Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights

Dmitri Kavetski[1] and Fabrizio Fenicia[2,3]

Received 3 April 2011; revised 19 September 2011; accepted 28 September 2011; published 17 November 2011.

[1] In this article's companion paper, flexible approaches for conceptual hydrological modeling at the catchment scale were motivated, and the SUPERFLEX framework, based on generic model components, was introduced. In this article, the SUPERFLEX framework and the "fixed structure" GR4H model (an hourly version of the popular GR4J model) are applied to four hydrologically distinct experimental catchments in Europe and New Zealand. The estimated models are scrutinized using several diagnostic measures, ranging from statistical metrics, such as the statistical reliability and precision of the predictive distribution of streamflow, to more process-oriented diagnostics based on flow-duration curves and the correspondence between model states and groundwater piezometers. Model performance was clearly catchment specific, with a single fixed structure unable to accommodate intercatchment differences in hydrological behavior, including seasonality and thresholds. This highlights an important limitation of any "fixed" model structure. In the experimental catchments, the ability of competing model hypotheses to reproduce hydrological signatures of interest could be interpreted on the basis of independent fieldwork insights. The potential of flexible frameworks such as SUPERFLEX is then examined with respect to systematic and stringent hypothesis-testing in hydrological modeling, for characterizing catchment diversity, and, more generally, for aiding progress toward a more unified formulation of hydrological theory at the catchment scale. When interpreted in physical process-oriented terms, the flexible approach can also serve as a language for dialogue between modeler and experimentalist, facilitating the understanding, representation, and interpretation of catchment behavior.

Citation: Kavetski, D., and F. Fenicia (2011), Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, *Water Resour. Res., 47*, W11511, doi:10.1029/2011WR010748.

1. Introduction

[2] Despite a current proliferation of conceptual hydrological models, the presently dominant philosophy of model development can arguably be described as pursuing a "one model fits all" solution, a "universal" *fixed* model structure that is adequate over a large range of applications, and, in particular, is transposable in both space and time (e.g., as advocated by *Linsley* [1982] and *Andréassian et al.* [2009]). However, a case can also be made that *flexible* frameworks are a powerful tool for a more systematic hypothesis testing in hydrology, allowing the formulation and appraisal of alternative representations of hydrological systems of interest [e.g., *McDonnell*, 2003; *Clark et al.*, 2008, 2011a; *Fenicia et al.*, 2008b; *Savenije*, 2009, and others].

[3] In the companion paper [*Fenicia et al.*, 2011], the field of conceptual hydrological modeling was reviewed, motivating a flexible framework, termed SUPERFLEX, for building models from generic components intended to represent distinct hydrological functions at the catchment scale (e.g., see discussions by *Wagener et al.* [2007] and *Sivapalan* [2005]). The "fixed" and "flexible" approaches can be compared and contrasted with respect to several scientific and operational criteria, including their ability to provide adequate representations of dominant hydrological processes, to explore the model complexity supported by available data and prior knowledge, and to serve as a language of exchange between modelers and experimentalists. These issues, which have appreciable overlap, are briefly reviewed next (for fuller discussion, see the companion paper *Fenicia et al.* [2011]).

1.1. Hypothesis-Testing and "Uniqueness of Place"

[4] The lack of hydrological theories at the catchment scale has been noted by several commentators [e.g., *Sivapalan*, 2005; *McDonnell et al.*, 2007; *Troch et al.*, 2009]. An important related question is whether a single lumped model structure could feasibly characterize the diversity of catchments ("uniqueness of place," [*Beven*, 2000]), or, alternatively, whether the heterogeneities of natural systems require specific model modifications. For example,

[1]Environmental Engineering, University of Newcastle, Callaghan, New South Wales, Australia.
[2]Department of Environment and Agro-Biotechnologies, Centre de Recherche Public–Gabriel Lippmann, Belvaux, Grand-Duchy of Luxembourg.
[3]Water Resources Section, Delft University of Technology, Delft, Netherlands.

0043-1397/11/2011WR010748

does "appropriate" model complexity depend on the catchment characteristics [*Wagener et al.*, 2001; *Atkinson et al.*, 2002]? Can differences in inferred model structure and complexity be motivated and/or interpreted based on experimental understanding of catchment behavior? We note that the fixed modeling philosophy would appear to be in conflict with the perception that appropriate structure and complexity may depend on the catchment characteristics.

[5] Finally, while one could argue that model structure and complexity should be dictated solely by the physical laws governing the system of interest, in practice it appears difficult to avoid model development becoming dependent on a multitude of pragmatic factors, such as the availability and quality of data, computing budgets, perceptions of the dominant characteristics of the hydrologic system, and the modeling objectives [e.g., *Butts et al.*, 2004; *Refsgaard and Henriksen*, 2004; and others].

1.2. Model "Realism"

[6] Although it is not necessarily clear-cut in conceptual hydrological modeling where considerable spatio-temporal lumping and process abstraction takes place, we define "realism" in terms of the following two criteria: (1) the model contains components representing, in some way, processes that are deemed important in the particular catchment (e.g., as suggested by independent experimental fieldwork), and, importantly, (2) these components are behaving in a way that is qualitatively or quantitatively consistent with the processes they are intended to represent. Such correspondence between model and "reality," often described as "working for the right reasons" [*Kirchner*, 2006], is essential if the model is to be used as a tool for improving the understanding of a hydrological system, and/or used for prediction and extrapolation, such as simulating the impacts of land use change, variability in climatological forcings, etc.

1.3. Dialogue Between Modeler and Experimentalists

[7] The dialogue between the modeler and experimentalist has been described as an essential aspect of hydrological science [*Seibert and McDonnell*, 2002; *Fenicia et al.*, 2008a]. The experimentalist perspective is, arguably, irreplaceable in the interpretation of the data and in the quantitative assessment of the model "realism." The modeler perspective is needed to implement the range of competing model hypotheses, to quantitatively test them, and, ultimately, use them to produce predictions, uncertainty estimates, etc. In line with the view of science as a cycle of theory and experiments, where new theories motivate new experiments and vice versa, robust exchanges between modelers and experimentalists can help identify limitations in the current understanding of hydrological processes, motivate additional fieldwork investigations, and inform improved model representations. Yet establishing and maintaining such a dialogue is hampered by the absence of a common language of communication, which could be used by both sides to exchange both quantitative information and qualitative insights.

[8] Whether fixed or flexible model approaches are the most appropriate learning tools is debatable. In favor of the fixed model philosophy, *Linsley* [1982, p. 14] suggests that "a new model for every application would eliminate the opportunity for learning that comes with repeated applications of the same model." On the other hand, a flexible framework is a powerful instrument for comparing and testing alternative hypotheses and for systematically organizing and interpreting modeling results. This paper continues our discussions in the companion paper and in other work [*Clark et al.*, 2011a] and argues that, while fixed model structures play an important role in hydrological modeling, there are major benefits in using flexible frameworks to more directly represent current uncertainties and ambiguities in environmental process representation, and facilitate, whenever possible, the use of site-specific experimental insights as part of the model interpretation process.

2. Aims and Scope

[9] This paper discusses the application of the flexible framework SUPERFLEX, within which several specific model structures are hypothesized and implemented, and compares it to the "fixed structure" GR4H model (an hourly version of the popular GR4J model), on four catchments in Europe and New Zealand. Our intention is not to critique the GR4H and GR4J models per se, as these are valuable and popular models used in research worldwide and employed operationally in France [*Berthet et al.*, 2009]. Rather, the aims of the study are (1) to investigate the general limitations of restricting ourselves to a single fixed model structure, especially when facing the challenge of diversity and "uniqueness of place" [*Beven*, 2000], (2) to illustrate how alternative model structures can be hypothesized and implemented using the generic components, and (3) to argue that a flexible framework can, when applied carefully, overcome many of the limitations of the fixed model paradigm.

[10] As part of the hypothesis-based model evaluation, we compare the model results obtained under different inference assumptions, examine the statistical reliability and precision of the models' predictive distributions in reproducing the observed streamflow data, use flow-duration curves to more stringently diagnose model deficiencies in reproducing "data signatures" [*Gupta et al.*, 2008; *Yilmaz et al.*, 2008], and make further inroads into using groundwater piezometer measurements to appraise the "physical realism" of conceptual models [*Seibert*, 2000; *Seibert and McDonnell*, 2002; *Freer et al.*, 2004; *Fenicia et al.*, 2008b].

[11] Given the range of outstanding issues in hydrological modeling, and because of length constraints, we must necessarily limit the scope of the analyses and discussion. In particular, evaluations over large numbers of experimental and operational catchments [e.g., *Perrin et al.*, 2001; *Le Moine et al.*, 2007; *Merz et al.*, 2009] are beyond our scope here and will be carried out separately. Detailed recommendations and recipes on how best to pursue model development with different levels of prior information, from ungaged to well-instrumented catchments, are left to future investigations. The inclusion of rainfall uncertainty [e.g., *Kavetski et al.*, 2002; *Vrugt et al.*, 2008; *Götzinger and Bardossy*, 2008], more sophisticated structural error analysis [e.g., *Kuczera et al.*, 2006; *Reichert and Mieleitner*, 2009; *Doherty and Welter*, 2010], more comprehensive application of multiple model diagnostics [e.g., *Gupta et al.*, 2008; *Yilmaz et al.*, 2008], and the extension to semidistributed and fully distributed modeling systems [e.g., *Ivanov et al.*, 2004; *Immerzeel and Droogers*, 2008] are also substantial

further developments in their own right, and hence deferred to separate investigations.

3. Case Study Setup

3.1. Methodology

[12] The case study methodology was designed as follows. Four catchments with contrasting climatology and/or physical attributes were selected on the basis of prior experimental insights and data availability (section 3.2 and Table 1). Seven distinct model structures, partly motivated by previous modeling experience in these catchments, were then hypothesized and built using the SUPERFLEX framework (section 3.3). The GR4H model [*Le Moine*, 2008], which is a modified hourly version of the popular GR4J model [*Perrin et al.*, 2003], was included as a benchmark and as a representative of an a priori fixed "compromise" model structure.

[13] The parameters of the hypothesized model structures were calibrated using two regression schemes, namely the standard and weighted least squares methods (section 3.4). A range of statistical and process-oriented diagnostics was used to scrutinize the model performance in the calibration and validation periods (section 3.5). Several plausible interpretations of differences in performance across the different model structures and catchments were then proposed, based on the fieldwork understanding available in these catchments.

[14] Although one of the motivations of SUPERFLEX is to facilitate the process of iterative model improvement, the case study is formulated to explore the representation of catchment diversity using fixed versus flexible models. To keep the analyses and presentation manageable, the case study is carried out using a set of model structures hypothesized a priori. While previous modeling experience played a role in selecting these hypotheses (section 3.3.2), we do not attempt to refine them further as part of this application. For illustrations of such "learning from model improvement" using the earlier FLEX model, see *Fenicia et al.* [2008a].

3.2. Experimental Catchments

3.2.1. Maimai Catchment (New Zealand)

[15] The Maimai study area is located in the South Island of New Zealand (see *McGlynn et al.* [2002], for a review of hydrological investigations in this basin). In this work, we use the Maimai M8 catchment, which has a long history of fieldwork [e.g., *Mosley*, 1979; *Pearce et al.*, 1986; *McDonnell*, 1990] and modeling [e.g., *Seibert and McDonnell*, 2002; *Freer et al.*, 2004; *Vaché and McDonnell*, 2006; *Fenicia et al.*, 2010].

Table 1. Basic Summary of Catchments Used in the Case Studies

	Maimai	Wollefsbach	Useldange	Pfaffenthal
Area (km^2)	0.038	4.61	250	385
Average annual precipitation (mm)	2400	840	840	840
Average annual discharge (mm)	1400	200	160	400
Average annual potential ET (mm)	840	660	660	660
Forest (%)	100	7	32	27
Cropland (%)	0	28	29	25
Grassland (%)	0	65	35	28
Urban (%)	0	0	4	20

[16] The topography is characterized by steep slopes and deeply incised channels; its soils are shallow and underlain by a nearly impermeable cemented conglomerate. The climatology is humid, with little seasonal variation. The catchment is highly responsive to rainfall and its soils are normally at more than 90% saturation [*Mosley*, 1979].

[17] The response dynamics of the Maimai have been described as "strikingly simple" [*Vaché and McDonnell*, 2006]. The simplicity of its behavior has been attributed to (1) a wet climate with little seasonality, leading to a chronically wet catchment state and narrowing its range of hydrological regimes, and (2) a combination of steep slopes and an impermeable underlying substratum, resulting in quick forcing-response characteristics.

3.2.2. The Alzette Catchments: Wollefsbach, Useldange, and Pfaffenthal (Luxembourg)

[18] Three Luxembourgish catchments (Wollefsbach, Useldange, and Pfaffenthal) form part of the larger Alzette system (Figure 1), which is monitored by a dense hydrological observation network [*Pfister et al.*, 2002]. While their precipitation shows little seasonality, potential evaporation varies significantly across the seasons, resulting in the catchment discharge being generally higher in winter than in summer [*Pfister et al.*, 2004; *Hellebrand et al.*, 2008]. The Alzette region is characterized by a heterogeneous geological substratum, which strongly influences its response behavior. A summary of the catchments' climatology and hydrogeology is as follows (see Table 1 and *van den Bos et al.* [2006a, 2006b], for further details).

[19] The Wollefsbach catchment is located on marls formations. Its topsoils are shallow and loamy, and are underlain by low-permeability clay deposits. Since deep percolation is likely to be minor, the main storage unit of the catchment is provided by its top layer overlaying the clay bedrock. While the Wollefsbach catchment is highly responsive during winter, its summer streamflow is generally low. Since the rainfall is quite uniform throughout the year, streamflow variations can be attributed to seasonal variations of evaporation associated with the low storage capacity of this catchment [*Pfister et al.*, 2004].

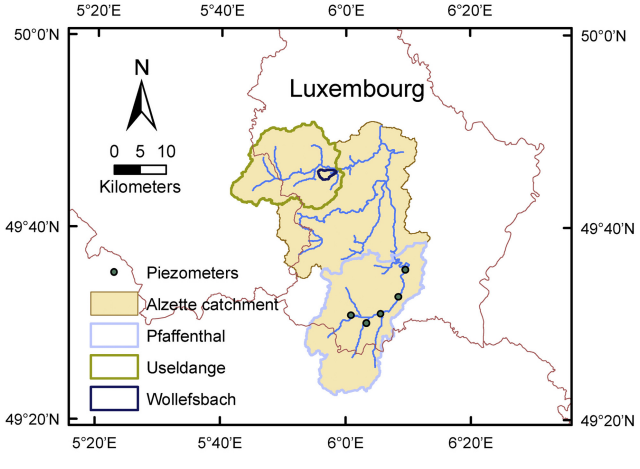

Figure 1. Luxembourgish catchments within the Alzette region. The location of the piezographs used in the model evaluation process for the Pfaffenthal catchment are depicted with circles.

[20] The Useldange catchment contains the Wollefsbach as one of its subcatchments. However, its geology is markedly more heterogeneous. While the main geological unit is the largely impermeable marls (a clay) (55%), the basin also contains significant schist (30%) and some sandstone (15%). Although the schist is fractured (foliated), it appears essentially impermeable in the vertical direction, since the fractures are obstructed by clay deposits. Hence, its contribution to groundwater flows is generally low. However, at the soil-bedrock interface, where a well-developed weathered zone is present, the fractured schist formation, in combination with the irregular shape of the bedrock topography, generates local reservoirs where water can be stored. When saturated, this flow network can support subsurface lateral flow [e.g., *van den Bos et al.*, 2006b]. The sandstone formation, in contrast to marls and schist, is highly permeable and supplies a groundwater component during dry weather periods. However, this formation occupies a relatively small fraction of this catchment.

[21] The Pfaffenthal catchment is the largest catchment in our analysis (385 km^2). Its geology is composed of sandstone (70%) and marls (30%). The large and highly permeable sandstone formation provides a stable groundwater component, sustaining streamflows during the summer season.

3.2.3. Rationale Behind the Catchment Selection

[22] The catchments were selected to include a range of different hydrological dynamics, as well as on the basis of fieldwork insights into their climatological and physical characteristics (e.g., *McGlynn et al.* [2002] for the Maimai, and *van den Bos et al.* [2006a] for the Alzette). Some considerations are listed below:

[23] 1. Maimai: A Small catchment described as an "end member" on a simplicity-complexity scale, with rapid response and little seasonality. It is included to explore model complexity issues (including internal scrutiny using piezometric data), and to provide a contrast to the more complex Luxembourgish catchments.

[24] 2. Wollefsbach: It is slightly larger than the Maimai, but still a comparatively small catchment. Although its geology and soils are different from the Maimai, they are still fairly uniform and deep percolation appears negligible. Hence, hydrological differences between the Wollefsbach and the Maimai appear driven primarily by climatological differences: in the Wollefsbach, rainfall is still uniformly distributed, but evaporation varies seasonally. As a result, the discharge regime is quite different in summer ("dry") versus winter ("wet").

[25] 3. Useldange: A larger basin comprising the Wollefsbach as one of its headwater catchments. Since their climatology is the same, the hydrological differences between the Useldange and Wollefsbach basins can be attributed to differences in physical attributes. In particular, the Useldange catchment is composed of a wider variety of soils; its lithology is also more heterogeneous, comprising three main units. However, similar to the Maimai and Wollefsbach, hydrogeological studies suggest that groundwater flows are minor.

[26] 4. Pfaffenthal: A heterogeneous catchment in terms of geology and land use, which could therefore be a priori classified as moderately complex. Unlike the previous catchments, the importance of any catchment compartment cannot be excluded a priori. Importantly, there is a substantial contribution to base flow from the sandstone formation, leading to significant groundwater activity.

[27] The range of selected catchments allows exploring the representation of diverse hydrological dynamics using lumped conceptual models (the "uniqueness of place" argument [*Beven*, 2000]), while the availability of fieldwork-based insights helps interpret the model results.

3.2.4. Data Used in the Analysis

[28] The inference and validation data are composed of consecutive 2-year calibration and 1-year validation periods, selected as 1 September 2005 – 1 September 2007 – 1 September 2008 for the Luxembourgish basins, and 1 January 1985 – 1 January 1987 – 1 January 1988 for the Maimai catchment. In the Alzette, approximately 30 storm events were included in the calibration period and about 20 events in validation. In the Maimai, about 100 storm events were included in the calibration and about 50 events in validation. The analysis periods were selected to avoid changes in instrumentation in the catchments and to have the same total duration. Hourly resolution precipitation, discharge, and potential evaporation data are available in all four catchments. An additional 1-yr warm-up period was included prior to the calibration period. The validation analysis used the preceding calibration period as a warm-up.

[29] In the Maimai basin, the rainfall is measured with a recording rain gage within the catchment, and potential evaporation is estimated as described by *Rowe et al.* [1994] and *Vaché and McDonnell* [2006]. In the Wollefsbach, Useldange, and Pfaffenthal catchments, the rainfall was measured using tipping bucket rain gages (2, 5, and 4 rain gages, respectively), while evaporation was estimated using the Penman equation.

[30] In the Maimai and Pfaffenthal basins, in addition to diagnostic measures derived solely from streamflow data, the internal dynamics of the inferred models are evaluated against observed groundwater levels not used in the calibration. In the Maimai catchment, 20-min resolution piezometer time series are available at two locations: one in the proximity of the stream, and another on a steep upslope area. A detailed description of this data and its uncertainties is given by *Freer et al.* [2004], where the measurement locations correspond to the "NS" and "P5" sites, respectively. For the Pfaffenthal catchment, hourly resolution data is available from five piezometers, located in roughly equal spacing in the proximity of the main stream channel, as shown in Figure 1. Some of the important limitations of using the piezometer data for model evaluations are reviewed in section 4.3.

3.3. Hydrological Models

3.3.1. Model Hypotheses Explored in the Case Study

[31] Eight alternative model hypotheses of varying complexity are considered in this study, including seven SUPERFLEX configurations and the GR4H model (the latter used as a representative fixed model structure). The SUPERFLEX model schematics are shown in Figure 2 and summarized in Table 2. The model equations are detailed in the appendix (Tables A1 and A2), with constitutive functions already introduced in Table 1 of the companion paper. Section 3.3.2 outlines the perceptual motivation behind these hypotheses.

[32] All models simulate total discharge Q_t, given observed precipitation P_t, and pre-estimated potential evaporation E_p. The seven SUPERFLEX models are a combination of, at most, three reservoirs, which are for convenience

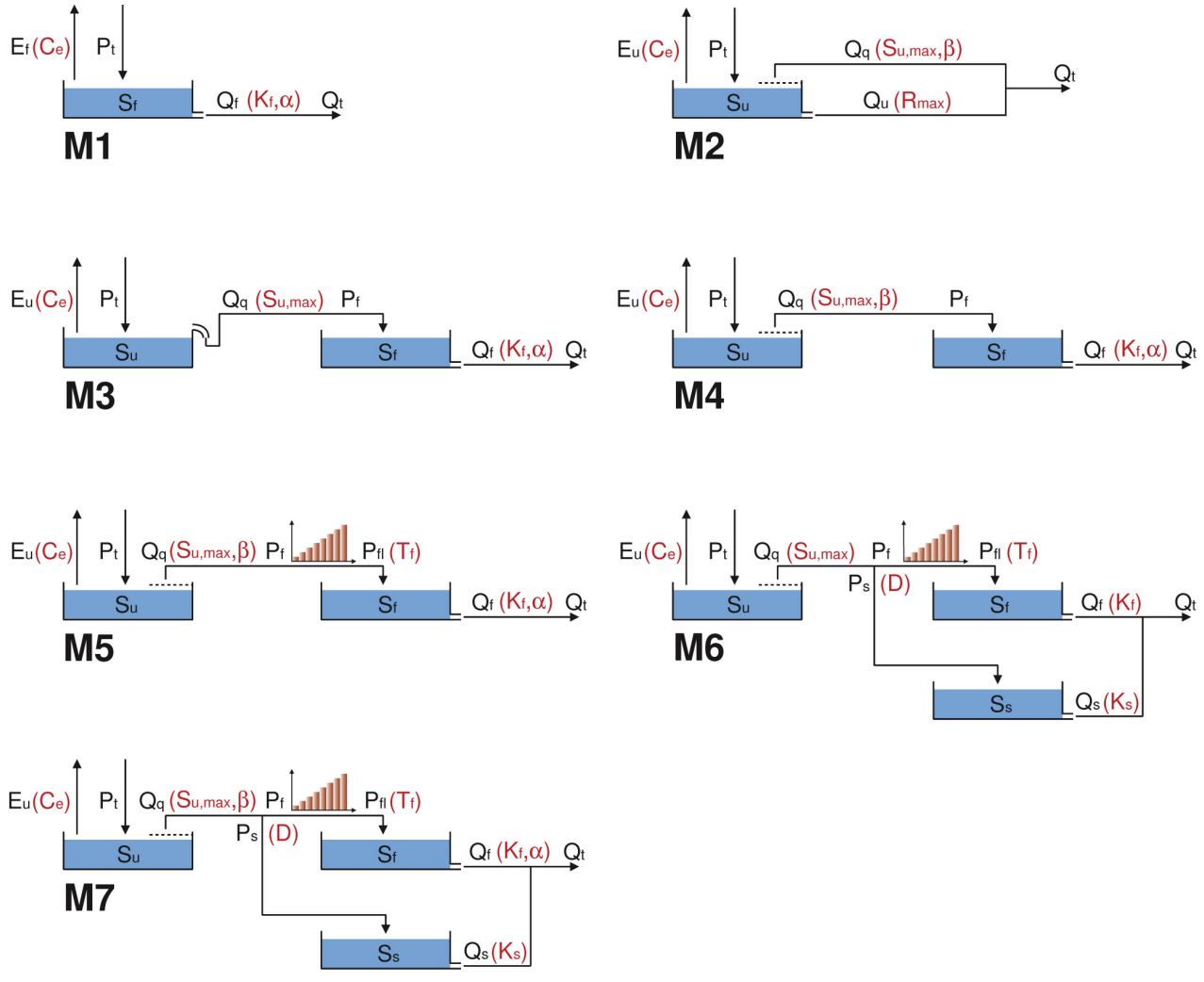

Figure 2. Schematic representations of the hydrological model structures analyzed in this study. The states and fluxes are noted in black and the associated parameters are in red.

labeled "fast" (FR), "slow" (SR), and "unsaturated" (UR). The subscripts $x = f, s, u$, which are used to index the reservoir storages (S_x) and fluxes (P_x, E_x, Q_x) refer to the fast, slow, and unsaturated reservoirs, respectively. In addition, three of the models employ a lag function (LF) to represent flow network routing delays.

3.3.2. Process-Oriented Rationale for the A Priori Model Hypotheses

[33] The selection of model hypotheses in this study is motivated by several considerations, including prior insights into the catchment characteristics (section 2.3.3), previous modeling experience in Luxembourg and the Maimai [e.g.,

Table 2. Components and Parameters of Model Structures M1–M7

Model			Components				Parameters								
	N_θ[a]	N_s[b]	UR[c]	FR[d]	SR[e]	LF[f]	C_e (−)	$S_{u,\max}$ (mm)	β (−)	R_{max} (mm h^{-1})	T_f (h)	K_f (mm$^{1-\alpha}$ h^{-1})	α (−)	D (−)	K_s (1 h^{-1})
M1	3	1	-	✓	-	-	✓	-	-	-	-	✓	✓	-	-
M2	4	1	✓	-	-	-	✓	✓	✓	✓	-	-	-	-	-
M3	4	2	✓	✓	-	-	✓	✓	-	-	-	✓	✓	-	-
M4	5	2	✓	✓	-	-	✓	✓	✓	-	-	✓	✓	-	-
M5	6	3	✓	✓	-	✓	✓	✓	✓	-	✓	✓	✓	-	-
M6	6	4	✓	✓	✓	✓	✓	✓	-	-	✓	✓	-	✓	✓
M7	8	4	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓

[a]N_θ is the number of parameters.
[b]N_s is the number of states.
[c]UR, unsaturated reservoirs.
[d]FR, fast reservoirs.
[e]SR, slow reservoirs.
[f]LF, lag function.

Fenicia et al., 2008a, 2010], and to at least partially reflect the range of complexities and configurations typically encountered in conceptual hydrological modeling [e.g., as reviewed by *Beven*, 2001; *Singh and Woolhiser*, 2002, and others]. Some of the rationale for the selected models is outlined next.

[34] Model M1 is composed of a single nonlinear reservoir (FR) with three parameters. The outflow is a power function of the storage, while the predicted evaporation is proportional to the potential rate (with a smoothing function for near-zero storage values). This model can be considered an "end member" on the simplicity scale.

[35] Model M2 resembles the core soil moisture accounting component of TOPMODEL and VIC [*Beven*, 1997; *Wood et al.*, 1992]. It is composed of a single reservoir (UR) with four parameters and two outputs: quick flow Q_q and slow flow Q_u. The evaporation E_u is assumed to be proportional to its potential value, with smoothing for near-zero storage.

[36] Model M3 characterizes flow generation as a threshold process, which is typical of small scale systems [*Spence*, 2010]. It is composed of two reservoirs (UR and FR) and has four parameters. In this model, UR is a bucket with a (smoothed) threshold (Table A2), where outflow Q_q occurs when the reservoir storage reaches the threshold.

[37] Model M4 differs from M3 solely in the outflow from UR, which in M4 and M2 is a power function of the storage (whereas in structure M3, the unsaturated soil store has a threshold). It has two reservoirs and five parameters.

[38] Model M5 differs from M4 by the addition of a lag function to the flow that connects UR and FR. This model has six parameters, two reservoirs, and one lag function.

[39] Model M6 represents a special case of a multistore model where all fluxes are linear with respect to the storages. It differs from M5 by the inclusion of an additional reservoir for the simulation of slow flow (SR, representing a groundwater component). The output from UR is linearly partitioned between FR and SR. In contrast to M5, this model is characterized by linear constitutive functions, i.e., both FR and UR are linear reservoirs. This structure is characterized by six parameters, three reservoirs, and one lag function.

[40] Model M7 also contains a groundwater reservoir, but differs from M6 in that the constitutive relationships of UR and FR are nonlinear, as in model M5. This structure has eight parameters, three reservoirs, and one lag function. Hypotheses M6 and M7, which contain several interconnected nonlinear reservoirs and a lag function, resemble more complex models such as the HBV model [*Lindström et al.*, 1997].

[41] Finally, the GR4H model is used as a representative "fixed" model structure that provided a best "average" performance in extensive empirical trials over hundreds of catchments with different physical and climatological properties [*Perrin et al.*, 2001, 2003; *Le Moine et al.*, 2007]. It is characterized by four state variables and four parameters [see *Perrin et al.*, 2003, and others].

3.4. Model Inference Methods

3.4.1. Bayesian Inference Equations and Error Models

[42] The hydrological parameters are inferred from observed rainfall-runoff data $\left(\tilde{\boldsymbol{P}}, \tilde{\boldsymbol{Q}}\right)$ using Bayes equation,

$$p\left(\boldsymbol{\theta}, \boldsymbol{\Xi} \middle| \tilde{\boldsymbol{P}}, \tilde{\boldsymbol{Q}}\right) = p\left(\tilde{\boldsymbol{Q}} \middle| \tilde{\boldsymbol{P}}, \boldsymbol{\theta}, \boldsymbol{\Xi}\right) p(\boldsymbol{\theta}, \boldsymbol{\Xi}), \tag{1}$$

where $p\left(\boldsymbol{\theta}, \boldsymbol{\Xi} \middle| \tilde{\boldsymbol{P}}, \tilde{\boldsymbol{Q}}\right)$ is the posterior distribution of the parameters $\boldsymbol{\theta}$ of the hydrological model and the parameters $\boldsymbol{\Xi}$ of the residual error model, $p\left(\tilde{\boldsymbol{Q}} \middle| \tilde{\boldsymbol{P}}, \boldsymbol{\theta}, \boldsymbol{\Xi}\right)$ is the likelihood function, and $p(\boldsymbol{\theta}, \boldsymbol{\Xi})$ 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 noninformative priors for $\boldsymbol{\theta}$ and $\boldsymbol{\Xi}$ [*Box and Tiao*, 1992].

[43] The error model is based on the weighed least squares (WLS) scheme, which assumes zero-mean Gaussian errors and allows for heteroscedasticity. Here we hypothesize that the standard deviation of individual residuals increases linearly with the corresponding simulated streamflows,

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

$$\sigma_n = a + b\hat{Q}_n, \tag{3}$$

where σ_n is the standard deviation of the residual errors at time step n, $\tilde{Q}_n$ and $\hat{Q}_n\left[\tilde{\boldsymbol{P}}, \boldsymbol{\theta}\right]$ 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, and N_t is the number of observations.

[44] The standard least squares (SLS) scheme, which can be viewed as a special case of WLS with $b = 0$, makes the assumption of constant variance of residual errors. It corresponds closely to widely used objective functions such as the Nash-Sutcliffe (NS) index and the root-mean-square error (RMSE). Despite its considerable limitations [e.g., *Schaefli and Gupta*, 2007], the Nash-Sutcliffe index remains widely used in hydrological communication. Including the SLS scheme into the analysis allows us to appraise the influence of the likelihood function and diagnostic metrics on model results and interpretation.

[45] The least squares regression in equations (2) and (3) provides only an approximate description of the heteroscedasticity of the total data and structural errors. It also ignores autocorrelation [e.g., *Sorooshian and Dracup*, 1980; *Schoups and Vrugt*, 2010, and many others], which generally leads to an underestimation of the parametric uncertainty. Hence, several posterior diagnostics are employed to appraise the predictive performance of the inferred models. Additional "process-oriented" diagnostics are also included to provide a tentative evaluation of the physical "realism" of the models. These diagnostics are detailed in section 3.5.

3.4.2. Multistart Quasi-Newton Parameter Optimization

[46] Parameter calibration was carried using a multistart quasi-Newton method (e.g., *Kavetski and Clark* [2010], see also *Moore and Clarke* [1981] and *Skahill and Doherty* [2006] for hydrological applications of the related Gauss-Newton method). Here we initiated independent local quasi-Newton searches (with trust-region safeguards) from 1000 random seeds in the feasible parameter space. The large number of initial seeds was used to thoroughly examine the parameter space; the CPU runtimes were on the order of a few hours on a single standard desktop CPU (i.e., without using any multicore parallel computation).

3.4.3. Markov Chain Monte Carlo Analysis of Posterior Parameter Distributions

[47] The posterior distributions of the model parameters were approximated using the Markov chain Monte Carlo (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 [*Gelman et al.*, 2004] and the final 10,000 samples were used to analyze and report the parameter distributions.

[48] The stationarity (convergence) of the MCMC chains was evaluated using the Gelman-Rubin statistic [*Gelman et al.*, 2004]. Given the vulnerability of common convergence diagnostics to the entrapment of MCMC chains on local optima of the target distribution [e.g., as shown by *Schoups et al.*, 2010], randomly seeded multistart quasi-Newton optimization analyses of the Bayesian posteriors were carried out to explore their macroscale multimodality structure [*Kavetski and Clark*, 2010]. Their termination points were used as starting seeds for the MCMC chains, in an attempt to further guard against false convergence.

3.5. Posterior Diagnostics

3.5.1. Statistical Diagnostic Tests of the Predictive Distribution of Streamflow

[49] The statistical reliability of the predictions, i.e., the consistency of the observed data and the predictive distribution generated by the model is examined using quantile-quantile (QQ) plots [e.g., *Gneiting et al.*, 2007; *Thyer et al.*, 2009]. In particular, the uniformity of the distribution of the quantiles of the observed data within the predictive distribution can be viewed as a measure of consistency of the model and the observed data. Departures from linearity of the QQ plot can be interpreted in specific terms, such as underestimation of predictive uncertainty, systematic underestimation of responses, etc. (see detailed Figure 3 by *Renard et al.* [2010]). Statistical reliability can also be quantified using numerical measures [e.g., see *Renard et al.*, 2010]. Here the area $\alpha(-)$ between the QQ curve and the 1:1 diagonal is reported. Larger values of α indicate lower reliability.

[50] In addition to the statistical reliability, the precision (or "sharpness") of the predictive distribution is reported. The distinction is important: two predictive distributions can both be reliable (implying the distributions of actual errors are suitably approximated by the error models), yet one can be more precise and hence more useful to a modeler [see also *Gneiting et al.*, 2007]. Here the predictive precision is quantified using the average standard deviation of the predictions (σ_a), where the averaging is over the time steps within the period of interest (e.g., calibration and/or validation). This quantity is directly related to the widely used Nash-Sutcliffe (NS) measure Φ_{NS}, which is also reported for the case of SLS calibration.

[51] Note that the case of high precision but low reliability represents a prediction that is misleading with respect to its accuracy, which is clearly undesirable. Hence, in our hypothesis-testing analyses, the predictive reliability of a calibrated model takes some priority over its predictive precision.

3.5.2. Process-Oriented Model Diagnostic Tests

[52] The model's ability to reproduce characteristic "signatures" of the catchment response is of clear importance from a hydrological perspective and should be appraised as part of posterior diagnostics [*Gupta et al.*, 2008].

[53] The ability of the models to approximate the observed flow duration (FD) curves [e.g., see *Yilmaz et al.*, 2008] is reported. This is important because flow duration curves tend to reflect climatological and geomorphological catchment attributes [e.g., *Linsley et al.*, 1949; *Wagener and Wheater*, 2006].

[54] In addition to the FD curve criteria, which are based on the same data as used in the calibration, available data on groundwater levels in the Maimai and Pfaffenthal basins are exploited as an independent data source. The use of groundwater data for the evaluation of a conceptual hydrological model has proven useful in several previous studies. These include investigating model inadequacies [*Seibert*, 2000], improving model realism [*Seibert and McDonnell*, 2002; *Son and Sivapalan*, 2007], reducing equifinality in the parameter space [*Freer et al.*, 2004], and supporting model improvement [*Fenicia et al.*, 2008a].

[55] In this study, because of commensurability limitations (section 4.3), the piezometer data is used as a largely qualitative check of the overall trends in the internal model dynamics. By not calibrating the model to these time series, the latter can be used to provide an independent assessment of the fidelity of model components to the physical processes they are intended to represent. This then contributes to the comparative evaluation of the physical realism of the different model structures [*Seibert and McDonnell*, 2002].

4. Results of Empirical Case Studies

[56] Section 4 presents the modeling results, with a particular emphasis on comparing model performances in the various catchments using a range of diagnostic metrics, and on interpreting the differences in model performance based on known differences in catchment characteristics.

[57] We report the results for the WLS inference, unless indicated otherwise. The inclusion of SLS results allows the appraising of the sensitivity of the results to the likelihood function. In our opinion, although consistent behavior across different likelihood functions is not a strictly necessary criterion for hydrological hypothesis testing (because it is unreasonable to expect even a "good" hypothesis to perform well when calibrated under poor error model assumptions), it can provide some additional indication of whether the inference is robust.

4.1. Comparison of Model Structures

[58] Figure 3 compares numerical measures of reliability, precision, and accuracy of all model structures on the four catchments. The hydrographs and the associated 95% prediction limits are shown in Figure 4 for selected model structures, and the corresponding predictive QQ plots are shown in Figure 5.

[59] In the Maimai catchment, the single-reservoir configuration M1 is competitive with, and indeed often superior to, the more complex models. Although its average precision is somewhat lower than more complex models, this model ranks best with respect to the critical metric of predictive reliability in the validation period (Figures 3 and 5). The

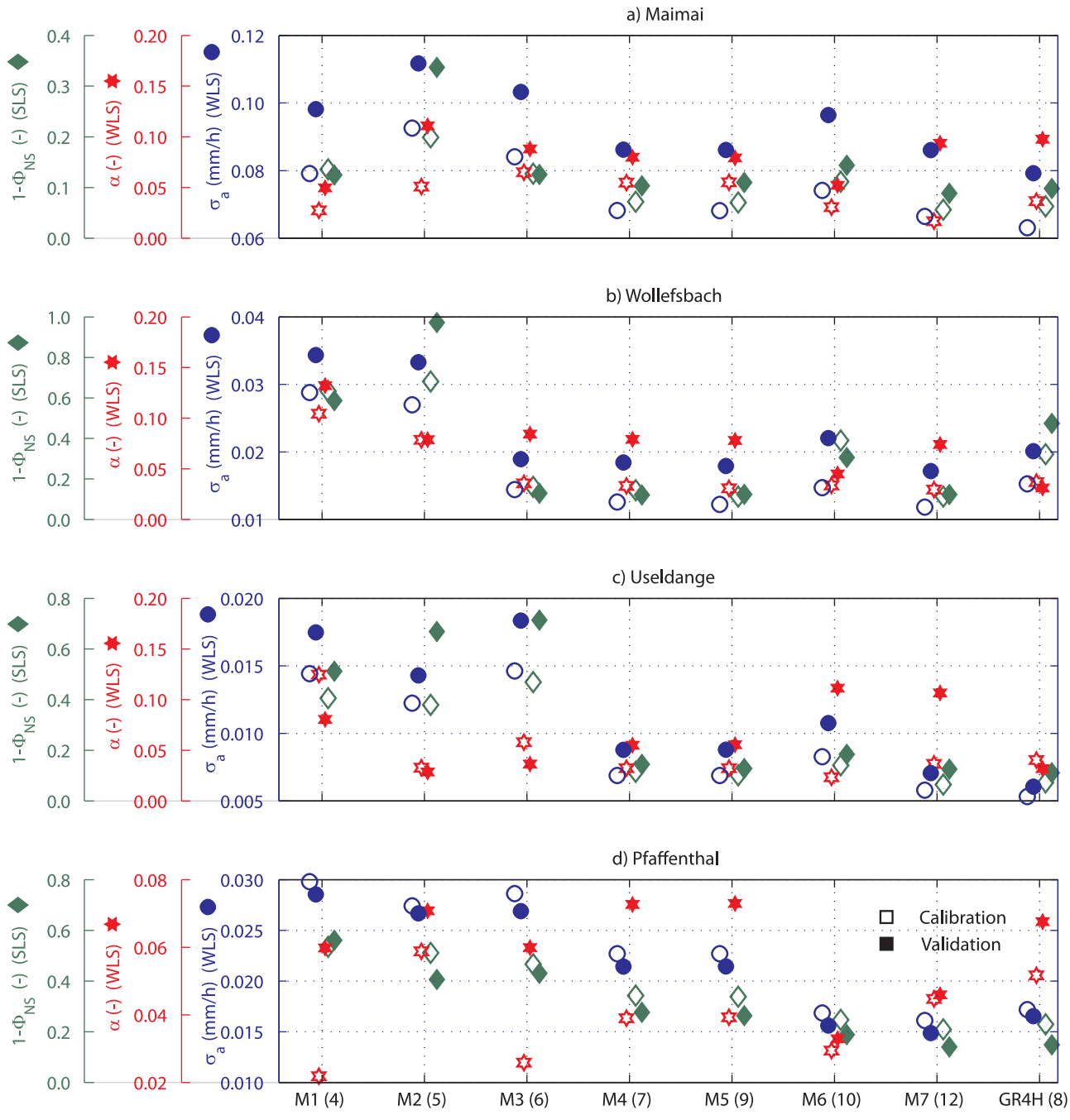

Figure 3. Statistical metrics used to evaluate the predictive distributions of streamflow against the observed data, applied separately to the calibration and validation periods. The x-axis lists the model structures, with an approximate measure of complexity ($N_\theta + N_s$, see Table 2) indicated in parentheses. The statistical reliability is quantified by the area between the predictive QQ curves and the 1:1 diagonal (α, see section 3.5.1), while the predictive precision is quantified using the average standard deviation of the predictions (σ_a). The metrics were computed for the predictive distributions estimated using a WLS regression. In addition, the Nash-Sutcliffe measure Φ_{NS} obtained using SLS is reported. While in some cases (e.g., the Wollefsbach) the metrics were broadly consistent with one another, in other cases (e.g., the Pfaffenthal) considerable tradeoffs arise.

hydrograph comparison of M1 and M7 shows that M1 is better at capturing the peaks, while M7 tends to underestimate them, in favor of a better representation of low flows (Figure 4). With respect to the NS index, calculated for the SLS calibration, the performance of M1 is comparable to more complex models, with a value above 0.88 in the validation period (Figure 3). The ability to represent the Maimai catchment dynamics by a single nonlinear reservoir corroborates that the behavior of this catchment is relatively simple (section 3.2.1).

[60] The Wollefsbach catchment, which is larger in size, though still relatively homogeneous in terms of its geology and land cover, displays a more complex hydrological behavior than the Maimai catchment. M1 has a relatively poor performance on this catchment, while M3, which contains a threshold-type saturated area function, appears to be the simplest configuration that captures the essential response dynamics. The average precision (WLS) and the NS efficiency (SLS) are comparable to other, more complex models (Figure 3). For example, M3 has an NS index of 0.87 in the validation period, exceeding the value of 0.53 obtained using the more complex GR4H. In terms of hydrograph behavior, Figure 4 shows that, despite being more complex (Figure 3), GR4H is unable to capture the seasonal behavior of this catchment, underestimating the flow during winter and overestimating it during summer.

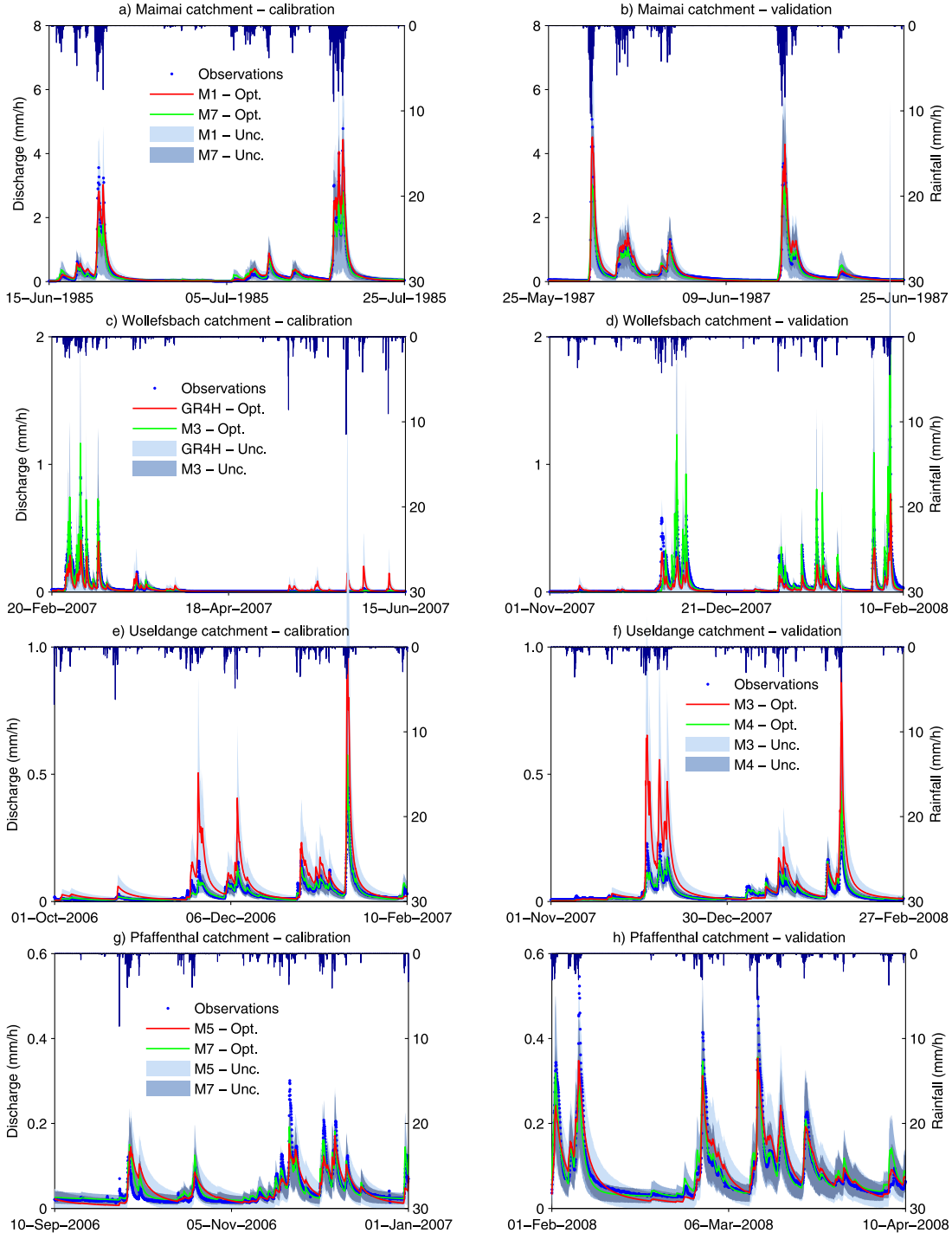

Figure 4. Model predictions and associated uncertainty bounds for selected model structures and representative portions of the calibration and validation periods. Lines denote near-optimal predictions, while bands indicate 95% prediction limits.

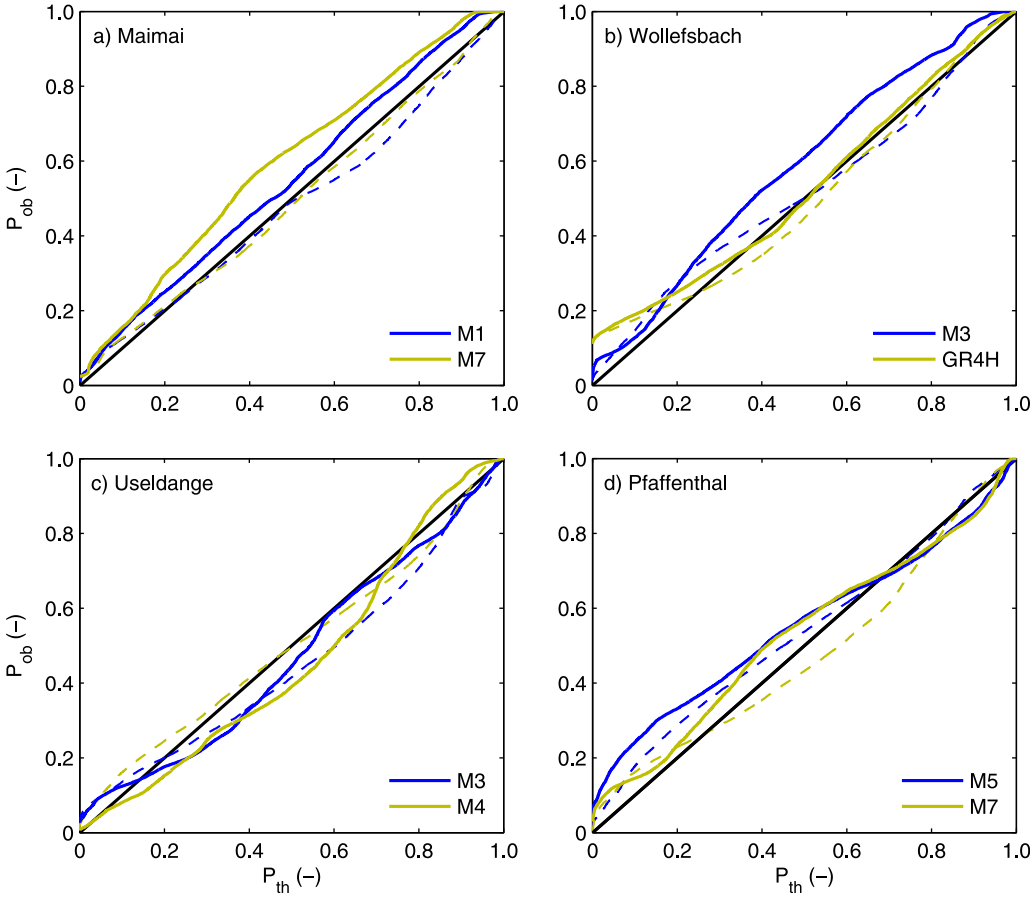

Figure 5. Quantile-quantile plots for appraising the statistical reliability of the predictive distributions of streamflow for selected models, in calibration (dashed lines) and validation (solid lines). Departures from the 1:1 diagonal pattern represent a measure of the discrepancy between the predictive distribution and the observed data. For example, when validation and calibration performances were compared in the Maimai basin, model M7 suffered a notably larger loss of predictive reliability than model M1.

As will be discussed in section 4.4.2, capturing this seasonality requires particular features in the model structure, such as the stronger thresholds present in hypothesis M3.

[61] It is worth noting that the QQ plots in Figure 5 indicate that predictive *reliability* of GR4H is high despite its relatively poor accuracy. The predictive reliability, here defined as the consistency of the predictive distribution and the observed data, is a property not just of the (deterministic) model, but of the error models. Although the error of the calibrated GR4H model is relatively large, it appears well described by the heteroscedastic error model in equation (3). Hence, while less precise, its predictive distributions appear statistically trustworthy.

[62] The Useldange catchment contains the Wollefsbach as one of its headwater subcatchments. Here M1–M3 have a relatively low performance, while M4 performs notably better (Figure 3). Hypotheses M4 and M3 differ solely in the form of the saturated area function, which in M4 allows for smoother saturation dynamics. Figure 4 contrasts the hydrographs predicted by models M3 and M4 for this catchment. M3 tends to overpredict the flows, while M4 has a more balanced performance. Figure 5 also suggest that M4 is somewhat more reliable for this catchment (e.g., in the center of the plot, the M4 QQ line deviates notably from the expected diagonal pattern). The GR4H performs relatively well and is even superior to other models of similar complexity.

[63] The Pfaffenthal catchment is the largest catchment in this study (385 km^2) and, unlike the previous catchments, its geological formation comprises a large groundwater reservoir. Models M1–M5 fail to adequately represent the behavior of this catchment. In particular, the shapes of their recessions were qualitatively different from the observed recessions. Conversely, hypotheses M6 and M7, which include a groundwater reservoir, perform considerably better. From these two models, M7, which includes nonlinearities, performs better than M6 in terms of average precision and in terms of the Nash-Sutcliffe index Φ_{NS}, while M6 is slightly better in terms of the reliability index α (Figure 3). The suitable performance of the linear multistore model M6 for the Pfaffenthal can be contrasted to its behavior on the other catchments, where this model ranked as one of the worst with respect to several criteria. Figure 4 compares the hydrographs of M5 and M7. While M5 misses hydrograph recessions, M7 is able to capture the recession signature quite well. GR4H performs similarly to M6 and M7 with respect to Φ_{NS}, however, it has poorer statistical reliability.

[64] The performance of the GR4H model, both in absolute terms and relative to competing models, was clearly catchment-specific. In the Maimai and Useldange basins, GR4H ranks among the best models with respect to most of the performance criteria analyzed. Yet, for example, in the Wollefsbach its performance is markedly poor: it achieves a NS value of about 0.5, whereas the simpler M3 structure, which contains a threshold runoff mechanism, reaches NS values of about 0.9 and has a clearly superior precision. In the Pfaffenthal catchment, GR4H has a low statistical reliability, with models M6 and M7 performing better overall.

4.2. Simulation of Catchment Signatures: Flow Duration Curves

[65] Figure 6 shows the flow duration curves, including the uncertainty limits of the predicted curves, for all catchments and selected model structures in the validation period. Figure 6a shows the FD curves for the Maimai catchment, contrasting the simplest (M1) versus the most complex (M7) hypotheses. It can be seen that structure M7 (which includes a groundwater component) has a higher uncertainty in the low flows, while structure M1 (which lacks a groundwater store) has larger uncertainty in the high flow. Hence, despite its additional complexity, M7 does not appear to improve performance across the full range of responses.

[66] Figure 6b compares the GR4H and M3 models applied to the Wollefsbach catchment. The limitations of the GR4H model for this catchment are apparent, with a simulated flow duration curve that is significantly different from the observed, and a large uncertainty, especially in the low flows. Conversely, the simpler hypothesis M3 provides a notably better and tighter approximation of the range of responses of this catchment.

[67] Figure 6c compares the flow duration curves of models M3 and M4 in the Useldange catchment. Here M3 has significantly larger uncertainty and is in poorer agreement with the observation than M4.

[68] Finally, Figure 6d shows the effect of the addition of a groundwater reservoir for the modeling of the Pfaffenthal catchment (which is known to have a considerable groundwater store). Model M7 performs better than M5 on low flows, while the two models have a more or less similar performance on high flows.

4.3. Simulation of Internal Catchment Function: Groundwater Levels

[69] Figure 7 compares the evolution of the internal states of model hypotheses M1 and M7 to the observed water level dynamics (measured using the piezometers, see section 3.2.4). The top row indicates the observations, where the measurement locations have been ordered from upstream to downstream. Although there are substantial differences between individual piezometer readings, their overall behavior does follow a common pattern in response to the recharge arising from the rainfall input [*Seibert and McDonnell*, 2002; *Freer et al.*, 2004]. The second and third rows of Figure 7 show the corresponding time series of storages predicted using model structures M1 (which appears to work reasonably for Maimai and poorly for Pfaffenthal) versus the storages predicted using M7 (which works well for Pfaffenthal, but may be overly complex for Maimai). The second rows show inferences using SLS, and the third row shows the inferences using WLS.

[70] The single reservoir hypothesis M1 appears to capture reasonably well the qualitative groundwater dynamics of the Maimai catchment. Storage dynamics are similar for SLS and WLS. On the Pfaffenthal catchment, the storage of M1 appears to react fast when SLS is used, mimicking the fast reaction of the catchment, while it has a more damped behavior when WLS is used. This is likely because SLS favors the fitting of high flows, while WLS, by approximating the heteroscedasticity of the errors, can also give considerable weight to low flows.

[71] Conversely, the more complex model M7 performs poorly on Maimai when SLS is used. The storage S_s, which in hypothesis M7 is intended to represent a groundwater store, exhibits a continuously increasing trend that is in clear disagreement with the piezometer measurements. Yet when WLS is used, the storages are in a markedly better agreement with the observations. In the Pfaffenthal, the storage dynamics of M7 are similar for both inference schemes, with storages S_u and S_f appearing to follow some of the observed piezograph patterns.

[72] These findings can be used to scrutinize the physical realism of the model hypotheses, subject to important limitations with respect to the scale commensurability of the data and the model. In particular, a piezometer measurement time series can be used to explore the degree to which the internal model states approximated the observed groundwater dynamics, and to detect cases of unreasonable behavior. However, the insights from this comparison are limited by the fact that model states within a lumped model are, at best, "effective" average representations over an entire catchment, while piezometer observations apply over a much smaller (essentially, "point") scale at specific locations. Hence, the "scale representativeness" of the piezometer time series can be called into question [e.g., *Wagener and Gupta*, 2005; *Freer et al.*, 2004; *Liu and Gupta*, 2007].

[73] In our (somewhat pragmatic) opinion, the scale mismatch between groundwater piezometer measurements and a state variable in a catchment-scale model does not negate the utility of inspecting trends and patterns in this data to scrutinize the internal performance of a lumped model hypothesis. Indeed, observed groundwater levels have been very useful in appraising model "realism" and in discriminating between competing lumped model structures in a series of previous studies [e.g., *Seibert*, 2000; *Seibert and McDonnell*, 2002; *Freer et al.*, 2004; *Fenicia et al.*, 2008a; *Son and Sivapalan*, 2007]. Here they were interpreted as supporting the perception, that hypothesis M7 was poorly identifiable from rainfall-runoff data alone for the Maimai catchment, but that it appeared physically plausible for the Pfaffenthal basin. Conversely, comparison of the groundwater time series patterns suggested that hypothesis M1 is too simplistic for the Pfaffental but, within the limits of the data used in this study (see section 4.4.1), appears adequate for the Maimai. These findings highlight the benefits of independent complementary sources of information to assess model performance [e.g., *Seibert and McDonnell*, 2002; *Fenicia et al.*, 2008a, and others]. They also highlight the poor discriminative power of any single performance indicator, e.g., the Nash-Sutcliffe error measure applied solely

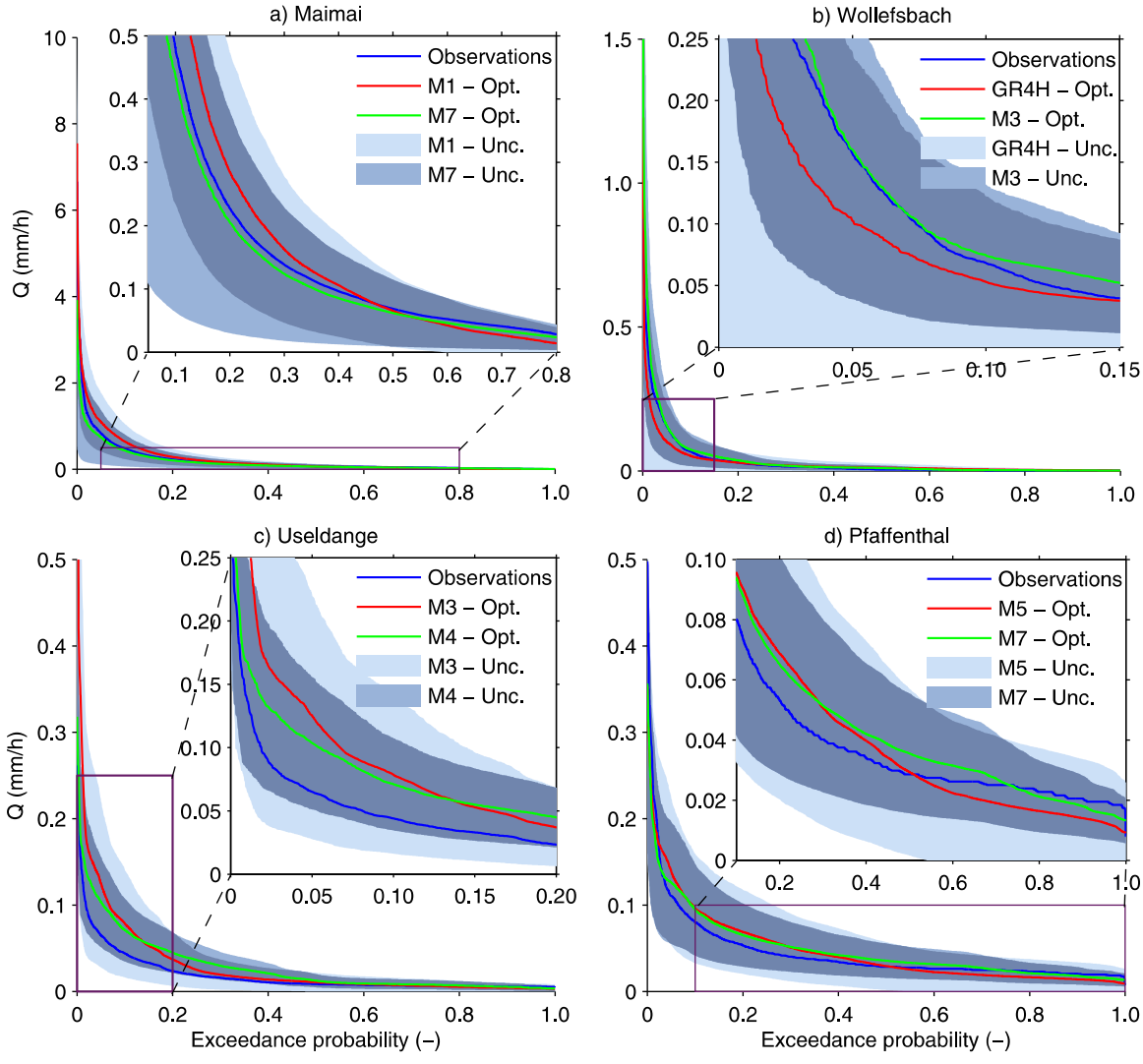

Figure 6. Flow duration curves for selected model structures in all catchments over the validation period. The 95% uncertainty limits are shown. Overall, the best model structure for reproducing flow duration characteristics appears to be catchment-specific. For example, GR4H struggles to capture the streamflow distribution in the Wollefsbach, and the uncertainty limits are particularly wide. Similarly, model M3 does not adequately reproduce the flow duration signature of the Useldange catchment, while M5 does poorly in the Pfaffenthal.

to the streamflow time series [e.g., *Schaefli and Gupta*, 2007; *Gupta et al.*, 2008].

4.4. Relations Between Catchment Attributes and Model Structures

[74] Section 4.4 attempts to provide a physically oriented interpretation of the model hypotheses. Such analysis can improve our understanding of the correspondence between catchment properties and model structure, which is an important, yet poorly understood aspect of contemporary conceptual hydrology [*McDonnell and Woods*, 2004]. Although it is currently difficult to a priori translate experimental findings into a perceptual model of catchment behavior, such insights can help provide a posteriori plausible explanations and interpretations of differences in the performance of different models. The main insights, and associated caveats, are as follows.

4.4.1. Maimai Catchment

[75] Despite the perceived functional simplicity of the Maimai, previous models used in this catchment differed substantially in terms of structure, complexity, and level of spatial discretization. For example, *Vaché and McDonnell* [2006] developed a distributed grid-based model, *Freer et al.* [2004] applied TOPMODEL, *Beven* [1997] and *Seibert and McDonnell* [2002] developed a three-reservoir model, and *Fenicia et al.* [2010] used a single reservoir model. In our opinion, such a broad range of models highlights that limited guidance is available for model-building even in experimental basins (see also the critique by *McDonnell and Woods* [2004]).

[76] In this study, the hydrograph dynamics of the Maimai catchment were adequately captured by a simple single-reservoir nonlinear model [see also *Fenicia et al.*, 2010]. More complex models have a slightly better performance,

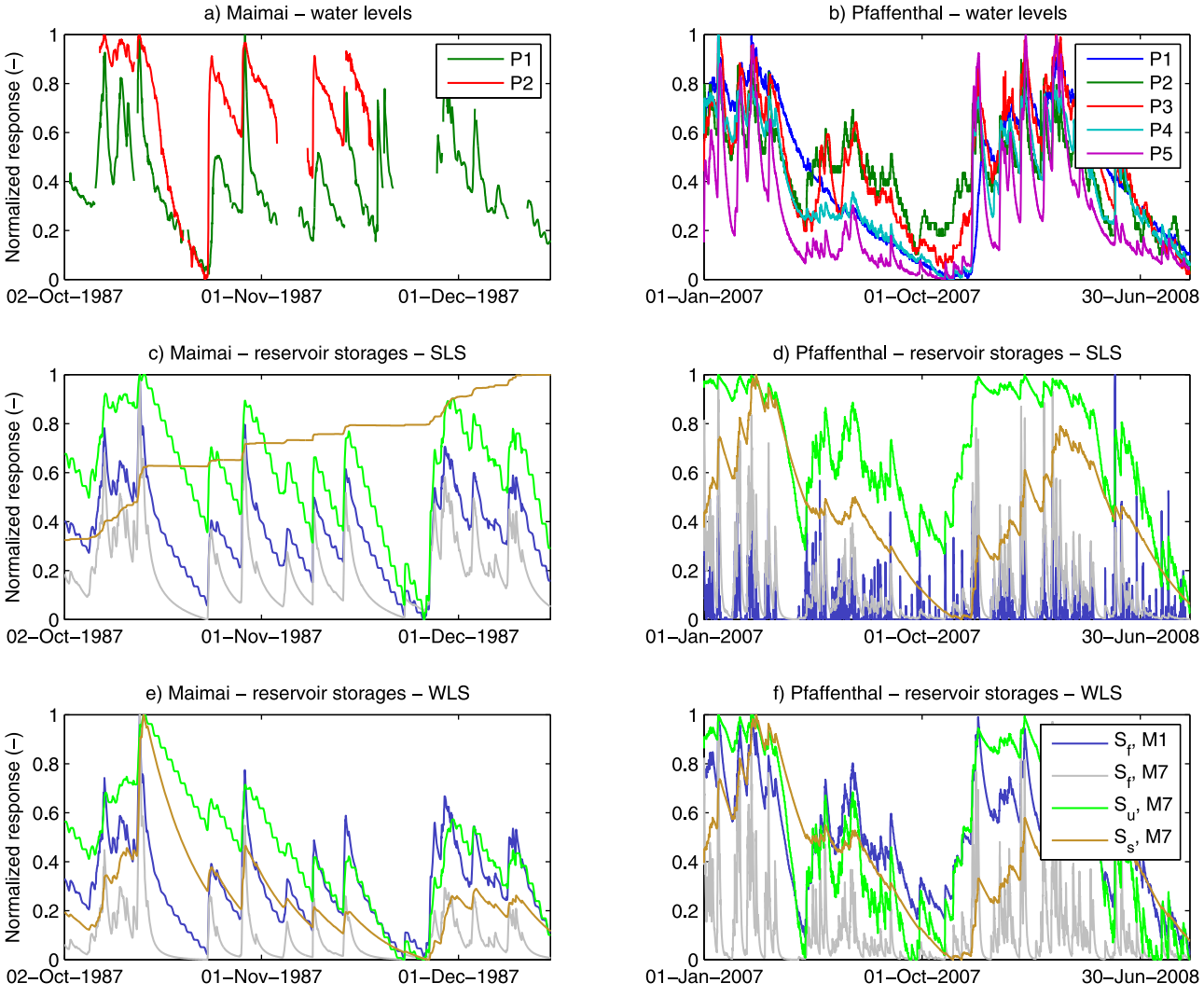

Figure 7. Correspondence between groundwater piezometer levels in the Maimai and Pfaffenthal catchments with the storages in models M1 and M7. The results of calibrating solely to runoff using SLS versus WLS are contrasted. When model M7 is calibrated to the Maimai catchment using SLS, its internal states, intended to represent groundwater storage, exhibit a continuous rising trend, casting doubt on its physical realism.

yet their additional complexity may result in overfitting when calibrating solely to rainfall-runoff data. For example, inspection of internal model states and their comparison with groundwater levels suggests that model M7 is overfitted and provides a poor approximation of internal dynamics when SLS calibration is used (Figure 7). Under this model estimation scenario, the inference of hypothesis M7 is ill-posed: It may appear to work well for some performance indicators, such as the Nash-Sutcliffe index, but "it works for the wrong reasons" [*Kirchner*, 2006]. This is suggested by the deficiencies in reproducing the internal dynamics of the system, and, tellingly, by the considerable loss of predictive reliability in the validation period (see the QQ plots in Figure 5).

[77] On the other hand, the internal dynamics of hypothesis M1 appear in much better agreement with the groundwater levels, its (single) storage variable followed a similar overall pattern of behavior as the piezometer heads (despite not being calibrated to them). However, further analysis would be needed before hypothesis M1 can be accepted as a physically realistic description of the Maimai catchment. In particular, it is markedly simpler than the perceptual models suggested by *McGlynn et al.* [2002]. This may be explained by the (relative) simplicity of the model development methodology used in this study. For example, the groundwater piezometer data were used solely for posterior diagnostics: it was exploited neither to guide model development or improvement nor to constrain model calibration. Other experimental insights were also used only a posteriori in the interpretation of model results, rather than throughout model development. A more comprehensive strategy would exploit different sources of information, including groundwater and tracer data, hydrogeological insights, etc., within an iterative process of model development, which could help elucidate

additional system complexity and inform more physically realistic model hypotheses [e.g., *Seibert and McDonnell*, 2002; *Uhlenbrook et al.*, 2004; *Vaché and McDonnell*, 2006; *Fenicia et al.*, 2008a; *Birkel et al.*, 2010, and others].

4.4.2. Wollefsbach Catchment

[78] The Wollefsbach catchment displays a more complex hydrological behavior than the Maimai catchment. Since it still relatively homogeneous in terms of its geology and land cover, a key difference from the Maimai is arguably in the climatic forcing. While the Maimai catchment is permanently humid and wet, the Wollefsbach catchment is characterized by a clear switch between a "dry" summer regime versus a "wet" winter regime. This behavior can be interpreted if, following the fieldwork based perceptions of *van den Bos et al.* [2006a], we view the topsoil as the main storage reservoir of the catchment, and represent it using the unsaturated reservoir (UR). When the catchment is dry, its soil profile can store water, resulting in little or no streamflow response to rainfall events. When the soil profile saturates, excess water cannot be accommodated and saturation overland flow occurs, resulting in a strong and rapid response to rainfall events.

[79] These (relatively simple) process perceptions suggest that the seasonal switch in behavior could be captured using a threshold in the storage-discharge function of the unsaturated store (hypothesis M3). Under this hypothesis, high evaporation during summer keeps the storage below the runoff-generating threshold, while the lower evaporation during winter results in a saturation of the store beyond the threshold and determines a pronounced streamflow response. The transition between "dry" (nonresponsive) and "wet" (responsive) must be reasonably sharp to reflect the distinct seasonal switch in the streamflow dynamics of this basin.

[80] These considerations may provide an intuitively plausible interpretation of why model hypothesis M3, which is characterized by a threshold-like saturated area (overflow) function in the unsaturated zone store, is the simplest configuration that appears able to capture the seasonality of the Wollefsbach catchment. Interestingly, the more complex GR4H model is not able to capture the seasonal switch, underestimating the flow during winter and overestimating it during summer. This is probably because of the shape of the response function in the unsaturated soil ("production") store of the GR4H model, which is fixed to be quadratic [*Edijatno and Michel*, 1989; *Perrin et al.*, 2003]. Also note that the more complex models M4 and M7, which do not fix the shape of the $A(S|\boldsymbol{\theta})$ function and include M3 as a special case (Table A2), do not perform much better than structure M3. Hence, in terms of reproducing rainfall-streamflow dynamics, their additional complexity appears unwarranted.

[81] Finally, with respect to interpreting catchment function in terms of traits and patterns [e.g., *Sivapalan*, 2005; *McDonnell et al.*, 2007], we may question if the threshold behavior would be apparent in the Wollefsbach if its climatic conditions were closer to the Maimai. Similarly, we could speculate that even a small and "simple" catchment such as the Maimai could produce a more complex response under different climatic settings.

4.4.3. Useldange Catchment

[82] The Useldange catchment is larger and more heterogeneous than the Wollefsbach. Indeed, it contains the Wollefsbach as one of its headwater subcatchments. As seen in Figures 3 and 4, model M3, which contains a threshold-type saturated area function, yields a poor performance for this catchment, while model M4, which allows for a smoother behavior, performs substantially better.

[83] The higher performance of M4 relative to M3 could be interpreted as suggesting that the threshold-like soil saturation dynamics at the small scale (Wollefsbach catchment, 4.6 km^2) becomes increasingly smoother at larger spatial scales (Useldange catchment, 250 km^2). Such behavior is perhaps unsurprising: even processes with sharp threshold-type behavior at the local scale, when aggregated over larger spatial (and temporal) areas, produce an increasingly smooth overall system response (e.g., see the classic demonstration by *Moore* [1985], where local threshold dynamics give rise to a power law behavior in the overall reservoir). Indeed, while experimental work has often detected threshold behavior at the plot and hillslope scale, threshold behavior at the catchment scale appears more rare [*Spence*, 2010].

4.4.4. Pfaffenthal Catchment

[84] The Pfaffenthal catchment is the largest (385 km^2) and most complex catchment in this study (section 3.2.2). Here the simpler models M1–M5, despite spanning a wide range of structures and complexities, perform markedly worse than model M7, in particular, with respect to representing the flow recessions. This can be reconciled with independent experimental knowledge, which has highlighted the important contribution of the sandstone formation to groundwater flow [*van den Bos et al.*, 2006a]. In particular, the inability to adequately represent the behavior of this catchment using models M1–M5 could be attributed to their lack of a groundwater component. This proposition is supported by inspecting Figure 4, which shows that the recession shapes simulated using the M5 model were qualitatively different from the shapes of the observed recessions.

[85] Adding a model component to represent the groundwater store makes an immediate difference: model hypothesis M6, although including only linear reservoirs, outperforms the (nonlinear) models M1–M5, which lack a groundwater component. Model hypothesis M7, which includes nonlinearities in the unsaturated zone and the fast routing reservoir, further improves the model predictions. As can be seen in Figure 4, while model M5 misses hydrograph recessions, model M7 captures the recession signature quite well.

[86] The physically oriented interpretation is further supported by the finding that the storage dynamics of hypothesis M7 are in reasonable agreement with the groundwater level observations within the Pfaffenthal catchment (Figure 7). This lends further support to the physical realism of hypothesis M7 for the Pfaffenthal Basin. This contrasts with the case of the Maimai catchment, where the internal dynamics of this same model were in clear disagreement with the observed dynamics of the processes they are designed to represent.

5. General Discussion

5.1. Hypothesis Testing: Fixed Versus Flexible Model Structures

[87] The variability in the performance of individual model structures in different catchments illustrates the difficulties in developing generic model structures (hypotheses) that are valid across different hydrological regimes.

For example, as a result of extensive testing in a large number of catchments, the GR4H model was able to provide a good compromise performance in the Maimai, Useldange, and Pfaffenthal catchments, which are generally, wet catchments with little seasonality or threshold behavior. However, its ability to approximate different hydrological regimes is poorer, both in absolute terms and relative to the alternative SUPERFLEX-derived model hypotheses. For example, GR4H was unable to capture the seasonal dynamics and the switch in hydrograph response from wet to dry conditions occurring in the Wollefsbach catchment.

[88] These findings are representative of difficulties typically encountered when attempting to cross-validate a fixed model structure on different catchments. For example, as part of a regionalization study, *Merz and Blöschl* [2004] calibrated the popular HBV model to 308 catchments in Austria. For some of the catchments, the calibrated performance was very poor (e.g., with Nash-Sutcliffe efficiencies below 0.5, volume errors exceeding 25%, etc.), which *Merz and Blöschl* [2004] attributed to data and/or structural error problems. Given analogous findings in our (much smaller) case study applying another fixed model, GR4H, to four diverse catchments, may be indicative that "pursuing model transposability in space" [e.g., as advocated by *Klemes*, 1986; *Andréassian et al.*, 2009] as a *condition* for evaluating "model adequacy" and testing model hypotheses may be an unduly stringent, or perhaps an even unattainable, objective. We stress that we do not imply that hydrologically similar catchments cannot be modeled using similar (or identical) model structures, but rather that catchments with wide differences in climatology, hydrogeology, stream network geometry, vegetation, and land use, etc., may require different model structures, especially when modeled at lumped spatio-temporal scales.

[89] Not only the mechanistic behavior of different catchments may require different structural representations, but "appropriate" model complexity also appears to be quite catchment-specific. This held despite all models being evaluated using data of the same type, length, and resolution. In particular, the behavior of the Maimai catchment appeared remarkably simple, and was effectively captured using a single-reservoir model. At the other extreme, the Pfaffenthal catchment exhibited a much more complex behavior, and its simulation required the inclusion of multiple states, including a representation of a groundwater store. A fixed structure with predefined complexity may provide an adequate compromise for some, perhaps many, catchments, but it will generally be either too complex, or too simple, for many other applications. This is especially undesirable in studies that attempt to physically interpret model complexity and its internal parameters and dynamics [e.g., *Gupta et al.*, 2008; *Schoups et al.*, 2008].

[90] In addition, the application of multiple models allows an assessment of model performance in relative terms, which facilitates the evaluation of the relative merits and limitations of individual model structures. For example, *Perrin et al.* [2001], when trialing multiple competing model structures as part of seeking a single "best" fixed model, relate the performance of each candidate to the best-performing model structure. Such comparison can provide more hydrologically insightful benchmarks than metrics such as the Nash-Sutcliffe index [*Schaefli and Gupta*, 2007]. See also the discussion by *Savenije* [2009], advocating a search for a "better" model in lieu of a "good" model. However, we stress that, for a comparison to be meaningful and to provide maximum guidance for subsequent model improvement, the competing models must differ from each other in a controlled way. Otherwise, differences in performance cannot be reliably attributed to specific hypotheses, and can be masked by interactions of model components, unaccounted differences in overall model philosophy, numerical implementation, etc. [*Clark et al.*, 2011a]. A well-designed flexible framework addresses this, by giving the modeler fine-grain control over individual model components as well as over the overall model architecture.

[91] We also stress that the pursuit of more rigorous hypothesis testing using flexible model frameworks does not imply disregarding the considerable insights embodied in many existing models. Models such as GR4J, TOPMODEL, HBV, and others have themselves emerged after a process of extensive refinement and adjustment, even if this process is not always fully detailed in the published literature. Such models can inspire further model development, and can serve both as initial points for further work and as widely accepted benchmarks in model comparison studies. For example, the M1–M7 models used here have been developed to be generally representative of operational forecasting models [*Berthet et al.*, 2009] and of the conceptual models used in major hydrological investigations [e.g., *Duan et al.*, 2006]. In addition, we have used GR4H, which represents a "best-compromise" model identified in previous studies, as a benchmark for comparison.

[92] We also note that the flexibility of a modular modeling framework such as SUPERFLEX, if applied without adequate scrutiny, can result in unwarranted model "customization" that is not supported by empirical data or other evidence. In extreme scenarios, this could lead to new models being developed indiscriminately for every catchment, failing to recognize the commonality of hydrological function [*Wagener et al.*, 2007]. Hence, the use of multiple diagnostics, both of the overall model predictions and also of its internal components, is an essential aspect of hydrological model development, as eloquently noted by *Gupta et al.* [2008] and others. These principles hold equally strongly for flexible models as they do for any particular fixed model structure.

5.2. Toward More "Realistic" Models: Flexible Frameworks as a "Language" for Dialogue Between Modeler and Experimentalist

[93] The realism of conceptual hydrological models, if defined in terms of the criteria in section 1, is far from achieved in current practice. It is not even clear how to evaluate the "realism" of a lumped model where "soil" is represented by a single reservoir [e.g., see discussions by *Wagener*, 2003; *Liu and Gupta*, 2007; *Beven*, 2008, and others]. Yet, such limitations notwithstanding, in our opinion, there remains much room for improvement through a better dialogue between the modeler and experimentalist [*Seibert and McDonnell*, 2002]. Such exchanges can occur: (1) a priori, during the development of a perceptual model based on an initial understanding of processes [e.g., *Seibert and McDonnell*, 2002]; (2) a posteriori, during the scrutiny of model structures using a range of diagnostics and

independent data [e.g., *Gupta et al.*, 2008]; or, preferably, (3) continuously across several or all stages of model development [e.g., *Fenicia et al.*, 2008a]. In all cases, the fruitfulness of the dialogue depends on the ability to meaningfully communicate, robustly implement, and rigorously compare alternative hypotheses [*Fenicia et al.*, 2008a]. These may involve both major and minor (often subtle) changes in the type and number of model elements, their connectivity, and constitutive functions.

[94] We argue that a flexible framework can facilitate the dialogue between the modeler and experimentalist [*Seibert and McDonnell*, 2002] by providing (1) a more precise language for exchanging perceptual/conceptual information (e.g., see section 3.3, which illustrates some of the reasoning behind the construction of the model hypotheses used in this study), and (2) a robust platform to systematically generate, implement, and compare model hypotheses. By quantitatively accommodating the experimentalist's perceptual understanding [e.g., *Vaché and McDonnell*, 2006], importantly, using robust mathematical techniques [e.g., see *Clark and Kavetski*, 2010], it can facilitate model development and improvement, both in research and operation.

[95] While it is beyond our scope to provide detailed "how-to" recipes for using the flexible modeling approaches for different applications, we note that flexible frameworks have already facilitated more in-depth investigations of catchment behavior. For example, *Fenicia et al.* [2008b] illustrated an iterative approach of model development, where different types of data are progressively introduced to constrain the hypothesis space, and the model is updated as part of a modeler-experimentalist dialogue [e.g., *Seibert and McDonnell*, 2002]. Similarly, *Clark et al.* [2011b] detail the application of a range of diagnostics within the FUSE framework to test competing hypotheses using field data and improve process representation in an experimental catchment.

6. Conclusions

[96] The two-part paper proposed and illustrated a flexible model framework for conceptual hydrological modeling at the catchment scale. Within this framework, model structures are hypothesized and constructed using generic components such as reservoirs and lag functions, assembled (connected) into a coupled-system model using junctions and fluxes, and parameterized using constitutive functions relating internal states and fluxes.

[97] The potential of the flexible framework is explored in a four-catchment study, where seven a priori selected model hypotheses of varying structure and complexity are contrasted with the fixed-GR4H structure representative of a "best-compromise" model identified in previous studies. The key conclusions of the study are as follows:

[98] (1) Catchments with distinct hydrological dynamics appear best characterized using distinctly different lumped model structures. A fixed model structure struggles to accommodate the wide range of possible behavior encountered even in just the four catchments considered in this study. For example, the GR4H model is unable to satisfactorily reproduce the seasonal signature of the Wollefsbach catchment (arguably because of the absence of thresholds in its runoff production function). An alternative four-parameter hypothesis of the catchment function, including a threshold-like production store, notably improves the model performance. Importantly, the threshold-based model hypothesis is in better qualitative agreement with independent experimental insights available in this catchment, including the shallowness of the soil store and flashy response of the schistose geology. Similar findings held in the Pfaffenthal catchment, where a groundwater store component is needed to accommodate the sandstone aquifer dynamics.

[99] (2) A flexible structure can be exploited to incorporate independent experimental understanding, such as the presence or absence of groundwater storage, threshold dynamics, etc. It can also readily accommodate nonstandard storage-recession relationships derived from recession analysis, saturation-response functions derived from topographic analysis, etc. This offers better prospects for exploiting spatial analysis, both within semidistributed and fully distributed contexts. Conversely, the fixed model paradigm is poorly suited to the incorporation of independent data analysis insights from a specific catchment.

[100] (3) A structure with a predefined complexity, while perhaps representing a "best compromise" over a broad range of catchments, can be either too complex, or too simple for a specific catchment. In the former case, nonidentifiability will result, whereas in the latter case the best performance will not be achieved. In yet other cases, a fixed model structure could be both too simplistic for some aspects of catchment behavior (e.g., it may be inherently unable to represent a threshold) and too complex for other aspects (e.g., it could include reservoirs and lag functions not needed for the particular basin). Conversely, in conjunction with a set of stringent diagnostics, the appropriate complexity of a flexible model structure can be explored more thoroughly and transparently, taking into account aspects such as catchment size, the spatial and temporal resolution of the available data, the purpose of the modeling application, etc.

[101] (4) The fixed model structure, on the other hand, may be advantageous in many operational contexts, in particular, when there are insufficient time and/or human resources to setup and trial multiple model structures, etc. A fixed structure may also be easier to interpret in terms of parameter differences across multiple applications, and, potentially, may serve as a useful initial point for iterative model improvement exploiting site-specific insights. As such, even for proponents of seeking a fixed model structure, a flexible framework offers a more systematic platform for the ongoing identification and refinement of such models.

[102] (5) A flexible model structure offers the potential for interpreting differences in model performance, understanding catchment behavior, and relating it to independent experimental insights. A well-designed flexible modeling framework allows a careful systematic generation and comparison of system-level hypotheses and their finer-grain components under a common framework. On the other hand, previous model comparison experiments may have been obscured by major unaccounted for differences in model conceptualization, numerical implementation, and model evaluation. Such uncontrollable differences may obscure hypothesis testing and result in important process-understanding insights being missed and/or misinterpreted.

Appendix A: Details of the Hydrological Models

[103] Table A1 details the water balance equations of the SUPERFLEX model structures. Table A2 describes the constitutive relationships used in the model components. The functions f_x are defined in Table 1 of the companion paper [*Fenicia et al.*, 2011].

Table A1. Water Balance Equations of the Models Used in the Experiments[a]

Water Balance Equations	M1	M2	M3	M4	M5	M6	M7
$dS_f/dt = P_t - Q_f - E_f$	✓	-	-	-	-	-	-
$dS_u/dt = P_t - Q_q - Q_u - E_u$	-	✓	-	-	-	-	-
$dS_u/dt = P_t - Q_q - E_u$	-	-	✓	✓	✓	✓	✓
$dS_f/dt = P_f - Q_f$	-	-	✓	✓	-	-	-
$dS_f/dt = P_{fl} - Q_f$	-	-	-	-	✓	✓	✓
$dS_s/dt = P_s - Q_s$	-	-	-	-	-	✓	✓
$Q_q = P_f + P_s$	-	-	-	-	-	✓	✓
$Q_t = Q_f$	✓	-	✓	✓	✓	-	-
$Q_q = Q_q + Q_u$	-	✓	-	-	-	-	-
$Q_t = Q_f + Q_s$	-	-	-	-	-	✓	✓

[a]The ✓ and "-" indicate presence or absence, respectively.

Table A2. Constitutive Relationships of the Model Structures M1–M7[a]

Constitutive Relationships	M1	M2	M3	M4	M5	M6	M7
$\bar{S}_u = S_u/S_{u,\max}$	-	✓	✓	✓	✓	✓	✓
$Q_q = P_t f_p(\bar{S}_u \mid \beta)$	-	✓	-	✓	✓	-	✓
$Q_q = P_t f_h(\bar{S}_u \mid m_1)$	-	-	✓	-	-	-	-
$Q_q = P_t \bar{S}_u$	-	-	-	-	-	✓	-
$E_u = C_e E_p f_m(\bar{S}_u \mid m_2)$	-	✓	✓	✓	✓	✓	✓
$E_f = C_e E_p f_e(S_f \mid m_3)$	✓	-	-	-	-	-	-
$P_{fl} = (P_f^* h_f)(t)$	-	-	-	-	✓	✓	✓
$h_f = \begin{cases} t/T_f^2 & t \le T_f \\ 0, & t > T_f \end{cases}$	-	-	-	-	✓	✓	✓
$P_s = DQ_q$	-	-	-	-	-	✓	✓
$Q_u = R_{\max} \bar{S}_u$	-	✓	-	-	-	-	-
$Q_f = k_f S_f^{\alpha}$	✓	-	✓	✓	✓	-	✓
$Q_f = k_f S_f$	-	-	-	-	-	✓	-
$Q_s = k_s S_s$	-	-	-	-	-	✓	✓

[a]The lag function was smoothed using the method by *Kavetski and Kuczera* [2007]. The ✓ and "-" indicate presence or absence, respectively. The operator * in the equation for P_{fl} denotes the convolution operator.

[104] **Acknowledgments.** We thank Martyn Clark for his feedback on an earlier version of this manuscript. We also thank our colleagues, including Vazken Andréassian, Keith Beven, Jim Freer, Hoshin Gupta, George Kuczera, Jeff McDonnell, Charles Perrin, Laurent Pfister, Jérôme Juilleret, Gerrit Schoups, Hubert Savenije, Murugesu Sivapalan, Mark Thyer, Peter Young, and others, for valuable discussions on the topic of this work. Finally, we thank the reviewers for their constructive comments, criticisms and suggestions, which have significantly improved the paper. This work was funded by the National Research Fund of Luxembourg (grant FOR 1598 – CAOS "From Catchments as Organized Systems to Models based on Dynamic Functional Units") and AM2c grants for the Mobility of Researchers. We also thank the STITPRO Foundation for financial support.

References

Andréassian, V., C. Perrin, L. Berthet, N. Le Moine, J. Lerat, C. Loumagne, L. Oudin, T. Mathevet, M.-H. Ramos, and A. Valery (2009), Crash tests for standardized evaluation of hydrological models, *Hydrol. Earth Syst. Sci.*, *13*, 1757–1764.

Atkinson, S. E., R. A. Woods, and M. Sivapalan (2002), Climate and landscape controls on water balance model complexity over changing landscapes, *Water Resour. Res.*, *38*(12), 1314, doi:10.1029/2002WR001487.

Berthet, L., V. Andréassian, C. Perrin, and P. Javelle (2009), How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments, *Hydrol. Earth Syst. Sci.*, *13*, 819–831.

Beven, K. (1997), TOPMODEL: A critique, *Hydrol. Processes*, *11*(9), 1069–1085.

Beven, K. (2008), On doing better hydrological science, *Hydrol. Processes*, *22*, 3549–3553.

Beven, K. J. (2000), Uniqueness of place and process representations in hydrological modelling, *Hydrol. Earth Syst. Sci.*, *4*(2), 203–213.

Beven, K. J. (2001), *Rainfall—Runoff Modelling: The Primer*, 360 pp., John Wiley, Chichester, U. K.

Birkel, C., D. Tetzlaff, S. M. Dunn, and C. Soulsby (2010), Towards a simple dynamic process conceptualization in rainfall-runoff models using multi-criteria calibration and tracers in temperate, upland catchments, *Hydrol. Processes*, *566*(24), 260–275.

Box, G. E. P., and G. C. Tiao (1992), *Bayesian Inference in Statistical Analysis*, John Wiley, New York, 588 pp.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, *298*, 242–266.

Clark, M. P., and D. Kavetski (2010), Ancient numerical daemons of conceptual hydrological modeling. Part 1: Fidelity and efficiency of time stepping schemes, *Water Resour. Res.*, *46*, W10510, doi:10.1029/2009WR008894.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.

Clark, M. P., D. Kavetski, and F. Fenicia (2011a), Pursuing the method of multiple working hypotheses in hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.

Clark, M. P., H. K. McMillan, D. B. G. Collins, D. Kavetski, and R. A. Woods (2011b), Hydrological field data from a modeller's perspective. Part 2: Process-based evaluation of model hypotheses, *Hydrol. Processes*, *25*, 523–543.

Doherty, J., and D. Welter (2010), A short exploration of structural noise, *Water Resour. Res.*, *46*, W05525, doi:10.1029/2009WR008377.

Duan, Q., et al. (2006), Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*(1–2), 3–17.

Edijatno,and C. Michel (1989), Un modele pluie-debit journalier a trois parametres, *La Houille Blanche*, *2*, 113–121.

Fenicia, F., J. J. McDonnell, and H. H. G. Savenije (2008a), Learning from model improvement: On the contribution of complementary data to process understanding, *Water Resour. Res.*, *44*(6), W06419, doi:10.1029/2007WR006386.

Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2008b), Understanding catchment behavior through stepwise model concept improvement, *Water Resour. Res.*, *44*, W01402, doi:10.1029/2006WR005563.

Fenicia, F., S. Wrede, D. Kavetski, L. Pfister, L. Hoffmann, H. Savenije, and J. J. McDonnell (2010), Impact of mixing assumptions on mean residence time estimation, *Hydrol. Processes* (*Special Issue on Residence Times and Preferential Flows*), *24*(12), 1730–1741.

Fenicia, F., D. Kavetski, and H. H. G. Savenjie (2011), Elements of a flexible approach for conceptual hydrological modeling: Part 1. Motivation and theoretical development, *Water Resour. Res.*, *47*, W11510, doi:10.1029/2010WR010174.

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

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004), *Bayesian Data Analysis*, 2nd ed., 696 pp., Chapman and Hall.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and sharpness, *J. R. Stat. Soc. Ser. B.*, *69*, 243–268.

Götzinger, J., and A. Bardossy (2008), Generic error model for calibration and uncertainty estimation of hydrological models, *Water Resour. Res.*, *44*, W00B97, doi:10.1029/2007WR006691.

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*, 3802–3813.

Hellebrand, H., R. van den Bos, L. Hoffmann, J. Juilleret, and L. Pfister (2008), The potential of winter stormflow coefficients for hydrological regionalization purposes in poorly gauged basins of the middle Rhine region, *Hydrol. Sci. J.*, *53*(4), 773–788.

Immerzeel, W. W., and P. Droogers (2008), Calibration of a distributed hydrological model based on satellite evapotranspiration, *J. Hydrol.*, *349*, 411–424.

Ivanov, V. Y., E. R. Vivoni, R. L. Bras, and D. Entekhabi (2004), Catchment hydrologic response with a fully distributed triangulated irregular network model, *Water Resour. Res.*, *40*(11), W11102, doi:10.1029/2004WR003218.

Kavetski, D., and G. Kuczera (2007), Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration, *Water Resour. Res.*, *43*, W03411, doi:10.1029/2006WR005195.

Kavetski, D., and M. P. Clark (2010), Ancient numerical daemons of conceptual hydrological modeling. Part 2: Impact of time stepping scheme on model analysis and prediction, *Water Resour. Res.*, *46*, W10511, doi:10.1029/2009WR008896.

Kavetski, D., S. Franks, and G. Kuczera (2002), Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models, Water Science and Application Series 6*, edited by Q. Y. Duan, et al., pp. 49–68, American Geophysical Union, Washington, D. C.

Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, *42*(3), W03S04, doi:10.1029/2005WR004362.

Klemes, V. (1986), Operational testing of hydrologic simulation models, *Hydrol. Sci. J.*, *31*, 13–24.

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, *331*(1–2), 161–177.

Le Moine, N. (2008), Le bassin versant de surface vu par le souterrain: Une voie d'amélioration des performances et du réalisme des modèles pluie-débit? PhD thesis, Université Pierre et Marie Curie, Paris, 324 pp.

Le Moine, N., V. Andrassian, C. Perrin, and C. Michel (2007), How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments, *Water Resour. Res.*, *43*, W06428, doi:10.1029/2010WR009827.

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

Linsley, R. K. (1982), Rainfall-runoff models: An overview, in *Proceedings of the International Symposium of Rainfall-Runoff Modelling*, edited by V. P. Singh, pp. 3–22, Water Resource Publications, Littleton, Colorado.

Linsley, R. K., M. A. Kohler, and J. L. H. Paulhus (1949), *Applied Hydrology*, 689 pp., McGraw-Hill, New York.

Liu, Y. Q., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, *43*(7), W07401, doi:10.1029/2006WR005756.

McDonnell, J. J. (1990), A rationale for old water discharge through macropores in a steep, humid catchment, *Water Resour. Res.*, *26*, 2821–2832, doi:10.1029/WR026i011p02821.

McDonnell, J. J. (2003), Where does water go when it rains? Moving beyond the variable source area concept of rainfall-runoff response, *Hydrol. Processes*, *17*(9), 1869–1875.

McDonnell, J. J., and R. Woods (2004), On the need for catchment classification, *J. Hydrol.*, *299*(1–2), 2–3.

McDonnell, J. J., et al. (2007), Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology, *Water Resour. Res.*, *43*, W07301, doi:10.1029/2006WR005467.

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

Merz, R., and G. Blöschl (2004), Regionalisation of catchment model parameters, *J. Hydrol.*, *287*(1–4), 95–123.

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

Moore, R. J. (1985), The probability-distributed principle and runoff production at point and basin scales, *Hydrol. Sci. J.*, *30*(2), 273–297.

Moore, R. J., and R. T. Clarke (1981), A distribution function approach to rainfall runoff modeling, *Water Resour. Res.*, *17*(5), 1367–1382, doi:10.1029/WR017i005p01367.

Mosley, M. P. (1979), Streamflow generation in a forested watershed in New Zealand, *Water Resour. Res.*, *15*, 795–806, doi:10.1029/WR015i004p00795.

Pearce, A. J., M. K. Stewart, and M. G. Sklash (1986), Storm runoff generation in humid headwater catchments 1. Where does the water come from?, *Water Resour. Res.*, *22*, 1263–1272, doi:10.1029/WR022i008p01263.

Perrin, C., C. Michel, and V. Andréassian (2001), Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, *J. Hydrol.*, *242*(3–4), 275–301.

Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, *279*(1–4), 275–289.

Pfister, L., J. F. Iffly, and L. Hoffmann (2002), Use of regionalized stormflow coefficients with a view to hydroclimatological hazard mapping, *Hydrol. Sci. J.*, *47*(3), 479–491.

Pfister, L., J. Kwadijk, A. Musy, A. Bronstert, and L. Hoffmann (2004), Climate change, land use change and runoff prediction in the Rhine-Meuse basins, *River Res. Appl.*, *20*(3), 229–241.

Refsgaard, J. C., and H. J. Henriksen (2004), Modelling guidelines—terminology and guiding principles, *Adv. Water Resour.*, *27*, 71–82.

Reichert, P., and J. Mieleitner (2009), Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters, *Water Resour. Res.*, *45*, W10402, doi:10.1029/2009WR007814.

Renard, B., D. Kavetski, M. Thyer, G. Kuczera, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.

Rowe, L. K., A. J. Pearce, and C. L. O'Loughlin (1994), Hydrology and related changes after harvesting native forest catchments and establishing *Pinus radiata* plantations. Part 1. Introduction to study, *Hydrol. Processes*, *8*(3), 263–279.

Savenije, H. H. G. (2009), The art of hydrology, *Hydrol. Earth Syst. Sci.*, *13*, 157–161.

Schaefli, B., and H. V. Gupta (2007), Do Nash values have value?, *Hydrol. Processes*, *21*(15), 2075–2080.

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.

Schoups, G., N. C. van de Giesen, and H. H. G. Savenije (2008), Model complexity control for hydrologic prediction, *Water Resour. Res.*, *44*, W00B03, doi:10.1029/2008WR006836.

Schoups, G., J. A. Vrugt, F. Fenicia, and N. C. van de Giesen (2010), Corruption of accuracy and efficiency of Markov chain Monte Carlo simulation by inaccurate numerical implementation of conceptual hydrologic models, *Water Resour. Res.*, *46*, W10350, doi:10.1029/2009WR008648.

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 J. J. McDonnell (2002), On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Resour. Res.*, *38*(11), 1241, doi:10.1029/2001WR000978.

Singh, V. P., and D. A. Woolhiser (2002), Mathematical modeling of watershed hydrology, *J. Hydrol. Eng.*, *7*(4), 270–292.

Sivapalan, M. (2005), Pattern, process and function: Elements of a new unified hydrologic theory at the catchment scale, in *Encyclopaedia of Hydrologic Sciences*, Vol. 13(1/1), edited by M. G. Anderson, pp. 193–219, John Wiley, New York.

Skahill, B. E., and J. Doherty (2006), Efficient accommodation of local minima in watershed model calibration, *J. Hydrol.*, *329*(1–2), 122–139.

Son, K., and M. Sivapalan (2007), Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, *Water Resour. Res.*, *43*, W01415, doi:10.1029/2006WR005032.

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrological rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resour. Res.*, *16*(2), 430–442, doi:10.1029/WR016i002p00430.

Spence, C. (2010), A paradigm shift in hydrology: Storage thresholds across scales influence catchment runoff generation, *Geography Compass*, *4*(7), 819–833.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modelling: A case study using Bayesian total

error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.

Troch, P. A., G. A. Carrilo, I. Heidbuchel, R. Seshadri, M. Switanek, T. H. M. Volkman, and M. Yaeger (2009), Dealing with landscape heterogeneity in watershed hydrology: A review of recent progress toward new hydrological theory, *Geography Compass*, *3*, 395–392.

Uhlenbrook, S., S. Roser, and N. Tilch (2004), Hydrological process representation at the meso-scale: The potential of a distributed, conceptual catchment model, *J. Hydrol.*, *291*(3–4), 278–296.

Vaché, K. B., and J. J. McDonnell (2006), A process-based rejectionist framework for evaluating catchment runoff model structure, *Water Resour. Res.*, *42*, W02409, doi:10.1029/2005WR004247.

van den Bos, R., L. Hoffmann, J. Juilleret, P. Matgen, and L. Pfister (2006a), Conceptual modelling of individual HRU's as a trade-off between bottom-up and top-down modelling: A case study, paper presented at *Conf. Environ. Modell. Software, Proc. 3rd Biennial Meeting of International Environ. Modell. Software Society*, Vermont.

van den Bos, R., L. Hoffmann, J. Juilleret, P. Matgen, and L. Pfister (2006b), Regional runoff prediction through aggregation of first-order hydrological process knowledge: A case study, *Hydrol. Sci. J.*, *51*(6), 1021–1038.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Wagener, T. (2003), Evaluation of catchment models, *Hydrol. Processes*, *17*(16), 33753378.

Wagener, T., and H. V. Gupta (2005), Model identification for hydrological forecasting under uncertainty, *Stochastic Environmental Research and Risk Assessment*, *19*(6), 378–387.

Wagener, T., and H. S. Wheater (2006), Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, *J. Hydrol.*, *320*(1–2), 132–154.

Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian (2001), A framework for development and application of hydrological models, *Hydrol. Earth Syst. Sci.*, *5*(1), 13–26.

Wagener, T., M. Sivapalan, P. Troch, and R. A. Woods (2007), Catchment classification and hydrologic similarity, *Geography Compass*, *1*(4), 901–931.

Wood, E. F., D. P. Lettenmaier, and V. G. Zartarian (1992), A land-surface hydrology parameterization with subgrid variability for general circulation models, *J. Geophys. Res. Atmospheres*, *97*(D3), 2717–2728.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, *44*, W09417, doi:10.1029/2007WR006716.

F. Fenicia, Department of Environment and Agro-Biotechnologies, Centre de Recherche Public–Gabriel Lippmann, rue du Brill 41, L-4422 Belvaux, Grand-Duchy of Luxembourg. (fenicia@lippmann.lu)

D. Kavetski, Environmental Engineering, University of Newcastle, University Dr., Callaghan, NSW 2308, Australia.