Gabriel Lippmann, Belvaux, Luxembourg*
[2] *Civil, Environmental, and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia*
[3] *Water Resources Section, Delft University of Technology, Delft, Netherlands*
[4] *Research Applications Laboratory, National Centre for Atmospheric Research (NCAR), Boulder, Colorado, USA*
[5] *School of Geographical Sciences, University of Bristol, Bristol, UK*

Abstract:

This study investigates the possible correspondence between catchment structure, as represented by perceptual hydrological models developed from fieldwork investigations, and mathematical model structures, selected on the basis of reproducing observed catchment hydrographs. Three Luxembourgish headwater catchments are considered, where previous fieldwork suggested distinct flow-generating mechanisms and hydrological dynamics. A set of lumped conceptual model structures are hypothesized and implemented using the SUPERFLEX framework. Following parameter calibration, the model performance is examined in terms of predictive accuracy, quantification of uncertainty, and the ability to reproduce the flow–duration curve signature. Our key research question is whether differences in the performance of the conceptual model structures can be interpreted based on the dominant catchment processes suggested from fieldwork investigations. For example, we propose that the permeable bedrock and the presence of multiple aquifers in the Huewelerbach catchment may explain the superior performance of model structures with storage elements connected in parallel. Conversely, model structures with serial connections perform better in the Weierbach and Wollefsbach catchments, which are characterized by impermeable bedrock and dominated by lateral flow. The presence of threshold dynamics in the Weierbach and Wollefsbach catchments may favour nonlinear models, while the smoother dynamics of the larger Huewelerbach catchment were suitably reproduced by linear models. It is also shown how hydrologically distinct processes can be effectively described by the same mathematical model components. Major research questions are reviewed, including the correspondence between hydrological processes at different levels of scale and how best to synthesize the experimentalist's and modeller's perspectives. Copyright © 2013 John Wiley & Sons, Ltd.

KEY WORDS conceptual models; perceptual models; catchment form; function; hypothesis testing; model interpretation; top-down; bottom-up; SUPERFLEX

Received 16 April 2012; Accepted 7 January 2013

INTRODUCTION

Conceptual hydrological models are typically developed through a series of steps, including the formulation of a qualitative perceptual model, the development of a conceptual model comprising the mathematical relationships between system forcing, states and responses, and the implementation of the computational model that solves or approximates the model equations (e.g. see Clark and Kavetski, 2010; Beven, 2012; Gupta *et al.*, 2012 – note that different authors may use different definitions), generally followed by parameter calibration and posterior scrutiny. Each step of this modelling chain requires careful attention (e.g. Clark *et al.*, 2011).

The perceptual model is often the starting point of the model development process (McGlynn *et al.*, 2002; Weiler *et al.*, 2005; Bracken and Croke, 2007; e.g. Jencso *et al.*, 2009; Graham *et al.*, 2010; Savenije, 2010; see also Beven 2012 for a review of perceptual models). It aims to reflect the experimentalist understanding of catchment functioning based on the interpretation of field data and visual observations. Considering that field data can seldom be used directly in lumped catchment-scale modelling due to commensurability limitations (e.g. Freer *et al.*, 2004), the perceptual model represents valuable information for the modelling process, both to inform conceptual model development and to evaluate model realism (e.g. Ambroise *et al.*, 1996; Pinol *et al.*, 1997; Seibert and McDonnell, 2002; Uhlenbrook and Leibundgut, 2002; Freer *et al.*, 2003; Graham and McDonnell, 2010; Clark *et al.*, 2011; McMillan *et al.*, 2011).

*Correspondence to: Fenicia, Fabrizio, Department of Environment and Agro-Biotechnologies, Centre de Recherche Public Gabriel Lippmann, Belvaux, Luxembourg
Email: fenicia@lippmann.lu

Much work is still needed to understand how to make the best use of different measurements and visual observations to characterize the dominant processes in a catchment. Although some studies have detailed the fieldwork analyses that have motivated specific interpretations of certain environments (e.g. McGlynn *et al.*, 2002; Tromp-van Meerveld and McDonnell, 2006; Blume *et al.*, 2009), the formulation of the perceptual model has seldom been made explicit and is generally poorly documented. An additional difficulty in taking advantage of those experimental studies is that they have often been focused on a few individual locations, and results have been difficult to generalize to other locations (McDonnell *et al.*, 2007).

The relationship between perceptual and conceptual models is also not straightforward. Perceptual models tend to be based on small-scale observations, whereas the modelling objective is to represent catchment-scale processes (Sivapalan, 2003). Model development must then recognize the scale dependencies of hydrological processes, and that small-scale understanding may not be representative of large-scale behaviour - the 'Paradox of the Ant' as described by Savenije (2009) (see also Young and Beven, 1994; McDonnell *et al.*, 2007, and others).

So how to analyse catchment-scale hydrological behaviour, and relate it to catchment characteristics and fieldwork-based understanding? When treated as tentative hypotheses of catchment dynamics, hydrological models can serve as powerful instruments for investigating catchment behaviour (e.g. Atkinson *et al.*, 2002; Fenicia *et al.*, 2008; Savenije, 2009; Buytaert and Beven, 2010; Krueger *et al.*, 2010; Clark *et al.*, 2011; Kavetski and Fenicia, 2011; McMillan *et al.*, 2011). In particular, the analysis and comparison of different model variants can help interpret dominant processes, suggest improved representations, and approximate structural uncertainties. Meaningful application of such multi-hypotheses approaches, however, requires controlled model development, analysis, and comparison, where individual differences between models are isolated and, whenever possible, scrutinized using multiple diagnostics (Clark *et al.*, 2011).

The understanding of hydrological behaviour at the catchment scale also benefits from intercomparison studies, where similar models, experiments, and analyses are applied at different locations (e.g. Refsgaard and Knudsen, 1996; Perrin *et al.*, 2001; Kavetski and Fenicia, 2011). As noted by Sivapalan (2009), such studies remain quite rare, and the power of comparative hydrology, which aims to learn from the similarities and differences between catchments, remains largely unexploited.

The aims of this study are to: (1) investigate the possible correspondence between catchment structure, as represented by perceptual hydrological models developed from process-oriented fieldwork insights, and conceptual model structures, selected on the basis of reproducing observed catchment hydrographs, (2) illustrate how controlled hypothesis testing with conceptual hydrological models can improve the synthesis of modelling and fieldwork perspectives, and (3) illustrate how comparative hydrology can generate useful insights into catchment behaviour. Model development is carried out using the recently introduced flexible framework SUPERFLEX (Fenicia *et al.*, 2011), which provides a versatile and computationally robust platform for modelling and hypothesis testing in catchment-scale hydrological applications.

While previous studies used the perceptual model to inform the mathematical model, we take an alternative approach. In this study, the correspondence between perceived catchment structure and fitted model structure is a hypothesis rather than an underlying assumption. Therefore, the perceptual model is only used a posteriori to appraise model realism and interpret model results, rather than for detailed a priori guidance in model development. Our intention is to avoid an a priori mapping of small-scale understanding to the large scale, yet to still allow potential correspondence to emerge a posteriori *if* independently derived modelling results support it.

This study also makes further inroads into the joint use of top-down and bottom-up approaches (Klemes, 1983; Sivapalan *et al.*, 2003). Since the proposed perceptual models synthesize the experimental understanding acquired at small scales (e.g. plot, hillslope), they can be viewed as pursuing the 'bottom-up' strategy for system conceptualization (Sivapalan, 2003). The present application of multi-hypotheses frameworks to characterize catchment responses can be considered as a 'top-down' route to processes conceptualization.

The case study is based on three headwater catchments of the Attert basin in the Grand Duchy of Luxembourg, which have been studied during several previous fieldwork campaigns (van den Bos *et al.*, 2006; Pfister *et al.*, 2009; Martinez-Carreras *et al.*, 2010; Pfister *et al.*, 2010; Juilleret *et al.*, 2012). Although these catchments are closely spaced and hence subject to a similar climatological regime, they are characterized by different physical attributes (shape, morphology, geology, land cover) and behave hydrologically differently. The field experiments have been used to form a set of perceptual models, which represent the dominant processes for the three catchments.

The paper is structured as follows. The Study Area section reviews the fieldwork knowledge and corresponding perceptual models. The Methods section describes the formulation, inference, and evaluation of conceptual model hypotheses. The Results section presents the modelling results. The Discussion section interprets the differences in the performance of different model hypotheses and relates them to process-based insights available in the three experimental catchments. The conclusions are summarized in the Conclusions section.

STUDY AREA

Summary of experimental insights

Three headwater catchments of the Attert basin in Luxembourg are considered: the Huewelerbach, Weierbach, and Wollefsbach catchments (Figure 1). These catchments have been selected because they share the same climatology (due to close proximity to each other) yet differ in their physical attributes, including geology and land use. Hence, differences in the hydrological behaviour ('function') can be more confidently attributed to differences in catchment structure ('form') (Wagener *et al.*, 2007). Importantly, appreciable experimental insights are available due to ongoing fieldwork (van den Bos *et al.*, 2006; Pfister *et al.*, 2009; Martinez-Carreras *et al.*, 2010; Pfister *et al.*, 2010; Juilleret *et al.*, 2012). This section summarizes the key aspects relevant to the present work.

Previous fieldwork has used diverse experimental techniques to characterize the three catchments, including drills and pits, analysis of soil samples, Electrical Resistivity Tomography, as well as analysis of flow and tracer responses. The three catchments appear to have different dominant runoff-generating processes, which are depicted schematically in the perceptual models shown in Figure 2. Since these models are based on small-scale experimental data, they represent a bottom-up route to catchment conceptualization (Sivapalan, 2003). Previous experimental investigations at these locations and the development processes of the perceptual models will be detailed elsewhere.

The Huewelerbach catchment (Figure 2a), with an area of 2.7 km^2, is located in the south of the Attert. Its hillslopes and plateaus are forested, whereas its near-stream areas are used for agriculture. The geology is characterized by a permeable sandstone formation overlying an effectively impermeable (due to the low hydraulic conductivity of clay (Freeze and Cherry, 1979)) marly formation. The dominant hydrological process in the sandstone is deep percolation to the water table level. Sources with relatively stable flow develop at the contact zone with the underlying marls, which dominate the near-stream areas. On marls, deep percolation is impeded, and the dominant runoff generating

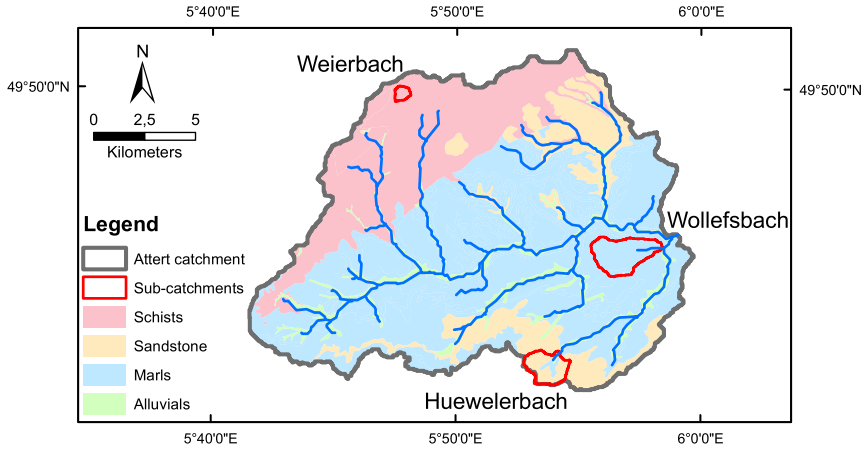

Figure 1. Experimental watersheds on different geological substrata in the Attert catchment in Luxembourg

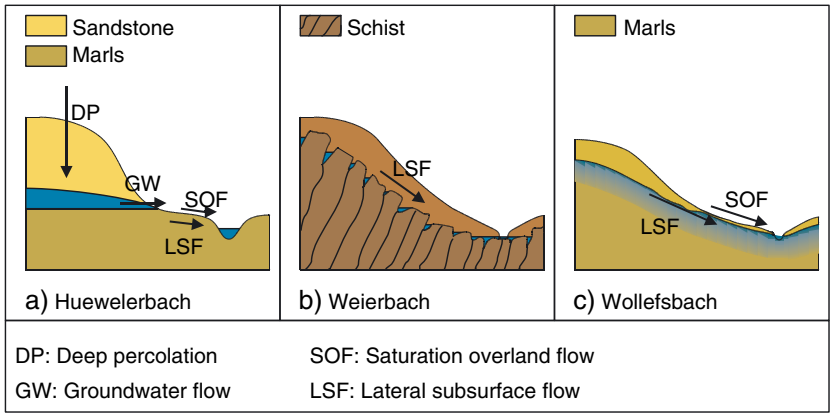

Figure 2. Perceptual models of dominant runoff processes in the Huewelerbach, Weierbach, and Wollefsbach catchments

processes are perceived to be subsurface or overland flow triggered by saturation excess.

The Weierbach catchment (Figure 2b), with an area of 0.42 km^2, is located in the north Attert and is fully forested. Its geology is dominated by schists. Previous fieldwork has suggested that the schist formation is generally compacted, yet its rock layers tend to disaggregate (foliate) towards the land surface. Hence, the bedrock is assumed to be impermeable, and the dominant runoff generating process is hypothesized to be lateral flow at the contact zone between soil and underlying bedrock. The forested soil is relatively permeable, and surface runoff has not been observed.

The Wollefsbach catchment (Figure 2c), with an area of 4.5 km^2, is located in the central part of the Attert basin. Primary land use is agriculture with pasture and crops. The catchment is located on marls formations, which are considered to be impermeable (due to the low hydraulic conductivity of clay). Similar to the marly portion of the Huewelerbach catchment, the dominant process is assumed to be saturated subsurface flow and saturation-excess overland flow.

Although the Weierbach and Wollefsbach are characterized by impermeable bedrock, the limited response to rainfall in the summer season has suggested that these catchments are able to store water. In the Weierbach, water is assumed to be stored in the irregular topographic relief of weathered bedrock (Figure 2). In the Wollefsbach, water is assumed to reside in the soil, which is considerably cracked as a result of repeated expansion/compaction of the clayey material during wetting–drying cycles.

Overall, despite sharing the same climatology, the three catchments exhibit strikingly different hydrological behaviour and dynamics (see Table I and Figure 5 later in the text). The Huewelerbach catchment exhibits little seasonality, with a relatively stable baseflow sustained by the sandstone formation and with peaks generated by runoff from the near-stream marly area. The Weierbach catchment has a markedly seasonal behaviour: winter responses are two peaked (with the first peak near concomitant with the rainfall event and the delayed peak occurring several hours or days later), while in the summer season, only the first peak is present (e.g. see Figure 6 in Kavetski *et al.*, 2011). Given the small size of the Weierbach, such delayed response was unexpected and, at least a priori, could not be understood based on experimental evidence alone (this will be discussed in the Weierbach Catchment section). The Wollefsbach catchment has a relatively fast response with appreciable seasonality: the winter months flows are relatively high, whereas the summer flows are low or absent.

Table I further illustrates important differences in the key hydrograph characteristics of the three catchments. For example, the lowest observed flows and runoff coefficients are much more seasonal in the Weierbach and Wollefsbach than in the Huewelerbach. The seasonality of the Weierbach manifests itself particularly strongly in the time-to-peak values in summer and winter.

The flow–duration curves (FDCs) of the catchments are also markedly different (e.g. see Figure 6 later in the text). In general, the Huewelerbach catchment has higher baseflow than the Weierbach and Wollefsbach catchments, but the latter have generally higher peak flow. The Weierbach differs from the Wollefsbach mainly in the middle range of flow, which is higher in the Weierbach. This is reflects the 'flashier' behaviour of the Wollefsbach compared to the Weierbach.

Hydrological data used in model calibration and validation

A 5-year period of hourly discharge, rainfall, and potential evaporation data from 1-Sept-2004 to 31-Aug-2009 was used in all catchments. The first year was used for model initialization (warm-up), the following 2 years were used for calibration, and the last 2 years were used for validation in a classic split-sample evaluation framework.

METHODS

Model hypotheses under consideration

A total of 12 alternative model structures were hypothesized and implemented using the SUPERFLEX framework (Fenicia *et al.*, 2011). The hypotheses differ in their representation of flow paths through a catchment, which in this study are conceptualized as combinations of different storage elements. We argue that this is an appropriate level of model complexity for studies concerned primarily with integrated catchment-scale responses such as discharge and

Table I. Summary of observed hydrograph characteristics. The subscripts *w* and *s* denote winter and summer seasons, respectively. Q_{hi} and Q_{lo} are the maximum and minimum discharge. R_c is the runoff coefficient (calculated as the cumulative discharge divided by the cumulative rainfall). T_{lag} is the time to peak (calculated as the lag time that maximizes the cross-correlation of the rainfall and runoff time series, see also Kavetski *et al.* (2011))

Catchment	$Q_{hi,w}$ (mm h^{-1})	$Q_{hi,s}$ (mm h^{-1})	$Q_{lo,w}$ (mm h^{-1})	$Q_{lo,s}$ (mm h^{-1})	$R_{c,w}$ (−)	$R_{c,s}$ (−)	$T_{lag,w}$ (h)	$T_{lag,s}$ (h)
Huewelerbach	$3.9\times\ 10^{-1}$	1.2×10^{-1}	1.2×10^{-2}	1.1×10^{-2}	3.2×10^{-1}	2.5×10^{-1}	3	2
Weierbach	1.1	2.9×10^{-1}	1.7×10^{-2}	0.0	9.8×10^{-1}	9.5×10^{-2}	29	0
Wollefsbach	1.5	1.4×10^{-1}	3.1×10^{-3}	7.8×10^{-4}	6.4×10^{-1}	3.0×10^{-2}	4	4

that it provides a basis for investigating differences in dominant flow path behaviour across the three catchments considered in this study.

The model domain is discretized into a set of reservoirs, which are named according to the processes they are broadly intended to represent: UR = unsaturated soil reservoir, FR = fast reservoir, SR = slow reservoir, RR = riparian zone reservoir, and IR = interception reservoir. The states (storages) in these reservoirs are labelled S_u, S_f, S_s, S_r, and S_i, respectively. The models are illustrated in Figure 3, and their key differences are outlined next (see also Tables AI–AIV in Appendix A).

An important distinction relevant to this study is between 'serial' *versus* 'parallel' model architectures, reflecting different hypothesized connectivities of the flow pathways:

Single-reservoir structures: M01 is a single nonlinear reservoir model, arguably one of the simplest hydrological models. M02 is a single-reservoir model that resembles the core block of the VIC model (Wood *et al.*, 1992).

Serial structures: M03 comprises two reservoirs in series: precipitation enters the unsaturated reservoir (UR), and any storage in excess of a threshold overflows into a downstream 'fast' reservoir (FR). M04 differs from M03 in that the outflow from UR is a power function of the storage, rather than a threshold function. M05 differs from M04 by including a transfer function between the UR and FR reservoirs. M06 differs from M05 by including an interception reservoir (IR).

Parallel structures: M07 differs from M05 by including a riparian zone reservoir (RR), which receives a constant fraction of the total precipitation. Note that this inclusion results in two parallel flowpaths in the M07 model, albeit with a somewhat different overall connectivity than in the M08–M12 structures listed next. M08 is one of the simplest parallel structures, with precipitation partitioned between two linear reservoirs (FR and SR). M09 also includes a linear UR, with its outflow partitioned between FR and SR based on a linear function of S_u. M10 differs from M09 by including a transfer function between UR and FR. M11 differs from M10 in that the outflow from UR is split between FR and SR based on a nonlinear (power) function of S_u. Finally, M12 differs from M11 by the inclusion of the IR. M08, M09, and M10 are linear in the states.

We note that all model hypotheses considered in this study are lumped. Though lumped models are widely used both in research and operations, their chief limitation is that the behaviour of different runoff mechanisms cannot be allocated to different areas within the catchment. While semi-distributed and fully distributed models in principle can overcome this limitation, they have considerably larger data and

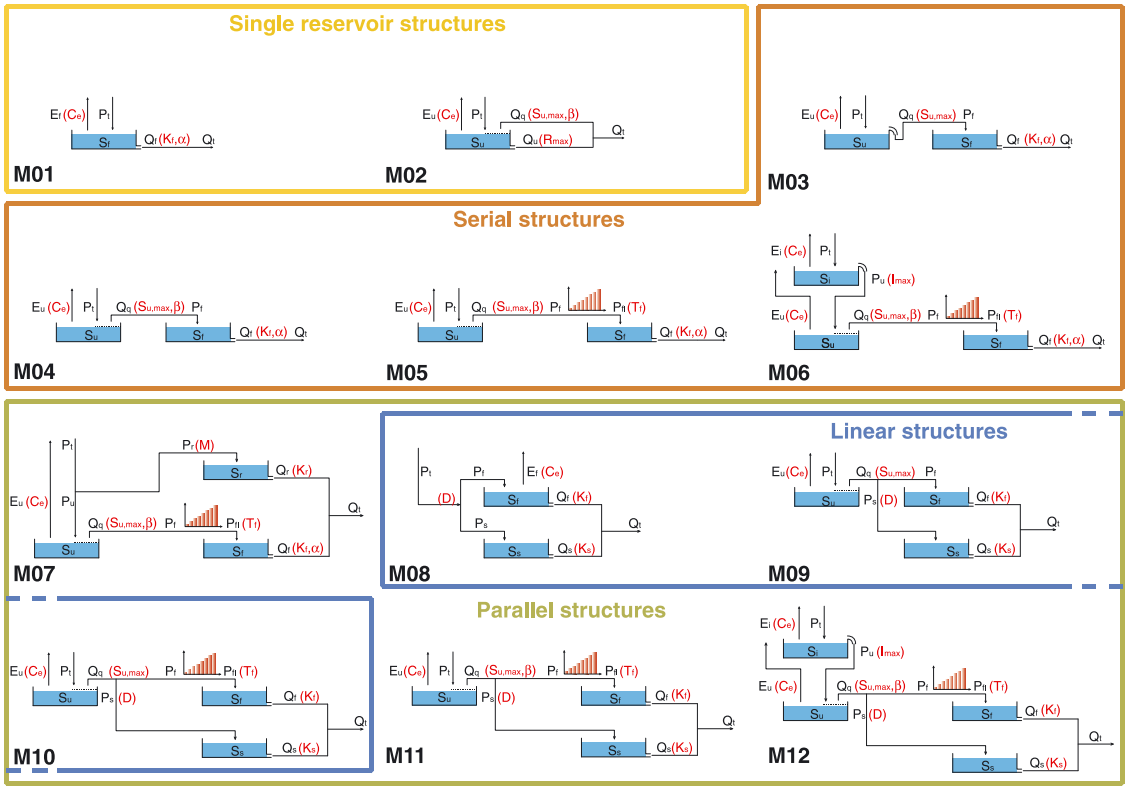

Figure 3. Multiple model hypotheses considered in this study. All hypotheses were implemented using the SUPERFLEX model development framework

computational requirements, and were hence not considered in this study.

All model equations were implemented numerically using the implicit Euler time stepping scheme with fixed hourly steps and a tight iteration tolerance (Clark and Kavetski, 2010). Using the same numerically robust algorithm in all models is essential to avoid obscuring and corrupting genuine aspects and differences in model behaviour by unnecessary numerical differences and artefacts.

Motivation of model hypotheses

The selection of the model hypotheses M01–M12 was motivated by several considerations. A key consideration is the comparison of serial *versus* parallel flowpath hypotheses. The models also span a range of complexities, from single-reservoir models, such as M01 and M02 (sometimes sufficient to characterize catchment behaviour, see Kavetski and Fenicia (2011)), to more complex models containing multiple interconnected reservoirs and flow pathways. Importantly, the models were constructed to differ in a controlled way, so that differences in model performance could be more directly attributed to differences in the model structure (Freer *et al.*, 2003; Clark *et al.*, 2011; Kavetski and Fenicia, 2011).

As the correspondence between perceived catchment structure and suitable model structure is the working hypothesis of this study rather than its underlying assumption, we have included a relatively broad range of a priori model hypotheses. This avoids unduly restricting the model space under consideration, which could happen if we over-relied on small-scale understanding to dictate and constrain the structure of the large-scale catchment model (Sivapalan, 2003; McDonnell *et al.*, 2007). In general, the use of multiple model hypotheses also reduces biases due to modellers' personal preferences (and 'parental affection', in the words of Chamberlin, 1965) and over-reliance on a few commonly used model structures (which is also a motivation behind the Data Based Mechanistic modelling approach (Young and Beven, 1994)). The model structures can be scrutinized to appraise whether they provide physically plausible descriptions of the catchment-scale response and to test a range of physically oriented hypotheses of catchment behaviour (e.g. dominant flow pathways, thresholds, etc.). For example:

i. hypotheses regarding multiple aquifers acting relatively independently can be tested by comparing the performance of serial structures (M03–M06, where the reservoirs are connected in series) *versus* parallel structures (M07–M12, which include connections in parallel);
ii. the 'on-off' runoff production mechanism, hypothesized based on the fieldwork perceptions in some of the catchments, is represented using the threshold in the UR reservoir of model M03;
iii. some model structures test the effect of a lag function, which introduces delays into the fluxes and the hydrograph response (e.g. M05 *vs* M04; M10 *vs* M09);
iv. the significance of a certain flow process. For example, contribution of the interception component can be tested by comparing M06 *versus* M05 and M12 *versus* M11;
v. process linearity. For example, models with linear reservoirs (e.g. M08, M09, and M10) can be compared to models with nonlinear reservoirs.

Model inference framework

The hydrological model parameters are inferred from observed rainfall-evaporation-runoff data $\left(\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{E}}, \widetilde{\boldsymbol{Q}}\right)$ using Bayes equation,

$$p(\boldsymbol{\theta}, \boldsymbol{\Xi}, |\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{E}}, \widetilde{\boldsymbol{Q}}, M) = p(\widetilde{\boldsymbol{Q}}|\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{E}}, \boldsymbol{\theta}, \boldsymbol{\Xi}, M) p(\boldsymbol{\theta}, \boldsymbol{\Xi}|M) \quad (1)$$

where $p(\boldsymbol{\theta}, \boldsymbol{\Xi}, |\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{E}}, \widetilde{\boldsymbol{Q}}, M)$ is the posterior distribution of the parameters $\boldsymbol{\theta}$ of the hydrological model M and the parameters $\boldsymbol{\Xi}$ of the residual error model, $p(\widetilde{\boldsymbol{Q}}|\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{E}}, \boldsymbol{\theta}, \boldsymbol{\Xi}, M)$ is the likelihood function, and $p(\boldsymbol{\theta}, \boldsymbol{\Xi}\,|\,M)$ is the prior. The tilde indicates quantities that are observed and hence subject to sampling and measurement uncertainties. In the absence of additional knowledge, we used non-informative priors for $\boldsymbol{\theta}$ and $\boldsymbol{\Xi}$ (Box and Tiao, 1992).

The error model is based on the Weighed Least Squares (WLS) scheme, which assumes zero-mean Gaussian errors and hypothesizes that the standard deviation of individual residuals increases linearly with the corresponding simulated streamflows,

$$p(\widetilde{\boldsymbol{Q}}|\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{E}}, \boldsymbol{\theta}, a, b, M) = \prod_{n=1}^{N_t} N(\widetilde{Q}_n - \hat{Q}_n\left[\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{E}}, \boldsymbol{\theta}\right] | 0, \sigma_n^2, M) \quad (2)$$

$$\sigma_n = a + b\hat{Q}_n \quad (3)$$

where σ_n is the standard deviation of the residual errors at time step n, $\widetilde{Q}_n$ and $\hat{Q}_n\left[\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{E}}, \boldsymbol{\theta}\right]$ are the observed and predicted streamflows, respectively, $N(z|m,s^2)$ is the probability density of a Gaussian deviate z with mean m and variance s^2, N_t is the number of observations, and a and b are the parameters of the error model. The WLS scheme relaxes the assumption of constant variance made by Standard Least Squares scheme and other calibration methods based on (un-weighted) sums-of-squared errors objective functions, such as root mean square error, Nash–Sutcliffe efficiency, etc. However, it has the limitation of ignoring the autocorrelation of the residual errors. We also refer the reader to Beven *et al.* (2012) and Clark *et al.* (2012) for a further discussion of challenges facing current model estimation, diagnostic, and prediction methods.

Parameter calibration was carried out using a multi-start quasi-Newton method (see Kavetski and Clark, 2010 for applications in hydrology), with 100 independent local searches initiated from random seeds in the feasible parameter space.

The posterior parameter distributions were explored using the MCMC sampling strategy described by Thyer *et al.* (2009) with a total of 60 000 model runs and five parallel chains. During the first 10 000 samples, the jump distribution was tuned one parameter at a time. During the next 10 000 samples, the jump distribution was tuned by scaling its entire covariance matrix. The jump distribution was then fixed and 40 000 samples collected. The first 30 000 samples were treated as a burn-in and therefore discarded from the computation of the statistics (Gelman *et al.*, 2004), and the final 10 000 samples were used to analyse and report the parameter distributions.

We note that both parameter optimization and uncertainty analysis benefit from the numerical model implementation approach used in SUPERFLEX, which results in a smooth model and objective function (Fenicia *et al.*, 2011; Kavetski and Clark, 2011).

Model evaluation metrics and diagnostics

In this study, the quality of the probabilistic model predictions is gauged using the Continuous Rank Probability Score (CRPS) (Φ_{CRPS}) (Hersbach, 2000), which is defined as follows:

$$\Phi_{CRPS} = \frac{1}{N_t}\sum_{n=1}^{N_t}\int \left(F_n(Q) - H\left\{Q \geq \widetilde{Q}_n\right\}\right)^2 dQ \qquad (4)$$

where $F_n(Q)$ is the cumulative distribution function of the model predictions for the time step n, and $H\{Q \geq \widetilde{Q}_n\}$ is the Heaviside step function that takes the value 1 if $Q \geq \widetilde{Q}_n$ and 0 otherwise. For perfect predictions, $\Phi_{CRPS} = 0$.

For deterministic predictions, the Φ_{CRPS} reduces to the Mean Absolute Error. For probabilistic predictions, Φ_{CRPS} provides a measure of the difference between the predictive distribution and the observations, and hence reflects both the reliability and precision of the predictive distribution (Renard *et al.*, 2010). It is therefore quite powerful as an error measure. In contrast, traditional error measures such as the Nash–Sutcliffe index focus entirely on the goodness-of-fit of a single deterministic prediction.

Despite its appeal, the Φ_{CRPS} is a single metric that cannot, by itself, provide a complete and nuanced comparison of multiple models. Hence, in addition to the Φ_{CRPS}, we explore the model's ability to reproduce different aspects (characteristic 'signatures') of the catchment response.

Here, we use: (1) visual inspection of model predictions and (2) FDCs (calculated for the calibration and validation periods), which describe the distributional properties of streamflow. These diagnostics can provide insights not apparent from aggregate performance metrics alone.

RESULTS

Hydrograph representation

Figure 4 compares the Φ_{CRPS} calculated for different model structures in the calibration and validation periods. To facilitate the interpretation of the performance of different models, the bottom panel indicates serial, parallel, and linear models. The hydrographs simulated using selected model structures are shown in Figures 5b–d. The forcing data, similar for all the catchments due to their mutual proximity, are shown in Figure 5a.

Figure 4 shows that single-reservoir models have a poor performance in all catchments. It is also apparent that, when considering more complex models, some structures work well in particular catchments and poorly in others. The main findings of this top-down analysis are summarized below for each of the three catchments.

Huewelerbach catchment. In the Huewelerbach catchment, models M01–M06 perform poorly compared to other models (Figure 4a), even in the calibration period. Models M01–M06 are either single-reservoir models or serial structures. As soon as parallel connections are introduced (i.e. models M07–M12), the performance improves considerably. Interestingly, the best performing models in the validation period are the linear models M09 and M10. More complex models, such as M11 and M12, which include nonlinearities, suffer from a loss of performance in the validation period, suggesting they were over-parameterized with respect to the calibration data set.

Figure 5b compares the calibrated hydrographs of M06, M09, and M12. M06 is a complex model with reservoirs connected in series. Yet, in both the calibration and validation periods, this model is unable to fit high and low flows simultaneously. Models M01–M05 have similar weaknesses. Models M09 and M12 have similar performance as each other, meaning that the increased complexity of M12 does not translate into improvements in predictive performance. In fact, M12 performs slightly worse than M09 in the validation period.

Weierbach catchment. Here, in contrast to the Huewelerbach catchment, serial structures perform much better than parallel structures in terms of Φ_{CRPS} (Figure 4b). For example, the addition of a groundwater reservoir (SR), which represents a parallel flowpath, led to a deterioration in model performance (Figure 4b).

The double-peaked wet-season response of the Weierbach catchment (Summary of Experimental Insights

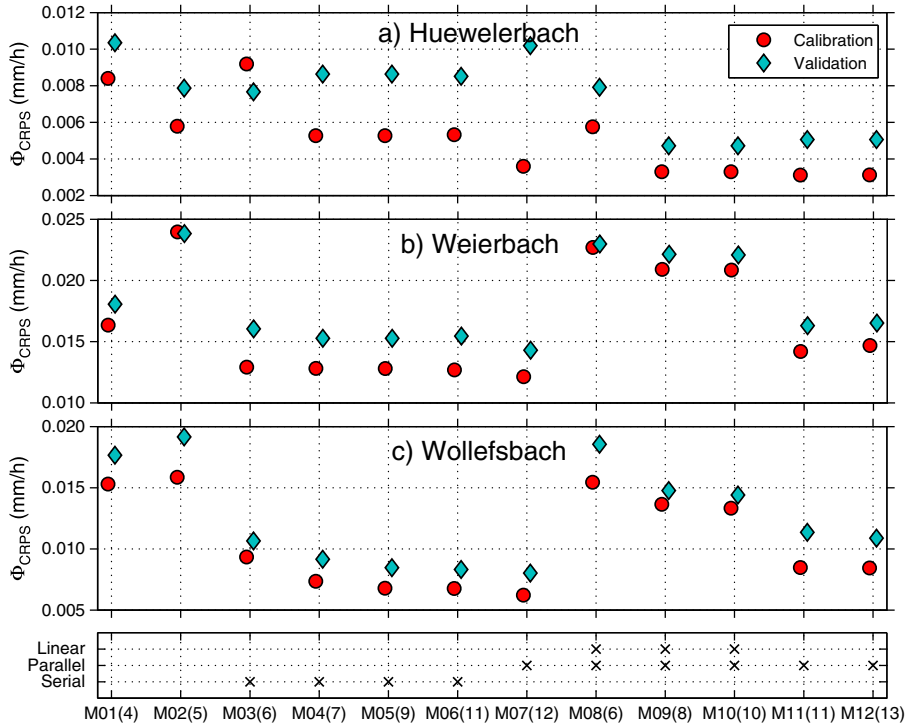

Figure 4. CRPS performance of the 12 model structures in the three catchments during calibration (circle) validation (square). The x-axis lists the model structures, with the brackets indicating the total number of parameters and states to give a sense of the 'complexity' of these lumped models. The bottom panel summarizes common characteristics of the models

section) is not an easy feature to reproduce. For example, models M04 and M05, which differ in the presence of a lag function, produce similar time series of model predictions (and hence a similar Φ_{CRPS}). Inspection of the calibrated parameters shows that M05 matches the timing of the first peak at the expense of the second peak, by reducing the lag parameter T_f so that M05 converges to M04.

The only model that could reproduce the dual response of the catchment, in particular the timing of the second peak, is M07. When calibrated, this model appears to use the quick response produced by RR for the first peak and the slower response produced by FR for the second peak.

Figure 5c shows the hydrographs of M04, M07, and M10. M10, which is a linear model, struggles to simulate the difference between winter and summer conditions. M04 and M05 do not correctly capture the timing of the second peak. In addition, they poorly approximate the summer dynamics, producing a 'flat' response. M07, which differs from M05 by the presence of RR, is able to capture both the delayed response in the winter season and the quick response in the summer season.

It is also noted that threshold models, even the very simplistic M03 model, appear to capture the dynamics of this catchment considerably better than linear models, including the quite complex multi-reservoir models M08-M10. A comparison can also be made between models M03 (which has a threshold-like UR reservoir, meaning that incoming precipitation flows to downstream reservoirs only when UR is full) and M04 (which partitions the precipitation based on a power function of UR). The performance of these models is quite similar, indicating the threshold-like response of this catchment.

Wollefsbach catchment. The pattern of model performance in the Wollefsbach is similar to that in the Weierbach, with the best predictions provided by serial models (Figure 4c). Furthermore, similar to the Weierbach results, models that include threshold dynamics clearly outperformed models with linear storage–discharge relationships.

Figure 5d illustrates these differences in performance, contrasting M05, M07, and M10. The linear model M10 has a visibly poor performance, underestimating the hydrograph during wet (winter) conditions, and overestimating it during dry (summer) conditions. This suggests strong nonlinearities in the storage–discharge relationship of this catchment. The nonlinearities of M05 and M07 appear able to effectively represent the seasonality of this catchment, which switches from particularly high streamflows in the wet season to quite low streamflows in the dry season. While M07 performs slightly better than M05 in terms of Φ_{CRPS}, the predictions of these models look very similar. Hence, unlike in the Weierbach, the additional complexity of M07 may not be supported.

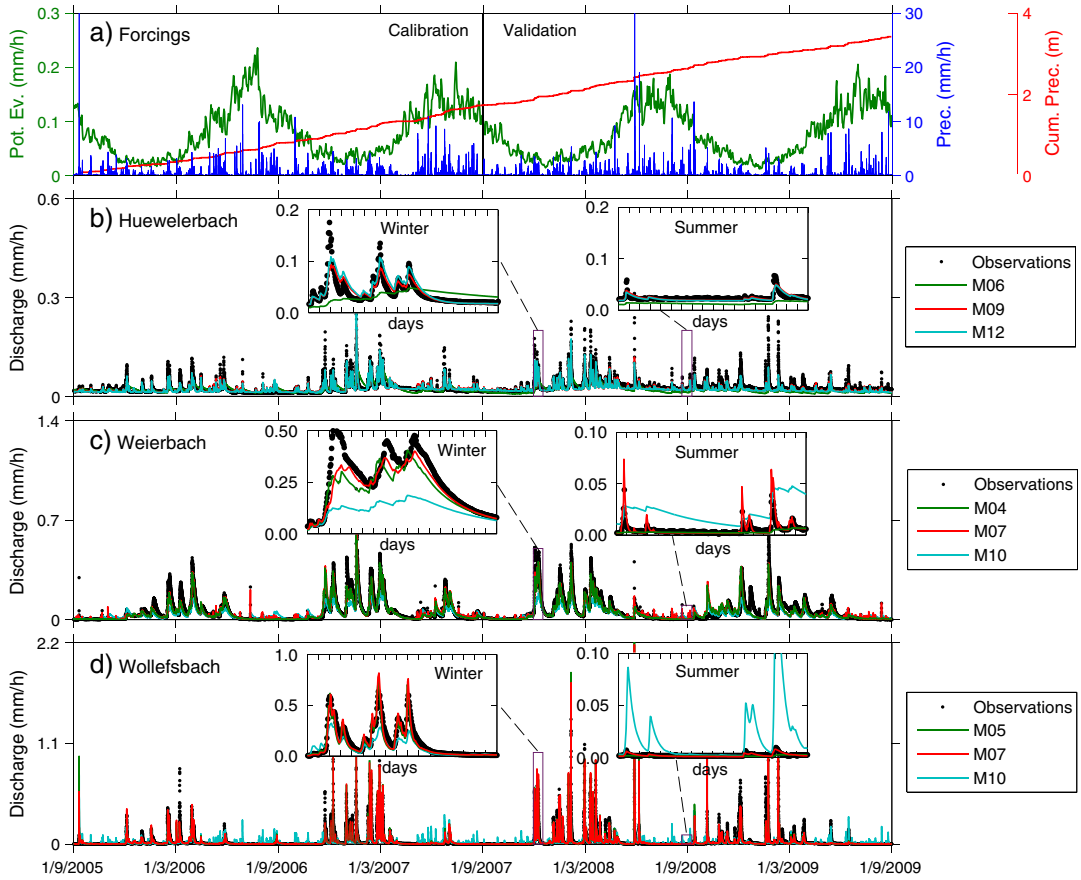

Figure 5. Hydrograph representation of selected model structures on the three catchments. Panel (a) shows the observed forcing data. Note the relatively uniform rainfall regime and the strong seasonal variations in evaporation and discharge regimes (panels b,c,d). Panel (b): M06: complex 'serial' structure; M09: simple 'linear' and 'parallel' structure; M12: complex 'parallel' structure. Panel (c): M04: simple 'serial' structure; M07: complex structure allowing for double-peak and delayed response. M10: 'linear' structure. Panel (d): M05: 'serial' structure. The insets zoom in on the same time period for the three catchments, demonstrating their different response dynamics, including different seasonality patterns

FDCs

Figure 6 illustrates the FDCs for the three catchments in calibration and validation. Figures 6a–b clearly show the differences in behaviour between different model structures in the Huewelerbach catchment. M06 under-predicts the high flow both in calibration and in validation, and under-predicts the baseflow in validation. As discussed in the section on Hydrograph Representation, we attribute this poor performance to the absence of parallel reservoir connections in the hypothesized model architecture. M10 and M12 have similar performance in calibration and validation. However, it is also apparent that the two models do not correctly match the FDC in the validation period, underestimating both the high and low flows. On the other hand, it can be seen that the observed FDCs differ substantially in the calibration and validation period, with the validation curve lying much higher than the calibration curve. This suggests that the calibration period provided only a limited sample of the complete range of behaviour of this slow-dynamic catchment and that longer calibration time series may be needed.

In the Weierbach catchment, Figures 6c–d confirm that the linear model M10 is not able to adequately reproduce the simulations. M04 and M07 provide a better fit, although high flows are underestimated. Interestingly, M04 and M07 have similar performance in terms of FDCs, in spite of the clear differences in the hydrographs shown in Figure 5c. This finding emphasizes that certain differences in hydrograph behaviour are not apparent in the FDC. In particular, while the FDC reflects the (marginal) distributional properties of streamflow, it suppresses timing ('frequency') information (Kavetski *et al.*, 2011). The hydrographs simulated by M04 and M07 differ mainly in frequency aspects, due to: (1) the presence/absence of a lag function component and (2) presence/absence of a riparian zone component.

In the Wollefsbach catchment, Figures 6e–f indicate that the linear model M09 performs poorly: the slope of its FDC is notably different from the slope of the observed FDC. Models M05 and M07 perform much better than M09.

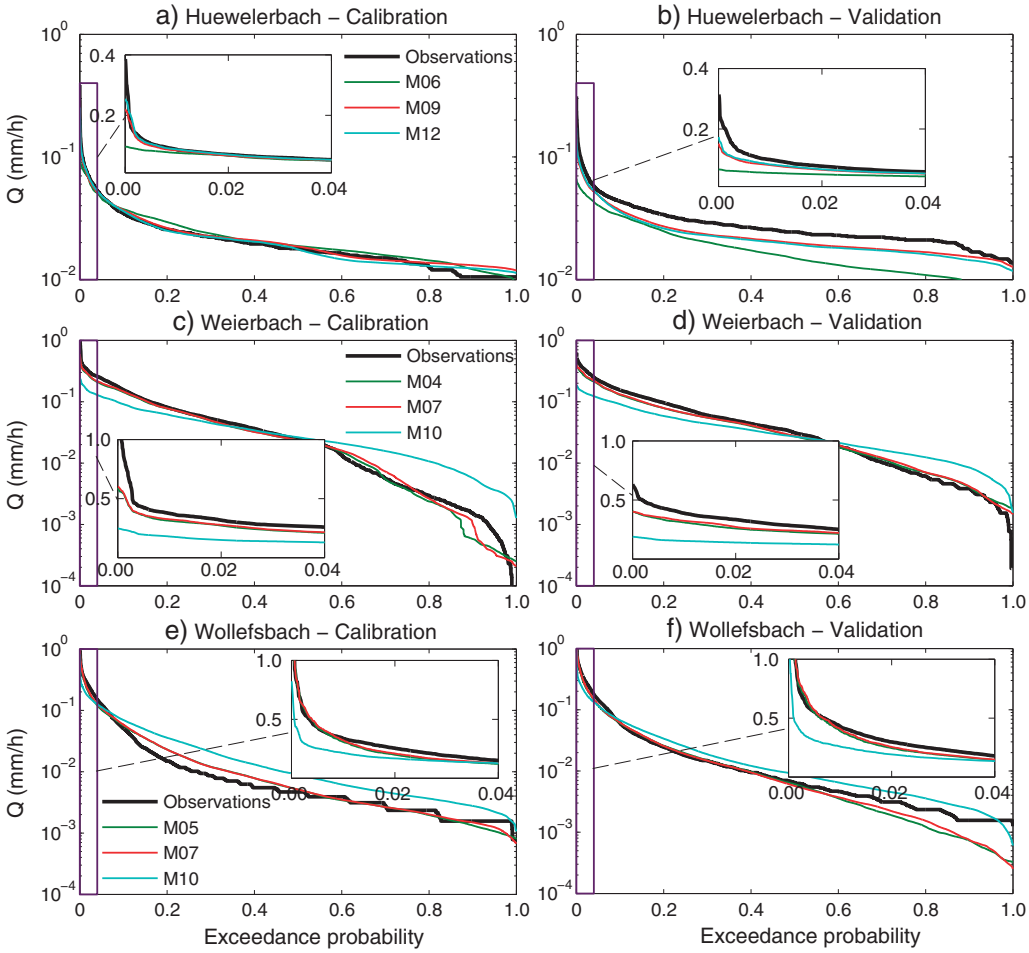

Figure 6. Observed and predicted flow–duration curves (FDCs) for selected models (y-axis in the insets is in natural scale, to emphasize the high flows). Also note the differences in FDCs for the three catchments. The Huewelerbach catchment, with higher baseflow, is best described by 'parallel' structures (M09, M12). In Weierbach catchment, M04 and M07 have similar FDCs, although their underlying streamflow time series are markedly different (Figure 5); M10 (linear) performs poorly. In the Wollefsbach catchment, M05 and M07 perform similarly, while M10 performs poorly.

Model structure differences through parameter inference

This section considers what can be learnt by comparing the parameter values inferred for different catchments. We restrict the discussion in this section to model M07, which works quite well in the Weierbach and Wollefsbach catchments. We exclude the Huewelerbach catchment from this analysis because hypothesis M07 was a poor description of this catchment (as shown in Figure 4), and hence its parameter estimates are unlikely to be meaningful. We focus on selected aspects of the parameter distributions, which appear to be interpretable based on the similarities and differences between these two catchments.

Figure 7a shows the inferred storage–discharge relationship in the UR reservoir. Both catchments exhibit strong nonlinearities, with a convex saturation–area function characterizing a pronounced threshold-like response. The parameter distributions appear to be similar for these catchments, with stronger nonlinearities for the Weierbach catchment. This causes UR to operate as a threshold reservoir and explains why even simplistic threshold models (such as M03) perform comparatively better than more highly parameterized linear models (such as M09 and M10).

Figure 7b shows the inferred distribution of parameter T_f, which controls the width of the lag function and hence the delay in the catchment streamflow response. In the Weierbach, T_f is much larger than in the Wollefsbach. The values of T_f are generally consistent with the time-to-peak values shown in Table I. The different response times of the catchments are also evident from Figure 5: for the same rainfall event, the response of the Wollefsbach is relatively fast, while the response of the Weierbach occurs after a notable delay.

Finally, Figure 7c shows the storage–discharge relation of FR in the two catchments, which depends on the values of parameters K_f and α. The storage–discharge relation of the Wollefsbach is estimated to be more nonlinear than the storage–discharge relation in the Weierbach.

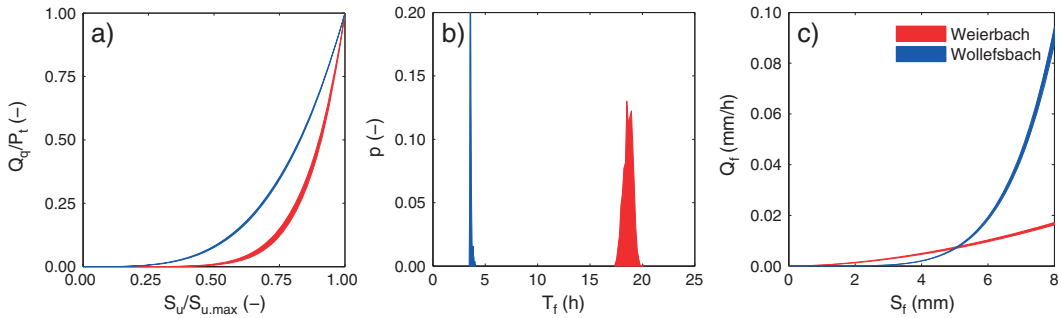

Figure 7. Selected aspects of inferred model structure M07 in the Weierbach and Wollefsbach basins (hypothesis M07 is unsuitable for the Huewelerbach as it lacks a groundwater store component). Panel (a) depicts the constitutive relationships inferred for the UR store, which are similar in the two catchments. The uncertainty bands reflect the 95% quantiles derived from parameter distributions inferred under the WLS regression assumptions (where the model structural error is captured in the additive error term rather than in the parametric uncertainty). Panel (b) shows the distribution of parameter T_f. Although smaller, the Weierbach catchment has a much longer routing time. Panel (c) shows the constitutive functions inferred for the FR store, which are estimated to be more nonlinear in the Wollefsbach than in the Weierbach

DISCUSSION

Model structure and fieldwork insights: is there a correspondence?

This section discusses potential correspondences between the top-down insights from the model comparison (Discussion section) and the bottom-up insights gained from experimental fieldwork (Study Area section). The simultaneous analysis of multiple models on different catchments also illustrates the contribution of comparative hydrology to processes understanding and representation.

Huewelerbach catchment. As noted in the Discussion section, models with serial reservoir connections performed poorly compared to models with a parallel connection. This modelling result could be interpreted based on the following experimental justification. Previous fieldwork has suggested that this catchment has an essentially 'vertical' structure, with a permeable sandstone formation lying on top of an effectively impermeable marly layer. The two zones act almost independently, with the sandstone formation providing a stable groundwater component, and with the marl formation responsible for the quick reaction to rainfall. Representing these distinct flow-generating mechanisms requires distinct model compartments and state variables.

The partitioning of precipitation between the two zones (a fast-reacting zone, represented by the marls, and a slow-reacting zone, represented by the sandstone) is not constant. Indeed, model M08, which hypothesizes a constant partitioning, performs comparatively poorly. M09, which partitions the flow according to the wetness of the catchment, performs clearly better. This is likely due to the behaviour of the marly zone, which, as it appears more clearly from the analysis of the Wollefsbach catchment, has a strongly saturation-dependent response.

The Results section also shows that linear models tend to work comparatively well in this catchment. This could be due to the response of this basin being dominated by a strong groundwater component, which is sustained by the sandstone formation. The slow dynamics of groundwater reservoirs can often be well represented by linear models (e.g. Fenicia *et al.*, 2006).

Weierbach catchment. In the Weierbach catchment, models with serial reservoir connections are able to represent catchment behaviour much better than models with parallel connections. This suggests that the catchment behaves as a 'horizontal', serial system. Indeed, experimental investigations have indicated that the bedrock is essentially impermeable (preventing deep percolation) and that water flows are predominantly lateral and take place at the soil–bedrock interface.

The experimental knowledge available at the Weierbach catchment indicates that the bedrock has an irregular topography. The schist formation tends to be fractured towards the surface, forming a system of local reservoir where water can be stored (Figure 2). Lateral flow can therefore be interpreted as a movement of water across multiple reservoirs. This hypothesis is consistent with experimental work and explains the large delay of this catchment from a modelling perspective.

This catchment is characterized by marked threshold behaviour, as evidenced by the relatively good performance of threshold models (e.g. M03) with respect to linear models, and by the strong nonlinearity of the UR storage discharge relationship of M07 in Figure 7a. This threshold response can be motivated by the 'fill-and-spill' hypothesis (Tromp-van Meerveld and McDonnell, 2006). Runoff is generated from the spilling of water from the reservoirs developed by the irregular bedrock topography (which, when disconnected, empty solely due to evaporation).

The fill-and-spill hypothesis is strongly related to the concept of connectivity of flow pathways (Bracken and Croke, 2007; Jencso *et al.*, 2009), which explains the

state-dependent response of the catchment. The connectivity increases during wet conditions, resulting in a larger fraction of the catchment contributing as one or more connected units. Conversely, connectivity decreases during dry conditions, preventing the transmission of flow and resulting in lower flow volumes.

The transfer function was found to be an important model component in the Weierbach catchment, in order to reproduce the delays of the streamflow response with respect to the rainfall forcing. There is a close mathematical correspondence between reservoirs and transfer function elements. For example, the impulse–response function associated with a sequence of N identical linear reservoirs (the 'Nash cascade') is the Gamma function. It appears that transfer functions (corresponding to a sequence of multiple reservoirs) provide a good approximation to the delays in the Weierbach system (which as discussed above could be related to the fill-and-spill mechanism).

In terms of process understanding, the double-peaked response of this catchment is of interest. While the delayed peak is present only during wet (winter) conditions, the first peak, which is near concomitant to the rainfall event, is present during both winter and summer. During winter the response of this catchment is dominated by a delayed response, while during summer the quick reaction is dominant (see also Kavetski *et al.*, 2011). M07 provides a tentative mechanistic interpretation of the 'double personality' of this catchment, with a RR that is permanently active and reacts promptly to rainfall, and the other model components conceptualizing the delayed and threshold-like response. Conceptually, RR can be interpreted as a model component that represents saturation overland flow and 'rain on water' in the impervious zone and in the near-stream saturated zones. Since these zones do not vary much in size due to the pronounced topographic relief of the catchment, this may explain why this fast-reacting mechanism is always quite active.

Wollefsbach catchment. Similar to the Weierbach catchment, the Wollefsbach catchment also appears to function as a serial system. This is suggested by the relatively good performance of models with serial reservoir connections and finds an experimental justification as the marly bedrock has been classified as effectively impermeable and the water flows are predominantly lateral.

This catchment also shows a marked threshold-like behaviour. This is apparent from the low performance of linear models compared to nonlinear models (e.g. M03, with a threshold structure), and also from the nonlinearities of the storage–discharge relationships shown in Figure 7.

The threshold-like response of this catchment can be reasonably interpreted based on experimental evidence. The main stormflow generation mechanisms of this catchment are saturation subsurface flow and saturation overland flow (Figure 2). The soil becomes progressively saturated from below as the catchment wetness increases (with evaporation being the only appreciable removal mechanism). Runoff is triggered once the soil cannot store any more water. Overall, the perceived behaviour of this catchment corresponds well to the 'variable source area' conceptualization (Hewlett and Hibbert, 1967), which assumes that runoff is generated on areas that are saturated from below and which vary in size depending on the wetness of the catchment.

In contrast to the Weierbach, the Wollefsbach basin has a much faster response (Table I). This is also reflected by the comparison of parameter estimates of M07 in Figure 7. Indeed, the dominant processes are completely different: in the Weierbach, the flow is generated primarily at the soil–bedrock interface, whereas in the Wollefsbach, it is generated primarily at the land surface.

Similar mathematical representations of different hydrological processes?

Although hydrologically different, some of the key processes of the Weierbach and Wollefsbach catchments can be represented using the same mathematical model concepts and components. In particular, the pattern of model performance in these two catchments is quite similar (Figure 4). However, while model M07 is the best performing model in both catchments, the experimental interpretation given to the individual model components can be quite different.

For example, consider the UR reservoir in M07, and its estimated parameter distributions shown in Figure 7. In both catchments, the parameter β is much larger than 1.0, resulting in mathematically similar behaviour. However, while the UR component is critical for representing the threshold-like behaviour of these catchments, the processes that UR is intended to represent are considerably different. In the Weierbach catchment, UR is perceived to represent the concept of connectivity of flow pathways and a fill-and-spill flow-generating mechanism, while in the Wollefsbach catchment, it is perceived to represent the variable source area concept.

More generally, many models adopt similar mathematical formulations even if based on markedly different hydrological conceptualizations. For example, the UR component, which in this study forms part of most other more complex models (Figure 3), is the core component of many widely used existing conceptual models. These include TOPMODEL (Beven and Kirkby, 1979), HBV (Lindstrom *et al.*, 1997), PDM (Moore, 2007), VIC (Wood *et al.*, 1992) HYMOD (Vrugt *et al.*, 2003), and many others. These models are often developed based on distinctly different

experimental justifications. For example, TOPMODEL was initially introduced as a variable source area model (Beven and Kirkby, 1979), while PDM is developed based on a 'fill and spill' concept (Moore, 2007). However, they are mathematically very similar, as discussed by Kavetski *et al.* (2003); Moore (2007), and Clark *et al.* (2008).

The similarity of the mathematical representation of hydrologically distinct flow mechanisms indicates that the correspondence between lower to higher spatial scales is generally a many-to-one relationship (Klemes, 1983). The resulting ambiguities highlight the limitation of the top-down modelling approach in decomposing a hydrological system into its constitutive components based on system-averaged responses and stresses the importance of the bottom-up, fieldwork-based perspective to constrain and/or interpret model results.

CONCLUSIONS

This paper investigated the correspondence between 'catchment structure', estimated based on insights gained from experimental fieldwork, and 'model structure', inferred from the hydrological response of the catchment using inverse modelling. This question is relevant for several major research themes in hydrology, including catchment classification and prediction in ungauged basin, which try to understand the relationships between catchment form and function.

The case study analysed the hydrological behaviour of three headwater catchments in Luxembourg and related it to the performance of 12 competing model hypotheses, implemented within the flexible modelling framework SUPERFLEX. This methodology allowed comparing the 'top-down' modelling perspective with 'bottom-up' experimental insights represented by a set of perceptual models of the three catchments. Our conclusions are as follows:

1. Single-reservoir structures performed poorly in all catchments, indicating that these models are too simplistic for these catchments and that multiple dominant flowpath representations are necessary.
2. In the Huewelerbach catchment, underlain by a permeable sandstone formation, we found that parallel model structures (where reservoirs are connected in parallel) perform better than serial model structures (where reservoirs are connected in series). This is consistent with existing fieldwork information on the presence of a large sandstone aquifer.
3. The Weierbach and Wollefsbach catchments were well represented by serial model structures, which is again consistent with fieldwork perceptions of the absence of deep groundwater flow and the dominance of lateral flow.
4. Linear models performed poorly in the Weierbach and Wollefsbach catchment in contrast to the Huewelerbach catchment. This helped the interpretation of threshold mechanisms in these catchments.
5. The strong delay in the winter response of the Weierbach catchment was interpreted as corresponding to a 'fill-and-spill' runoff-generating mechanism and appeared best represented using a transfer function equivalent to a system of cascading reservoirs.
6. The modelling application considerably augmented the experimental insights into the dynamics of the Weierbach catchment. While the double-peak delayed dynamics were viewed as unexpected from an experimental perspective (given the small size of the catchment and its impermeable bedrock), modelling suggested that this delay and its threshold response could be represented by a system of serially connected reservoirs and, mechanistically, correspond to a fill-and-spill process.

These findings suggest that meaningful correspondences between catchment structure and model structure can exist even for quite simple representations of dominant flowpaths using lumped conceptual models. The conceptualizations developed through small-scale fieldwork investigation and the insights about catchment-scale hydrological behaviour generated by inverse modelling were compatible and could be tentatively related to each other, at least for the catchments analysed in this study. However, although different dominant processes may require different model structures, it was also apparent that hydrologically distinct flow mechanisms could also be represented using similar mathematical formulations.

The study also illustrated the complementarity of modeller and experimentalist perspectives, and the complementarity of the top-down *versus* bottom-up modelling approaches. We noted the difficulties in inferring catchment behaviour based on modelling results alone, due to poor model identifiability from input–output data alone, especially considering that hydrologically distinct mechanisms could be modelled using similar or identical mathematical formulations. A synthesis of model-based and fieldwork-based perspectives is hence of clear interest for future work.

This study also highlighted the limitations of individual lumped model structures in encompassing the diversity of hydrological processes operating in different catchments. Importantly, the more systematic and controlled comparison of different model hypotheses, which is facilitated in flexible modelling frameworks such as SUPERFLEX, can reveal important differences and similarities between catchments, thus helping the interpretation of observed catchment behaviour.

In terms of future work, analogous investigations of the mapping between catchment features and hydrological

response using multi-hypothesis frameworks can be applied to a larger range of catchments. In particular, such mapping remains particularly problematic in larger catchments, where process heterogeneity with a variety of lag effects (both within and outside channels) and complex networks of tributary streams make it more difficult to interpret hydrograph response to precipitation (e.g. Ward and Robinson, 1990). In this respect, the incorporation of distributed information in the modelling process is of interest for future research. Here, there is scope to use multi-hypothesis frameworks such as SUPERFLEX to explore the organizing principles that might be emerging at larger scales and hence improve our understanding and predictions of hydrological function at the larger catchment scale.

ACKNOWLEDGEMENTS

We thank the reviewers for their constructive comments, criticisms, and suggestions, which significantly improved the paper. We acknowledge the financial support of the National Research Fund of Luxembourg through Grant INTER/DFG/11/01 – CAOS 'From Catchments as Organized Systems to Models based on Dynamic Functional Units'. Jim Freer's time in co-writing this paper was funded by NERC grant number NE/I005366/1.

REFERENCES

Ambroise B, Freer J, Beven K. 1996. Application of a generalized TOPMODEL to the small Ringelbach catchment, Vosges, France. *Water Resources Research* **32**: 2147–2159.

Atkinson SE, Woods RA, Sivapalan M. 2002. Climate and landscape controls on water balance model complexity over changing timescales. *Water Resources Research* **38**: 1314. DOI: 10.1029/2002wr001487

Beven K, Kirkby MJ. 1979. A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin – Bulletin Des Sciences Hydrologiques* **24**: 43–69.

Beven KJ. 2012. *Rainfall-runoff modelling : the primer*. Wiley-Blackwell: Chichester; **xxix**, 457.

Beven KJ, Smith PJ, Westerberg IK, Freer J. 2012. Comment on Clark et al., Pursuing the method of multiple working hypotheses for hydrological modeling, W09301, 2011. *Water Resources Research*. DOI:10.1029/2012WR012282, in press.

Blume T, Zehe E, Bronstert A. 2009. Use of soil moisture dynamics and patterns at different spatio-temporal scales for the investigation of subsurface flow processes. *Hydrol. Earth Syst. Sci.* **13**: 1215–1233, 10.5194/hess-13-1215-2009.

Box GEP, Tiao GC. 1992. *Bayesian inference in statistical analysis*. Wiley: New York (NY); 588.

Bracken LJ, Croke J. 2007. The concept of hydrological connectivity and its contribution to understanding runoff-dominated geomorphic systems. *Hydrological Processes* **21**: 1749–1763. DOI: 10.1002/Hyp.6313.

Buytaert W, Beven K. 2010. Models as multiple working hypotheses: hydrological simulation of tropical alpine wetlands. *Hydrological Processes*, 10.1002/hyp.7936.

Chamberlin TC. 1965. The Method of Multiple Working Hypotheses. *Science, New Series* **148**(3671): 754–759.

Clark MP, Kavetski D. 2010. Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. *Water Resources Research* **46**: W10510. DOI: 10.1029/2009wr008894.

Clark MP, Kavetski D, Fenicia F. 2011a. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research* **47**: W09301, 10.1029/2010wr009827.

Clark MP, Kavetski D, Fenicia F. 2012. Response to the comment by Keith Beven et al. on "Pursuing the method of multiple working hypotheses for hydrological modeling". *Water Resources Research* **48**: DOI:10.1029/2012WR012547.

Clark MP, McMillan HK, Collins DBG, Kavetski D, Woods RA. 2011b. Hydrological field data from a modeller's perspective: Part 2: process-based evaluation of model hypotheses. *Hydrological Processes* **25**: 523–543. DOI: 10.1002/Hyp.7902.

Clark MP, Slater AG, Rupp DE, Woods RA, Vrugt JA, Gupta HV, Wagener T, Hay LE. 2008. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research* **44**: W00b02. DOI: 10.1029/2007wr006735.

Fenicia F, Kavetski D, Savenije HHG. 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research* **47**: W11510, 10.1029/2010wr010174.

Fenicia F, McDonnell JJ, Savenije HHG. 2008. Learning from model improvement: On the contribution of complementary data to process understanding. *Water Resources Research* **44**: DOI:10.1029/2007WR006386, W06419. DOI: 10.1029/2007wr006386.

Fenicia F, Savenije HHG, Matgen P, Pfister L. 2006. Is the groundwater reservoir linear? Learning from data in hydrological modelling. *Hydrology and Earth System Sciences* **10**: 139–150.

Freer J, Beven K, Peters N. 2003. Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure. In *Calibration of Watershed Models*. AGU: Washington, DC; 69–87.

Freer JE, McMillan H, McDonnell JJ, Beven KJ. 2004. Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology* **291**: 254–277. DOI: 10.1016/j.jhydrol.2003.12.037.

Freeze RA, Cherry JA. 1979. *Groundwater*. Prentice-Hall: Englewood Cliffs, N.J.; vol. **xvi**, 604.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. *Bayesian data analysis*. Chapman & Hall/CRC: Boca Raton, Fla.; vol. **xxv**, 668.

Graham CB, McDonnell JJ. 2010. Hillslope threshold response to rainfall: (2) Development and use of a macroscale model. *Journal of Hydrology* **393**: 77–93. DOI: 10.1016/j.jhydrol.2010.03.008.

Graham CB, Woods RA, McDonnell JJ. 2010. Hillslope threshold response to rainfall: (1) A field based forensic approach. *Journal of Hydrology* **393**: 65–76. DOI 10.1016/j.jhydrol.2009.12.015.

Gupta HV, Clark MP, Vrugt JA, Abramowitz G, Ye M. 2012. Towards a comprehensive assessment of model structural adequacy. *Water Resources Research* **48**: W08301, 10.1029/2011wr011044.

Hersbach H 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting* **15**: 559–570.

Hewlett JD, Hibbert AR. 1967. Factors affecting the response of small watersheds to precipitation in humid areas. In *International Symposium on Forest Hydrology*, Sopper WE, Lull HW (eds). Pergamon, Oxford, UK; 275–290.

Jencso KG, McGlynn BL, Gooseff MN, Wondzell SM, Bencala KE, Marshall LA. 2009. Hydrologic connectivity between landscapes and streams: Transferring reach-and plot-scale understanding to the catchment scale. *Water Resources Research* **45**: W04428; DOI: 10.1029/2008wr007225.

Juilleret J, Iffly JF, Hoffmann L, Hissler C. 2012. The potential of soil survey as a tool for surface geological mapping: a case study in a hydrological experimental catchment (Huewelerbach, Grand-Duchy of Luxembourg). *Geologica Belgica* **15**: 36–41.

Kavetski D, Clark MP. 2010. Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on

model analysis and prediction. *Water Resources Research* **46**: W10511. DOI: 10.1029/2009wr008896.
Kavetski D, Clark MP. 2011. Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis testing. *Hydrological Processes* **25**: 661–670. DOI: 10.1002/Hyp.7899.
Kavetski D, Fenicia F. 2011. Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research* **47**: W11511, 10.1029/2011wr010748.
Kavetski D, Fenicia F, Clark MP. 2011. Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment. *Water Resources Research* **47**: W05501, 10.1029/2010wr009525.
Kavetski D, Kuczera G. 2007. Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration. *Water Resources Research* **43**: W03411. DOI: 10.1029/2006wr005195.
Kavetski D, Kuczera G, Franks SW. 2003. Semidistributed hydrological modeling: A "saturation path'' perspective on TOPMODEL and VIC. *Water Resources Research* **39**: 1246. DOI: 10.1029/2003wr002122.
Klemes V 1983. Conceptualization and Scale in Hydrology. *Journal of Hydrology* **65**: 1–23.
Krueger T, Freer J, Quinton JN, Macleod CJA, Bilotta GS, Brazier RE, Butler P, Haygarth PM. 2010. Ensemble evaluation of hydrological model hypotheses. *Water Resources Research* **46**: W07516. DOI: 10.1029/2009wr007845.
Lindstrom G, Johansson B, Persson M, Gardelin M, Bergstrom S. 1997. Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology* **201**: 272–288.
Martinez-Carreras N, Krein A, Udelhoven T, Gallart F, Iffly JF, Hoffmann L, Pfister L, Walling DE. 2010. A rapid spectral-reflectance-based fingerprinting approach for documenting suspended sediment sources during storm runoff events. *Journal of Soils and Sediments* **10**: 400–413. DOI 10.1007/s11368-009-0162-1.
McDonnell JJ, Sivapalan M, Vache K, Dunn S, Grant G, Haggerty R, Hinz C, Hooper R, Kirchner J, Roderick ML, Selker J, Weiler M. 2007. Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research* **43**: W07301. DOI: 10.1029/2006wr005467.
McGlynn BL, McDonnel JJ, Brammer DD. 2002. A review of the evolving perceptual model of hillslope flowpaths at the Maimai catchments, New Zealand. *Journal of Hydrology* **257**: 1–26.
McMillan HK, Clark MP, Bowden WB, Duncan M, Woods RA. 2011. Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure. *Hydrological Processes* **25**: 511–522. DOI: 10.1002/Hyp.7841.
Moore RJ. 2007. The PDM rainfall-runoff model. *Hydrology and Earth System Sciences* **11**: 483–499.
Perrin C, Michel C, Andreassian V. 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology* **242**: 275–301.
Pfister L, McDonnell JJ, Hissler C, Hoffmann L. 2010. Ground-based thermal imagery as a simple, practical tool for mapping saturated area connectivity and dynamics. *Hydrological Processes* **24**: 3123–3132. DOI: 10.1002/Hyp.7840.
Pfister L, McDonnell JJ, Wrede S, Hlubikova D, Matgen P, Fenicia F, Ector L, Hoffmann L. 2009. The rivers are alive: on the potential for diatoms as a tracer of water source and hydrological connectivity. *Hydrological Processes* **23**: 2841–2845. DOI: 10.1002/Hyp.7426.
Pinol J, Beven K, Freer J. 1997. Modelling the hydrological response of Mediterranean catchments, Prades, Catalonia. The use of distributed models as aids to hypothesis formulation. *Hydrological Processes* **11**: 1287–1306.
Refsgaard JC, Knudsen J. 1996. Operational validation and intercomparison of different types of hydrological models. *Water Resources Research* **32**: 2189–2202.
Renard B, Kavetski D, Kuczera G, Thyer M, Franks SW. 2010. Understanding predictive uncertainty in hydrologic modelling: the challenge of identifying input and structural errors. *Water Resources Research*, **46**: W05521. DOI: 10.1029/2009wr008328.
Savenije HHG. 2009. HESS Opinions "The art of hydrology". *Hydrology and Earth System Sciences* **13**: 157–161.
Savenije HHG. 2010. HESS Opinions "Topography driven conceptual modelling (FLEX-Topo)". *Hydrology and Earth System Sciences* **14**: 2681–2692. DOI: 10.5194/hess-14-2681-2010.
Seibert J, McDonnell JJ. 2002. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research* **38**: 1241. DOI: 10.1029/2001wr000978.
Sivapalan M 2003. Process complexity at hillslope scale, process simplicity at the watershed scale: is there a connection? *Hydrological Processes* **17**: 1037–1041. DOI: 10.1002/Hyp.5109.
Sivapalan M 2009. The secret to 'doing better hydrological science': change the question! *Hydrological Processes* **23**: 1391–1396, 10.1002/hyp.7242.
Sivapalan M, Bloschl G, Zhang L, Vertessy R. 2003. Downward approach to hydrological prediction. *Hydrological Processes* **17**: 2101–2111.
Thyer M, Renard B, Kavetski D, Kuczera G, Franks SW, Srikanthan S. 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research* **45**: W00b14. DOI: 10.1029/2008wr006825.
Tromp-van Meerveld HJ, McDonnell JJ. 2006a. Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope. *Water Resources Research* **42**: W02410. DOI: 10.1029/2004wr003778.
Tromp-van Meerveld HJ, McDonnell JJ. 2006b. Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis. *Water Resources Research* **42**: W02411. DOI: 10.1029/2004wr003800.
Uhlenbrook S, Leibundgut C. 2002. Process-oriented catchment modelling and multiple-response validation. *Hydrological Processes* **16**: 423–440.
van den Bos R, Hoffmann L, Juilleret J, Matgen P, Pfister L. 2006. Conceptual modelling of individual HRU's as a trade-off between bottom-up and top-down modelling: A case study In *3rd Biennial Meeting of International Environmental Modelling and Software Society*. Burlington.
Vrugt JA, Gupta HV, Bouten W, Sorooshian S. 2003. A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research* **39**: 1201. DOI: 10.1029/2002wr001642.
Wagener T, Sivapalan M, Troch PA, Woods R. 2007. Catchment Classification and Hydrologic Similarity. *Geography Compass* **1**: 901–931.
Ward RC, Robinson M. 1990. *Principles of hydrology*. McGraw-Hill: London; New York; **xvi**, 365.
Weiler M, McDonnel JJ, Tromp-van Meerveld HJ, Uchida T. 2005. Subsurface Stormflow. In *Encyclopedia of Hydrological Sciences*, Anderson M (ed). John Wiley & Sons, Ltd; 1719–1732.
Wood EF, Lettenmaier DP, Zartarian VG. 1992. A Land-Surface Hydrology Parameterization with Subgrid Variability for General-Circulation Models. *Journal of Geophysical Research-Atmospheres* **97**: 2717–2728.
Young PC, Beven KJ. 1994. Data-Based Mechanistic Modeling and the Rainfall-Flow Nonlinearity. *Environmetrics* **5**: 335–363.

APPENDIX A

Details of the Hydrological models

Details of the 12 model structures are provided in Tables AI–AIV. Table AI summarizes the components and parameters present in each model structure. Table AII details the water balance equations of the various model components. Table AIII describes the constitutive functions relating storages and fluxes. Table AIV provides further details of the functions used in Table AIII.

Table AI. Components and parameters of model structures M01-M12. N_θ is the number of parameters and N_s is the number of states. IR, UR, FR, SR and LF denote the interception, unsaturated, fast, slow reservoirs and lag function respectively.

Model			Components						Parameters											
	N_s	N_θ	IR	UR	FR	SR	RR	LF	C_e (−)	I_{max} (mm)	$S_{u,max}$ (mm)	β (−)	M (−)	K_r (1/h)	R_{max} (mm/h)	T_f (h)	K_f ($mm^{1-\alpha}$/h)	α (−)	D (−)	K_s (1/h)
M01	1	3	-	-	✓	-	-	-	✓	-	-	-	-	-	-	-	✓	✓	-	-
M02	1	4	-	✓	-	-	-	-	✓	-	✓	✓	-	-	✓	-	-	-	-	-
M03	2	4	-	✓	✓	-	-	-	✓	-	✓	-	-	-	-	-	✓	✓	-	-
M04	2	5	-	✓	✓	-	-	-	✓	-	✓	✓	-	-	-	-	✓	✓	-	-
M05	3	6	-	✓	✓	-	-	✓	✓	-	✓	✓	-	-	-	✓	✓	✓	-	-
M06	4	7	✓	✓	✓	-	-	✓	✓	✓	✓	✓	-	-	-	✓	✓	✓	-	-
M07	4	8	-	✓	✓		✓	✓	✓	-	✓	✓	✓	✓	-	✓	✓	✓	-	-
M08	2	4	-	-	✓	✓	-	-	✓	-	-	-	-	-	-	-	✓	-	✓	✓
M09	3	5	-	✓	✓	✓	-	-	✓	-	✓	-	-	-	-	-	✓	-	✓	✓
M10	4	6	-	✓	✓	✓	-	✓	✓	-	✓	-	-	-	-	✓	✓	-	✓	✓
M11	4	7	-	✓	✓	✓	-	✓	✓	-	✓	✓	-	-	-	✓	✓	-	✓	✓
M12	5	8	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	-	-	-	✓	✓	-	✓	✓

Table AII. Water balance equations of the models used in the experiments (✓ and "-" indicate presence or absence respectively).

Water balance equations:	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
$\frac{dS_f}{dt} = P_f - Q_f - E_f$	✓	-	-	-	-	-	-	✓	-	-	-	-
$\frac{dS_f}{dt} = P_f - Q_f$	-	-	✓	✓	-	-	-	-	✓	-	-	-
$\frac{dS_f}{dt} = P_{fl} - Q_f$	-	-	-	-	✓	✓	✓	-	-	✓	✓	✓
$\frac{dS_u}{dt} = P_u - Q_q - Q_u - E_u$	-	✓	-	-	-	-	-	-	-	-	-	-
$\frac{dS_u}{dt} = P_u - Q_q - E_u$	-	-	✓	✓	✓	✓	✓	-	✓	✓	✓	✓
$\frac{dS_s}{dt} = P_s - Q_s$	-	-	-	-	-	-	-	✓	✓	✓	✓	✓
$\frac{dS_i}{dt} = P_t - P_u - E_i$	-	-	-	-	-	✓	-	-	-	-	-	✓
$\frac{dS_r}{dt} = P_r - Q_r$	-	-	-	-	-	-	✓	-	-	-	-	-
$P_t = P_u + P_r$	-	-	-	-	-	-	✓	-	-	-	-	-
$P_t = P_f + P_s$	-	-	-	-	-	-	-	✓	-	-	-	-
$P_t = P_f$	✓	-	-	-	-	-	-	-	-	-	-	-
$P_t = P_u$	-	✓	✓	✓	✓	-	-	-	✓	✓	✓	-
$Q_q = P_f + P_s$	-	-	-	-	-	-	-	-	✓	✓	✓	✓
$Q_t = Q_f$	✓	-	✓	✓	✓	✓	-	-	-	-	-	-
$Q_t = Q_f + Q_r$	-	-	-	-	-	-	✓	-	-	-	-	-
$Q_t = Q_q + Q_u$	-	✓	-	-	-	-	-	-	-	-	-	-
$Q_t = Q_f + Q_s$	-	-	-	-	-	-	-	✓	✓	✓	✓	✓

Table AIII. Constitutive functions of the models used in the experiments (✓ and "-" indicate presence or absence respectively). The operator * in the equation for P_{fl} denotes the convolution operator*. The parameters m_1, m_2 and m_3 are "smoothing" parameters (Kavetski and Kuczera, 2007) and they are fixed (to a value of 10^{-2}).

Constitutive functions	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
$\bar{S}_i = S_i/S_{i,max}$	-	-	-	-	-	✓	-	-	-	-	-	✓
$P_u = P_t f_h(\bar{S}_i \mid m_1)$	-	-	-	-	-	✓	-	-	-	-	-	✓
$E_i = C_e E_p f_m(\bar{S}_i \mid m_2)$	-	-	-	-	-	✓	-	-	-	-	-	✓
$\bar{S}_u = S_u/S_{u,max}$	-	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓
$Q_q = P_u f_p(\bar{S}_u \mid \beta)$	-	✓	-	✓	✓	✓	✓	-	✓	✓	✓	✓
$Q_q = P_u f_h(\bar{S}_u \mid m_1)$	-	-	✓	-	-	-	-	-	-	-	-	-
$E_u = C_e E_p f_m(\bar{S}_u \mid m_2)$	-	✓	✓	✓	✓	-	✓	-	✓	✓	✓	-
$E_u = C_e E_p (1 - f_m(\bar{S}_i \mid m_2)) f_m(\bar{S}_u \mid m_2)$	-	-	-	-	-	✓	-	-	-	-	-	✓
$E_f = C_e E_p f_e(S_f \mid m_3)$	✓	-	-	-	-	-	-	✓	-	-	-	-
$P_{fl} = (P_f * h_f)(t)$	-	-	-	-	✓	✓	✓	-	-	✓	-	✓
$h_f = \left\{ \begin{matrix} 2t/T_f^2, t<T_f \\ 0, \quad t>T_f \end{matrix} \right\}$	-	-	-	-	✓	✓	✓	-	-	✓	-	✓
$P_r = MP_t$	-	-	-	-	-	-	✓	-	-	-	-	-
$P_s = DQ_q$	-	-	-	-	-	-	-	-	✓	✓	✓	✓
$P_s = DP_t$	-	-	-	-	-	-	-	✓	-	-	-	-
$Q_u = R_{max}\bar{S}_u$	-	✓	-	-	-	-	-	-	-	-	-	-
$Q_r = k_r S_r$	-	-	-	-	-	-	✓	-	-	-	-	-
$Q_f = k_f S_f$	-	-	-	-	-	-	-	✓	✓	✓	✓	✓
$Q_f = k_f S_f^a$	✓	-	✓	✓	✓	✓	✓	-	-	-	-	-
$Q_s = k_s S_s$	-	-	-	-	-	-	-	✓	✓	✓	✓	✓

*Lag function smoothed using the method in Kavetski and Kuczera (2007).

Table AIV. Constitutive functions

Functions	Name
$f_p(x \mid m) = x^m$	Power function
$f_r(x \mid m) = 1 - (1-x)^m$	'Reflected' power function
$f_m(x \mid m) = \frac{x(1+m)}{x+m}$	Monod-type kinetics, adjusted so that $f_m(1 \mid m) = 1$
$f_h(x \mid m) = 1 - \frac{(1-x)(1+m)}{1-x+m}$	'Reflected' hyperbolic function, scaled to the unit square
$f_e(x \mid m) = 1 - e^{-x/m}$	Tessier function (note that $f_e(x \mid m) \to 1$ as $x \to \infty$)