AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2014WR015484

Process consistency in models: The importance of system signatures, expert knowledge, and process complexity

M. Hrachowitz[1], O. Fovet[2,3], L. Ruiz[2,3], T. Euser[1], S. Gharari[1], R. Nijzink[1], J. Freer[4], H. H. G. Savenije[1,5], and C. Gascuel-Odoux[2,3]

[1]Water Resources Section, Faculty of Civil Engineering and Applied Geosciences, Delft University of Technology, Delft, Netherlands, [2]INRA, UMR 1069, Sol Agro et hydrosystème Spatialisation, Rennes, France, [3]Agrocampus Ouest, UMR 1069, Sol Agro et hydrosystème Spatialisation, Rennes, France, [4]School of Geographical Sciences, University of Bristol, Bristol, UK, [5]UNESCO-IHE Institute for Water Education, Delft, Netherlands

Abstract Hydrological models frequently suffer from limited predictive power despite adequate calibration performances. This can indicate insufficient representations of the underlying processes. Thus, ways are sought to increase model consistency while satisfying the contrasting priorities of increased model complexity and limited equifinality. In this study, the value of a systematic use of hydrological signatures and expert knowledge for increasing model consistency was tested. It was found that a simple conceptual model, constrained by four calibration objective functions, was able to adequately reproduce the hydrograph in the calibration period. The model, however, could not reproduce a suite of hydrological signatures, indicating a lack of model consistency. Subsequently, testing 11 models, model complexity was increased in a stepwise way and counter-balanced by "prior constraints," inferred from expert knowledge to ensure a model which behaves well with respect to the modeler's perception of the system. We showed that, in spite of unchanged calibration performance, the most complex model setup exhibited increased performance in the independent test period and skill to better reproduce all tested signatures, indicating a better system representation. The results suggest that a model may be inadequate despite good performance with respect to multiple calibration objectives and that increasing model complexity, if counter-balanced by prior constraints, can significantly increase predictive performance of a model and its skill to reproduce hydrological signatures. The results strongly illustrate the need to balance automated model calibration with a more expert-knowledge-driven strategy of constraining models.

Key Points:
- Expert-knowledge-based prior information has strong constraining power
- Complex models with prior constraints are consistent and robust
- Uncalibrated but constrained complex models outperform standard models

Supporting Information:
- Readme
- Figure S1
- Figure S2
- Figure S3

Correspondence to:
M. Hrachowitz,
m.hrachowitz@tudelft.nl

Citation:
Hrachowitz, M., O. Fovet, L. Ruiz, T. Euser, S. Gharari, R. Nijzink, J. Freer, H. H. G. Savenije, and C. Gascuel-Odoux (2014), Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, *Water Resour. Res.*, *50*, doi:10.1002/2014WR015484.

Received 20 FEB 2014
Accepted 29 AUG 2014
Accepted article online 2 SEP 2014

1. Introduction

In recent years, an increased awareness has developed that an improved consistency of landscape-scale conceptual environmental models, i.e., more plausible representations of observed system dynamics, is required [e.g., *Wagener and Gupta*, 2005; *Kirchner*, 2006; *Martinez and Gupta*, 2011; *Gupta et al*., 2012; *Hrachowitz et al*., 2013a] to increase the predictive power of models, in the sense of constraining their responses "to achieve the least uncertainty for forecasts" [*Kumar*, 2011]. The ability of models to adequately reproduce relevant system dynamics is typically undermined not only by aleatory and epistemic uncertainties in data [*Beven and Westerberg*, 2011; *Beven*, 2013] but also by ontological uncertainties, i.e., our limited knowledge of process heterogeneity together with our often rudimentary understanding of mechanisms of nonstationarity in real-world systems. This implies that all models are simplifications, no matter their physical and spatial complexity [*Gupta et al*., 2012]. The above uncertainties and the limited number of observations in a continuous spatial domain typically make such models ill-posed inverse problems. Being overdetermined and frequently insufficiently constrained lends such problems elevated degrees of freedom, resulting in nonunique (i.e., equifinal) and instable solutions that can be highly sensitive to already small errors [*Yeh*, 1986; *Beven*, 2000; *Neuman*, 2003]. To ensure robust predictions, modelers have to find a suitable level of model complexity that allows an adequate reproduction of functional characteristics of a system with a minimum of parameter and architectural uncertainty. However, due to insufficient model selection and testing procedures [e.g., *Klemes*, 1986; *Wagener*, 2003; *Jakeman et al*., 2006; *Gupta et al*., 2008; *Andréassian et al*., 2009; *Coron et al*., 2012], deceptively good calibration performances can frequently be mere reflections of the mathematical fitting process in a typically overparameterized domain and may

generate undesirable model internal dynamics, thereby potentially limiting the predictive capability in independent test periods [*Wagener and Gupta*, 2005; *Beven*, 2006; *Kirchner*, 2006; *Andréassian et al*., 2012; *Gharari et al*., 2013a].

In the hypothetical most ideal and general sense, it is the modeler's task to sufficiently constrain the theoretical complete set of "all possible models" to the subset of models that can, in the perception of the modeler, reproduce several signatures of the observed response dynamics of a system to facilitate as reliable predictions as possible given the available data. More specifically, models are typically selected according to a hierarchy of different types of constraints whose effectiveness is dependent on the range of observations and knowledge available. These constraints include (1) model *architectural constraints* (hereafter interchangeably used with the term "model structure") that limit the possible response dynamics of the modeled variable(s) by the choice of the model structure(s) [e.g., *Leavesley et al*., 1996; *Wagener et al*., 2001; *Neuman*, 2003; *Clark et al*., 2008, 2011; *Fenicia et al*., 2011]. For example, a model consisting of one simple linear reservoir is to be rejected where discharge is the combination of contributions from storage components with different response times; (2) model *parameterization constraints* that define the subset of possible responses of a given model architecture by the choice of constitutive functions [cf. *Kavetski and Fenicia*, 2011]; (3) *modeling objective constraints* that limit the feasible model space to those models that can reproduce the chosen modeling objectives and criteria. In the past, the use of multiple objective functions [e.g., *Gupta et al*., 1998; *Freer et al*., 2004; *Wagener et al*., 2009; *Efstratiadis and Koutsoyiannis*, 2010; *Hrachowitz et al*., 2013b] and/or criteria such as catchment signatures [e.g., *Yadav et al*., 2007; *Wagener and Montanari*, 2011; *Westerberg et al*., 2011; *Euser et al*., 2013; *Coxon et al*., 2014] has proven valuable for the selection of robust subsets of feasible models; and finally (4) model *prior constraints*, i.e., prior information that helps identifying models which do not contradict physical necessities and/or the modeler's perception of how the system functions [e.g., *Ambroise et al*., 1996a].

The value of *prior constraints*, e.g., using prior knowledge on parameters, uncertainties in forcing or spatial heterogeneities, to better pose inverse problems has been successfully demonstrated in the past [e.g., *Yeh*, 1986; *Renard et al*., 2010] using regularization [e.g., *Tonkin and Doherty*, 2005; *Pokhrel et al*., 2008; *Kumar et al*., 2010, 2013; *Samaniego et al*., 2010] and other techniques [e.g., *Carrera and Neuman*, 1986; *Refsgaard et al*., 2006; *Jafarpour*, 2011; *Jafarpour and Tarrahi*, 2011; *Shi et al*., 2014]. Alternatively, but rarely fully exploited in hydrology, where not enough data are available to warrant the above methods, prior information can also be incorporated based on explicit hydrological reasoning and semiquantitative or anecdotal expert knowledge [*Seibert and McDonnell*, 2002; *Gao et al*., 2014; *Gharari et al*., 2013b; *Hughes*, 2013]. Following this approach, a model will only be retained as feasible if it can satisfy these *prior constraints* imposed on parameters and modeled processes.

In this paper, we will describe a procedure for systematically applying anecdotal and expert-knowledge-based, semiquantitative prior information to better pose overdetermined inverse problems. Specifically, the novel contribution of this paper is to test and illustrate the potential of above procedure for reconciling increased process heterogeneity, i.e., hydrological consistency, and limited uncertainty in hydrological models to iteratively obtain models that better reflect the understanding of overall catchment functioning, therefore providing higher predictive power for forecasting, as illustrated in a proof-of-concept case study.

2. Study Site and Hydrological Data

The study site is the 11 ha Kerrien experimental catchment, which is part of the ORE-AgrHys observatory (www.inra.fr/ore_agrhys) in French Brittany (Figure 1).

In the oceanic climate with mild winters and relatively cool summers, the long-term mean annual temperature reaches 11.4°C with a mean annual precipitation of 1035 mm yr^{-1}. Average potential Penman evaporation is estimated at 690 mm yr^{-1}. An intermittent first-order stream, typically running dry between August and October, drains the catchment with a long-term mean runoff of about 270 mm yr^{-1}.

Daily precipitation totals and parameters to estimate the daily potential evaporation (Penman-Monteith) were obtained from an automated weather station located at around 700 m from the catchment outlet for the 11 year period from 1 October 2001 to 30 September 2012 (Figures 2a and 2b). Daily runoff values for the same period were obtained from a V notch weir equipped with a shaft encoder (OTT Thalimedes;

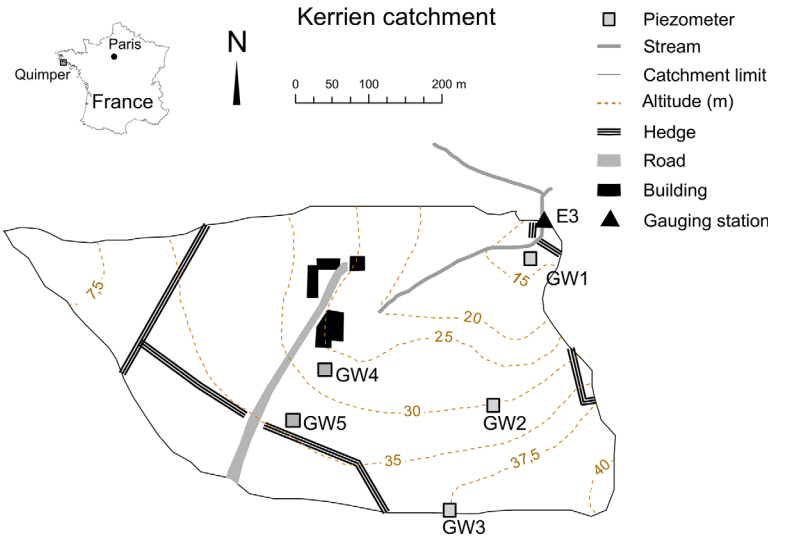


Figure 1. The Kerrien study catchment, which is part of the ORE-AgrHys network.

Figure 2b). Groundwater level data from a total of 5 piezometers (depth $>$ 5 m) located along two hillslope transects were used (GW1–GW5; Figure 2c).

Covering an elevation range between 15 and 40 m, the catchment is underlain by a series of differently weathered regolith facies on top of unweathered granite at a depth of around 20 m [*Legchenko et al*., 2004]. The well-drained district cambisols are mostly sandy loam, rich in organic matter, and on average 80 cm deep. In correspondence with the high-permeability regolith, the groundwater table responds quickly and exhibits intraannual amplitudes of up to 7 m [*Martin et al*., 2006; *Legout et al*., 2007]. Based on soil maps, topography and expert knowledge, around 10% of the catchment can be characterized as wetlands with very shallow groundwater tables. The recharge period typically extends from November through March.

Approximately 40% of the catchment is used for the cultivation of maize, while a further 40% is used as grassland. For a more detailed description of the study catchment, the reader is referred to *Ruiz et al*. [2002] and *Molenat et al*. [2008].

3. Modeling Strategy

Adopting a flexible modeling strategy, in which a model should reflect the characteristics of a given catchment [*Fenicia et al*., 2013], each model structure in this study represents a testable hypothesis [*Beven*, 2001; *Clark et al*., 2011; *Fenicia et al*., 2011]. Following the principle of model parsimony [e.g., *Jakeman et al*., 2006],

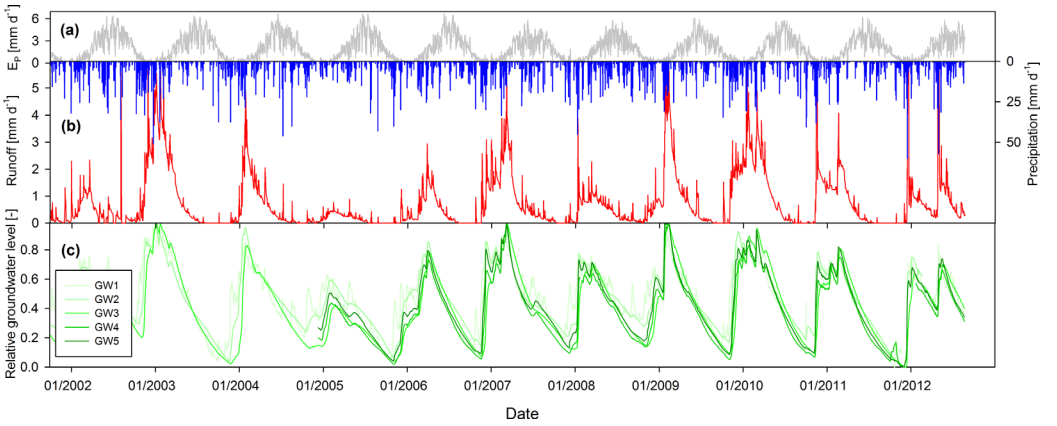


Figure 2. Observed time series of (a) daily potential evaporation, (b) daily precipitation and streamflow, and (c) the groundwater levels of 5 piezometers located along two transects (see Figure 1), normalized by their respective range between minimum and maximum observed values. Note that GW1 is located in the wetland zone.

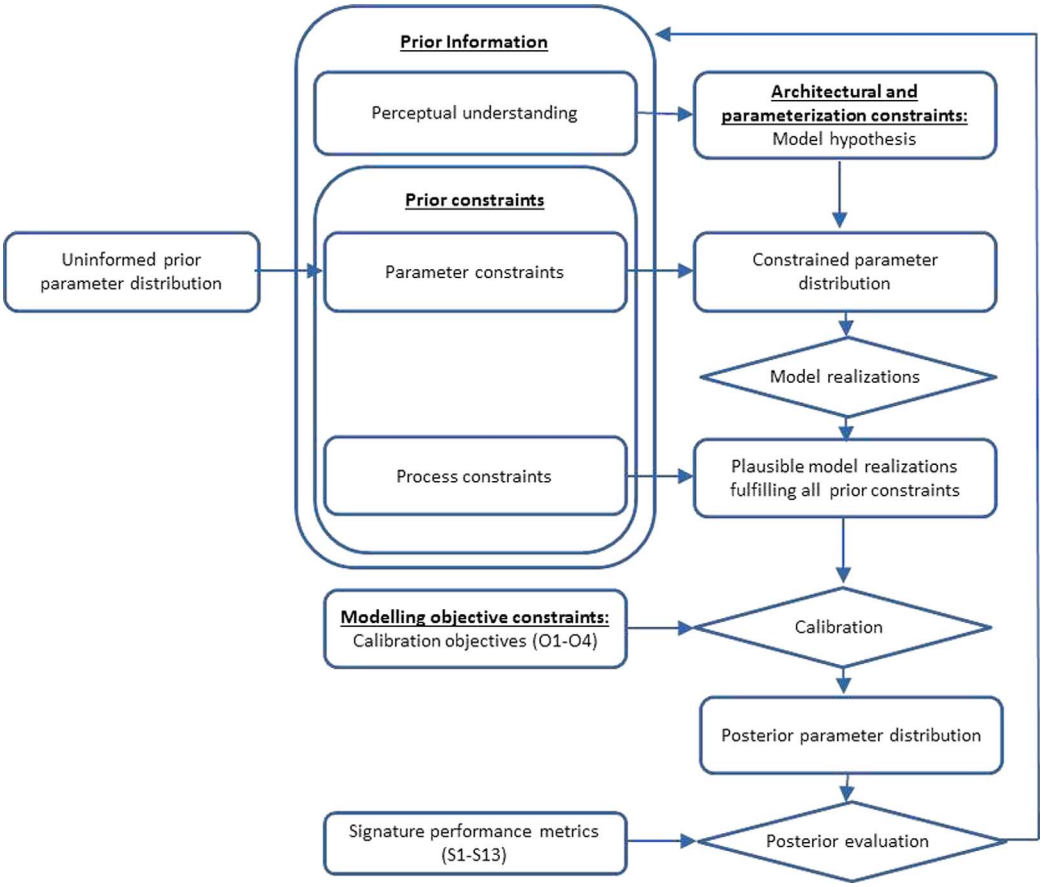


Figure 3. Flow chart of the modeling strategy.

thus keeping to a minimum the necessary parameters and thereby the adverse effects of equifinality [e.g., *Schoups et al*., 2008], a suite of model structure hypotheses, i.e., *architectural* and *parameterization constraints*, of increasing process complexity was thus designed. This suite of models was developed based on both expert knowledge of the underlying processes and additional data and then tested in a stepwise effort [e.g., *Fenicia et al*., 2008a, 2008b]. To counterbalance model equifinality, introduced by additional parameters, suitable *prior constraints* were imposed on the model after each iteration, i.e., adaptation of the model structure. The adequacy of the individual models was subsequently assessed not only based on calibration performance (*modeling objective constraints*) but also based on the associated predictive power post calibration [cf. *Klemes*, 1986] and, more importantly, on their skill to reproduce a range of functional characteristics, i.e., hydrological signatures [e.g., *Gupta et al*., 2014].

The general procedure as outlined also in Figure 3 can in principal be applied to any catchment using any type of model setup. Therefore, the model structure development will in the following only be roughly outlined (see also Table 1). Emphasis will rather be given to a detailed rationale of the different types of constraints and signatures used in this study.

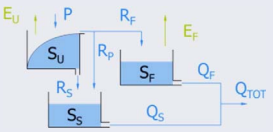
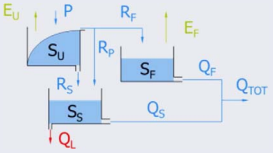
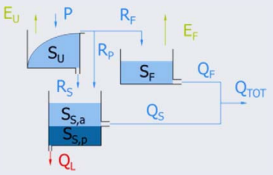
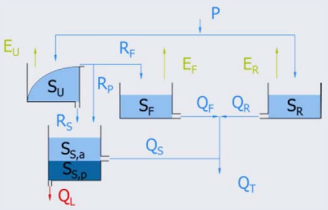
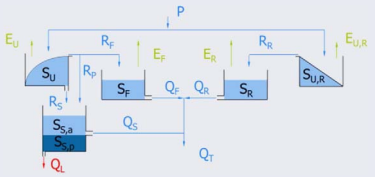
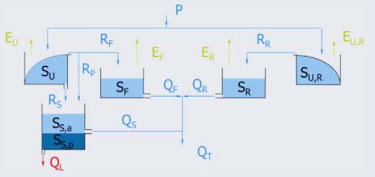
3.1. Stepwise Model Hypothesis Development

Making use of the recently developed conceptual DYNAMIT flexible modeling framework [*Hrachowitz et al*., 2013b], a total of 11 model hypotheses, i.e., different *architectural* and *parameterization constraints*, were tested in a stepwise development in this study (M1–M11; Table 1). Details, including the relevant state and flux equations, are given in Table 2.

3.1.1. Defining the Benchmark Model

As a starting point, a simple three box model (M1; seven parameters) was tested and compared, as a benchmark, to three models commonly used in many scientific studies and in operational hydrology: NWS Sacramento [*Burnash*, 1995], HBV [*Bergström*, 1995], and HyMod [*Moore*, 1985; *Wagener et al*., 2001].

Table 1. Model Architectural Constraints (Model Structures), Parameterization Constraints (Parameters and Equations W1–W44), and Prior Constraints (C1–C8) Used for Model Formulations M1–M11[a]

Model Structure	Model	Parameters	Equations	Prior Constraints	Par. No.	Constr. No.
	M1	C_P, k_F, k_S, L_P, P_{max}, $S_{Umax,H}$, β_H	W1–W12, W20–W21, W41, W43	C1	7 (7)	1
	M2	C_P, k_F, k_L, k_S, L_P, P_{max}, $S_{Umax,H}$, β_H	W1–W11, W13, W20, W22–W25, W41, W43	C1	8 (8)	1
	M3	C_P, k_F, k_L, k_S, L_P, P_{max}, $S_{Umax,H}$, β_H	W1–W11, W13, W20, W22–W25, W40, W42	C1, C4, C5	8 (8)	3
	M4	C_P, k_F, k_L, k_S, L_P, P_{max}, $S_{S,p,max}$, $S_{Umax,H}$, β_H	W1–W11, W14–W16, W24–W29, W41, W43	C1, C4, C5	9 (9)	3
	M5	C_P, k_F, k_L, k_S, L_P, P_{max}, $S_{S,p,max}$, $S_{Umax,H}$, β_H	W1–W11, W14–W16, W24–W29, W41, W43	C1, C4, C5, C6	9 (9)	4
	M6	C_P, k_F, k_S, L_P, $Q_{L,const}$, P_{max}, $S_{Umax,H}$, β_H	W1–W11, W17–W19, W25–W26, W30–W31, W41, W43	C1, C4, C5, C6	8 (6)	4
	M7	C_P, f, k_F, k_R, k_S, L_P, $Q_{L,const}$, P_{max}, $S_{Umax,H}$, β_H	W1–W11, W17–W19, W25–W26, W30–W31, W37–W40, W42, W44	C1, C2, C4, C5, C6	10 (7)	5
	M8	C_P, f, k_F, k_R, k_S, L_P, $Q_{L,const}$, P_{max}, $S_{Umax,H}$, $S_{Umax,R}$, β_H	W1–W11, W17–W19, W25–W26, W30–W35, W37–W40, W42, W45	C1, C2, C3, C4, C5, C6	11 (8)	6
	M9	C_P, f, k_F, k_R, k_S, L_P, $Q_{L,const}$, P_{max}, $S_{Umax,H}$, $S_{Umax,R}$, β_H	W1–W11, W17–W19, W25–W26, W30–W35, W37–W40, W42, W45	C1, C2, C3, C4, C5, C6, C7	11 (8)	7
	M10	C_P, f, k_F, k_R, k_S, L_P, $Q_{L,const}$, P_{max}, $S_{Umax,H}$, $S_{Umax,R}$, β_H	W1–W11, W17–W19, W25–W26, W30–W35, W37–W40, W42, W45	C1, C2, C3, C4, C5, C6, C7, C8	11 (8)	8
	M11	C_P, f, k_F, k_R, k_S, L_P, $Q_{L,const}$, P_{max}, $S_{Umax,H}$, $S_{Umax,R}$, β_H, β_R	W1–W11, W17–W19, W25–W26, W30–W34, W36–W40, W42, W45	C1, C2, C3, C4, C5, C6, C7, C8	12 (9)	8

[a]Underlined parameters are free calibration parameters, not underlined parameters are fixed. A list of symbols is given in the notation section.

3.1.2. Stepwise Adaptation of Architectural Constraints

3.1.2.1. Deep Infiltration Loss Constraints

Analyzing the long-term water balance in the study catchment revealed a significant deficit (see section 3.2.2.2). Although this cannot be fully verified, as pointed out by *Beven* [2001], there is evidence that such deficits are in many catchments—at least partly—caused by significant intercatchment groundwater flow [e.g., *Le Moine et al.*, 2007; *Schaller and Fan*, 2009]. Nevertheless, such deep infiltration losses are rarely accounted for in standard formulations of common conceptual models [e.g., *Bergström*, 1995; *Burnash*, 1995], with a rare exception being the GR4J model [*Perrin et al.*, 2003]. In the absence of detailed process knowledge, the apparent deep infiltration losses from the study catchment were here conceptualized with *architectural* and *parameterization constraints* of varying degrees of complexity and tested in model setups M2–M6 (Table 1). In the least complex conceptualizations (M2, M3; eight parameters), a second outlet was incorporated into the groundwater storage. Subsequently, a threshold ($S_{S,p,max}$) was added to allow for continued loss when the stream runs dry, representing a hydraulically passive storage volume below the level of the stream bed [e.g., *Fenicia et al.*, 2008a; M4–M5; nine parameters).

Table 2. Model Parameterization Constraints, i.e., Water Balance, State, and Flux Equations of the Models Used in the Analysis[a]

Process	Water Balance	Equation	Flux and State Equations, Constitutive Relationships	Equation
Unsaturated zone	$dS_U/dt = P - E_U - R_F - R_P - R_S$	(W1)	$E_U = E_P min\left(1, \frac{S_U}{S_{Umax,H}} \frac{1}{L_P}\right)$	(W2)
			$R_U = (1 - C_R)P$	(W3)
			$R_F = C_R(1 - C_P)P$	(W4)
			$R_P = C_R C_P P$	(W5)
			$R_S = P_{max}\left(\frac{S_U}{S_{Umax,H}}\right)$	(W6)
			$C_R = \frac{1}{\left(1 + exp\left(\frac{-S_U/S_{Umax,H} + 0.5}{\beta_H}\right)\right)}$	(W7)
Fast reservoir	$dS_F/dt = R_F - Q_F - E_F$	(W8)	$S_{F,in} = S_F + R_F dt$	(W9)
			$Q_F = S_{F,in}\left(1 - e^{-k_F t}\right) dt^{-1}$	(W10)
			$E_F = min\left(E_P - E_U, \frac{S_{F,in}}{dt} - Q_F\right)$	(W11)
Slow reservoir	$dS_S/dt = R_S + R_P - Q_S$	(W12)	$S_{S,in} = S_S + R_S dt + R_P dt$	(W20)
	$dS_S/dt = R_S + R_P - Q_S - Q_L$	(W13)	$Q_S = S_{S,in}\left(1 - e^{-k_S t}\right) dt^{-1}$	(W21)
	$dS_{S,a}/dt = \begin{cases} \left(S_{S,a} - S_{S,tot,out} + S_{S,p,max}\right) dt^{-1} & \\ 0, \quad S_{S,tot,out} < S_{S,p,max} & \end{cases}$	(W14)	$Q_{S,tot} = S_{S,in}\left(1 - e^{(-k_S - k_L)t}\right) dt^{-1}$	(W22)
	$dS_{S,p}/dt = \begin{cases} 0, \quad S_{S,tot,out} \geq S_{S,p,max} & \\ \left(S_{S,tot,in} - S_{S,tot,out}\right) dt^{-1}, \quad S_{S,tot,out} < S_{S,p,max} & \end{cases}$	(W15)	$\frac{Q_S}{Q_L} = \frac{k_S}{k_L}$	(W23)
	$dS_S/dt = dS_{S,a}/dt + dS_{S,p}/dt = R_S + R_P - Q_S - Q_L$	(W16)	$Q_L = \frac{Q_{S,tot}}{\left(\frac{Q_S}{Q_L} + 1\right)}$	(W24)
	$dS_{S,a}/dt = \begin{cases} \left(S_{S,a} - max\left(0, S_{S,tot,out}\right)\right) dt^{-1}, \quad S_{S,tot,in} > 0 & \\ 0, \quad S_{S,tot,in} \leq 0 & \end{cases}$	(W17)	$Q_S = max\left(0, Q_{S,tot} - Q_L\right)$	(W25)
Slow reservoir	$dS_{S,p}/dt = \begin{cases} \left(S_{S,p} + min\left(0, S_{S,tot,out}\right)\right) dt^{-1}, \quad S_{S,tot,in} > 0 & \\ \left(S_{S,p} + S_{S,tot,out}\right) dt^{-1}, \quad S_{S,tot,in} \leq 0 & \end{cases}$	(W18)	$S_{S,tot,in} = S_{S,a} + S_{S,p} + R_S dt + R_P dt$	(W26)
	$dS_S/dt = dS_{S,a}/dt + dS_{S,p}/dt = R_S + R_P - Q_S - Q_{L,const}$	(W19)	$S_{S,tot,out} = \frac{k_S S_{S,p}}{k_S + k_L} + S_{S,tot,in} e^{(-k_S - k_L)t} - \frac{k_S S_{S,tot,in} e^{(-k_S - k_L)t}}{k_S + k_L}$	(W27)
			$Q_{S,tot} = \frac{S_{S,tot,in} - S_{S,tot,out}}{dt}$	(W28)
			$\frac{Q_S}{Q_L} = max\left(0, \frac{k_S\left(S_{S,tot,in} - S_{S,p}\right)}{k_L S_{S,tot,in}}\right)$	(W29)
			$S_{S,tot,out} = \begin{cases} S_{S,tot,in} e^{-k_S t} - \frac{Q_{L,const}}{k_S}\left(1 - e^{-k_S t}\right), & S_{S,tot,in} > 0 \\ S_{S,tot,in} - Q_{L,const}, & S_{S,tot,in} \leq 0 \end{cases}$	(W30)
			$Q_{L,const} = const.$	(W31)
Unsaturated riparian zone	$dS_{U,R}/dt = P - E_{U,R} - R_R$	(W32)	$E_{U,R} = E_P min\left(1, \frac{S_{U,R}}{S_{Umax,R}} \frac{1}{L_P}\right)$	(W33)
			$R_R = C_{R,R} P$	(W34)
			$C_{R,R} = min\left(1, \left(\frac{S_{U,R}}{S_{U,max,R}}\right)\right)$	(W35)
			$C_{R,R} = min\left(1, \left(\frac{S_{U,R}}{S_{Umax,R}}\right)^{\beta_R}\right)$	(W36)
Riparian reservoir	$dS_R/dt = R_R - Q_R - E_R$	(W37)	$S_{R,in} = S_R + R_R dt$	(W38)
			$Q_R = S_{Rin}\left(1 - e^{-k_R t}\right) dt^{-1}$	(W39)
			$E_R = min\left(E_P - E_{U,R}, \frac{S_{R,in}}{dt} - Q_R\right)$	(W40)
Total runoff	$Q_T = Q_F + Q_S$	(W41)		
	$Q_T = (1 - f)(Q_F + Q_S) + f Q_R$	(W42)		
Total evaporative fluxes	$E_A = E_U + E_F$	(W43)		
	$E_A = (1 - f)(E_U + E_F) + f E_R$	(W44)		
	$E_A = (1 - f)(E_U + E_F) + f\left(E_{U,R} + E_R\right)$	(W45)		

[a]A list of symbols is given in the notation section.

Analysis of the available piezometer data (Figure 2) did not only reveal that the groundwater table fluctuations are rather homogeneous along the two observed transects and may therefore serve as indicators for catchment storage but also that under no-flow conditions the groundwater levels continued to fall at an approximately constant rate, which can be directly estimated from the available data (see Appendix A). Thus, in model setup M6 the deep percolation loss was treated as known ($Q_{L,const}$), rather than as a process determined by calibration parameters. In addition, with an a priori estimation of $Q_{L,const}$, the storage coefficient k_S could be determined from a Master Recession Curve [e.g., *Lamb and Beven*, 1997; *Hrachowitz et al.*, 2011] after subtracting $Q_{L,const}$ from Q_{obs} at each recession time step ($k_S = 0.041$ d^{-1}). This resulted in a reduction to six free calibration parameters in M6.

3.1.2.2. Riparian Zone Constraints

In many catchments, distinct process dynamics characterize the responses of hillslopes and riparian zones [e.g., *Seibert et al*., 2003b; *Freer et al*., 2004; *Detty and McGuire*, 2010]. This can be attributed to reduced storage capacities in riparian areas, leading to faster storm flow generation, in particular during relatively dry conditions [e.g., *McGlynn et al*., 2004; *Molenat et al*., 2005]. In spite of calls to explicitly account for these differences [e.g., *Savenije*, 2010; *Gharari et al*., 2011], only a minority of models makes use of individual parallel components for hillslopes and riparian areas in (semi)distributed setups [e.g., *Knudsen et al*., 1986; *Beven and Freer*, 2001; *Peters et al*., 2003; *Seibert et al*., 2003a; *Birkel et al*., 2010; *Gao et al*., 2014; *Gharari et al*., 2013b]. Here model complexity was thus further increased by incorporating a parallel model component representing wetlands. As the proportion f of wetland (or riparian area) in the catchment is estimated at 10%, based on soil data and long-term on-site experience [*Martin et al*., 2004], the simplest conceptualization required only one additional free calibration parameter: the storage coefficient of the wetland reservoir (M7; seven free calibration parameters). More complex model structures to be tested subsequently allowed for an unsaturated riparian zone with different flow generation functions (M8–M11; eight to nine free calibration parameters). The three benchmark models (NWS Sacramento, HBV, and HyMod) were adapted to allow a process representation equivalent to M11, including constant deep infiltration and a parallel wetland component.

3.2. Prior Constraints

Model equifinality and predictive uncertainty, inherently linked to the ill-posed nature of many inverse problems, are strongly associated with an elevated number of degrees of freedom in combination with insufficient model constraints [*Yeh*, 1986; *Gupta et al*., 2008]. In other words, the model space (i.e., the combined space of model structures and parameters) is insufficiently constrained by typically uninformative prior distributions within frequently rather loose limits and with an insufficient number of objective functions. This may, in spite of adequate calibration performances, yield model internal dynamics that contradict the modeler's perception of how the system works, that do not correspond with available observations or that potentially contradict physical necessities, such as the requirement that forests are characterized by higher interception capacities than grassland. It is therefore critical to identify and exclude these combinations from the feasible model space, thereby better posing the problem by confining the model space and enforcing the selection of models that better reflect our perception of the real system.

An obvious candidate for potentially powerful model constraints are often overlooked and not systematically exploited plausibility checks using prior information, such as semiquantitative or indicative information from expert knowledge and the long-term water balance. Note that only few, if any, of these *prior constraints* introduced hereafter are particularly new and there is no claim to originality or to particular creativity made here. Some of them were already successfully used in previous studies [e.g., *Seibert and McDonnell*, 2002; *Winsemius et al*., 2009; *Gao et al*., 2014; *Gharari et al*., 2013b]. However, such frequently anecdotal "bits and pieces" of information are rarely used systematically in an explicit and exhaustive way, i.e., a significant number of them at one time to improve the overall modeled system behavior, thereby increasing model consistency [e.g., *Hughes*, 2013].

Following *Gharari et al*. [2013c], two classes of *prior constraints* were distinguished in this study. On the one hand, there are *parameter constraints*. These purely relational constraints ensure that only parameter combinations that follow the logic of the modeler's perception of the system are selected as feasible. This approach takes a middle way between the use of a priori fixed narrow parameter ranges [e.g., *Ambroise et al*., 1996a; *Anderson et al*., 2006; *Koren et al*., 2008] and automated calibration using the complete parameter space defined by uninformed prior distributions, largely avoiding the disadvantages of either method. On the other hand, the class of *process constraints* ensures that, in the absence of observed time series to be directly evaluated against, some summary metric of individual modeled fluxes (e.g., their modeled mean value) falls within an expectation interval. This can be defined within limits of acceptability [*Beven*, 2006] inferred from data or expert judgment-based hydrological reasoning. The details of the *prior constraints* investigated in this study are given below and in Table 1.

3.2.1. Parameter Constraints

3.2.1.1. Storage Coefficients (C1, C2)

The storage coefficients k_S (d^{-1}), k_F (d^{-1}), and k_R (d^{-1}) control the outflow from the respective reservoirs S_S (mm), S_F (mm), and S_R (mm) (see Table 2). Due to the intrinsic definition of S_S as slow responding reservoir,

representing groundwater dynamics, its storage coefficient k_S needs to be smaller than the storage coefficient k_F of the fast responding reservoir S_F [e.g., *Seibert and Vis*, 2012] (constraint C1):

$$k_S < k_F. \tag{1}$$

Although this is implicit in many studies (albeit not all!) by the choice of nonoverlapping prior parameter ranges for k_S and k_F, it is nevertheless explicitly included here to underline its nature as a hydrologically critical *prior constraint*. To acknowledge its common use, this *prior constraint* will here be used in all model setups (M1–M11).

Likewise, the storage coefficient k_F, representing the outflow dynamics of a fast responding reservoir, frequently conceptualized as flow through higher permeability matrix or preferential flow path networks on hillslopes, can be argued to be lower than k_R, characterizing saturation overland flow from a riparian zone. The rationale being that there is evidence that, once connected, effective flow velocities of riparian zones exceed those of preferential flow networks from hillslopes [cf. *Anderson et al*., 2009] (C2):

$$k_F < k_R. \tag{2}$$

Constraint C2 was applied to models that distinguish wetland areas (M7–M11).

3.2.1.2. Soil Moisture Capacity (C3)

The parameters defining the soil moisture capacity of hillslopes and riparian zones, $S_{U,max,H}$ (mm) and $S_{U,max,R}$ (mm), respectively, define the dynamic part of the unsaturated zone. They are limited by the rooting depth or the depth of the groundwater table. Due to generally shallower ground water levels, the unsaturated storage capacity in riparian zones is assumed to be lower than the storage capacity of hillslopes (C3; M8–M11):

$$S_{U,max,R} < S_{U,max,H}. \tag{3}$$

3.2.2. Process Constraints

3.2.2.1. Long-Term Mean Annual Actual Evaporation (C4)

In the absence of detailed actual evaporation (E_A) data, already monthly or annual actual evaporation rates can be valuable to identify inadequate models, as for example shown by *Winsemius et al*. [2008]. Alternatively, the Budyko framework [*Budyko*, 1974] may be a useful source of information especially in the potential presence of intercatchment groundwater flow [e.g., *Andréassian and Perrin*, 2012]. Assuming that over a multiyear period, catchment storage changes are negligible, upper and lower limits of long-term average annual E_A can be estimated [e.g., *Arora*, 2002] (Appendix B). The modeled long-term average annual evaporation ($E_{A,m}$ (mm yr^{-1})) had to fall in between these limits for a model to be retained as feasible. In absence of more information and in contrast to more detailed limits of acceptability approaches [e.g., *Liu et al*., 2009], all values within the limits were considered equally likely. Thus, constraint C4, applied in M3–M11, was here formulated as

$$515 \text{ mm yr}^{-1} < E_{A,m} < 654 \text{ mm yr}^{-1}. \tag{4}$$

3.2.2.2. Long-Term Mean Annual Loss Rates (C5)

Bounds on the long-term mean annual loss rate $Q_{L,mean}$ due to intercatchment groundwater flow can be estimated by closing the water balance using the upper and lower bounds of mean annual $E_{A,m}$. Constraint C5, applied in M3–M11, was here formulated as

$$106 \text{ mm yr}^{-1} < Q_{L,mean} < 246 \text{ mm yr}^{-1}. \tag{5}$$

3.2.2.3. Long-Term Mean Base Flow Contribution (C6)

To enhance process plausibility, it was deemed desirable that the model component S_S, representing the groundwater storage, generates flow dynamics that reflect what is commonly referred to as base flow. The long-term average relative base flow contribution $\overline{C_B}(-)$ can be estimated from stream flow data with the use of digital low-pass filters [e.g., *Chapman and Maxwell*, 1996; *Eckhardt*, 2005; *Merz et al*., 2006; *Hrachowitz et al*., 2011]:

$$\overline{C_B} = \frac{\sum_{t=1}^{T} \frac{a}{2-a} Q_S(t-\Delta t) + \frac{1-a}{2-a} Q_T(t)}{\sum_{t=1}^{T} Q_T(t)}, \quad Q_S(t) \le Q_T(t), \tag{6}$$

$$a = e^{-k_S \Delta t}, \tag{7}$$

where $Q_S(t)$ and $Q_T(t)$ are base flow and total runoff at time $t = 1, \ldots, T$, respectively, Δt is the observation time step, and a is the base flow recession constant, derived from storage coefficient k_S. For the determination of a, k_S was estimated from a preliminary Master Recession Curve. To account for uncertainties introduced by the effects of unknown loss rates Q_L in M5 and for the fact that base flow is frequently not completely generated from S_S, upper and lower bounds of the long-term average relative base flow contribution were established using values of k_S that range between 0.75 and 1.25 times the k_S obtained from the MRC. Constraint C6, applied in M5–M11, was therefore here formulated as

$$0.36 \; < \; \overline{C_B} \; < 0.54. \tag{8}$$

3.2.2.4. Long-Term Mean Fast Wetland Contributions (C7, C8)

During dry periods (here: June–October), wetlands or riparian areas can be expected to contribute a higher average proportion of fast flow ($Q_{R,dry,peak}$ (mm d^{-1})) in response to storm events than hillslopes ($Q_{F,dry,peak}$ (mm d^{-1})), which are frequently characterized by a more pronounced soil moisture deficit as discussed in section 3.1.2.2. Constraint C7, applied in models that distinguish wetlands (M9–M11) was formulated as

$$\sum_{t=1}^{T} Q_{R,dry,peak}(t) > \sum_{t=1}^{T} Q_{F,dry,peak}(t). \tag{9}$$

Conversely, during wet periods (here: December–March) the average fast flow generated from the hillslope ($Q_{F,wet}$ (mm d^{-1})) can be expected to exceed the total flow from wetlands ($Q_{R,wet}$ (mm d^{-1})) as both landscape units are connected to the stream for most of the time but hillslopes occupy around 90% of the catchment area, while wetlands only account for 10%. Constraint C8 was applied in M10–M11:

$$\sum_{t=1}^{T} (1-f) Q_{F,wet}(t) > \sum_{t=1}^{T} f Q_{R,wet}(t). \tag{10}$$

Note that the above presented *prior constraints* are mere suggestions. Additional and/or different constraints may be applicable in other catchments, depending on the modeler's understanding of the system and available data, leaving the necessity of choices on an ad hoc basis [e.g., *Ambroise et al.*, 1996a, 1996b; *Seibert and McDonnell*, 2002; *Hughes*, 2013]. It is also emphasized that prior information is not only applicable in natural systems, but potentially even more so in anthropogenically influenced environments, where more information about disturbances may be available. Note that the setup of the study mainly aims at reducing the probability of Type I errors (false positives). Due to the conservative formulation of constraints and the modular modeling progression (i.e., more complex models encapsulate essentially all processes of less complex models), the method is not expected to increase the probability of Type II errors (false negatives).

3.3. Model Evaluation

3.3.1. Performance Metrics

The ability of the model to reproduce the time series of flow and several other catchment signatures was quantified by different metrics, such as the Nash-Sutcliffe efficiency (E_{NS}) [*Nash and Sutcliffe*, 1970], the volume error (E_V) [*Criss and Winston*, 2008], and the relative error (E_R) [e.g., *Euser et al.*, 2013], depending on the type of signature, as given in Table 3. As a more global summary metric in the presence of model uncertainty, the Continuous Rank Probability Score (Φ_{CRPS}) [*Hersbach*, 2000; *Fenicia et al.*, 2013], a distance measure between observations and predictive distributions, reflecting both precision and reliability of a model, was additionally used:

$$\Phi_{CRPS} = \frac{1}{N_t} \sum_{n=1}^{N_t} \int_{-\infty}^{\infty} \left| F_n(x_m) - H_n\{x_m \geq x_o\} \right| dx, \tag{11}$$

where N_t is the number of observations, x_o and x_m are the observed and modeled variables (e.g., Q), respectively, $F_n(x_m)$ is the cumulative distribution function of the model predictions at (time) step n, and H_n is the Heaviside step function with $H_n = 1$ if $x_m \geq x_o$ and 0 otherwise. For a perfect model fit, Φ_{CRPS} reduces to 0 and for deterministic predictions it equals the mean absolute error. Furthermore, assessment of model

Table 3. Hydrological Variables and Signatures With the Corresponding Performance Metrics Used for Calibration and Postcalibration Model Evaluation[a]

	Variable/Signature	Abbreviation	ID	Performance Metric	Reference
Calibration	Time series of flow	Q	O1	$E_{NS,Q}$	*Nash and Sutcliffe* [1970]
			O2	$E_{NS,log(Q)}$	
			O3	$E_{V,Q}$	*Criss and Winston* [2008]
	Flow duration curve	FDC	O4	$E_{NS,FDC}$	*Jothityangkoon et al.* [2001]
				$D_{E,cal}$	*Schoups et al.* [2005]
Evaluation	Flow during low flow period	Qlow	S1	$E_{NS,Q,low}$	*Freer et al.* [2003]
	Groundwater dynamics[b]	GW	S2	$E_{NS,GW}$	*Fenicia et al.* [2008a]
	Flow duration curve low flow period	FDC,low	S3	$E_{NS,FDC,low}$	*Yilmaz et al.* [2008]
	Flow duration curve high flow period	FDC,high	S4	$E_{NS,FDC,high}$	*Yilmaz et al.* [2008]
	Groundwater duration curve[b]	GDC	S5	$E_{NS,GDC}$	
	Peak distribution	PD	S6	$E_{NS,PD}$	*Euser et al.* [2013]
	Peak distribution low flow period	PD,low	S7	$E_{NS,PD,low}$	*Euser et al.* [2013]
	Rising limb density	RLD	S8	$E_{R,RLD}$	*Shamir et al.* [2005]
	Declining limb density	DLD	S9	$E_{R,DLD}$	*Sawicz et al.* [2011]
	Autocorrelation function of flow[c]	AC	S10	$E_{NS,AC}$	*Montanari and Toth* [2007]
	Lag-1 autocorrelation of high flow period	AC1, Q10	S11	$E_{R,AC1,Q10}$	*Euser et al.* [2013]
	Lag-1 autocorrelation of low flow period	AC1,low	S12	$E_{R,AC1,low}$	*Euser et al.* [2013]
	Runoff coefficient[d]	RC	S13	$E_{R,RC}$	*Yadav et al.* [2007]
				D_E	*Schoups et al.* [2005]

[a]The performance metrics include the Nash-Sutcliffe efficiency (E_{NS}), the volume error (E_V), and the relative error (E_R). For all variables and signatures, except for Q, Qlow, and GW, the long-term averages were used.

[b]Averaged and normalized time series data of the 5 piezometer were compared to normalized fluctuations in model state variable S_S.

[c]Describing the spectral properties of a signal and thus the memory of the system, the observed and modeled autocorrelation functions with lags from 1 to 100d were compared.

[d]Note that in catchments without long-term storage changes and intercatchment groundwater flow, long-term average RC equals the long-term average 1-EA.

overall performance was gauged by the mean Euclidean distance D_E from the "perfect" model based on equally weighted performance metrics [e.g., *Schoups et al.*, 2005; *Hrachowitz et al.*, 2013b]:

$$D_E = \sqrt{\frac{\sum_{n=1}^{N}(1-E_n)^2}{N}}, \tag{12}$$

where E_n is the performance metric (E_{NS}, E_V, E_R) of the nth variable or signature and N is their complete number. Note that in the following "low flow" refers to the period 1 June to 31 October and "high flow" refers to the period 1 December to 28 February.

3.3.2. Model Calibration

Each model setup (M1–M11), run on daily time steps, was after a 1 year warm-up period calibrated for the period 1 October 2002 to 30 November 2007 using a Monte-Carlo sampling strategy (10^7 realizations), based on multiobjective evaluation [e.g., *Gupta et al.*, 1998], i.e., *modeling objective constraint*, in an attempt to identify sufficiently consistent parameterizations. The four calibration objective functions (O1–O4) are given in Table 3 and the uninformed prior parameter distributions are shown in Table 4. Considering that mathematically (pareto) optimal parameter sets are unlikely to be the hydrologically most suitable ones [e.g., *Beven*, 2006], all parameter sets that fell within the space spanned by the four-dimensional pareto fronts, approximated by the cloud of sample points, were retained as feasible. Note that here the objective functions were equally weighted as models that can best reproduce the overall system dynamics were sought. To construct model uncertainty intervals, the feasible parameter sets where thereafter weighted according to a likelihood measure $L = D_{E,cal}^{p}$ [cf. *Freer et al.*, 1996], where the exponent was set to $p = 10$ to emphasize models with good overall calibration performance and where $D_{E,cal}$ is the mean Euclidean distance (equation (12)) to the perfect model with respect to the four calibration objectives O1–O4.

3.3.3. Posterior Model Evaluation

The tested models were not only evaluated against their respective skill to predict the system response with respect to the calibration objectives during an independent test period, thereafter referred to as validation period (1 October 2007 to 30 November 2012). Rather, a range of catchment signatures (S1–S13), as described in Table 3, was used in a multiobjective/criteria posterior evaluation strategy based on the mean

Table 4. Prior Distributions and 5/95th Percentiles of Posterior Distributions for M1–M11[a]

		C_P	f	k_F (d^{-1})	k_L (d^{-1})	k_R (d^{-1})	k_S (d^{-1})	L_P (−)	$Q_{L,const}$ (mm d^{-1})	P_{max} (mm d^{-1})	$S_{S,p,max}$ (mm)	$S_{Umax,H}$ (mm)	$S_{Umax,R}$ (mm)	β_H	β_R
Prior distribution		0–1	0.1	0.025–1	0.0001–0.001	0.05–2	0.001–0.05	0–1	0.37	0–4	0–2000	0–1500	0–750	0–100	0–2
Posterior distribution	M1	0.12/0.63		0.042/0.094			0.031/0.049	0.00/0.07		0.03/0.29		637/1446		10.5/61.5	
	M2	0.08/0.62		0.039/0.175	0.027/0.098		0.031/0.047	0.00/0.15		0.21/3.19		569/1401		10.6/70.1	
	M3	0.10/0.59		0.046/0.500	0.029/0.0968		0.030/0.049	0.01/0.28		0.20/3.19		608/1427		4.4/68.2	
	M4	0.14/0.94		0.060/0.906	0.0002/0.0009		0.032/0.049	0.04/0.32		0.14/2.87	309/1803	706/1466		3.4/51.4	
	M5	0.15/0.56		0.055/0.648	0.0002/0.0009		0.031/0.049	0.05/0.32		0.18/2.02	361/1744	712/1461		3.1/35.5	
	M6	0.14/0.55		0.054/0.627			0.041[b]	0.05/0.34	0.37[b]	0.27/1.98		722/1461		2.4/36.9	
	M7	0.21/0.73	0.1[b]	0.052/0.674		0.336/1.891	0.041[b]	0.01/0.20	0.37[b]	0.19/2.37		638/1444		6.3/69.0	
	M8	0.18/0.66	0.1[b]	0.054/0.666		0.352/1.894	0.041[b]	0.04/0.28	0.37[b]	0.24/2.40		689/1444	88/720	4.3/66.9	
	M9	0.15/0.63	0.1[b]	0.054/0.604		0.324/1.894	0.041[b]	0.04/0.27	0.37[b]	0.32/2.27		677/1439	128/731	11.7/69.4	
	M10	0.15/0.64	0.1[b]	0.054/0.619		0.333/1.863	0.041[b]	0.04/0.27	0.37[b]	0.34/2.29		686/1442	132/725	13.6/69.7	
	M11	0.19/0.64	0.1[b]	0.054/0.466		0.318/1.857	0.041[b]	0.04/0.27	0.37[b]	0.29/2.18		683/1444	120/730	13.0/69.2	0.13/1.86
	M11a	0.27/0.98	0.1[b]	0.055/0.904		0.113/1.868	0.041[b]	0.02/0.40	0.37[b]	0.15/2.95		502/1429	71/721	7.2/69.8	0.13/1.91
	M11b	0.00/0.50	0.05/0.15	0.043/0.167		0.283/1.891	0.022/0.044	0.06/0.34	0.23/0.57	0.22/1.53		737/1462	110/736	17.8/73.1	0.10/1.86

[a]A list of symbols is given in the notation section.
[b]Fixed parameter values.

Euclidean distance D_E from the "perfect" model based on *all* equally weighted performance metrics (O1–O4; S1–S13), under the assumption that the more signatures can be adequately reproduced by a given model, the higher the confidence in this model being an adequate representation of the observed response dynamics in a catchment. Besides the distributions of D_E themselves, $\Phi_{CRPS,DE}$ was used as a robust distance measure to allow a clearer and mathematically more rigorous assessment of the overall performance differences for the individual models M1–M11. To further support decisions if one model is to be rejected in favor of another one, Wilcoxon Rank Sum Tests tested if the distributions of D_E of the individual models were significantly different from each other.

Further, the combined modeled consistency and performance was evaluated using the Framework to Assess the Realism of Models (FARM), introduced by *Euser et al*. [2013]. Briefly, FARM is based on principal component analysis for reduction of the 17 performance measure dimensions in this study, i.e., O1–O4 and S1–S13, to a more instructive two-dimensional representation. Specifically, the consistency of a model, i.e., the degree of direct correlation between values of the used performance measures and thereby a model's ability to reproduce several performance measures simultaneously, increases the more loading vectors (representing one signature each) point into the positive direction of PC1. In contrast, model performance is represented by scaling the loadings according to the median Euclidean distance to the perfect model D_E. The more observed response dynamics a model can reproduce and the better these individual performances, the smaller the plot and the smaller the spread between the loadings.

Note that in principle for all models, starting from the least complex one (M1), all *prior constraints* could have been simultaneously imposed and the models could have been calibrated to the full set of N performance metrics in an extended multiobjective strategy, i.e., a stricter *modeling objective constraint*. However, here we test how additional pieces of information can help to increase model consistency, i.e., to better reproduce the overall system response, characterized by the ensemble of signatures, in a stepwise a posterior evaluation.

4. Results and Discussion

4.1. Calibrated and Constrained Models

4.1.1. Benchmark Model

For the calibration period, the baseline model setup M1, quite well reproduced the general features of the hydrograph, while somewhat overestimating the amount of water in the system during wet conditions. Both the most balanced solution ($D_{E,cal,min}$) and the median of all retained feasible solutions ($D_{E,cal,med}$) with respect to the four calibration objectives (O1–O4) reached relatively high values (Figures 4a and 5). Inspection of the validation period, however, revealed a substantial overestimation of flow in wet periods for all retained models, resulting in a considerable decrease of all calibration objective functions during validation as shown in Figures 4a and 5. Thus, in spite of the reasonable performance during calibration, pointing

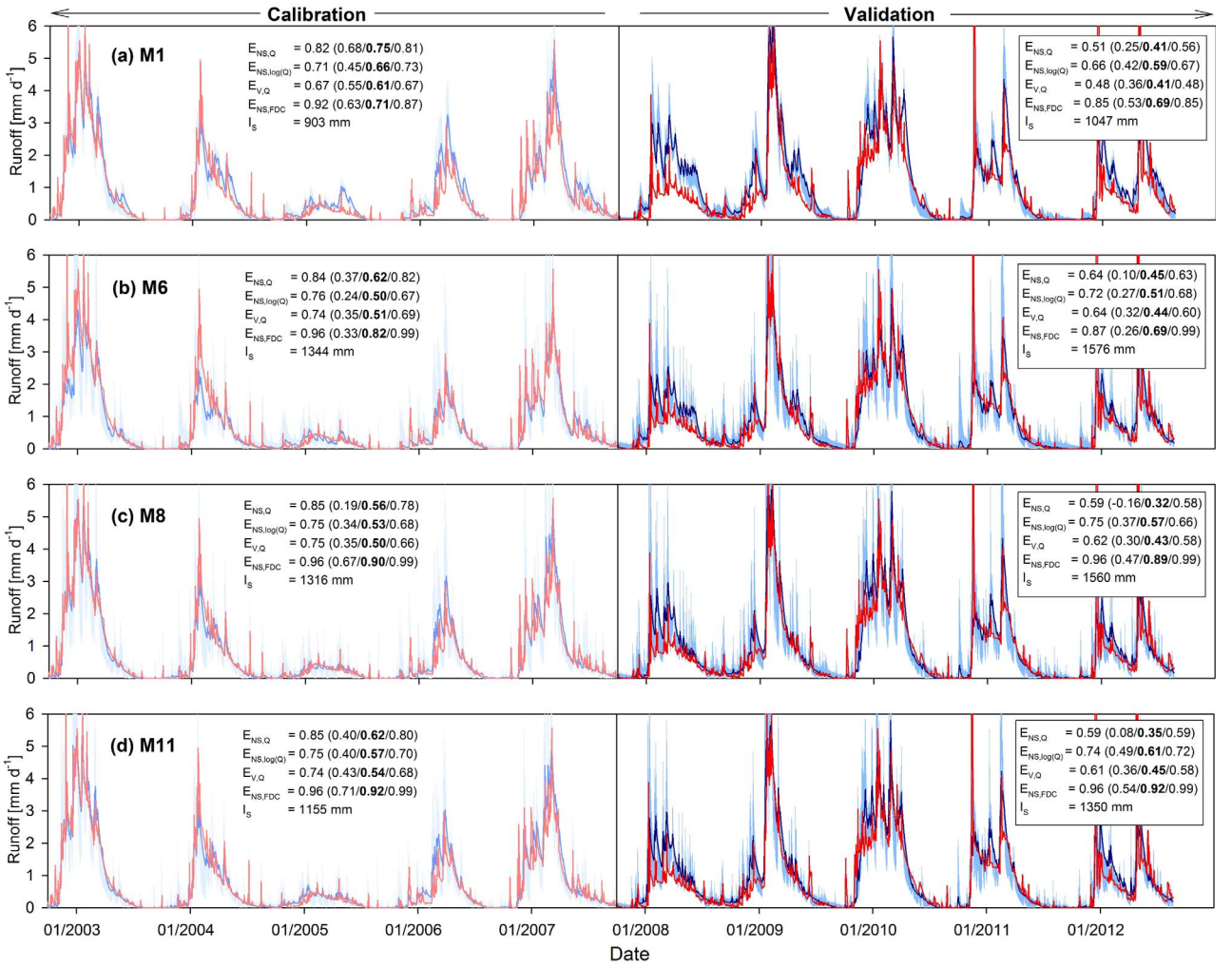


Figure 4. Observed (red line) and modeled runoff for model setups (a) M1, (b) M6, (c) M8, and (d) M11 in calibration and validation periods. Modeled runoff shown as most balanced solution (dark blue line) and the 5/95th uncertainty bounds (light blue shaded area). Values of the calibration objective functions $E_{NS,Q}$ (O1), $E_{NS,log(Q)}$ (O2), $E_{V,Q}$ (O3), and $E_{NS,FDC}$ (O4) in calibration and validation period are the performance of the most balanced solution and in brackets the 5/50 (bold)/95th percentiles of all retained feasible solutions. Model uncertainty is estimated by the area I_U spanned by the 5/95th percentiles of modeled runoff over the entire calibration and validation periods, respectively.

toward a well constrained, adequate and robust model, the validation results indicated serious limitations in the adequacy of M1. Analysis of the modeled catchment signatures clearly supported the assessment that M1 cannot be considered consistent (Figure 6). Some signatures could only be reproduced for either the calibration *or* the validation period (S4, S7, S8, and S10–S13; Figures 5 and 6a), highlighting their sensitivity to the distinct hydrometeorological conditions in these periods and the model's inability to sufficiently respond to these variations. Other signatures could be reproduced for neither period (S2–S3, S5, and S9), indicating a general deficiency of M1 to represent the processes controlling these signatures. This is also illustrated by a largely inadequate reproduction of the observed groundwater dynamics (Figure 7a): while the general dynamics are captured, the observed variance of the groundwater response substantially exceeds the modeled variance. The results suggest that inadequate process representations in M1 result in model internal dynamics that do not adequately reflect the system response dynamics but which are rather mere manifestations of the fitting process. Similar performances of the three benchmark models (NWS Sacramento, HBV, and HyMod; supporting information Figure S2) indicate little sensitivity of the analysis to the initial model choice. Following these results, model hypothesis M1 had to be rejected. Note that in the following assessment of the overall performance of the individual model setups is for brevity restricted to the median of all retained feasible solutions ($E_{NS,med}$, $E_{R,med}$, $D_{E,med}$). Due to the functionally equivalent results obtained, Φ_{CPRS}, provided for selected signatures in supporting information Figure S1, will not be individually discussed.

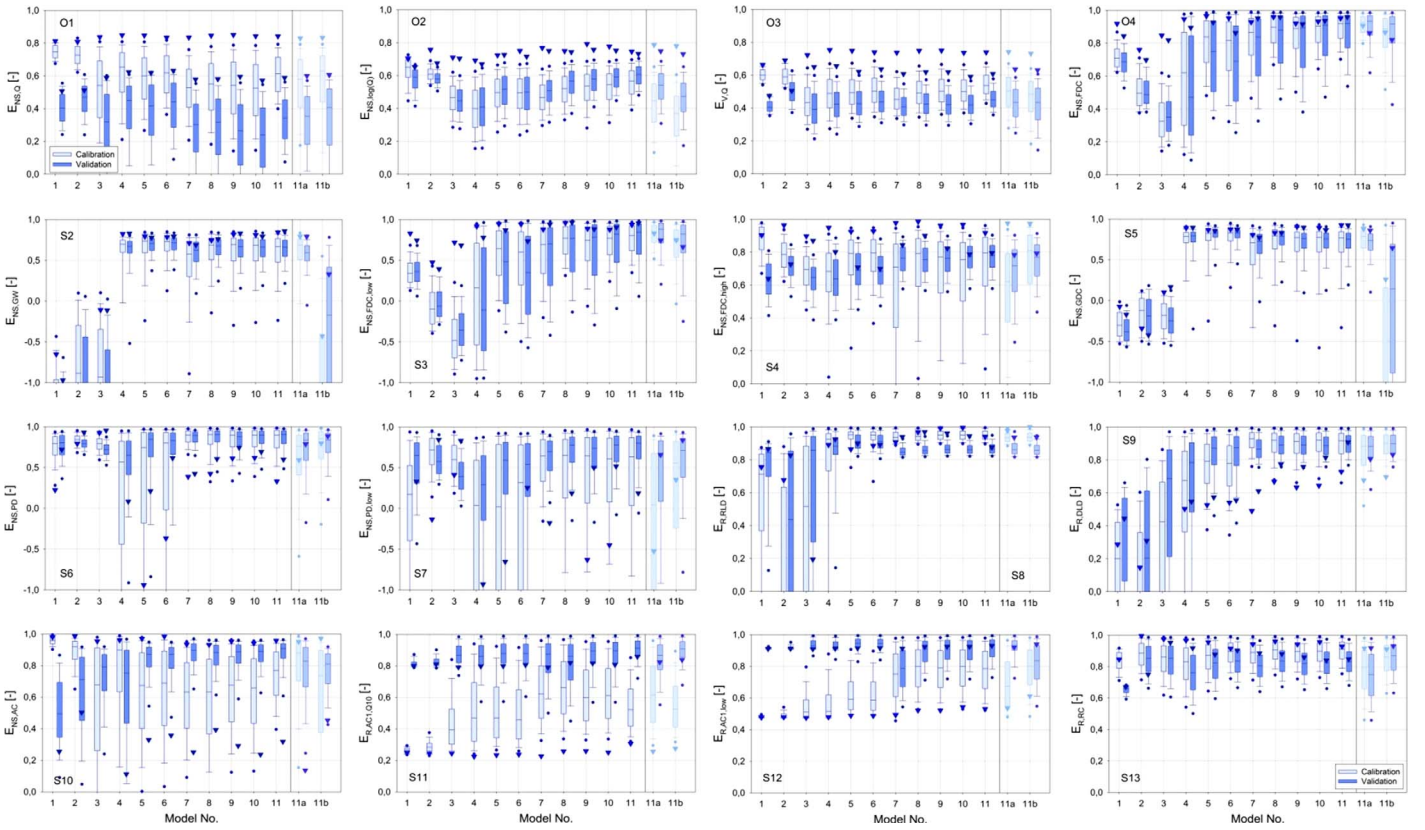


Figure 5. Performance of the most balanced solutions (triangles) and the complete sets of all feasible solutions (box plots; the dots indicate 5/95th percentiles, the whiskers 10/90th percentiles, and the horizontal middle line the median) with respect to the calibration objectives (O1–O4) and the catchment signatures (S2–S13) for calibration and validation periods for all model setups M1–M11.

4.1.2. Deep Infiltration Loss Constraints

Model setup M2, allowing for deep infiltration losses, resulted in a comparable calibration performance as M1 (Figure 5). The uncertainty in modeled stream flow, estimated as the total area spanned by the 5/95th percentiles of the uncertainty interval, however increased in M2 as a result of the additional free calibration parameter (k_L; Figure 8a). Although M2 could better reproduce the long-term annual runoff coefficient (S13), due to a more plausible partitioning of flows, as well as some low flow and groundwater related signatures (S2 and S5–S7), the overall model skill, as indicated by $D_{E,med}$ and $\Phi_{CPRS,DE}$, did not improve (Figure 8b).

Applying *prior constraints* C4 and C5, i.e., estimates of average annual actual evaporation and deep infiltration, model M3 experienced a deterioration in overall calibration performance (O1–O4 in Figure 5). This indicates that a range of well-performing parameterizations of the previous model (M2) were clear misrepresentations of the system as they could not reproduce the system response within the rather wide limits of acceptability of C4 and C5. This was also reflected in the modest overall skill of the model to reproduce other signatures (Figures 5 and 8b) except for S8 and S9 that were captured significantly better than in the preceding models.

As a result of their poor performances, M2 and M3 were therefore rejected. In M4, one additional free calibration parameter was added. The inclusion of a threshold in the groundwater reservoir proved highly beneficial for the model's ability to better reproduce the majority of signatures in this study (Figures 5 and 8b). Considerable improvements where in particular observed for groundwater related signatures (S2 and S5) as well as for the flow duration curves (O4 and S3–S4), all of which pointing toward the importance of the new threshold to adequately reproduce the system response. Yet as in the preceding models, not all signatures improved. The ability of M4 to reproduce O2 as well as S6–S7 decreased, underlining the distinct sensitivities of different parts of the system to different process representations. Owing to the relatively high

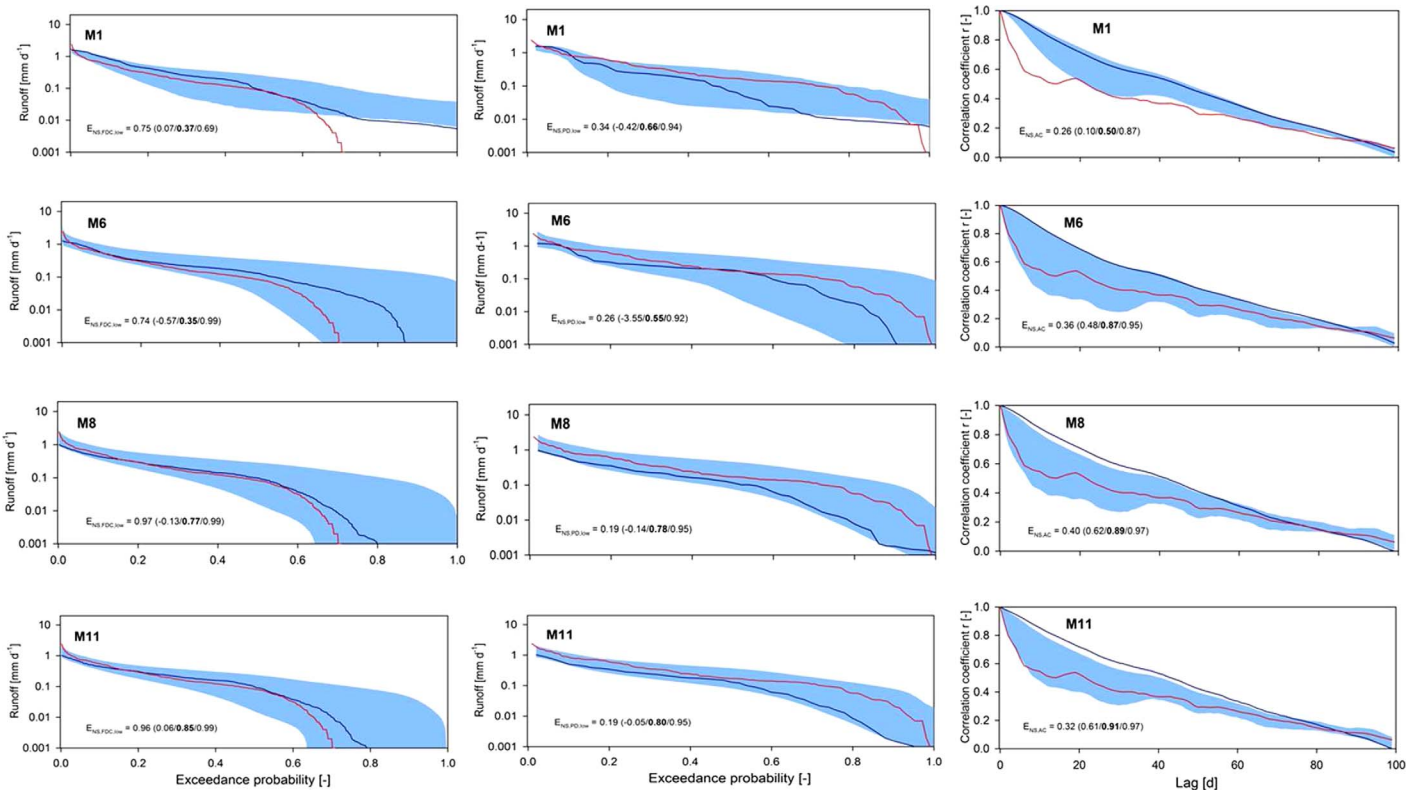


Figure 6. Selected signatures—FDC,low (S3; left), PD,low (S7; middle) and AC (S10; right)—for model setups (a) M1, (b) M6, (c) M8, and (d) M11 in the validation period. The red lines indicate signatures constructed from observed data, the dark blue lines represent the signature generated from the most balanced model solution, and the blue shaded area highlights the 5/95th uncertainty interval of the modeled signature. The performance metrics given are the most balanced solution and in brackets the 5/50 (bold)/95th percentiles of all retained feasible solutions.

number of free calibration parameters in M4, the model uncertainty for stream flow remains comparatively high. As discussed above, a high number of parameters can often result in good model performance, yet this is frequently offset by loss of internal consistency in the model. Closer inspection of model internal fluxes in M4 revealed that many parameter sets resulted either in the groundwater flux being effectively shut down for much of the time, i.e., when the modeled recharge was too low compared to the combined fluxes Q_S and Q_L or in an excessively high groundwater long-term average contribution Q_S (up to 98%) to stream flow Q_T. To limit such misrepresentations in favor of expert-based and empirical understanding of the system, constraint C6 was applied in the model setup M5.

In contrast to the previous set of *prior constraints* (C4 and C5), the additional application of C6 in M5 did result in significant improvements for virtually all signatures (Figure 5) and thus also for overall model performance (Figure 8b) both, during calibration and validation periods. Together with the significant reduction in uncertainty (Figure 8a), this clearly highlights the value and effectiveness of additional *prior constraints* in filtering out parameterizations that are mere artifacts of the calibration process.

In an attempt to keep the numbers of free calibration parameters to a minimum to preserve model parsimony, i.e., "as complex as necessary and as simple as possible," setting the deep infiltration loss rate Q_L to a constant value, as inferred from groundwater observations (see section 3.1.2.1) allowed the reduction of free calibration parameters from nine to six in M6. Like for M5, it was observed in M6 that although the calibration objective functions (O1–O4) did improve with respect to the preceding model setup, they could not quite reach the level of those of M1 (Figures 4b and 5). Further, low flow related signatures (S3, S7, S9, and S12; Figure 6b) still demonstrated relatively poor performance levels and considerable uncertainties. However, the overall skill of M6 to reproduce the ensemble of catchment signatures clearly exceeded the skill of the benchmark model M1 (Figure 8). This is also illustrated by the model's ability to more satisfactorily mimic the no-flow conditions related to the dynamics of the observed groundwater fluctuation as

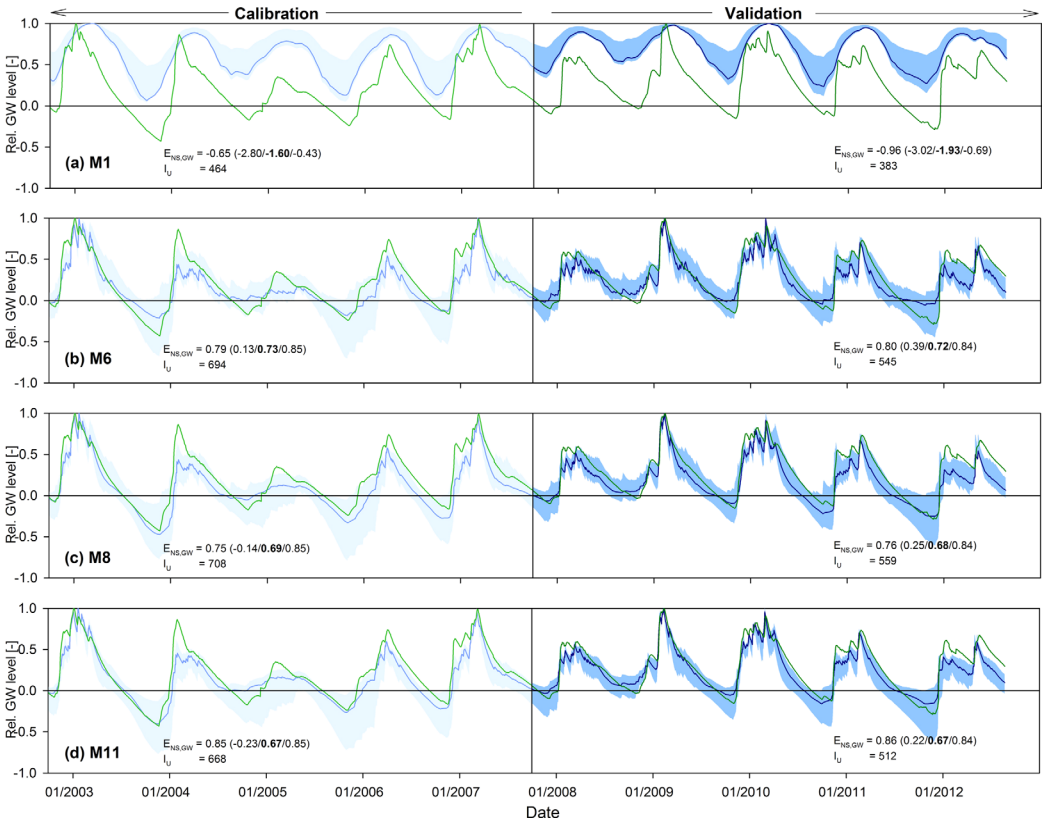


Figure 7. Observed (green line) and modeled average relative groundwater levels for model setups (a) M1, (b) M6, (c) M8, and (d) M11 in calibration and validation periods. Modeled relative groundwater levels are shown as most balanced solution (dark blue line) and the 5/95th uncertainty bounds (light blue shaded area). Values of the calibration objective functions $E_{NS,GW}$ (S2) in calibration and validation period are the performance of the most balanced solution and in brackets the 5/50 (bold)/95th percentiles of all retained feasible solutions. Uncertainty in the modeled relative groundwater level is estimated by the area I_U spanned by the 5/95th percentiles of modeled runoff over the entire calibration and validation periods, respectively.

compared to M1–M5 (Figures 5 and 7b). The results of M6 further suggested that the reduction of parameters was justified as it did not result in significant changes in the models ability to reproduce the overall system response in comparison to M5 (Figures 5 and 8b). In addition, the reduction of free calibration parameters slightly reduced model uncertainty (Figure 8a). Following these results, model setup M6 was used as the basis for further model development.

4.1.3. Riparian Zone Constraints

In M7, it was tested whether the inclusion of a separate model component representing the dynamics of the riparian zone, formulated as a simple linear reservoir, can improve the so far unsuitably reproduced response of the system to rain events during otherwise relatively dry periods, as indicated by, e.g., S3, S7, S9, or S12. It was observed that although a few signatures generated by M7 exhibited significantly improved performances (e.g., S3, S6, S7, and S9; Figure 5), many other signatures were characterized by considerable performance reductions (e.g., O1, O3, S2, and S5). On balance not significantly improving overall model skill compared to M6 (Figure 8b). The model setup M7 was therefore rejected as feasible model *architecture constraint*.

In M8, the representation of the riparian component in the model was adapted by adding an unsaturated zone. Together with an additional *prior constraint* (C3), M8 showed slightly increased performance with respect to O1–O4, in both calibration and validation period (Figures 4c and 5). In spite of two additional free calibration parameters as compared to M6, M8 was characterized by slightly reduced model uncertainty (Figures 4c and 8a). In addition to the clear improvement of the model's ability to reproduce S12, other signatures could be reproduced at similar performance levels as before but within reduced uncertainty (e.g., S3–S5 and S10; Figures 5, 6c, and 7c). It was found that as a result M8 exhibited a clearly

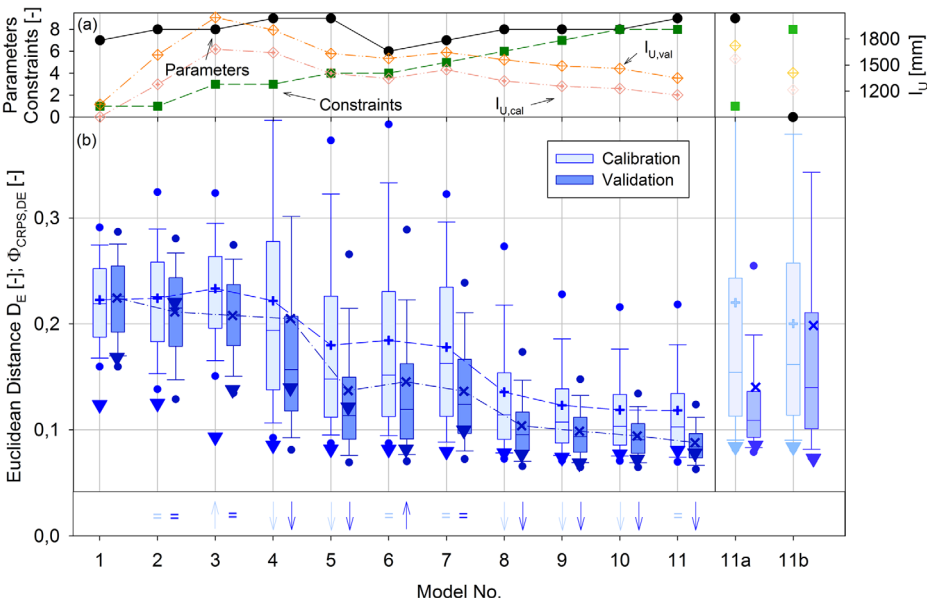


Figure 8. (a) Number of free calibration parameters (solid black line), number of *prior constraints* (green dashed line), and the area spanned by the 5/95th percentile uncertainty interval I_U for calibration (pink dash-dotted line) and validation periods (orange dash-dotted line) for model setups M1–M11. (b) Overall model performance for all model setups (M1–M11) expressed as Euclidean distance D_E from the "perfect model" computed from all calibration objectives (O1–O4) and signatures (S1–S13) with respect to calibration and validation periods. Triangles represent the most balanced solution, i.e., the solution obtained from the parameter set with the lowest Euclidean distance during calibration. Box plots represent the Euclidean distance for the complete sets of all feasible solutions (the dots indicate 5/95th percentiles, the whiskers 10/90th percentiles, and the horizontal middle line the median). The plus and cross symbols represent the $\Phi_{CPRS,DE}$ of the Euclidean distances of all feasible models with respect to the "perfect model." The results of the Wilcoxon Rank Sum Tests for pairwise comparisons of subsequent distributions of D_E (i.e., M1–M2, M2–M3, etc.) indicate if the distributions are significantly different from each other (↑, ↓) or not (=).

improved overall skill within narrower performance bounds in comparison to all preceding model setups in particular with respect to the validation period (Figure 8b).

Adding two further *prior constraints* (C7 and C8) in M9 and M10, respectively, slightly increased the models ability to reproduce objective function O2 as well as signatures S3 and S10–S11 in particular for the validation period by discarding parameter sets that could not simultaneously satisfy all *prior constraints*, thereby also gradually reducing model uncertainty in M9 and M10 (Figure 8a).

Replacing the linear flow generation mechanism in the unsaturated zone from the preceding model by a nonlinear mechanism, M11, with a total of nine free calibration parameters and eight *prior constraints*, was the most complex model structure tested in this study. The further increased ability of the model to reproduce relevant signatures (S3–S4 and S10; Figures 5 and 6d) and an significantly increased overall performance of M11 with respect to *all* performance measures (i.e., O1–O4 and S1–S13; Figure 8b) together with reduced uncertainty (Figures 4d and 8a) compared to M1–M10 give evidence that M11 is a clearly more adequate representation of the system response than the simpler models M1–M10 in spite of exhibiting essentially the same calibration performance (O1–O4; Figures 4d and 5).

Inspection of PCA plots generated within the FARM framework [*Euser et al.*, 2013] corroborates these findings. M1 (Figure 9a) is characterized by a star-shaped spread of the objective functions and signatures. This indicates that this model is not capable to simultaneously reproduce various signatures with the same parameter set. For example, it can be seen that parameters that result in high values for O1 show low values for O4 as they are plotting on roughly opposite ends of PC2, i.e., they are inversely correlated. The same is true, among others, for O2–S8/S9, O4/S3–O3/S13, or S4–S10/11/12. The PCA of M11's signatures, on the other hand, shows a clearly reduced spread, indicating a higher degree of model consistency, with only S10 showing significant loadings in the second and third quadrant Figure 9b. Conversely, O2 was now found to be directly correlated with S8/S9, while O4/S3 were uncorrelated with O3/S13 rather than inversely correlated as in M1. Similarly, S4 appeared to become more consistent with S11/12.

Note that M11 itself is clearly not to be seen as endpoint in the modeling efforts for this catchment, but rather as the, so far, best performing hypothesis of catchment function that was not yet falsified. Although

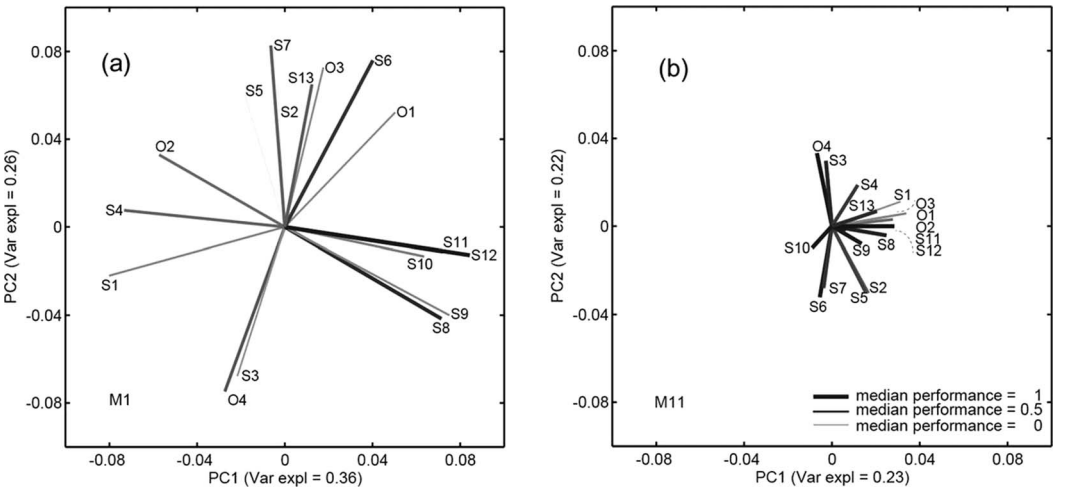


Figure 9. Overall performance-scaled PCA plots for (a) M1 and (b) M11 used to assess model realism in within the FARM framework [*Euser et al.*, 2013]. Analysis based on all objective functions (O1–O4) and signatures (S1–S13) of all solutions retained as feasible. Thin, light shaded lines indicate low average performances with respect to the individual signatures while dark, thick lines indicate high average performances.

it exhibits considerably higher process consistency than M1–M10, some performance metrics, such as O1 or S7, still offer some room for future improvement. Further model tests were not attempted here as the actual objective of this study was to demonstrate the value of anecdotal information to formulate *prior constraints* to bridge the gap between simplistic process representation and predictive uncertainty in more complex representations of the system, caused by model equifinality in an illustrative example. In spite of differences in the absolute performances of the three benchmark models whose complexity was adapted equivalently to M11, the similar pattern of performance changes from simple to more complex as compared to the progression of M1–M11 (supporting information Figure S2) indicates a certain level of general value of the suggested modeling approach, irrespective of the model used.

4.2. Calibrated but Unconstrained Model

In order to get a better understanding of the respective values of the *parameter* and *process constraints* used in this study, M11 was further evaluated in M11a both without these *prior constraints* and with each of them individually applied. Although the most balanced calibration and validation solutions remained in the range of what was found for M11 (Figure 8 and supporting information Figure S3a), the removal of all *prior constraints* resulted in a substantially reduced overall performance of M11a. This was caused by the inclusion of parameter sets that result in acceptable calibration performances but do not satisfy one or more of the previously applied *prior constraints*. It was also observed that these additional solutions, as unsuitable representations of the observed response dynamics and thus artifacts of the calibration process, tend to generate relatively low performance solutions for at least some of the signatures S1–S13. This further highlights the effectiveness of the combined use of signatures and *prior constraints*. In addition to an unconstrained model's inability to identify and discard a wide range of unsuitable parameter sets due to its ill-posed nature, omitting *prior constraints* also significantly increases model uncertainty as shown in Figure 10a, thereby negatively affecting its predictive capability.

To assess the impact of individual *prior constraints* on the model, C1–C8 were subsequently added one by one to M11a. It was found that no single *prior constraint* could significantly improve the overall model performance in both calibration and validation period (Figure 10). Thus, being dependent on the cumulative effects of all *prior constraints* applied, a significant effect will only be achieved once a certain number of *prior constraints* is applied. This is also supported by the results of *Winsemius et al.* [2009], who, applying three constraints, only obtained weak constraining power. Further evidence for this hypothesis is that although the individual reductions of model uncertainty ΔI_U due to C1–C8 are low (Figure 10), the sum of ΔI_U ($\sim$390 mm) approaches the difference of uncertainty between M11 and M11a ($\sim$420 mm). Note that the 30 mm difference is caused by the individual calibration runs for M11 and M11a.

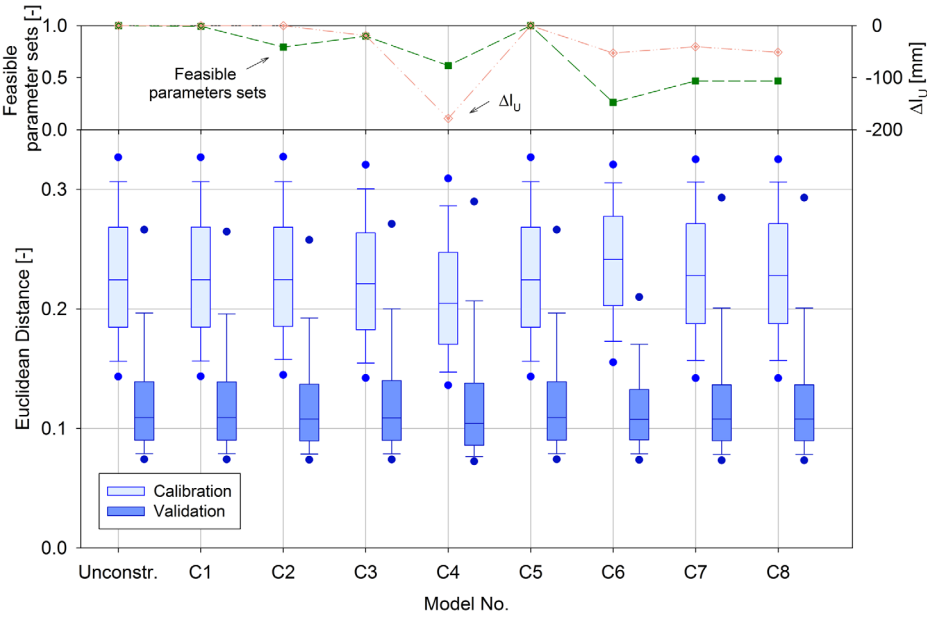


Figure 10. Evaluation of the influence of individual realism constraints on a calibrated but initially unconstrained model M11a. Parameter sets retained as feasible for each individual constraint (C1–C8) as proportion of the full set of feasible parameters obtained from calibrated but unconstrained M11a (green dashed line) and the change of the 5/95th uncertainty interval ΔI_U in the modeled hydrograph with respect to calibrated but unconstrained setup M11a (pink dash-dotted line). Box plots illustrate the influence of the individual realism constraints on the overall model performance (Euclidean distance to the perfect model) for calibrated but unconstrained M11a (see also Figure 8) and realism constraints C1–C8 for calibration and validation periods (the dots indicate 5/95th percentiles, the whiskers 10/90th percentiles and the horizontal middle line the median). Note that, being redundant with the use of $Q_{L,const}$, C5 has no effect in M11a.

In general, it became apparent that in this study *parameter constraints* (C1–C3) were less effective than *process constraints* (C4–C8). This is true for the reduction of feasible parameter sets but also for the reduction of uncertainty. While individual *parameter constraints* in this study could only reduce feasible parameter sets by up to ~20% (Figure 10), individual *process constraints* reduced the feasible parameter sets by up to ~65% (C6; Figure 10). This disparity becomes even more apparent for the reduction in uncertainty: *parameter constraints* reduce uncertainty by a total of ~40 mm, while *process constraints* are responsible for the remainder of ~350 mm (Figure 10). The by far most effective individual *prior constraint* in this study was C6, i.e., bounds on base flow contribution, underlining the importance of estimates of internal fluxes on model internal process consistency. Note that in more complex model formulations, where more *parameter constraints* can be applied [e.g., *Gao et al.*, 2014; *Gharari et al.*, 2013b], *parameter constraints* may become more important.

4.3. Uncalibrated but Constrained Model

As a final step to assess the value of *prior constraints* in complex models, model setup M11 was applied in an uncalibrated but constrained way. That is, feasible parameterizations in M11b were chosen exclusively on a rejectionist basis defined by the full set of *prior constraints*, i.e., all parameter sets that satisfied C1–C8 were retained as feasible.

The results were very encouraging in particular for predictions in ungauged basins (PUB) [e.g., *Sivapalan et al.*, 2003; *Hrachowitz et al.*, 2013a]: the overall performance of the uncalibrated but constrained complex model M11b clearly outperforms the relatively simple calibrated setups M1–M4 (Figure 8b and supporting information Figure S3b). Although model uncertainty is slightly higher for M11b than for benchmark setup M1, it is similar to that of M11, a calibrated and constrained setup, for both calibration and validation periods (Figure 8a). In other words, when little or unreliable calibration data are available, the use of a complex uncalibrated model setup in conjunction with a sufficient degree of prior information, expressed as prior *constraints* can prove valuable in comparison to a simple calibrated model.

5. Wider Implications

The results of this study in a small catchment do potentially have wider implications. It is frequently claimed that for many regions in the world not enough data are available to warrant the use of more complex

model structures than those of standard lumped conceptual models, arguably preventing the use of potentially more plausible process representations. This is in principle true when considering standard calibration techniques without efficient use of *prior constraints*. As it was shown here, although an unconstrained higher complexity model (M11a) can outperform a simple lumped model (M1), this comes at the price of substantially increased model uncertainty (Figure 8a). However, evaluating models not only on the basis of standard performance metrics but also on a wide variety of catchment signatures revealed that many models that perform well during the calibration period are effectively artifacts of the calibration process as they cannot reproduce a range of other signatures. It was shown that many models cannot simultaneously reproduce different aspects of the desired response variable when the model architecture is unsuitable for a given catchment and/or the feasible model parameter space is insufficiently constrained. Thus, the use of multiple hydrological signatures allowed for a falsification of models otherwise considered feasible, and it illustrated the importance of a suitable representation of the system response dynamics as a whole rather than the focus on one specific variable (here: the hydrograph). The use of very simple, anecdotal and/or expert-knowledge-based *prior constraints* subsequently proved critical for identifying and discarding parameterizations that resulted in unsuitable models, allowing that at least the following holds true: "[. . .] the best we can hope for is to demonstrate that the model does not violate our theoretical understanding of the system and is consistent with the available data [. . .]" [*Knutti*, 2008, p. 4651].

By following the stepwise development of the model and observing the interplay between model structure, *prior constraints*, and signatures, it became further clear that in the study catchment, increasingly complex model structures, i.e., *architectural and parameterization constraints*, are responsible for improving average model performances as illustrated by the performance jumps in M4, M8, and M11 (Figure 8b). Conversely, the choice of feasible parameter sets, controlled by *modeling objective* and *prior constraints* are rather linked to reducing the performance spread by discarding parameterizations at the low performance tail end of the performance distributions, i.e., the ill-posedness of the problem (M4–M11; Figure 8b).

In contrast to the prevailing notion that more complex models also entail increased model uncertainty, the results of this study suggest, in correspondence with the results of *Gharari et al*. [2013b], that simple, semi-quantitative prior *constraints* are highly valuable, arguably even necessary tools to reduce uncertainty of complex models, thereby allowing for increased model consistency within limited ranges of uncertainty. Depending on the model complexity and the choice of *prior constraints*, it was also shown that, in the study catchment, a complex uncalibrated but constrained model can on average reach the performance levels of calibrated but unconstrained simple model formulations (M1–M4), resembling standard models such as HBV (Figure 8b). This potentially opens a wide range of opportunities on how future studies in ungauged regions may be designed and it highlights the need for a paradigm shift away from pure automated calibration toward giving higher priority to constraining the feasible parameter space [e.g., *Martinez and Gupta*, 2011; *Gupta et al*., 2012], based on efficiently extracting information from widely available data and expert knowledge [*Seibert and McDonnell*, 2002; *Savenije*, 2009]. Although not capable of predicting so far unknown patterns emerging from multiscale process dynamics and feedbacks, i.e., ontological uncertainties, as new generations of models may do [e.g., *Ruddell and Kumar*, 2009a, 2009b], a paradigm shift in method, as presented here, could prove attractive and beneficial in a twofold way for forecasting and process understanding not only for scientific but also, and maybe more importantly for operational hydrology, as it would imply higher model consistency, and thereby potentially increased predictive power, as well as reduced need for calibration data at little additional computational cost. Future research will, however, need to test the general value of the suggested constraint-based framework for multiple catchment typologies and identify which types of expert judgment and constraints may be effective for different types of catchment function.

6. Conclusions

In this study, a simple conceptual baseline model for the study catchment was iteratively developed to allow for more process complexity and constrained by prior expert knowledge. It was found that

1. In spite of its relatively high calibration performance with respect to the four calibration objective functions, a simple baseline model (M1) was not able to adequately reproduce a range of catchment signatures, indicating an inadequate representation of the observed response dynamics.

2. Iteratively increasing model complexity while at the same time explicitly incorporating semiquantitative, expert-knowledge-based, constraining prior information resulted in a model (M11) that showed substantially higher skill to reproduce the overall system response with comparably limited uncertainty, highlighting the value of *prior constraints* in filtering out unsuitable models.
3. *Process constraints* were found to contain more constraining information than *parameter constraints* in this study.
4. A calibrated but unconstrained complex model (M11a) included a wide range of implausible parameterizations, resulting in increased uncertainty and an overall performance that was reduced compared to a calibrated and constrained model (M11).
5. An uncalibrated but constrained complex model (M11b) resulted in similar model performance and uncertainty as a calibrated but unconstrained standard lumped model (M1), further underlining the value of *prior constraints* and the balance automated model calibration with a more general expert-knowledge-driven and system understanding-based strategy of constraining and rejecting models if a certain level of model consistency wants to be achieved.

Appendix A: Estimating Constant Deep Infiltration Loss Rate

The long-term average constant loss rate of the normalized groundwater levels ($Q_{L,const,norm}$), representing catchment storage (S_{norm}), was estimated from a time-storage relation (T-$S_{0,norm}$) at a value of $Q_{L,const,norm} = 0.0022$ d^{-1} by concatenating the annual segments of $S_{0,norm}$ during no-flow conditions (Figures A1a and A1b). As this loss rate is related to normalized groundwater levels rather than to flow rates, it cannot be directly used in a model. For use in a model, the actual loss rate ($Q_{L,const}$) from catchment storage in mm d^{-1} had to be estimated. This was done by extrapolating the T-$S_{0,norm}$ relation backward in time to instants when stream flow was observed (Figure A1c). As field observations provide evidence for rather homogenous porosity at depth [*Legout et al.*, 2007], the decrease of S_{norm} caused by discharge only, i.e., $S_{Q,norm}$, could then be determined for each time step by subtracting $S_{0,norm}$ from the relative groundwater level (S_{norm}), during relatively dry recession periods (i.e., 2003 and 2006). From the observed flow (Q_{obs}) at time steps at which the gradients of the time-$S_{Q,norm}$ relationship are equal to $Q_{L,const,norm}$ (Figure A1d), the actual loss rate $Q_{L,const} = 0.37$ mm d^{-1} could be inferred (Figure A1e), as it was assumed that equal gradients in time-$S_{0,norm}$ and time-$S_{Q,norm}$ relationships imply equal flow rates for base flow (Q_S) and deep percolation losses ($Q_{L,const}$).

Appendix B: Limits on Average Annual E_A

The upper and lower limits on long-term mean annual E_A were established using data from 420 catchments belonging to the Model Parameter Estimation Experiment (MOPEX) [*Schaake et al.*, 2006]. Data from these catchments were plotted in the Budyko framework. The limits on E_A where then estimated from the 5/95th percentiles of E_A/P in the binned region of the aridity index $I_A = E_P/P$ that corresponded to the I_A of the study catchment ($I_A = 0.67$).

Notation

C_P	preferential recharge coefficient.
C_R	hillslope runoff generation coefficient.
$C_{R,R}$	riparian runoff generation coefficient.
E_F	transpiration fast responding reservoir, $L\ T^{-1}$.
E_P	potential evaporation, $L\ T^{-1}$.
E_R	transpiration from riparian reservoir, $L\ T^{-1}$.
E_U	transpiration from unsaturated reservoir, $L\ T^{-1}$.
$E_{U,R}$	transpiration unsaturated riparian reservoir, $L\ T^{-1}$.
f	proportion wetlands in the catchment.
k_F	storage coefficient of fast reservoir, T^{-1}.
k_L	storage coefficient for deep infiltration loss, T^{-1}.
k_R	storage coefficient of riparian reservoir, T^{-1}.
k_S	storage coefficient of slow reservoir, T^{-1}.
L_P	transpiration threshold.

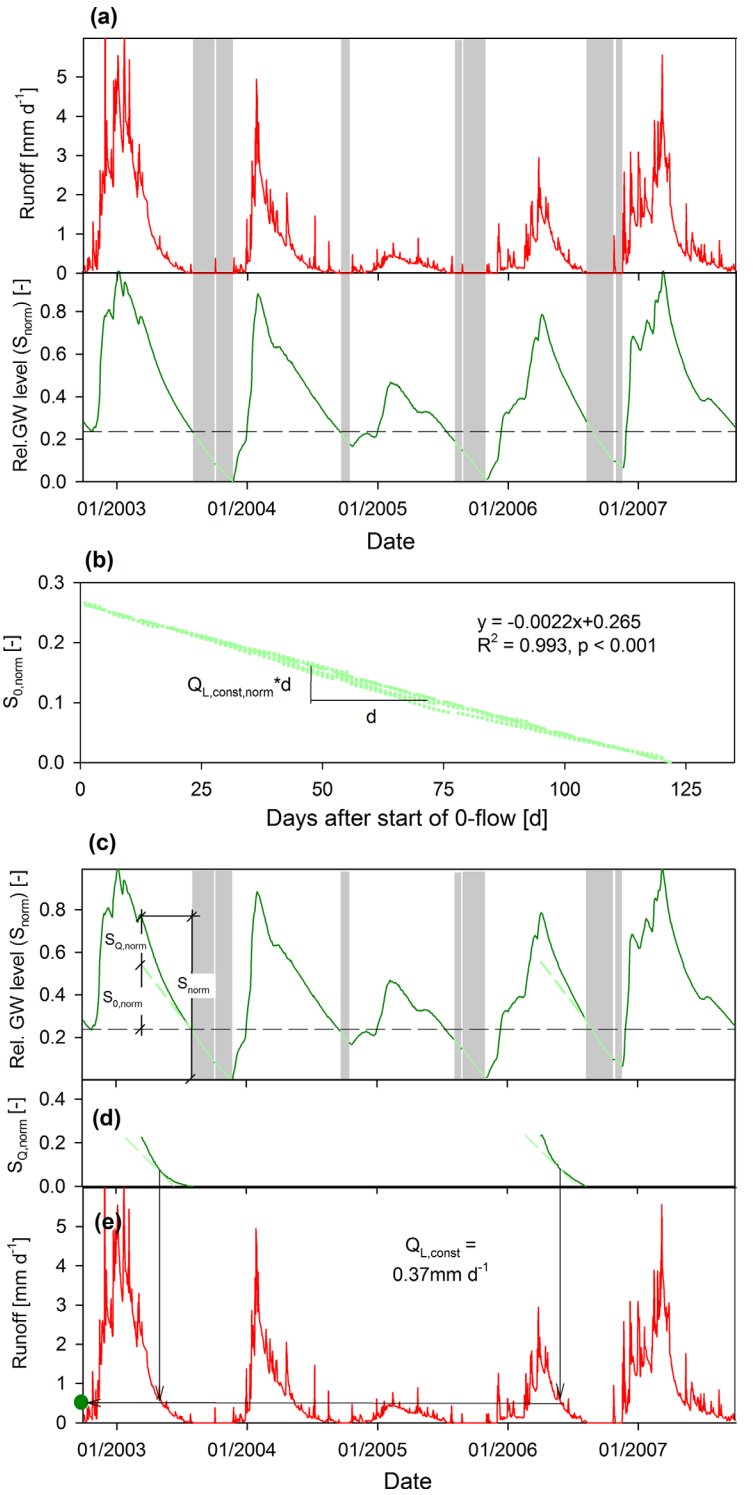


Figure A1. Illustrative example of how the observed averaged relative groundwater level was used to estimate a constant deep infiltration loss rate $Q_{L,const}$. (a) A selected period of the observed hydrograph (red) and the averaged relative groundwater level (green). No-flow periods are highlighted in gray. Light green lines for groundwater levels indicate the parts of the time series (no flow) used in to construct the time-storage relation T-$S_{0,norm}$. (b) Concatenation of segments of the relative groundwater levels during no-flow conditions to construct a "master" T-$S_{0,norm}$ relation. (c) Backward extrapolation of the T-$S_{0,norm}$ relation to instants when flow was observed to estimate $S_{Q,norm}$ in relatively dry recession periods. (d) At time t, when $S_{0,norm}$ (light green) is a tangent to $S_{Q,norm}$ (dark green) it was assumed that (e) $Q_{L,const}$ (dark green dot) is equal to Q_S (red line).

P	total precipitation, $L\ T^{-1}$.
P_{max}	percolation capacity, $L\ T^{-1}$.
Q_F	runoff from fast reservoir, $L\ T^{-1}$.
Q_L	deep infiltration loss, $L\ T^{-1}$.
$Q_{L,const}$	constant deep infiltration loss, $L\ T^{-1}$.
Q_R	runoff from riparian reservoir, $L\ T^{-1}$.
Q_S	runoff from slow reservoir, $L\ T^{-1}$.
R_F	recharge of fast reservoir, $L\ T^{-1}$.
R_P	preferential recharge of slow reservoir, $L\ T^{-1}$.
R_R	recharge of riparian reservoir, $L\ T^{-1}$.
R_S	recharge of slow reservoir, $L\ T^{-1}$.
R_U	infiltration into unsaturated reservoir, $L\ T^{-1}$.
S_F	storage in fast reservoir, L.
S_R	storage in riparian reservoir, L.
S_S	storage in slow reservoir, L.
$S_{S,a}$	active storage in slow reservoir, L.
$S_{S,p}$	passive storage in slow reservoir, L.
$S_{S,p,max}$	passive storage in slow reservoir, L.
$S_{S,tot}$	total storage in slow reservoir, L.
S_U	storage in unsaturated reservoir, L.
$S_{Umax,H}$	unsaturated hillslope storage capacity, L.
$S_{Umax,R}$	unsaturated riparian storage capacity, L.
β_H	hillslope shape parameter for C_R.
β_R	riparian shape parameter for $C_{R,R}$.

Acknowledgments

The effort of many staff and students at the ORE-AgrHys site Kerrien who collected and analyzed the samples in the data collected here is gratefully acknowledged. The data are provided online by the ORE AgrHys observatory (https://www6.inra.fr/ore_agrhys_eng/). The authors thank the editors as well as Christian Birkel and two further anonymous reviewers for highly constructive comments on an earlier version of the manuscript. In addition they thank Ronald Van Nooijen, Fabrizio Fenicia, Rohini Kumar, and Luis Samaniego for interesting discussions of the topic.

References

Ambroise, B., J. Freer, and K. Beven (1996a), Application of a generalized TOPMODEL to the small Ringelbach catchment, Vosges, France, *Water Resour. Res.*, *32*(7), 2147–2159.

Ambroise, B., K. Beven, and J. Freer (1996b), Toward a generalization of the TOPMODEL concepts: Topographic indices of hydrological similarity, *Water Resour. Res.*, *32*(7), 2135–2145.

Anderson, A. E., M. Weiler, Y. Alila, and R. O. Hudson (2009), Subsurface flow velocities in a hillslope with lateral preferential flow, *Water Resour. Res.*, *45*, W11407, doi:10.1029/2008WR007121.

Anderson, R. M., V. I. Koren, and S. M. Reed (2006), Using SSURGO data to improve Sacramento Model a priori parameter estimates, *J. Hydrol.*, *320*, 103–116.

Andréassian, V., and C. Perrin (2012), On the ambiguous interpretation of the Turc-Budyko nondimensional graph, *Water Resour. Res.*, *48*, W10601, doi:10.1029/2012WR012532.

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 a standardized evaluation of hydrological models, *Hydrol. Earth Syst. Sci.*, *13*, 1757–1764.

Andréassian, V., N. Le Moine, C. Perrin, M. H. Ramos, L. Oudin, T. Mathevet, J. Lerat, and L. Berthet (2012), All that glitters is not gold: The case of calibrating hydrological models, *Hydrol. Proc.*, *26*, 2206–2210.

Arora, V. K. (2002), The use of the aridity index to assess climate change effect on annual runoff, *J. Hydrol.*, *265*, 164–177.

Bergström, S. (1995), The HBV model, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 443–476, Water Resour. Publ., Highlands Ranch, Colo.

Beven, K. (2000), Uniqueness of places and process representations in hydrological modeling, *Hydrol. Earth Syst. Sci.*, *4*, 203–213.

Beven, K. (2001), On hypothesis testing in hydrology, *Hydrol. Proc.*, *15*, 1655–1657.

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

Beven, K. (2013), So how much of your error is epistemic? Lessons from Japan and Italy, *Hydrol. Proc.*, *27*, 1677–1680.

Beven, K., and J. Freer (2001), A dynamic TOPMODEL, *Hydrol. Proc.*, *15*, 1993–2011.

Beven, K., and I. Westerberg (2011), On red herrings and real herrings: Disinformation and information in hydrological inference, *Hydrol. Proc.*, *25*, 1676–1680.

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

Budyko, M. I. (1974), *Climate and Life*, Elsevier, N. Y.

Burnash, R. J. C. (1995), The NWS river forecast system—Catchment modeling, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 311–366, Water Resour. Publ., Highlands Ranch, Colo.

Carrera, J., and S. P. Neuman (1986), Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, *Water Resour. Res.*, *22*(2), 199–210.

Chapman, T. G., and A. I. Maxwell (1996), Baseflow separation—Comparison of numerical methods with tracer experiments, in *Hydrology and Water Resources Symposium*, pp. 539–545, Inst. of Eng., Barton, ACT, Australia.

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 (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.
Coron, L., V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx (2012), Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, *Water Resour. Res.*, *48*, W05552, doi:10.1029/2011WR011721.
Coxon, G., J. Freer, T. Wagener, N. A. Odoni, and M. Clark (2014), Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments, *Hydrol. Proc.*, doi:10.1002/hyp.10096, in press.
Criss, R. E., and W. E. Winston (2008), Do Nash values have a value? Discussion and alternate proposals, *Hydrol. Proc.*, *22*, 2723–2725.
Detty, J. M., and K. J. McGuire (2010), Topographic controls on shallow groundwater dynamics: Implications of hydrologic connectivity between hillslopes and riparian zones in a till mantled catchment, *Hydrol. Proc.*, *24*, 2222–2236.
Eckhardt, K. (2005), How to construct recursive digital filters for baseflow separation, *Hydrol. Processes*, *19*, 507–515.
Efstratiadis, A., and D. Koutsoyiannis (2010), One decade of multi-objective calibration approaches in hydrological modelling: A review, *Hydrol. Sci. J.*, *55*, 58–78.
Euser, T., H. C. Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H. G. Savenije (2013), A framework to assess the realism of model structures using hydrological signatures, *Hydrol. Earth Syst. Sci.*, *17*, 1893–1912.
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*, W06419, doi:10.1029/2007WR006386.
Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2008b), Understanding catchment behaviour through stepwise model concept improvement, *Water Resour. Res.*, *44*, W01402, doi:10.1029/2006WR005563.
Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, *Water Resour. Res.*, *47*, W11510, doi:10.1029/2010WR010174.
Fenicia, F., D. Kavetski, H. H. G. Savenije, M. P. Clark, G. Schoups, L. Pfister, and J. Freer (2013), Catchment properties, function, and conceptual model representation: Is there a correspondence?, *Hydrol. Proc.*, *28*, 2451–2467.
Freer, J., K. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resour. Res.*, *32*(7), 2161–2173.
Freer, J., K. Beven, and N. Peters (2003), Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure, in *Calibration of Watershed Models*, pp. 69–87, AGU, Washington, D. C.
Freer, J., 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*, 254–277.
Gao, H., M. Hrachowitz, F. Fenicia, S. Gharari, and H. H. G. Savenije (2014), Testing the realism of a topography driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China, *Hydrol. Earth Syst. Sci.*, *18*, 1895–1915.
Gharari, S., M. Hrachowitz, F. Fenicia, and H. H. G. Savenije (2011), Hydrological landscape classification: Investigating the performance of HAND based landscape classifications in a central European meso-scale catchment, *Hydrol. Earth Syst. Sci.*, *15*, 3275–3291.
Gharari, S., M. Hrachowitz, F. Fenicia, and H. H. G. Savenije (2013a), An approach to identify time consistent model parameters: Sub-period calibration, *Hydrol. Earth Syst. Sci.*, *17*, 149–161.
Gharari, S., M. Hrachowitz, F. Fenicia, H. Gao, and H. H. G. Savenije (2013b), Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration, *Hydrol. Earth Syst. Sci. Discuss.*, *10*, 14,801–14,855.
Gharari, S., M. Shafiei, M. Hrachowitz, F. Fenicia, H. V. Gupta, and H. H. G. Savenije (2013c), A strategy for "constraint-based" parameter specification for environmental models, *Hydrol. Earth Syst. Sci. Discuss.*, *10*, 14,857–14,871.
Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, *34*(4), 751–763.
Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Proc.*, *22*, 3802–3813.
Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resour. Res.*, *48*, W08301, doi:10.1029/2011WR011044.
Gupta, H. V., C. Perrin, R. Kumar, G. Blöschl, M. Clark, A. Montanari, and V. Andréassian (2014), Large-sample hydrology: A need to balance depth with breadth, *Hydrol. Earth Syst. Sci.*, *18*, 463–477.
Hersbach, H. (2000), Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecasting*, *15*(5), 559–570.
Hrachowitz, M., R. Bohte, M. L. Mul, T. A. Bogaard, H. H. G. Savenije, and S. Uhlenbrook (2011), On the value of combined event runoff and tracer analysis to improve understanding of catchment functioning in a data-scarce semi-arid area, *Hydrol. Earth Syst. Sci.*, *15*, 2007–2024.
Hrachowitz, M., et al. (2013a), A decade of predictions in ungauged basins (PUB)—A review, *Hydrol. Sci. J.*, *58*, 1198–1255.
Hrachowitz, M., H. Savenije, T. A. Bogaard, D. Tetzlaff, and C. Soulsby (2013b), What can flux tracking teach us about water age distribution patterns and their temporal dynamics?, *Hydrol. Earth Syst. Sci.*, *17*, 533–564.
Hughes, D. A. (2013), A review of 40 years of hydrological science and practice in southern Africa using the Pitman rainfall-runoff model, *J. Hydrol.*, *501*, 111–124.
Jafarpour, B. (2011), Wavelet reconstruction of geologic facies from nonlinear dynamic flow measurements, *IEEE Trans. Geosci. Remote Sens.*, *49*(5), 1520–1535.
Jafarpour, B., and M. Tarrahi (2011), Assessing the performance of the ensemble Kalman filter for subsurface flow data integration under variogram uncertainty, *Water Resour. Res.*, *47*, W05537, doi:10.1029/2010WR009090.
Jakeman, A. J., R. A. Letcher, and J. P. Norton (2006), Ten iterative steps in development and evaluation of environmental models, *Environ. Modell. Software*, *21*, 602–614.
Jothityangkoon, C., M. Sivapalan, and D. L. Farmer (2001), Process controls of water balance variability in a large semi-arid catchment: Downward approach to hydrological model development, *J. Hydrol.*, *254*(1), 174–198.
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.
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*, W03S04, doi:10.1029/2005WR004362.
Klemes, V. (1986), Operational testing of hydrological simulation models, *Hydrol. Sci. J.*, *31*, 13–24.
Knudsen, J., A. Thomsen, and J. C. Refsgaard (1986), WATBAL, *Nord. Hydrol.*, *17*(4–5), 347–362.
Knutti, R. (2008), Should we believe model predictions of future climate change?, *Philos. Trans. R. Soc. A*, *366*, 4647–4664.
Koren, V., F. Moreda, and M. Smith (2008), Use of soil moisture observations to improve parameter consistency in watershed calibration, *Phys. Chem. Earth*, *33*, 1068–1080.

Kumar, P. (2011), Typology of hydrologic predictability, *Water Resour. Res.*, *47*, W00H05, doi:10.1029/2010WR009769.
Kumar, R., L. Samaniego, and S. Attinger (2010), The effects of spatial discretization and model parameterization on the prediction of extreme runoff characteristics, *J. Hydrol.*, *392*, 54–69.
Kumar, R., L. Samaniego, and S. Attinger (2013), Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations, *Water Resour. Res.*, *49*, 360–379, doi:10.1029/2012WR012195.
Lamb, R., and K. Beven (1997), Using interactive recession curve analysis to specify a general catchment storage model, *Hydrol. Earth Syst. Sci.*, *1*, 101–113.
Le Moine, N., V. Andréassian, C. Perrrin, 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:10.1029/2006WR005608.
Leavesley, G. H., S. L. Markstrom, M. S. Brewer, and R. J. Viger (1996), The modular modelling system (MMS)—The physical process modelling component of a database-centered decision support system for water and power management, *Water Air Soil Pollut.*, *90*, 303–311.
Legchenko, A., J. M. Baltassat, A. Bobachevc, C. Martin, H. Robain, and J. M. Vouillamoz (2004), Magnetic resonance sounding applied to aquifer characterization, *Ground Water*, *42*, 363–373.
Legout, C., J. Molénat, L. Aquilina, C. Gascuel-Odoux, M. Faucheux, Y. Fauvel, and T. Bariac (2007), Solute transfer in the unsaturated zone-groundwater continuum of a headwater catchment, *J. Hydrol.*, *332*, 427–441.
Liu, Y. L., J. Freer, K. Beven, and P. Matgen (2009), Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error, *J. Hydrol.*, *367*(1–2), 93–103.
Martin, C., L. Aquilina, C. Gascuel-Odoux, J. Molenat, M. Faucheux, and L. Ruiz (2004), Seasonal and interannual variations of nitrate and chloride in stream waters related to spatial and temporal patterns of groundwater concentrations in agricultural catchments, *Hydrol. Proc.*, *18*, 1237–1254.
Martin, C., J. Molenat, C. Gascuel-Odoux, J. M. Vouillamoz, H. Robain, L. Ruiz, M. Faucheux, and L. Aquilina (2006), Modelling the effect of physical and chemical characteristics of shallow aquifers on water and nitrate transport in small agricultural catchments, *J. Hydrol.*, *326*, 25–42.
Martinez, G. F., and H. V. Gupta (2011), Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States, *Water Resour. Res.*, *47*, W12540, doi:10.1029/2011WR011229.
McGlynn, B. L., J. J. McDonnell, J. Seibert, and C. Kendall (2004), Scale effects on headwater catchment runoff timing, flow sources and groundwater-streamflow relations, *Water Resour. Res.*, *40*, W07504, doi:10.1029/2003WR002494.
Merz, R., G. Blöschl, and J. Parajka (2006), Spatio-temporal variability of event runoff coefficients, *J. Hydrol.*, *331*, 591–604.
Molenat, J., C. Gascuel-Odoux, P. Davy, and P. Durand (2005), How to model shallow water-table depth variations: The case of the Kervidy-Naizin catchment, France, *Hydrol. Proc.*, *19*, 901–920.
Molenat, J., C. Gascuel-Odoux, L. Ruiz, and G. Gruau (2008), Role of water table dynamics on stream nitrate export and concentration in agricultural headwater catchment (France), *J. Hydrol.*, *348*, 363–378.
Montanari, A., and E. Toth (2007), Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins?, *Water Resour. Res.*, *43*, W05434, doi:10.1029/2006WR005184.
Moore, R. J. (1985), The probability-distributed principle and runoff production at point and basin scales, *Hydrol. Sci. J.*, *30*(2), 273–297.
Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models: part 1—A discussion of principles, *J. Hydrol.*, *10*, 282–290.
Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environ. Res. Risk Assess.*, *17*(5), 291–305.
Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, *279*, 275–289.
Peters, N. E., J. E. Freer, and B. Aulenbach (2003), Hydrological dynamics of the Panola Mountain Research Watershed, Georgia, USA, *Groundwater*, *41*(7), 973–988.
Pokhrel, P., H. V. Gupta, and T. Wagener (2008), A spatial regularization approach to parameter estimation for a distributed watershed model, *Water Resour. Res.*, *44*, W12419, doi:10.1029/2007WR006615.
Refsgaard, J. C., J. P. Van der Sluijs, J. Brown, and P. Van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Adv. Water Resour.*, *29*(11), 1586–1597.
Renard, B., D. Kavetski, G. Kuczerac, M. Thyer, 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.
Ruddell, B. L., and P. Kumar (2009a), Ecohydrologic process networks: 1. Identification, *Water Resour. Res.*, *43*(3), W03419, doi:10.1029/2008WR007280.
Ruddell, B. L., and P. Kumar (2009b), Ecohydrologic process networks: 2. Analysis and characterization, *Water Resour. Res.*, *45*(3), W03420, doi:10.1029/2008WR007279.
Ruiz, L., S. Abiven, P. Durand, C. Martin, F. Vertes, and V. Beaujouan (2002), Effect on nitrate concentration in stream water of agricultural practices in small catchments in Brittany: I. Annual nitrogen budgets, *Hydrol. Earth Syst. Sci.*, *6*, 497–506.
Samaniego, L., R. Kumar, and S. Attinger (2010), Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, *Water Resour. Res.*, *46*, W05523, doi:10.1029/2008WR007327.
Savenije, H. H. G. (2009), The art of hydrology, *Hydrol. Earth Syst. Sci.*, *13*, 157–161.
Savenije, H. H. G. (2010), Topography driven conceptual modeling (FLEX-Topo), *Hydrol. Earth Syst. Sci.*, *14*, 2681–2692.
Sawicz, K., T. Wagener, M. Sivapalan, P. A. Troch, and G. Carrillo (2011), Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA, *Hydrol. Earth Syst. Sci.*, *15*, 2895–2911.
Schaake, J., S. Cong, and Q. Duan (2006), The US MOPEX data set, in *Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment (MOPEX)*, pp. 9–28, IAHS Press, Wallingford, U. K.
Schaller, M. F., and Y. Fan (2009), River basins as groundwater exporters and importers: Implications for water cycle and climate modelling, *J. Geophys. Res.*, *114*, D04103, doi:10.1029/2008JD010636.
Schoups, G., J. W. Hopmans, C. A. Young, J. A. Vrugt, and W. W. Wallender (2005), Multi-criteria optimization of a regional spatially-distributed subsurface water flow model, *J. Hydrol.*, *311*, 20–48.
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.
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.
Seibert, J., and M. J. P. Vis (2012), Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, *Hydrol. Earth Syst. Sci.*, *16*(9), 3315–3325.
Seibert, J., A. Rodhe, and K. Bishop (2003a), Simulating interactions between saturated and unsaturated storage in a conceptual runoff model, *Hydrol. Proc.*, *17*, 379–390.

Seibert, J., K. Bishop, A. Rodhe, and J. J. McDonnell (2003b), Groundwater dynamics along a hillslope: A test of the steady state hypothesis, *Water Resour. Res.*, *39*(1), 1014, doi:10.1029/2002WR001404.
Shamir, E., B. Imam, H. V. Gupta, and S. Sorooshian (2005), Application of temporal streamflow descriptors in hydrologic model parameter estimation, *Water Resour. Res.*, *41*, W06021, doi:10.1029/2004WR003409.
Shi, Y., K. J. Davis, F. Zhang, C. J. Duffy, and X. Yu (2014), Parameter estimation of a physically based land surface hydrologic model using the ensemble Kalman filter: A synthetic experiment, *Water Resour. Res.*, *50*, 706–724, doi:10.1002/2013WR014070.
Sivapalan, M., et al. (2003), IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences, *Hydrol. Sci. J.*, *48*(6), 857–880.
Tonkin, M. J., and J. Doherty (2005), A hybrid regularized inversion methodology for highly parameterized environmental models, *Water Resour. Res.*, *41*, W10412, doi:10.1029/2005WR003995.
Wagener, T. (2003), Evaluation of catchment models, *Hydrol. Proc.*, *17*, 3375–3378.
Wagener, T., and H. V. Gupta (2005), Model identification for hydrological forecasting under uncertainty, *Stochastic Environ. Res. Risk Assess.*, *19*, 378–387.
Wagener, T., and A. Montanari (2011), Convergence of approaches toward reducing uncertainty in predictions in ungauged basins, *Water Resour. Res.*, *47*, W06301, doi:10.1029/2010WR009469.
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., K. van Werkhoven, P. Reed, and Y. Tang (2009), Multiobjective sensitivity analysis to understand the information content in streamflow observations for distributed watershed modelling, *Water Resour. Res.*, *45*, W02501, doi:10.1029/2008WR007347.
Westerberg, I. K., J. L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, J. E. Freer, and C. Y. Xu (2011), Calibration of hydrological models using flow-duration curves, *Hydrol. Earth Syst. Sci.*, *15*, 2205–2227.
Winsemius, H. C., H. H. G. Savenije, and W. G. M. Bastiaanssen (2008), Constraining model parameters on remotely sensed evaporation: Justification for distribution in ungauged basins?, *Hydrol. Earth Syst. Sci.*, *12*, 1403–1413.
Winsemius, H. C., B. Schaefli, A. Montanari, and H. H. G. Savenije (2009), On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information, *Water Resour. Res.*, *45*, W12422, doi:10.1029/2009WR007706.
Yadav, M., T. Wagener, and H. V. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Adv. Water Resour.*, *30*, 1756–1774.
Yeh, W. W. G. (1986), Review of parameter identification procedures in groundwater hydrology: The inverse problem, *Water Resour. Res.*, *22*(2), 95–108.
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.