Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment

Dmitri Kavetski,[1] Fabrizio Fenicia,[2,3] and Martyn P. Clark[4]

Received 7 May 2010; revised 12 October 2010; accepted 3 November 2010; published 10 May 2011.

[1] This study presents quantitative and qualitative insights into the time scale dependencies of hydrological parameters, predictions and their uncertainties, and examines the impact of the time resolution of the calibration data on the identifiable system complexity. Data from an experimental basin (Weierbach, Luxembourg) is used to analyze four conceptual models of varying complexity, over time scales of 30 min to 3 days, using several combinations of numerical implementations and inference equations. Large spurious time scale trends arise in the parameter estimates when unreliable time-stepping approximations are employed and/or when the heteroscedasticity of the model residual errors is ignored. Conversely, the use of robust numerics and more adequate (albeit still clearly imperfect) likelihood functions markedly stabilizes and, in many cases, reduces the time scale dependencies and improves the identifiability of increasingly complex model structures. Parameters describing slow flow remained essentially constant over the range of subhourly to daily scales considered here, while parameters describing quick flow converged toward increasingly precise and stable estimates as the data resolution approached the characteristic time scale of these faster processes. These results are consistent with theoretical expectations based on numerical error analysis and data-averaging considerations. Additional diagnostics confirmed the improved ability of the more complex models to reproduce distinct signatures in the observed data. More broadly, this study provides insights into the information content of hydrological data and, by advocating careful attention to robust numericostatistical analysis and stringent process-oriented diagnostics, furthers the utilization of dense-resolution data and experimental insights to advance hypothesis-based hydrological modeling at the catchment scale.

Citation: Kavetski, D., F. Fenicia, and M. P. Clark (2011), Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment, *Water Resour. Res., 47*, W05501, doi:10.1029/2010WR009525.

1. Introduction

1.1. Data and Model Uncertainties in Hydrology: Time Scale Effects

[2] A major issue in hydrological and broader environmental modeling is the uncertainty in the observed data, in particular, the effects of sparse data sampling and averaging to temporal and spatial scales that may well exceed those of many hydrological dynamics of interest [e.g., *Blöschl and Sivapalan*, 1995]. For example, many rainfall-runoff models are calibrated and applied on daily steps, with rainfall averaged from sparse rain gauges and streamflow subject to random and systematic measurement errors [e.g., *Brath et al.*, 2004; *Thyer et al.*, 2009]. Consequently, the research and operational focus across all fields of environmental studies is increasingly shifting away from point estimation and toward probabilistic inference, which recognizes these data and model uncertainties [e.g., *Beven and Binley*, 1992; *Kuczera et al.*, 2006; *Reichert and Mieleitner*, 2009; *Cressie et al.*, 2009]. Currently, the statistical and computational complexity of inverse estimation of nonlinear models, especially when describing uncertainties using sampling methods such as Markov chain Monte Carlo (MCMC) schemes [e.g., *Kuczera and Parent*, 1998; *Vrugt et al.*, 2008], may favor computationally fast conceptual models, which in many situations can capture key hydrological dynamics given only limited data.

[3] The time and space resolution of the calibration data has a significant impact on model estimation and prediction. Even if the governing model equations are formulated in continuous time state-space form [e.g., *Clark et al.*, 2008], practical implementations generally operate in discrete time,

[1] Environmental Engineering, University of Newcastle, Callaghan, New South Wales, Australia.

[2] Department of Environment and Agro-Biotechnologies, Centre de Recherche Public–Gabriel Lippmann, Belvaux, Luxembourg.

[3] Water Resources Section, Delft University of Technology, Delft, Netherlands.

[4] National Center for Atmospheric Research, Boulder, Colorado, USA.

0043-1397/11/2010WR009525

producing predictions at a series of discrete time steps [*Liu and Gupta,* 2007] (see also *Young and Garnier* [2006] for a discussion of the effects of continuous versus discrete time implementations). In most cases, the temporal resolution is fixed by the data collection procedure (e.g., daily or hourly readings of rain and stream gauges) and controls the time step of the hydrological model. In principle, the resolution of the forcing versus response time series can differ, though in practice, it is usually the same. In addition, an increasing number of hydrological models use adaptive substepping within the outer "data resolution" time steps (see *Clark and Kavetski* [2010] for a review).

[4] A number of studies have explored the time scale dependencies of hydrological model parameters. For example, *Finnerty et al.* [1997] applied the Sacramento model over a range of time steps and concluded that its parameters are inherently tied to the calibration time scale. Analogous conclusions were reached by other authors [e.g., *Schaake et al.*, 1996; *Tang et al.*, 2007; *Cho et al.*, 2009]. In a recent case study based on the IHACRES model, *Littlewood and Croke* [2008] demonstrated a strong time scale dependency of model parameters and provided a methodology to relate parameter values to the modeling time step. Similarly, *Wang et al.* [2009] also found linear and nonlinear time scale dependencies in hydrological parameter estimates and analyzed them in the context of average rainfall intensities at different time scales.

[5] Despite these insights, the impact of data resolution and modeling time step on the inference of catchment structure, model parameters, and, more generally, catchment behavior remains poorly understood. While a general consensus exists that model parameters, simulation results, and process representations are inherently and strongly time scale dependent [e.g., *Duan et al.*, 2006; *Merz et al.*, 2009], there is insufficient quantitative understanding of the precise underlying causes and their mathematical representation and physical interpretation and a lack of conceptually sound and practically robust strategies to handle them. In the absence of an adequate mathematical framework explaining and predicting these dependencies and encompassing both conceptual and physically based models of different degrees of complexity, current treatments of parameter scaling are largely heuristic and empirical [*Littlewood and Croke*, 2008; *Wang et al.*, 2009]. Similarly, while it has been shown that increasingly complex models can be inferred from higher-resolution data [*Atkinson et al.*, 2002; *Farmer et al.*, 2003], the quantitative and qualitative understanding of the processes revealed by high-resolution data remains limited, especially at subdaily time scales [*Kirchner et al.*, 2004].

[6] A robust quantitative understanding of the (likely multiple) causes of time scale dependency of model parameters is critical for the advancement of several areas of hydrology, both in process understanding and in operational predictions. In particular, the aim of model identification and parameter estimation is not only obtaining suitable model performance (e.g., in a statistical sense), but, perhaps more importantly, gaining physically meaningful, interpretable, and transferable insights [e.g., *Gupta et al.*, 2008; *Bárdossy and Singh*, 2008]. The dependence of model parameters on factors such as time step size and calibration period obscures the physical interpretation of calibrated model structures, limits our ability to elucidate their connections to catchment attributes, and precludes their regionalization to ungauged catchments [*Wagener and Wheater*, 2006]. These challenges are among those at the forefront of contemporary hydrology and broader environmental sciences [e.g., *Sivapalan et al.*, 2003b].

1.2. Interplay Between Data Resolution and Statistical and Numerical Artifacts

[7] Considering their close relationship to the time approximation scheme, time scale dependencies of hydrological models may be highly susceptible to numerical and statistical artifacts [*Kavetski et al.*, 2003]. These issues are directly relevant to ongoing hydrological research and practice and hence form a key focus of this study.

[8] The general importance of adequate error models for reliable estimation and prediction is well known (e.g., see *Box and Tiao* [1992] for general theory; in hydrological calibration we can point to illustrations and discussions by *Sorooshian and Dracup* [1980], *Beven and Binley* [1992], *Thyer et al.* [2009], and *Schoups and Vrugt* [2010]). For example, *Thyer et al.* [2009] empirically illustrated that the consistency of parameter estimates in a rainfall-runoff model depends strongly on the adequacy of the hypothesized error models describing the uncertainties in the catchment-averaged rainfall and streamflow. In particular, ignoring rainfall uncertainty can introduce spurious and confounding dependence of inferred parameters on the rain gauge location and calibration period [see *Kavetski et al.*, 2002b].

[9] On another front, the importance of robust numerical design for meaningful estimation and prediction has emerged as a key finding in a series of studies, both in hydrology [e.g., *Clark and Kavetski*, 2010, and references therein] and in the broader environmental modeling discipline (e.g., see *Baker* [1995], *Miller et al.* [1998], and *Kavetski et al.* [2002a] for vadose zone modeling using Richards equation). It is increasingly apparent that models' predictions, and hence objective functions, are severely deformed by spurious artifacts when unreliable numerical implementations are used or when the model's constitutive relationships are exceedingly discontinuous [*Kavetski et al.*, 2006a]. These weaknesses have significant implications for sensitivity analysis, parameter estimation and interpretation, and operational prediction [*Kavetski and Clark*, 2010]. Several robust numerical approaches (unconditionally stable implicit integration and/or adaptive substepping) and model design recommendations (e.g., using smooth constitutive functions linking model storages and fluxes) were proposed to address these problems. In this paper, we exploit the numerical understanding gained in previous computationally oriented studies [e.g., *Kavetski et al.*, 2003; *Clark and Kavetski*, 2010] to support more robust process-oriented insights into the effects of environmental data resolution on parameter estimation and model identification of catchment-scale hydrological systems.

1.3. Toward a More Process-Oriented Hypothesis Testing in Hydrology

[10] Hydrological modeling is a multifaceted challenge, particularly in scientific contexts where a key objective is to improve the fidelity of the process conceptualizations. In addition to an adequate overall model structure that

represents the dominant hydrological processes, this also requires the parameters to take physically meaningful values [e.g., *Wagener and Wheater*, 2006; *Bárdossy and Singh*, 2008] and, ideally, for the model state variables to represent, at least approximately, the intended internal catchment dynamics [e.g., *Kuczera and Franks*, 2002]. Yet, especially when using generic "off-the-shelf" residual error models (objective functions) and little auxiliary prior knowledge, purely statistical methodologies, including traditional regression and related heuristic modifications, are often insufficiently discriminatory to identify physically realistic models. This is especially apparent when gauged against the process understanding gleaned from experimental catchments [e.g., *Seibert and McDonnell*, 2002; *Tromp-van Meerveld and McDonnell*, 2006a; *Vaché and McDonnell*, 2006].

[11] The "diagnostic" approach to hydrological model evaluation seeks to address these challenges using multiple "hydrological signatures" [*Gupta et al.*, 2008], which elucidate aspects of system behavior that may not be immediately apparent from the response time series alone. For example, *Yilmaz et al.* [2008] use the slope of the flow duration curve to infer model parameters that describe the vertical percolation of soil moisture. As such, diagnostics are a form of more incisive hypothesis testing, going beyond merely fitting the aggregate system response [e.g., *Kuczera and Franks*, 2002]. If applied to independently scrutinize and improve individual constituent hypotheses, they can also alleviate the identifiability problems that traditionally plague catchment-scale hydrology (the "model equifinality" problem in the language of *Beven* [2006]).

[12] Another important aspect of hypothesis testing in hydrology concerns the complexity of the identifiable models [e.g., *Jakeman and Hornberger*, 1993; *Schoups et al.*, 2008]. For example, in the data-based mechanistic (DBM) framework, models are constructed using linear transfer functions, with (possibly state-dependent) parameters estimated using instrumental variables techniques [e.g., *Young*, 1998, 2003]. In more traditional conceptual hydrology, the "top-down" strategy begins with a simple model and increments its complexity until some indicator(s) of model performance are judged to be adequate [e.g., *Sivapalan et al.*, 2003a]. In the absence of strong independent physical insights, the supported complexity of a model is intimately linked to the information content extracted from the calibration data. Therefore, understanding the interplays between data resolution, objective functions, and numerical approximation errors is of major significance for hypothesis testing.

1.4. Aims and Scope

[13] This work aims to use experimental data and process understanding to obtain deeper and more robust quantitative and qualitative insights into the time scale dependencies of hydrological parameters and their uncertainties. The focus is on both the origins and the impacts of time scale dependencies on both statistical and process-oriented hypothesis testing in catchment-scale hydrology. Our objective is to advance previous work on time scale dependencies [e.g., *Littlewood and Croke*, 2008; *Wang et al.*, 2009] by exploring structural complexity issues using a flexible model framework [*Fenicia et al.*, 2008], by illustrating the impacts of the numerical model implementation and the objective function used in the inference, and by investigating the behavior of parameter and streamflow distributions rather than point estimates. Importantly, we make a direct link between the quantitative mathematical analysis and process-oriented qualitative field insights available in this experimental catchment. We also examine the stability of the validation performance with respect to the resolution of the calibration data, diagnose the ability of the inferred models to reproduce hydrological signatures [*Gupta et al.*, 2008], and make inroads into understanding subhourly resolution effects. The implications of the findings for the community quest for stronger scrutiny of hydrological model hypotheses and the overall aim of more scientifically defensible and operationally reliable models are also discussed.

[14] The generality of the findings is further strengthened by interpreting the empirical results from a theoretical perspective of numerical error analysis and data-averaging considerations. An important emphasis of this presentation is on the key distinction between the time scale dependence of the governing model equations versus the time scale dependencies of their practical (numerical) implementations. Since insufficient attention to numerical aspects can create harmful numerical trends and artifacts, this distinction is of markedly underrated significance for meaningful advances in hydrological science and operations [e.g., *Kavetski and Clark*, 2010].

[15] In order to provide insights relevant to a wide audience of hydrologists, we focus on common calibration methods and time-stepping schemes. Assessments over large numbers of catchments [e.g., *Perrin et al.*, 2001; *Merz et al.*, 2009] are beyond our scope and will be pursued separately. Analysis of rainfall errors [e.g., *Kavetski et al.*, 2002b; *Vrugt et al.*, 2008; *Götzinger and Bárdossy*, 2008], more sophisticated structural error models [e.g., *Reichert and Mieleitner*, 2009; *Bulygina and Gupta*, 2009; *Renard et al.*, 2010; *Doherty and Welter*, 2010], and extensions to distributed models [e.g., *Ivanov et al.*, 2004; *Immerzeel and Droogers*, 2008] are deferred given their high data requirements and computational costs at the short time scales examined in this work.

[16] The paper is organized as follows. The study area is described in section 2, the hydrological models in section 3 (including the numerical algorithms in section 3.3), and the model analysis methods in section 4. The methodology is outlined in section 5, followed by empirical experiments presented in section 6. The findings are discussed in section 7 (including a theoretical interpretation and discussion of limitations). The discussion and conclusion summarize the key insights and outline our views on the current limitations facing conceptual catchment-scale modeling and on future research directions.

2. Experimental Catchment

[17] The study area for this investigation is the 0.47 km^2 Weierbach catchment in the Grand Duchy of Luxembourg. It has a history of experimental work that is providing both high-resolution rainfall-runoff-evapotranspiration time series and also valuable process-oriented insights into its geomorphology and dominant hydrological dynamics [e.g.,

van den Bos et al., 2006b; *Pfister et al.*, 2006; *Martinez-Carreras et al.*, 2010]. The Weierbach is largely forested (85%), with some agriculture on its plateaus (15%); its elevations range from 422 to 512 m above sea level. The hillslope soils are shallow because of continuous erosion, though deeper soils are present on the uphill plateaus. A well-developed weathered zone exists at the soil-bedrock interface and acts as a water store. The lithology is predominantly schistose, which is generally impermeable. This results in a relative absence of flow during dry periods because there is no significant bedrock aquifer storage. The ephemeral nature of uphill sources also suggests a limited storage depth. However, while largely impeding deep percolation, the schists are fractured in a preferential direction and, when saturated, form a complex flow network interspersed with local storage in the rock cracks.

[18] The Weierbach catchment is well suited for demonstrating and analyzing the time scale dependencies of model parameters and structures in a nontrivial and relatively general way. Despite its small size, the catchment has complex hydrologic dynamics, characterized by thresholds and delays operating over a spectrum of time scales. In particular, its streamflow response to a rainfall event in the wet (winter) season is often characterized by two distinct peaks with markedly different time scales [*van den Bos et al.*, 2006b]. The initial response is near concomitant with the rainfall event and, unless masked by rainfall variability, appears as a spiky first peak in the hydrograph. It has been attributed to rainfall over the near-stream riparian zone, which in the Weierbach catchment comprises rock outcrops and is generally saturated over most of the year. The delayed response generally takes the form of a broader second peak that reaches its maximum at a lag of approximately 48 h. It appears to be sustained by water flows at the soil-bedrock interface [*van den Bos et al.*, 2006b]. It can be hypothesized that this delay arises from a cascade of water reservoirs in cracks formed by the irregular bedrock topography. Chemical analyses have suggested that water flows during the two peaks have distinct chemical compositions and are likely to originate in different catchment compartments [*Pfister et al.*, 2006] (also see *Clark et al.* [2009] for similar experimental work on the Panola catchment).

[19] The delayed response of the catchment varies strongly depending on catchment wetness. This effect has been attributed to the connectivity of flow pathways [*Hopp and McDonnell*, 2009], which according to recent field investigations in this catchment, varies significantly across the seasons. The delayed runoff contributions are less evident during dry periods, when rainfall water remains trapped in the bedrock depressions and is lost through transpiration. During the wet season, interconnected saturated zones develop at the soil-bedrock interface and increase the areal fraction of the catchment contributing to runoff via saturation–excess flow mechanisms. The Weierbach catchment, including its spatial variability and hydrogeological complexity, remains the subject of ongoing field investigations [e.g., *van den Bos et al.*, 2006b; *Martinez-Carreras et al.*, 2010].

[20] Long-term annual average precipitation and discharge are about 900 and 480 mm, respectively. Discharge is measured at a 90° V notch weir. A single rain gauge about 3 km outside of the catchment was used. Potential evaporation was estimated using the Hamon equation [*Hamon and Belt*, 1973]. All time series were available at a 30 min resolution or shorter. This study uses 3 years of data in the period from 1 July 2005 to 1 July 2008. The first 2 years were used for calibration, and the final 1 year period was used for evaluation ("validation").

3. Hydrological Models

3.1. Overall Model Architecture and Governing Equations

[21] The multiple conceptual model structures analyzed in this study are obtained from the flexible model framework SUPERFLEX, which generalizes the original FLEX model of *Fenicia et al.* [2008]. In the SUPERFLEX framework, a complete model hypothesis is constructed by combining and configuring a number of generic components that approximate different mechanistic aspects of catchment dynamics, including partition, storage, release, and transmission of water [*Wagener et al.*, 2007]. In this study, reservoirs are used to represent storage and release of water, and transfer functions are used to represent the transmission and delay of flows. Each component is characterized by selected constitutive functions (e.g., relating fluxes to storages) and associated parameters. Multiple components are assembled according to a hypothesized connectivity using junction elements.

[22] Reservoir water balance dynamics are represented using a set of ordinary differential equations (ODEs):

$$\mathrm{d}\mathbf{S}(t)/\mathrm{d}t = \mathbf{g}_s[\mathbf{S}(t), \mathbf{X}(t)|\boldsymbol{\theta}], \tag{1a}$$

$$Q(t) = \mathbf{g}_Q[\mathbf{S}(t), \mathbf{X}(t)|\boldsymbol{\theta}], \tag{1b}$$

where $\mathbf{S}(t)$ are conceptual storage values at time t, $\mathbf{X}(t)$ is the (time-dependent) forcing (here rainfall P and potential evapotranspiration E_p), and $Q(t)$ is the streamflow response. In equation (1), $\mathbf{g}(\ )$ are the input-output fluxes connecting the model components, and $\boldsymbol{\theta}$ are the model parameters (see *Clark et al.* [2008] for the development of conceptual hydrological models in state-space form).

3.2. Specific Model Hypotheses Under Consideration

[23] In this study, we examine four models of increasing structural complexity (Figure 1 and Table 1). Their underlying hypotheses were selected on the basis of qualitative insights from previous field studies in the Weierbach catchment. This section summarizes the key qualitative features of these models; the equations are detailed in Tables 2 and 3.

[24] 1. The simplest structure (M1) is characterized by two reservoirs and five parameters. It includes an unsaturated soil reservoir (UR) and a fast-reacting reservoir (FR). UR represents the saturation–excess runoff response of this catchment (see section 2). Precipitation P_t is separated into a fast component Q_q routed through FR and a slow component routed through UR. Actual evaporation is proportional to the potential evaporation, with a smoothing function used for near-zero storage values. The unsaturated store UR is drained by the flux Q_u, which is linearly related to the storage S_u. The fast reservoir FR is linear. The streamflow is obtained by combining Q_u with the output flux Q_f from FR.

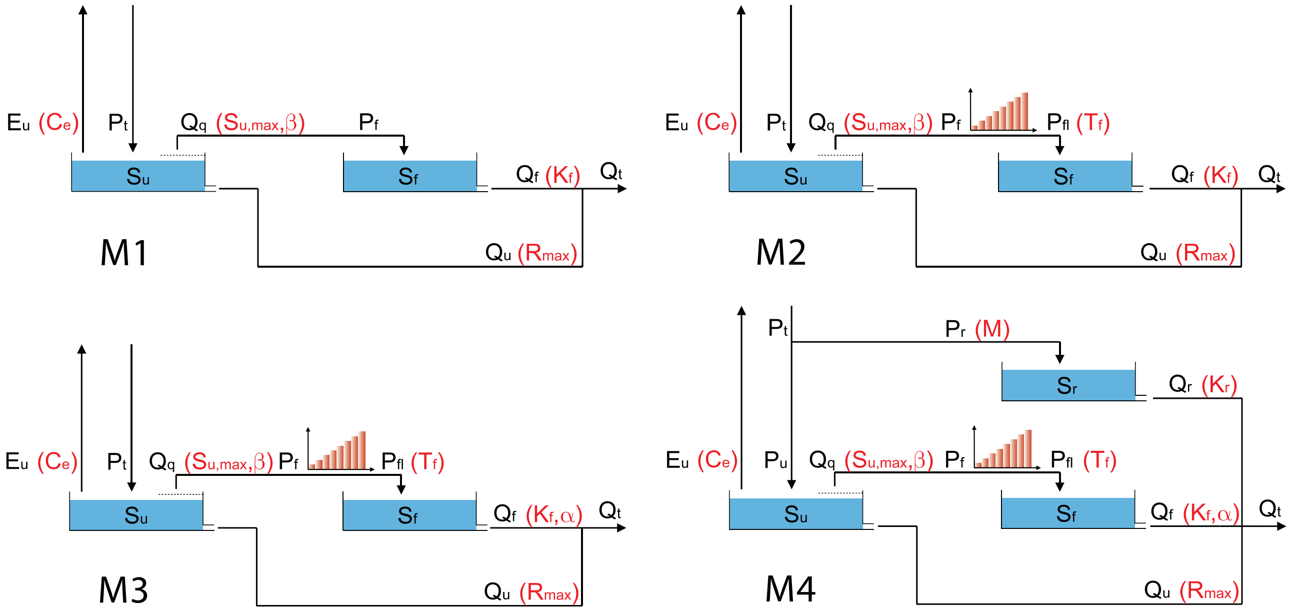

Figure 1. Schematic representation of the conceptual hydrological hypotheses M1–M4 analyzed in this study. The states and fluxes are in black, with associated parameters in red. The flexibility in selecting model structures within SUPERFLEX and similar modular frameworks [e.g., *Clark et al.*, 2008] is exploited here for systematic hypothesis testing with respect to data resolution, process representation, and numerical approximations.

[25] 2. The structure M2 differs from M1 by including a transfer function, which accounts for delays in the rainfall-runoff routing. The flux Q_q is convolved with a triangular transfer function, and the resulting flux P_{fl} is routed through the fast reservoir. The model has six parameters.

[26] 3. The structure M3 further generalizes M2 by making FR nonlinear. It has seven parameters.

[27] 4. The most complex structure (M4) differs from M3 by including a riparian zone reservoir (RR), which conceptualizes the contribution of the impervious zone of the catchment. The precipitation P_t is split into a component P_u that reaches UR and a component P_r that is routed through the (linear) reservoir RR. The streamflow is then obtained by combining Q_f, Q_u, and the output flux Q_r from RR. Configuration M4 is characterized by nine parameters.

[28] While fairly simple relative to physically based distributed models [e.g., *Ivanov et al.*, 2004], the model hypotheses M1–M4 are generally representative of operational forecasting models, such as GR4J [*Berthet et al.*, 2009], and of the conceptual models used in major hydrological investigations [e.g., *Duan et al.*, 2006].

3.3. Numerical Implementation of Model Hypotheses

[29] To prevent mathematical solution aspects from obscuring the comparison of physical process representations, the implementations of model hypotheses must be treated, and reported, separately from their conceptual development [*Kavetski et al.*, 2003] (see *Clark et al.* [2008] for a practical illustration).

[30] Since analytical solutions of ODE (1) do not exist for most flux formulations **g**( ), numerical approximations must be employed (e.g., see *Butcher* [2008] for numerical ODE theory and *Clark and Kavetski* [2010] for a discussion in hydrological contexts). In this study, we use the explicit and implicit Euler time-stepping schemes, applied over fixed discrete steps $\Delta t = t_{n+1} - t_n$, where the subscript n indexes the time step. For a better correspondence with rainfall and runoff measurement systems (which usually report accumulated totals), the hydrological models are forced with data $\overline{\mathbf{X}}_{n\to n+1}$ obtained by averaging $\mathbf{X}(t)$ over each data resolution step.

[31] The explicit Euler method (EE) evolves the approximation using the flux at the beginning of each time step,

Table 1. Components and Parameters of Model Structures M1–M4[a]

Model		Components				Parameters								
Structure	N_{par}	RR	UR	FR	LF	M	C_e	$S_{u,\max}$ (mm)	β	T_f (h)	K_r (1/h)	K_f (mm$^{1-\alpha}$ / h)	α	$R_{\max}$ (mm/h)
M1	5	-	√	√	-	-	√	√	√	-	-	√	-	√
M2	6	-	√	√	√	-	√	√	√	√	-	√	-	√
M3	7	-	√	√	√	-	√	√	√	√	-	√	√	√
M4	9	√	√	√	√	√	√	√	√	√	√	√	√	√

[a] N_{par} is the number of parameters. RR, UR, FR, and LF denote the riparian, unsaturated, and fast reservoirs and the transfer function, respectively. An √ indicates that a component or parameter is included in a structure, and a dash indicates that it is not included.

Table 2. Water Balance Equations of the Models Used in the Experiments[a]

Water Balance Equations	M1	M2	M3	M4
$P_t = P_u + P_r$	-	-	-	√
$P_t = P_u$	√	√	√	-
$\frac{\mathrm{d}S_u}{\mathrm{d}t} = P_u - Q_q - Q_u - E_u$	√	√	√	√
$\frac{\mathrm{d}S_r}{\mathrm{d}t} = P_r - Q_r$	√	√	√	√
$\frac{\mathrm{d}S_f}{\mathrm{d}t} = P_{fl} - Q_f$	-	√	√	√
$\frac{\mathrm{d}S_f}{\mathrm{d}t} = P_f - Q_f$	√	-	-	-
$Q_t = Q_r + Q_f + Q_u$	-	-	-	√
$Q_t = Q_f + Q_u$	√	√	√	-

[a]The √ and dash indicate presence and absence, respectively.

$$\mathbf{S}_{n+1}^{(\mathrm{EE})} = \mathbf{S}_n + \Delta t\mathbf{g}\left(\mathbf{S}_n, \overline{\mathbf{X}}_{n\to n+1}\right) . \tag{2}$$

[32] It remains widely used in hydrological applications because of its algorithmic simplicity and computational speed. However, explicit integration is conditionally stable, with instabilities developing in the EE scheme when $\Delta t > 2\ \mathrm{eigmax}|\partial\mathbf{g}/\partial\mathbf{S}|^{-1}$, where $\mathrm{eigmax}|\mathbf{z}|$ denotes the largest-magnitude eigenvalue of matrix $\mathbf{z}$. The explicit Euler scheme is therefore highly unreliable unless numerical error control is used [e.g., *Kahaner et al.*, 1989]. Here the EE scheme is implemented without error control but with a trivial check on the drainage fluxes to avoid negative storages as a result of "overdraining" the store in a single step. This avoids outright instabilities but is very ad hoc; it corresponds closely to common implementations of many current conceptual hydrological models (see *Clark and Kavetski* [2010] for a review). In this study, it is used only to illustrate the interplay between time scale effects and numerical implementations commonly encountered in conceptual hydrological modeling. Note that other heuristic approaches to time step size selection are possible, e.g., based on the estimated concentration time of a storm event (used in some engineering applications [*Maniak*, 1997]).

[33] In contrast, the implicit Euler method (IE), which uses the flux at the end of the time step,

$$\mathbf{S}_{n+1}^{(\mathrm{IE})} = \mathbf{S}_n + \Delta t\mathbf{g}\left(\mathbf{S}_{n+1}^{(\mathrm{IE})}, \overline{\mathbf{X}}_{n\to n+1}\right) \tag{3}$$

is unconditionally stable [e.g., *Kahaner et al.*, 1989]. Hence, despite requiring potentially expensive iterative solutions (e.g., using the Newton-Raphson root solver [see *Clark and Kavetski*, 2010]), the implicit Euler scheme is generally robust even for large step sizes. It is widely used in standard engineering software, e.g., the MODFLOW package for groundwater simulations, the ECLIPSE tool in the petroleum industry, and geotechnical consolidation codes [e.g., see *Clark and Kavetski*, 2010, and references therein].

Table 3. Constitutive Relationships of the Models Used in the Experiments[a]

Constitutive Relationships	M1	M2	M3	M4
$\overline{S}_u = S_u/S_{u,\max}$	√	√	√	√
$Q_q = P_u \frac{\left(1+e^{-\beta/2}\right)\left(e^{-\beta\overline{S}_u}-1\right)}{\left[1+e^{-\beta\left(\overline{S}_u-1/2\right)}\right]\left(e^{-\beta}-1\right)}$	√	√	√	√
$E_u = C_e E_p f_e\left(\overline{S}_u\right)$	√	√	√	√
$f_e\left(\overline{S}_u\right) = (1+m_e)\frac{\overline{S}_u}{\overline{S}_u+m_e}$	√	√	√	√
$P_{fl} = (P_f * h_f)\ (t)$	-	√	√	√
$h_f = \begin{cases} t/T_f^2, & t < T_f \\ 0, & t > T_f \end{cases}$	-	√	√	√
$Q_u = R_{\max}\overline{S}_u$	√	√	√	√
$Q_f = k_f S_f$	√	√	-	-
$Q_f = k_f S_f^{\alpha}$	-	-	√	√
$Q_r = k_r S_r$	-	-	-	√

[a]The operator * in the equation for P_{fl} denotes the convolution. The constant $e = 2.718$ denotes the natural logarithm base. The √ and dash indicate presence and absence, respectively.

[34] Importantly, when implemented using fixed steps and a tight Newton-Raphson tolerance, the implicit Euler solution is smooth with respect to its forcing and parameters (note the distinction between the temporal error tolerance τ of an (explicit or implicit) substepping algorithm versus the iteration tolerance τ_{NR} of the Newton-Raphson root solver for an implicit scheme) [*Clark and Kavetski*, 2010]. This results in a smooth objective function of the hydrological model, facilitates calibration, and leads to more stable prediction.

[35] Verifications against adaptive ODE solutions with tight error tolerances were undertaken for several model configurations and data resolutions. However, the use of strict near-exact solutions in this study was limited given the high cost of dense substepping for the large number of computationally demanding inference setups, while adaptive substepping with coarse tolerances was not used because of the roughness of the resulting objective functions complicating parameter analyses [*Kavetski and Clark*, 2010]. Therefore, the IE scheme in this work was not error controlled, which is an undeniable limitation that is elaborated on in sections 7.4.2 and 7.6. We also note that previous studies, including a broad evaluation over 6 distinct models and 13 basins with diverse hydroclimatic and physical properties, have suggested that even when applied with daily steps, the numerical errors of conceptual rainfall-runoff models approximated using the fixed step IE method were well below errors because of inaccurate forcing data and model structural defects [*Clark and Kavetski*, 2010].

4. Model Evaluation Methods

4.1. Bayesian Inference Formulations

[36] The hydrological parameters are inferred given observed rainfall-runoff data $\left(\widetilde{\mathbf{P}},\widetilde{\mathbf{Q}}\right)$ using Bayes' equation:

$$p(\boldsymbol{\theta}, \boldsymbol{\Xi}|\widetilde{\mathbf{P}},\widetilde{\mathbf{Q}}) = p(\widetilde{\mathbf{Q}}|\widetilde{\mathbf{P}}, \boldsymbol{\theta}, \boldsymbol{\Xi})p(\boldsymbol{\theta}, \boldsymbol{\Xi}) , \tag{4}$$

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

[37] This study considers two commonly used inference schemes based on distinct assumptions describing the residual errors ξ (the discrepancy between observed and simulated responses). The first approach is the standard least

squares (SLS) scheme, which assumes Gaussian residuals with zero mean and constant standard deviation (i.e., homoscedastic). The second approach is the weighted least squares (WLS) scheme, which also assumes zero-mean Gaussian errors but allows for heteroscedasticity. Here we hypothesize that the standard deviation of individual residuals increases linearly with the corresponding streamflows [e.g., *Thyer et al.*, 2009].

[38] The likelihood functions for SLS and WLS are as follows:

$$p(\widetilde{\mathbf{Q}}|\widetilde{\mathbf{P}}, \boldsymbol{\theta}, \boldsymbol{\Xi}) = \prod_{n=1}^{Nt} N[\xi_n(\widetilde{\mathbf{P}}, \boldsymbol{\theta}, \widetilde{Q}_n)|0, \sigma_n^2], \tag{5}$$

$$\begin{array}{ll} \text{SLS} & \sigma_n = \sigma, \\ \text{WLS} & \sigma_n = a + b\widetilde{Q}_n, \end{array} \tag{6}$$

where $N(x|m,s^2)$ is the probability density function of a Gaussian deviate x with mean m and variance s^2, N_t is the number of observations, and σ_n is the standard deviation of the nth residual error $\xi_n = \widetilde{Q}_n - Q_n(\widetilde{\mathbf{P}}|\boldsymbol{\theta})$. In the case of SLS, $\boldsymbol{\Xi}$ contains the (unknown) standard deviation of the residuals, σ (mm/h); in the case of WLS, $\boldsymbol{\Xi}$ comprises the (unknown) parameters a (mm/h) and b (dimensionless) controlling the heteroscedasticity of residual errors.

[39] Given the probability models (5) and (6), the total uncertainty in the predicted streamflow comprises the exogenous ("residual") error term (here, additive Gaussian noise with variance given by equation (6)) and also the effects of the posterior uncertainty in the model parameters. A notable omission from the inference equations (5) and (6) is a specialized treatment of data and structural uncertainties, e.g., in the catchment-averaged rainfall forcing. At best, these uncertainties are crudely lumped within the exogenous error term [*Renard et al.*, 2010]. While a more robust treatment of input and structural uncertainties is a major goal in hydrological modeling (e.g., as pursued by *Beven and Binley* [1992], *Kavetski et al.* [2002b, 2006b], *Bulygina and Gupta* [2009], *Reichert and Mieleitner* [2009], and *Vrugt et al.* [2005, 2008]), its generally large data analysis and computational requirements make it currently infeasible for the significant number of experiments required in this study, including those with multiple models and high-resolution (down to 30 min) calibration data. This limitation, its implications, and future remedies are discussed in section 7.6.

4.2. Analysis of Posterior Distributions

[40] The posterior distributions of parameters and streamflow predictive were explored using the MCMC sampling strategy described by *Thyer et al.* [2009] with a total of 40,000 model runs (five parallel chains). During the first 2000 samples, the jump distribution was tuned one parameter at a time. During the next 2000 samples, the jump distribution was tuned by scaling its entire covariance matrix. The jump distribution was then fixed, and 35,000 samples were collected. The first 25,000 samples were discarded as burn-in [*Gelman et al.*, 2004], and the final 10,000 "production" samples were used to analyze and report the parameter distributions.

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

[42] Given a focus on experimental data analysis using comparatively simple models and inference schemes, a more thorough development, exposition, and discussion of MCMC sampling techniques is beyond our scope here.

4.3. Hydrologically Oriented Model Diagnostics

[43] We consider two different diagnostic measures to scrutinize the inferred hydrological models: (1) the traditional flow duration curve for evaluating the overall streamflow distributions [*Linsley et al.*, 1949; *Dingman*, 1994; *Wagener and Wheater*, 2006] and (2) a diagnostic for elucidating the rainfall-runoff cross-correlation characteristics that may not be apparent solely from inspecting the time series. These analyses focus more directly on the key hydrological variables and signatures of interest and extend more traditional statistical diagnostics, such as evaluations of the marginal Gaussianity and autocorrelation of residuals (e.g., see *Box and Tiao* [1992] for general theory and *Thyer et al.* [2009] for an illustration in rainfall-runoff hydrology).

[44] The flow duration (FD) curves are constructed as the marginal distributions of streamflow. The lower, flatter sections of the FD curve generally correspond to base flow, whereas the higher, steeper sections generally correspond to quick flow. As discussed in the hydrological literature [e.g., *Wagener and Wheater*, 2006], FD curves provide useful qualitative insights into the catchment behavior. For example, steeper FD curves indicate higher variability in the streamflow, which in addition to climatic factors such as rainfall variability, is also tied to storage characteristics arising from geomorphology, topography, vegetation, land use, etc. [e.g., *Linsley et al.*, 1949]. Hence, FD curve analysis diagnoses not only the fidelity of a model's representation of base flow and quick-flow processes but also, indirectly, its correspondence to the physical catchment attributes listed in the previous sentence.

[45] The FD curve, being a marginal distribution diagnostic, suppresses timing ("frequency") information. Since time resolution (sampling frequency) aspects are a key focus of this study, we examine the cross-correlation pattern between streamflow and lagged rainfall, contrasting results for low- versus high-frequency components (corresponding, loosely, to the distinct components of the "double-peak" response; see section 2). The low-pass-filtered data are obtained using Gaussian kernel smoothing of the time series, while high-pass-filtered data are obtained as the difference between the raw data and the low-pass filter. At the expense of foregoing distributional information (which, instead, can be seen in the FD curves), rainfall-streamflow cross-correlation diagnostics highlight timing aspects of the streamflow response with respect to the rainfall input. In particular, the seasonality of the streamflow dynamics of the Weierbach catchment becomes more clearly visible.

[46] We also stress that evaluating the ability of the models to reproduce the double-peak streamflow signature is in itself a diagnostic technique. Given the experimental evidence that the streamflow dynamics in the Weierbach catchment are controlled by differences in the time scales of the riparian versus subsurface flow paths [*van den Bos et al.*, 2006b], this signature is used to scrutinize the physical realism of the conceptual models.

5. Methodology

[47] Model calibration was based on the 26 month period from 1 July 2005 to 31 August 2007, with the first 2 months used for a warm-up. Model validation was based on the 1 year period from 1 September 2007 to 31 August 2008. The issue of unknown initial conditions in validation, which is tangential to the focus of our experiments, was minimized by using the preceding calibration period as a warm-up. To further reduce potential biases, we used exactly the same initialization setup in all model runs: reservoirs with finite storage were initialized to 20% of their capacity, while other stores were initialized as empty. Identical initialization means that any differences in the results are due solely to the only factors that were varied as part of the research objectives: data resolution, model complexity, time-stepping approximation, and the residual error model for calibration. Empirical analysis further suggested that the sensitivity of the results to the initial conditions was low.

[48] The parameters of the four hypothesized model structures were inferred from calibration data with eight different time resolutions ranging from 30 min to 3 days. In these experiments, the total observation period was fixed, and only the temporal resolution of the time series data was varied. We stress that the fine-scale data were obtained directly from the rainfall-runoff observation network, rather than estimated by subscale disaggregation of coarser measurements [e.g., *Kandel et al.*, 2005].

5.1. Parameter Uncertainty Analysis

[49] The posterior distributions of the parameter estimates and streamflow predictions were examined, both in calibration and validation, with respect to (1) the resolution of the calibration data, (2) the numerical implementation of the hydrological model hypothesis, and (3) the likelihood function used in the inference.

[50] Since the calibration period is fixed, increasing the data resolution results in a larger number of observations being used in the inference. Strictly speaking, this affects both the accuracy and precision of the forcing response time series (e.g., aggregation over smaller scales leads to less averaging of measurement errors and may result in larger data uncertainty), as well as the characteristics of both data and model errors (e.g., stronger autocorrelation at finer time scales). However, in the absence of reliable prior understanding of either data or model errors, we simply use the SLS and WLS methods with unknown residual error model parameters. Since these inference methods are widely used in hydrological research and practice, illustrating and understanding their behavior is important for ensuring a better and broader recognition of their limitations and how these limitations are likely to affect previous, current, and future scientific and operational applications (see section 7.6 and *Renard et al.* [2010] for further analysis and discussion).

[51] In the SLS and WLS schemes, an increase in the number of observations N_t leads to a reduction of uncertainty in the inferred parameters (though not in the response predictions; see section 4.1). Asymptotically,

$$\mathrm{sdev}[\theta|N_t] \propto 1\Big/\sqrt{N_t}\,, \tag{7}$$

where $\mathrm{sdev}[\theta|N_t]$ is the posterior standard deviation of parameter θ given N_t observations. This behavior is a property of most statistical inferences [e.g., *Box and Tiao*, 1992]. However, while highly germane in the broader context of parameter estimation [e.g., *Mantovan and Todini*, 2006; *Beven et al.*, 2008; *Stedinger et al.*, 2008], these effects are tangential to this work, which focuses on the impact of data resolution rather than data length. We refer the interested reader to *Brath et al.* [2004] and *Merz et al.* [2009] for a thorough empirical exploration of the effects of calibration data length given a fixed resolution.

[52] Instead, our primary interest here is on how some processes, e.g., quick flow, that are obscured by the data averaging inherent at coarse time scales are gradually revealed (or "disclosed," in the language of *Kirchner et al.* [2004]) as the resolution of the calibration data is refined. In order to more clearly present how uncovering these finer-scale dynamics impacts on posterior uncertainty, we "standardize" all parameter uncertainty intervals by dividing the width around their means by factors of $\sqrt{N_{\mathrm{days}}}\Big/\sqrt{N_t}$. This scaling allows an examination of whether refining the data resolution improves the parameter precision beyond the reduction expected in SLS and WLS merely from an increased number of observations. For convenience and ease of visualization, the term $\sqrt{N_{\mathrm{days}}}$ is included to scale all intervals relative to the uncertainties obtained for daily data.

[53] We considered several alternative approaches for isolating these aspects. For example, varying the data length and its resolution could keep N_t constant yet would change the calibration period and potentially bias the results by including additional storm events into the coarser-resolution experiments. While other "effective sample size" concepts could be employed [e.g., *Brillinger*, 1989; *Hamed and Rao*, 1998], these formulas depend on subjectively defined statistical parameters and would make it difficult to maintain a methodological consistency across the experiments, which spanned four models and eight different time resolutions. In addition, they correspond to modified SLS and WLS schemes, whereas our interest here is on common inference setups.

[54] Finally, the issue of effective sample size is closely related to the autocorrelation in the response time series and the model residuals. For example, hydrograph recessions are generally highly autocorrelated, and hence, increasing the number of recession observations may add little information to the inference. These effects are outlined in section 6.1.2, though a detailed quantitative analysis, based on incorporating autocorrelation into the error models and the likelihood functions, is deferred to a separate study.

[55] In addition to the primary focus of this study on experimental data insights, section 7.3 briefly outlines a

mathematical context for interpreting the empirical findings. This will be elaborated in a separate investigation.

5.2. Analysis Over a Range of Model Complexities

[56] The topic of model identification and structural complexity appropriate for hydrological modeling at a range of time scales is explored by evaluating the model performance, both in calibration and validation, as a function of calibration data resolution and model complexity. The latter is assessed in terms of both the total number and connectivity of model components and the parametric complexity of the constitutive (flux) functions.

[57] In the empirical assessments, we build up the optimal combination of modeling and calibration decisions by first considering the impact of the numerical scheme and then examining the influence of the likelihood function. This approach is used because the numerical implementation of a model is a "lower-level" decision than the selection of a likelihood function for its calibration. In particular, the objective function is generally application specific (it depends on the data uncertainty, model performance requirements, etc.), whereas, once selected, the numerical implementation of a model would not normally be altered except perhaps in special circumstances (e.g., to boost computational speed at the expense of accuracy in some time-critical context).

6. Empirical Findings Using Experimental Catchment Data

6.1. Time Scale Trends in Inferred Parameter Distributions

[58] This section explores the time scale dependencies of parameter distributions inferred using various combinations of likelihood functions and time-stepping schemes. The parameter distributions of all hydrological models are shown in Figures 2 and 3 and are presented as a function of the calibration data resolution.

6.1.1. Impact of Numerical Approximation

[59] The SLS results for models solved using the IE and EE schemes are shown in Figures 2 and 3. In general, the EE approximation introduces markedly stronger time scale dependencies. For example, parameter β, which controls the nonlinearity of the saturated area with respect to storage, exhibits an increasing trend as Δt increases when the EE scheme is used but has stable values for the IE scheme. Similarly, parameter $R_{\max}$ sharply decreases as Δt increases for the EE scheme but remains fairly stable for IE solutions. Worrisomely, the contrasts between the time scale trends in EE versus IE approximations are particularly pronounced for the more complex model hypotheses M3 and M4, which as will be seen in section 6.2, appear to provide a better and more physically interpretable overall representation of the catchment dynamics. Unlike the unconditionally stable IE scheme, the numerical stability of the EE scheme is very sensitive to flux nonlinearities: the strong time scale sensitivity of parameter β in models implemented using the fixed step EE scheme is therefore unsurprising.

6.1.2. Impact of Likelihood Function

[60] Figures 2 and 3 also show that WLS generally provides more stable parameter estimates than SLS. This held across the full range of time scales explored in this study. For example, while K_f exhibits significant trends when estimated using SLS, it has fairly stable values when inferred using WLS. In general, most model parameters have stable values when WLS is used, even for quite coarse time resolutions. In other cases, such as the delay parameter T_f of the transfer function, there remained time scale dependencies at coarse calibration data resolutions, but they disappear with finer resolution. The parameter stabilization occurs for data resolutions that are finer than the time scale of their corresponding processes: beyond this point, additional resolution does not contribute new information, and the parameter estimates remain constant. For example, the estimates of T_f largely stabilize for $\Delta t \leq 6$ h, suggesting that the dynamics of this catchment are subdaily.

6.1.3. Uncertainty Estimation

[61] Figures 2 and 3 also show how parameter uncertainty varies across time scales. As discussed in section 5.1, the effects of the asymptotic reduction in uncertainty with the number of observations are avoided by standardizing the uncertainty limits. This helps isolate the effects of data resolution, which are of central interest here.

[62] For most parameters, as the data resolution was refined, there was little uncertainty reduction beyond that expected from asymptotic arguments. This was the case for both SLS and WLS. However, some important exceptions are noted, e.g., the quick-flow parameter K_f in model M4, the uncertainty of which varied significantly before stabilizing for $\Delta t \leq 12$ h. Hence, uncertainty reduction effects depend on the time scale of the process.

[63] Our primary interest here is on how some processes, e.g., quick flow, are obscured by data averaging at coarse time scales but are gradually "revealed" as the resolution of the calibration data is refined. In particular, this is the case for parameter K_r, which controls the "quick reaction" of the model. It is largely nonidentifiable when $\Delta t \geq 6$ h but converges to fairly stable values when $\Delta t \leq 1$ h. This, in itself, is indicative of the time scale of the process (section 6.3). The effect of data resolution on quick-flow processes in the context of model prediction (as opposed to parameter estimation and structure identification) is elaborated in sections 6.3 and 7.2.

6.2. Time Scale Trends in Supported Hydrological Model Complexity

[64] We now consider the impact of calibration data resolution on the trade-offs between model complexity and predictive performance.

6.2.1. How Do Complex Models Behave With Respect to Data Resolution?

[65] It has been suggested that the "appropriate" model complexity generally increases as the temporal resolution of the data is refined [*Atkinson et al.*, 2002], implying a higher information content in high-resolution data. Our results suggest that this intuitive hypothesis remains valid at subdaily scales. For example, parameters related to RR (M and K_r in M4), which simulates the catchment's quick response to rainfall, become progressively better identified as the data resolution is refined (Figure 3), thus supporting additional model complexity.

[66] Conversely, Figure 3 also shows that progressive averaging of the calibration data reduces the identifiability

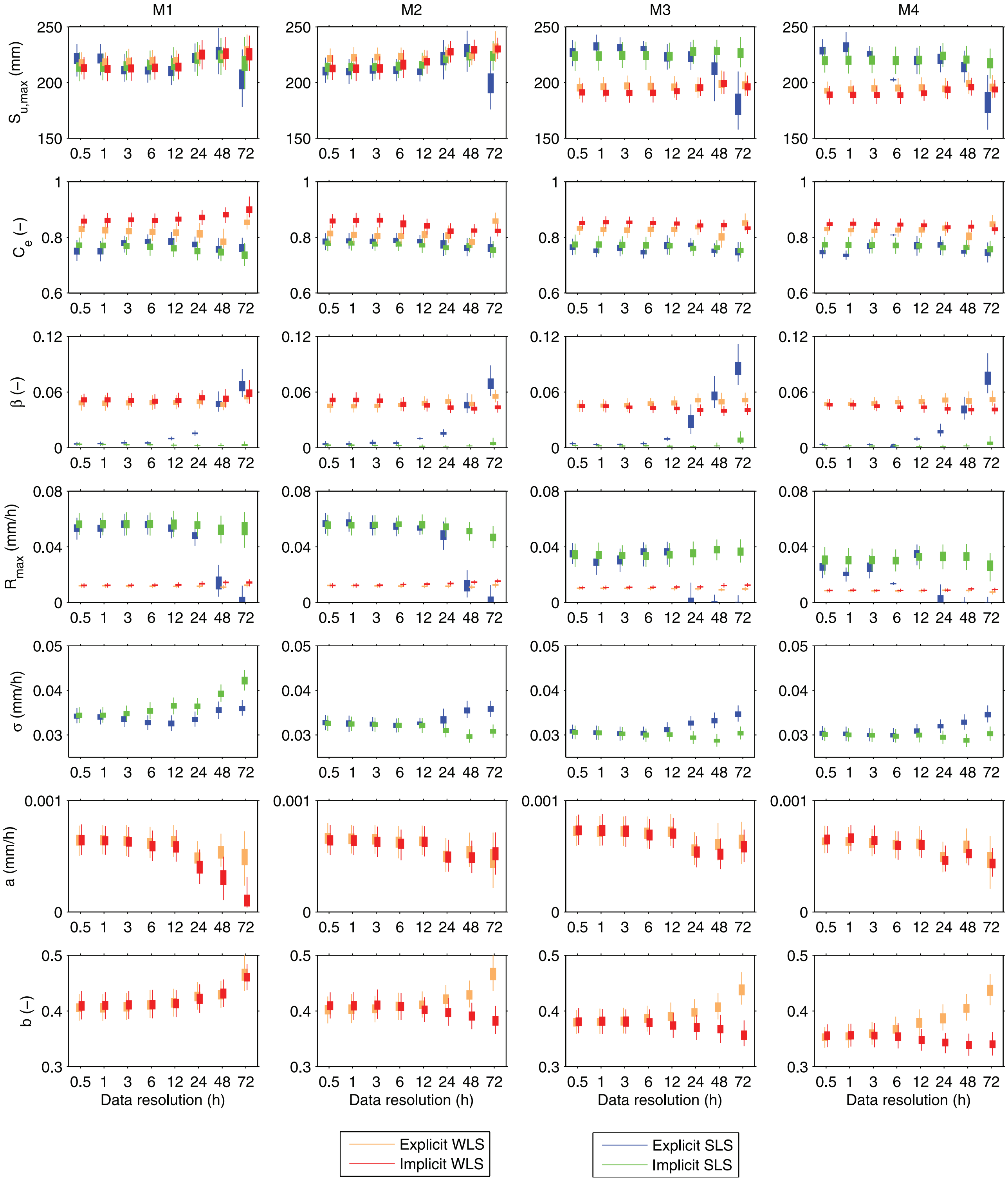

Figure 2. Inferred distributions of parameters common to model structures M1–M4. Box and whiskers denote the 50% and 95% quantiles, respectively. The uncertainty intervals are "standardized" by a factor of $\sqrt{N_{\text{days}}}/\sqrt{N_t}$ in order to partially remove the effects arising from the different number of observations and to use the results obtained with daily resolution as the basis for the comparison (see section 5.1). It can be seen that the fixed step explicit Euler time-stepping scheme and standard least squares (SLS) inference tend to introduce markedly strong time scale dependencies in many parameters, especially for fast-scale ("quick-flow") processes.

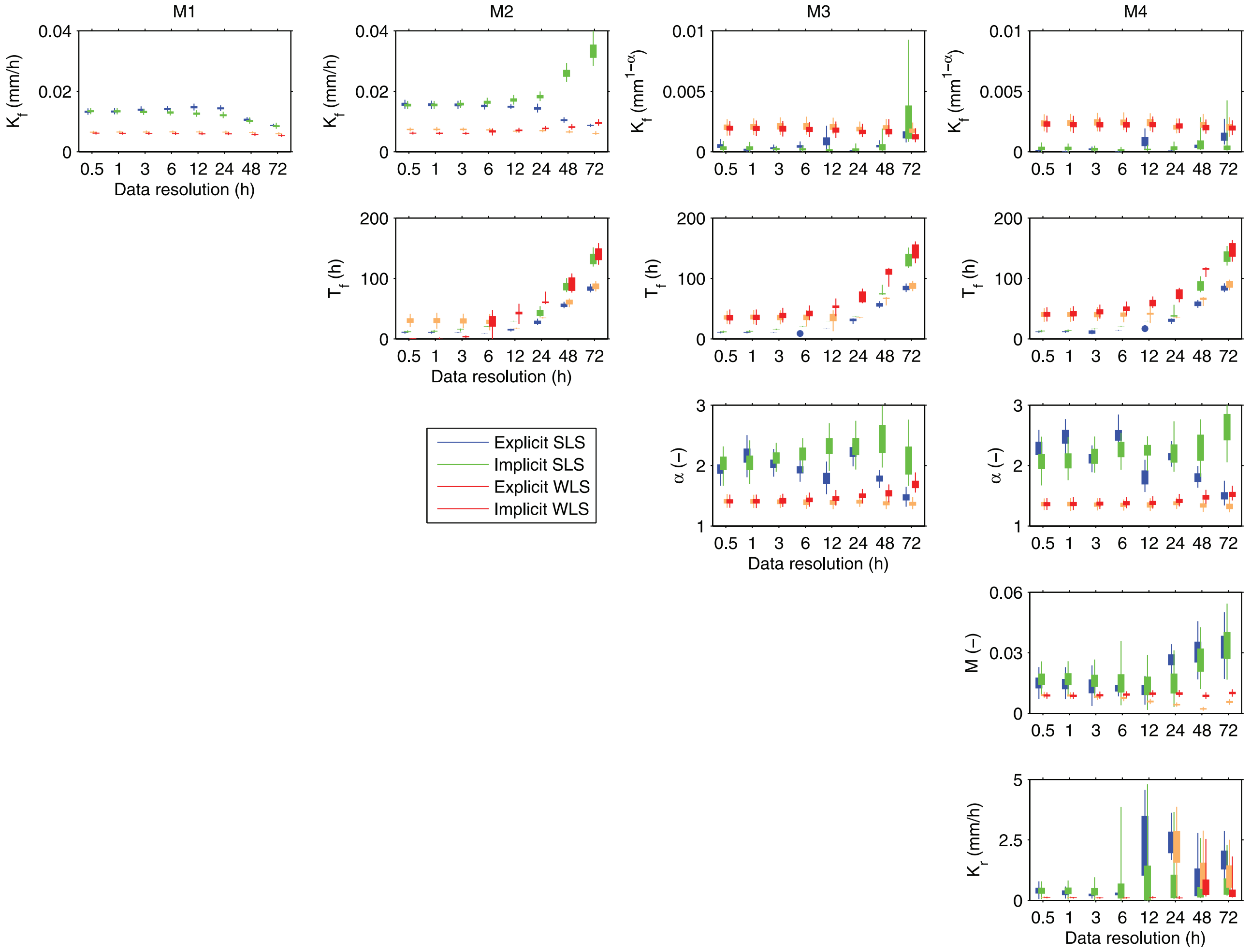

Figure 3. Inferred distributions of parameters specific to model structures M1–M4. Box and whiskers denote the 50% and 95% quantiles, respectively. Note that parameter K_f in structures M3 and M4 is distinct from K_f in M1 and M2 because it parameterizes a nonlinear reservoir (Table 3). The uncertainty intervals are "standardized" by a factor of $\sqrt{N_{\text{days}}}/\sqrt{N_t}$ in order to partially remove the effects arising from the different number of observations and to use the results obtained with daily resolution as the basis for the comparison (see section 5.1). It can be seen that the fixed step explicit Euler time-stepping scheme and SLS inference tend to introduce markedly strong time scale dependencies in many parameters, especially for fast-scale (quick-flow) processes.

of quick-flow parameters. This is particularly pronounced in this catchment, which responds quickly to precipitation falling on the near-stream impervious areas. The resulting hydrograph peak has a subhourly characteristic scale and disappears shortly after the rainfall event. The averaging of observed data above hourly scales smears this feature of the catchment response, making the quick-flow components of the model apparently "redundant."

[67] Validation performance may decrease with increasing model complexity if the model is overconditioned, i.e., with complex models using their degrees of freedom to fit data errors and noise and then performing poorly or unstably in prediction or extrapolation [e.g., *Mitchell*, 1997; *Kingston et al.*, 2008]. This was not observed in our experiments: provided a robust numerical approximation of the conceptual model was used, the relative ranking of performances during calibration and validation remained remarkably constant (Figure 4). This suggests that even the more complex models examined here are parsimonious with respect to the calibration data. Several previous studies have also suggested that physically motivated models are less prone to "overfitting" than purely data-driven models [e.g., *Dawdy*, 1983; *Schoups et al.*, 2008].

[68] Finally, note the effect of model structure on the time scale dependencies of the parameters. For example, exceedingly simple models may display strong time scale trends for parameters that compensate for gross structural errors. This can be seen for parameter K_f, which when fitted using SLS, is highly scale dependent in simple models (e.g., in hypothesis M1, which lacks a routing store), but becomes fairly stable in more complex models. The scale dependencies of the residual error parameters may also differ depending on the complexity of the hydrological model: e.g., parameters a and b fitted as part of WLS are highly scale dependent in the simplistic hypothesis M1 but become progressively more stable in more complex model structures (Figure 2).

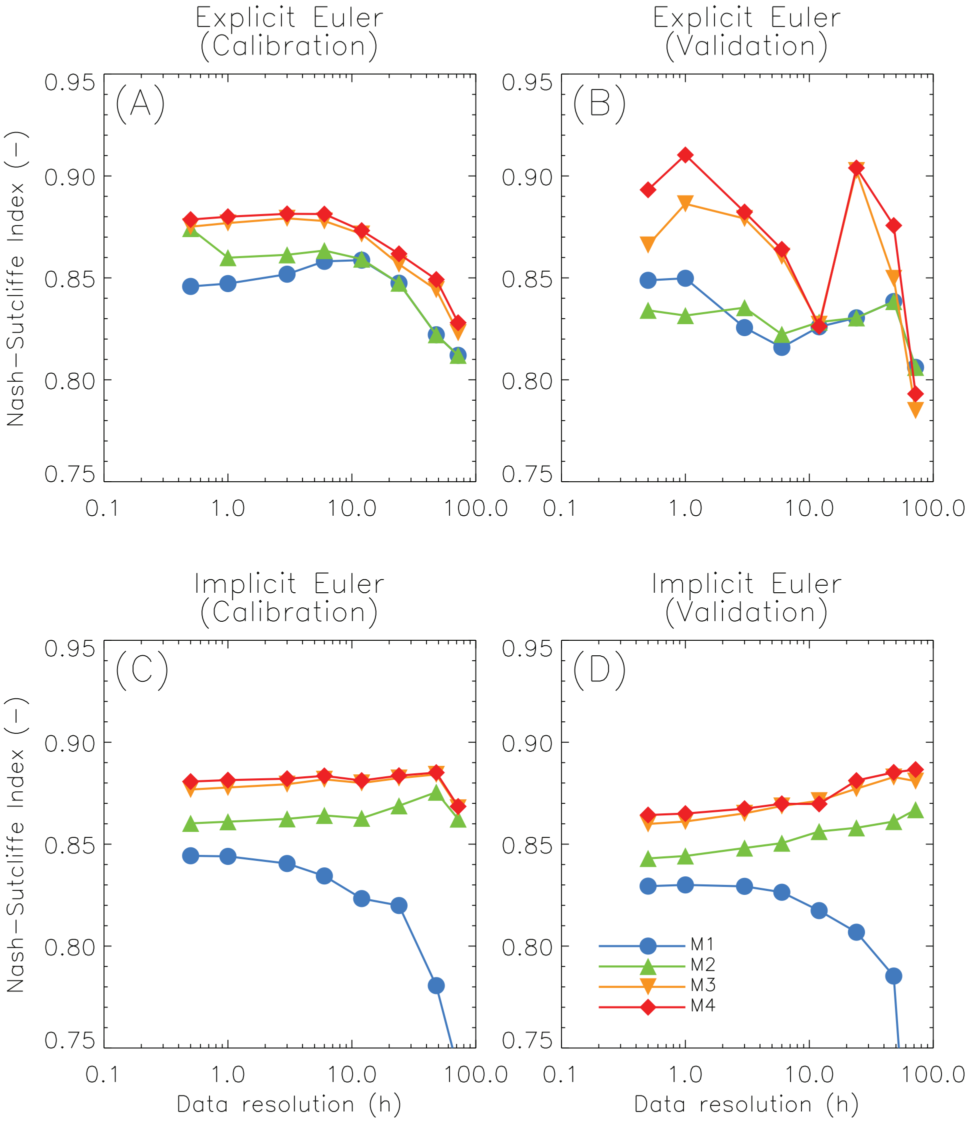

Figure 4. Effect of data resolution on hypothesis testing (including calibration and validation) of the models implemented using the explicit Euler (EE) and implicit Euler (IE) schemes. The runoff errors are quantified using the Nash-Sutcliffe (NS) index, and the results shown are therefore for the SLS inference (which maximizes the NS criterion). Faced with 10% swings in behavior when evaluating the model over a range of time scales in validation, a hydrologist relying on fixed step EE approximations as part of their hypothesis testing may be tempted to conclude that model structures M3 and M4 are overparameterized. Yet it is precisely these hypotheses that perform consistently best when implemented using numerically robust techniques. They are also the most realistic when judged against process understanding available in this experimental basin.

6.2.2. Numerical Approximation Effects

[69] Consider Figure 4, which contrasts the model performance, quantified using the Nash-Sutcliffe criterion in calibration versus validation, as a function of model complexity and calibration data resolution. In Figure 4 the IE versus EE approximations of the model equations calibrated using SLS and WLS are also contrasted.

[70] In terms of calibrated performance, EE-based models often have similar performance as IE-based models. Moreover, when calibrating the simplest model M1 at near-hourly resolutions, the EE approximation often has smaller overall errors than the IE approximation. For this specific model-data combination, this can be explained by the differences in the way the fluxes are approximated by these numerical techniques. While the IE scheme synchronizes all fluxes with the storage at the end point of the step, the EE scheme evaluates all the outflows before the inflows are added to the stores. This introduces a delay into the response dynamics, which turns out to be beneficial for model M1 because it lacks an explicit routing component. However, this is a delicate and unreliable interaction: when the routing component is included (configurations M2–M4), the implicit method outperforms the explicit approximation. More generally, we strongly oppose attempting to use numerical artifacts to compensate for structural weaknesses in the model [see *Kavetski and Clark*, 2010].

[71] A broader inspection shows that the predictive performances of the IE scheme in calibration and validation is much more stable than for the EE scheme. For example, at a 1 h resolution, the Nash-Sutcliffe performance of configuration M4 implemented using the EE scheme dropped by 5%, whereas for the IE implementation the deterioration was only 1%. Notably, this was the best performing model configuration, which was also favored on physical grounds because it includes the fast reservoir to represent the fast schistose response of the Weierbach catchment (sections 2 and 6.3). More generally, at near-daily scales, the deterioration of IE-based models was around 0.0%–1.6% for all hypotheses, worsening slightly (to 1.8%–2.0%) for validation on hourly and subhourly scales (which represents a particularly stringent test of predictive ability for a lumped conceptual model). Conversely, the degradation of the EE method in validation periods was largest at near-daily scales (e.g., see the 10% drop in performance of models M3 and M4 when switching from 24 h resolution to 12 h data).

[72] These inconsistencies are concerning, especially in a small experimental catchment where data uncertainty and structural errors are likely to be lower than in larger basins (e.g., Figure 3.7 of *Linsley and Kohler* [1958] indicates smaller rainfall errors over smaller areas, likely because of lower overall variability of the rain field). In an empirical investigation over 12 Model Parameter Estimation Experiment (MOPEX) basins [*Duan et al.*, 2006], *Kavetski and Clark* [2010] report a comparable or stronger degradation for models based on fixed step explicit time-stepping algorithms. Especially in the absence of independent information about data and structural errors, a modeler unaware of numerical errors could readily begin misattributing such deficiencies to poor model conceptualization, input data errors, etc.

[73] Similar results were obtained across most time scales and model complexities considered in our experiments. For example, models implemented using the EE scheme suffered, on average, a 50% larger degradation in predictive performance than those implemented using the IE scheme, as measured using Euclidean norms of the differences between the calibration and validation Nash-Sutcliffe performance.

6.3. Data Resolution and Model Complexity

[74] The hydrographs of models M1 and M4 calibrated at 1 and 24 h resolution are shown in Figure 5. Figure 5, in particular, the zoom plots in Figures 5h–5j, illustrates how high-frequency quick-flow processes are obscured by the smearing of the forcing and response data, which is inevitable when averaging over larger time scales [see also *Ostrowski et al.*, 2010]. As various features of hyetographs and hydrographs are smeared by data averaging, the model components intended for their representation become progressively nonidentifiable. For example, this may explain the high uncertainty in parameter K_r shown in Figure 3.

[75] While relatively coarse calibration data resolution could, perhaps, be tolerable in slowly responding catchments, any basin idiosyncrasies are liable to be obscured. This loss of physical realism can be detected using model diagnostics [*Gupta et al.*, 2008]. For example, the high-resolution streamflow data in Figure 5e shows that at short time scales the wet season response of the Weierbach catchment is characterized by a double peak (see section 2 for current experimental insights). Only the most complex configuration, M4, is able to even qualitatively capture this behavior because of the inclusion of a riparian submodel. However, deficiencies are still evident: the shape of the second peak is poorly captured, perhaps because of inaccurate antecedent conditions and/or remaining structural errors in slow-flow components.

[76] Also note that as seen in Figure 5, the second peak disappears in the summer season, when the saturated area of the catchment is reduced because of drier conditions [*Pfister et al.*, 2002]. All models are able to reproduce this seasonality behavior using the unsaturated store, which partitions precipitation into fast and slow components.

6.4. Model Diagnostics Based on Flow Duration Curves

[77] Flow duration curves diagnose a model's ability to represent the overall streamflow distribution [e.g., *Yilmaz et al.*, 2008]. Analysis of the flow duration curves in validation, shown in Figure 6, suggests the following:

[78] 1. The observed flow duration curves for hourly and daily data are quite similar, though the averaging of high-intensity, short-duration storm events does reduce the maximum streamflow values in the daily resolution data (when compared to hourly resolution data). The overall similarities arise because by its very construction, the flow duration curve is a purely distributional signature that discards all timing information (this particularly affects high-resolution data). Moreover, since high flows are rare in this catchment (e.g., the inset plots in Figure 6 show that flows exceeding 0.5 mm/h occur with a frequency of less than 0.5%), the flow duration curve necessarily focuses attention on low flows, when daily and hourly flows are fairly similar.

[79] 2. Overall, once calibrated, model structure appears to make little difference on the ability of the models to reproduce the observed FD curve. While structure M4 matches the flow duration curve better than M1 (consistent with the overall better performance of the M4 model) for a given objective function, the improvement in fit is less than when switching objective functions (see below). These findings are relatively unsurprising: the key challenge in the Weierbach catchment is reproducing the timing signatures and the seasonal dynamics, rather than flow volume distributions. For this, a different set of diagnostics is needed (section 6.5).

[80] 3. SLS-calibrated models fail to adequately represent base flow processes, while the linear heteroscedasticity assumption in the WLS inference, though forcing a better representation of low flows, is too lax for high flows (for which SLS-based models do better, though arguably, for the wrong reasons). This behavior is well known: absolute error criteria emphasize fitting high flows, whereas relative error criteria (or log transformations) emphasize fitting low flows [e.g., *Schaefli and Gupta*, 2007]. At least from general considerations, we expect the heteroscedastic model to be more realistic [e.g., *Sorooshian*, 1981]. However, what is missing here is the extension of the likelihood function to account for the autocorrelation of residual errors [e.g., *Sorooshian and Dracup*, 1980; *Schoups and Vrugt*, 2010], which is likely to increase strongly when fitting to high-resolution data.

[81] 4. Quite remarkably, at least for the case of high flows predicted using structure M1, changing the time-stepping scheme makes almost as much difference as changing

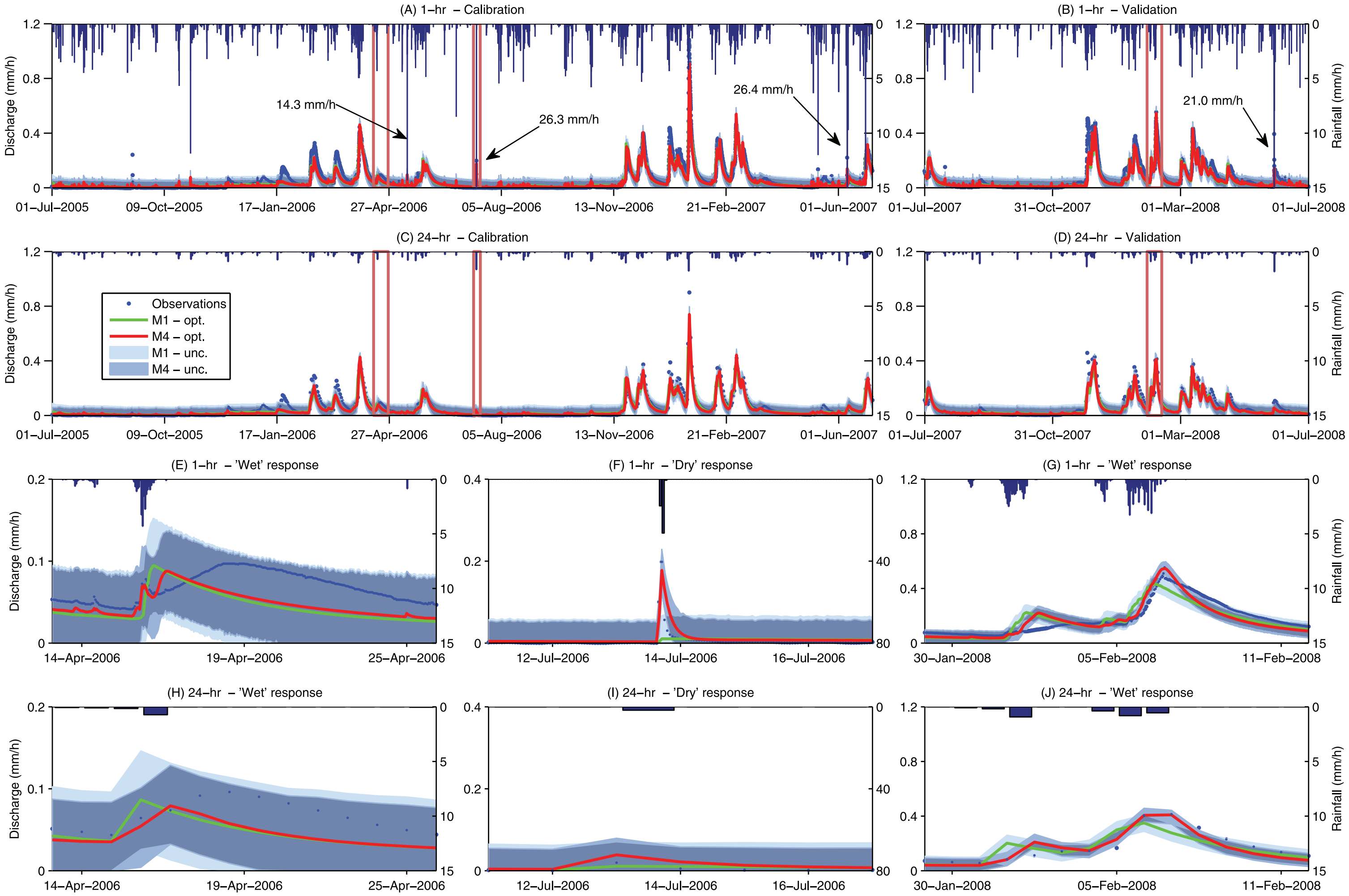

Figure 5. Streamflow predictions and total uncertainty (95% prediction limits) for model hypotheses M1 and M4, inferred from hourly versus daily resolution time series, for the calibration and validation periods. The smearing of the rainfall-streamflow time series erases the double-peak signature (Figure 5h) and other fast-scale features (Figure 5i). The models are "seeing" considerably different data, and this affects process identifiability: the improvement of model M4 over M1 is markedly reduced when daily data are used.

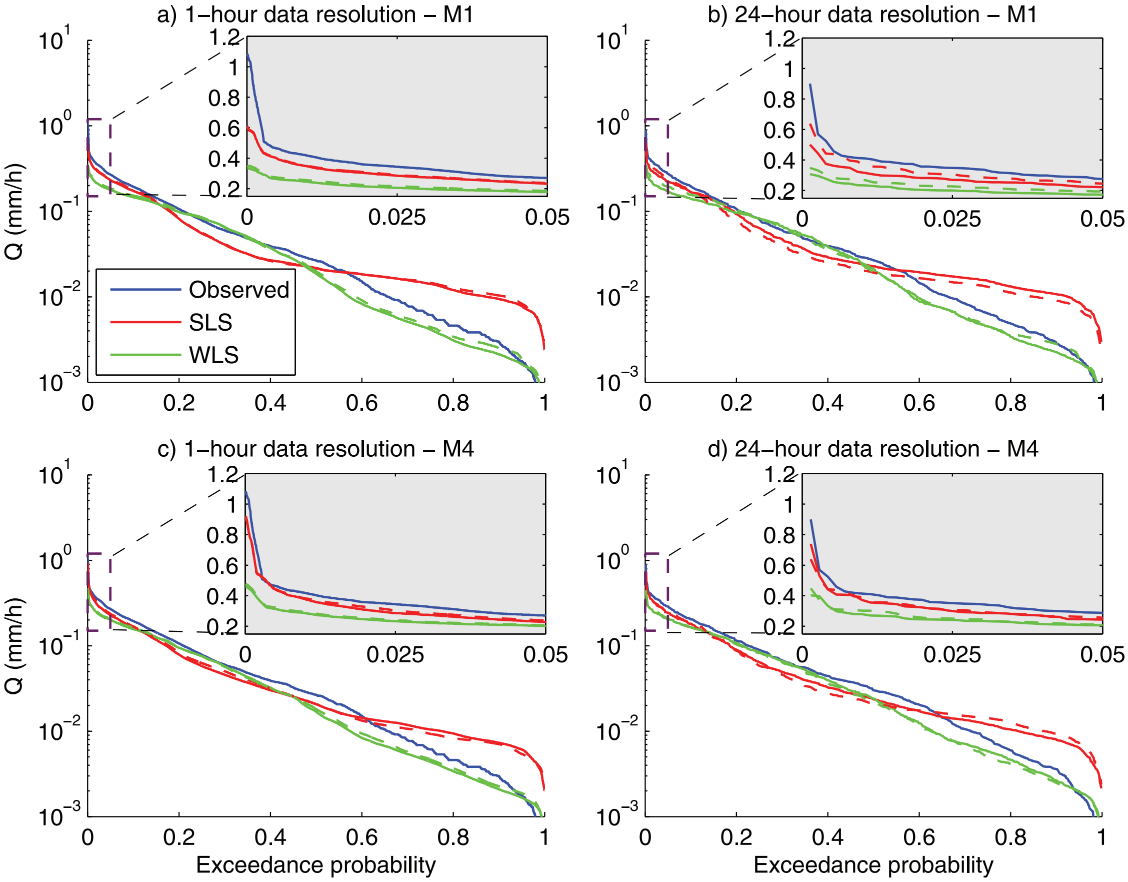

Figure 6. Flow duration curves over the validation period for selected model structures and time resolutions. Red and green lines contrast the results of standard least squares (SLS) versus weighted least squares (WLS) inferences, and solid and dashed lines contrast models solved using the implicit Euler (IE) and explicit Euler (EE) approximations. In the Weierbach catchment, model structure appears to make little difference on the flow duration curves, likely because the main modeling challenge lies in reproducing the timing of the storms, rather than their overall volumes. However, numericostatistical effects are still strong: SLS-calibrated models fail to adequately represent base flow processes, while the linear heteroscedasticity assumption in the WLS inference appears to overestimate the errors in the high flows. It can also be seen that at least for the case of high flows predicted using structure M1, changing the time-stepping scheme makes almost as much difference as changing the objective function.

the objective function. While the impact of objective functions on model identification is well known in conceptual hydrology, numerical approximation errors have traditionally been seldom, if ever, analyzed and reported during model development and hypothesis testing.

6.5. Model Diagnostics Based on Timing Analysis

[82] The cross-correlation diagnostics shown in Figure 7 evaluate the ability of the models to reproduce fast- and slow-scale timing dynamics, contrasting results obtained with hourly versus daily resolution. Similar to the FD curves, the cross-correlation diagnostics are applied to the validation period to provide more stringent scrutiny.

[83] Consider the analysis of hourly data (Figures 7a–7d). Although all models performed well for the slow-scale dynamics (consistently with the FD curve diagnostic), adequate representation of fast dynamics requires the explicit inclusion of a riparian zone reservoir. For example, structure M1, which lacks a routing component, performs badly overall and tends to produce the fast response too quickly during winter (Figure 7b). The addition of a routing component in structures M2 and M3 reproduces the delayed flow behavior (Figure 7a and, to a lesser extent, Figure 7c) and improves the fit of the fast response in winter (Figure 7b) but fails to capture the fast-scale dynamics in the summer season (Figure 7d). Only structure M4, which includes an additional reservoir specifically intended to represent riparian zone, qualitatively reproduces the lag correlation signatures of the observed data for both wet and dry and both slow- and fast-scale responses. We do note that although the slow-scale dynamics are reproduced well during summer in the calibration period (not shown), the slow-scale summer dynamics, although weak (Figure 5f), are underestimated in validation (Figure 7c). Also note that the cross correlations at high lags in Figures 7b and 7d are noisy and unlikely to be physically meaningful.

[84] Yet process component identifiability is strongly time scale dependent. This can be seen by comparing Figures 7a–7d with Figures 7e–7h. The use of lower-resolution daily data (Figures 7e–7h) notably degrades the capability for incisive hypothesis testing. Structure M1 still performs badly

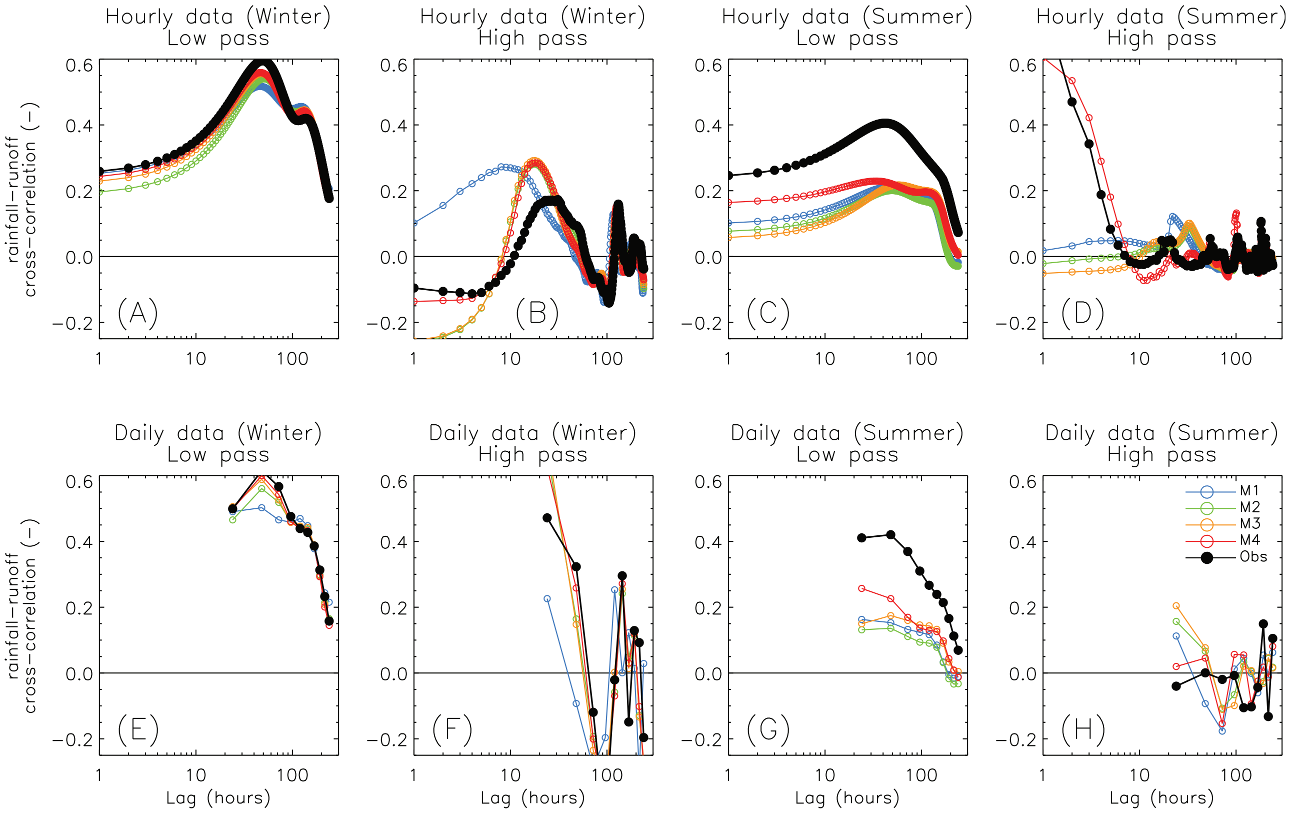

Figure 7. Rainfall-runoff cross-correlation diagnostics for evaluating the ability of the models to reproduce low- and high-frequency signatures in the streamflow response. The application to the validation period is shown, with stratification into wet (winter) and dry (summer) seasons. Only the most complex hypothesis, M4, which directly includes a riparian zone representation, is able to approximate the pattern, although not the precise numerical values, of the fast-response dynamics (isolated using a high-pass filter). Averaging the data to the daily scale smears the data signature, rendering the fast-scale riparian zone processes nonidentifiable. Note the comparatively poor representation of the lagged response in summer (dry season), when it is generally weak.

(because the routing time in this catchments appears to exceed the daily scale), but the other three structures perform similarly, with structure M4 not offering any worthwhile improvement in any of the daily signature aspects (compare, for example, Figures 7d and 7h). The riparian zone representation is simply not identifiable from daily data: its very signature for small lags below 10 h has been erased by data averaging.

[85] Finally, the fairly modest values of the cross correlations (0.4–0.6) suggest that the response complexity of this basin is only partially captured by a linear correlation analysis. The analysis can be obscured by temporal variability and errors in the rainfall and streamflow time series, variability in antecedent conditions, saturation threshold dynamics, nonlinearities in base flow behavior, etc. These issues impose limitations on what can be learned by purely statistical analysis and highlight the critical need for independent experimental insights.

7. Discussion

7.1. Parameter Inference: Spurious Trends Versus Genuine Scale Dependencies

[86] Strong time scale dependencies of hydrological model parameters are not a secret. Many previous studies have shown that model parameters are tied to the time scale and data length at which they are calibrated [*Blöschl and Sivapalan*, 1995; *Schaake et al.*, 1996; *Littlewood and Croke*, 2008; *Merz et al.*, 2009], pointing out that this confounds adequate selection and identification of the governing model equations and hampers the physical interpretation of model parameters. In turn, this undermines more ambitious hydrological undertakings, such as parameter regionalization and prediction in ungauged basins [*Littlewood and Croke*, 2008].

[87] This paper yields several practical insights into the origins and behavior of time scale trends. In particular, it illustrates the ease with which spurious time scale dependencies arise from nonrobust numerical design of the hydrological model and/or poor selection of the likelihood function in calibration. More generally, the complex interplay between data resolution, model complexity, numerical approximations, and parameter inference (including uncertainty estimation) is evident from the empirical analyses [see also *Kavetski and Clark*, 2010].

[88] In models implemented using unreliable approximations, such as the fixed step explicit Euler scheme, scale dependencies will often be dominated by confounding numerical artifacts. On the other hand, numerically robust models [e.g., *Vaché and McDonnell*, 2006; *Clark et al.*,

2008] may still have time scale dependencies because of the uncovering of finer-scale process dynamics by higher-resolution data (section 6.2). Finally, models employing heuristic time-stepping (e.g., the Sacramento model [*Burnash*, 1995], which uses operator splitting with substepping) could experience a mixture of effects, depending on the catchment, the calibration and design data, and the specifics of the numerical implementation.

[89] Our results also consolidate the findings of previous theoretical and empirical work that demonstrated the importance of a careful selection of error models in calibration [e.g., *Box and Tiao*, 1992; *Thyer et al.*, 2009]. In particular, ignoring the heteroscedasticity of residuals not only leads to poor prediction limits but also appears to make the inference more sensitive to the resolution of the forcing data. Conversely, the heteroscedasticity underlying WLS not only provides more intuitively meaningful uncertainty bands, but through a (still very incomplete) reduction in statistical artifacts due to a misspecified likelihood function, can also facilitate the translation of the increased information inherent in high-resolution data into hydrological models capable of improved representation of finer-scale catchment dynamics. Importantly, more complicated model configurations become better identified, which was not always the case when SLS was employed (Figures 2–4).

[90] Genuine time scale dependencies of the hypothesized governing equations may yield useful physical insights into process representation, especially with respect to correlations scales and threshold identification [e.g., *Western et al.*, 2005; *Tromp-van Meerveld and McDonnell*, 2006a], and characteristic mixing scales of internal catchment flows [e.g., *Fenicia et al.*, 2010] and, with additional analysis [*Littlewood and Croke*, 2008], may improve operational reliability. Conversely, numerical artifacts corrupt these dependencies, confounding any attempts at physical interpretation of the inferred model structure and causing erroneous and/or misleading predictions (see also the illustrations of *Kavetski and Clark* [2010]).

[91] While the results of the numerical experiments in section 6 are of interest and the reduction in time scale dependencies is encouraging, significant further analysis is warranted. Section 7.6 elaborates on major current limitations and reviews several of the possible strategies for overcoming them in future studies.

7.2. Hydrological Processes Revealed: The Information Content of High-Resolution Data

[92] A comparison of the parameter distributions of different models illustrates that in some cases, simpler models have lower parameter uncertainty than more complex models ($R_{\max}$ in M1 versus M2), whereas in other cases the converse was true (e.g., parameter $S_{u,\max}$ has a higher uncertainty in M1 than in the more complex models). Importantly, this variability is heavily influenced by the data resolution, which sets the time scale of the model.

[93] The time scale at which parameters become identifiable depends on the time scale of the processes they represent, i.e., on their "functionality." In particular, parameters related to slow processes such as base flow do not require such high-resolution data, and Figures 2 and 3 show that these parameters stabilize for time scales much larger than quick-flow parameters (unless numerical or statistical artifacts are present). Indeed, when calibrating base flow components to high-resolution data, the autocorrelation in the model residuals should be accounted for to avoid overconditioning the parameter estimates. On the other hand, meaningful identification of fine-scale quick-flow parameters requires higher-resolution data; otherwise, parameter inference is not very stable with respect to the time scale of application (Figures 2 and 3). As these processes are much "noisier" than base flow and generally tend to follow rainfall forcing patterns, autocorrelation effects may play a lesser role in their identification.

[94] These observations are logical and broadly consistent with previous empirical work on the relationship between data and parameter identifiability. For example, the results of a case study based on moving window calibration [*Wagener et al.*, 2003] indicate that capacity parameters of larger stores are sensitive to volume errors over longer time periods (e.g., total evaporation depends on the storage in the unsaturated zone), while time constants of faster-scale processes (e.g., vertical drainage) are sensitive over much shorter periods. As such, model parameters are sensitive over the time scales at which their respective processes are defined.

[95] The consistency of parameters across model structures, and the degree of consistency with respect to the resolution of the data, is also of interest. In general, given the high degree of conceptualization in current hydrological models, calibration often forces parameters to take unrealistic values to compensate for missing processes, for deficiencies in representing included processes, and for rainfall errors [e.g., *Clark and Vrugt*, 2006; *Beven*, 2006; *Kavetski et al.*, 2006b]. This is clearly undesirable because it implies that model components, or worse, the entire model, become infidelious to the processes they are intended to represent.

[96] In this study, we explore parameter consistencies with respect to the calibration data resolution. In some cases, when a model is calibrated to low-resolution data, its parameters may be able to compensate for missing components. Yet this can lead to either (1) increased uncertainty if a parameterized component is able to smoothly vary between representing different processes or (2) reduced uncertainty if the model component and associated parameter(s) discontinuously switch to represent a totally different process (this can also lead to multimodality). For example, this may explain changes in parameter $R_{\max}$ in model M2 (Figure 2).

[97] Similarly, when the model complexity exceeds the information content of the data (e.g., M4 calibrated to coarse-resolution data), parameter identifiability is generally poor (the classic "overparameterization" problem [e.g., *Dawdy*, 1983], which is often, though not always, detectable [e.g., *Renard et al.*, 2010]). Yet when the complexity of the model allows it to achieve seemingly very good fits to the data by mimicking spurious features (in particular, sampling and measurement errors), the inference scheme can also spuriously underestimate parameter uncertainty (e.g., echoing the concerns of *Beven* [2006], *Stedinger et al.* [2008], *Reichert and Mieleitner* [2009], and *Thyer et al.* [2009] in the (different) context of poorly selected likelihood functions).

[98] A key issue is that these parametric effects can exhibit strong interplays with the data resolution. For example, as fine-scale catchment features are revealed, the ability of simple models to represent increasingly complex response dynamics is reduced, and this will lead to case-specific

interactions between response features favored by the calibration (to the exclusion of other response dynamics), model components used to simulate them, parameter values, etc. Such interactions will also depend on the objective function: for example, the standard Nash-Sutcliffe measure will favor the fitting of the larger peaks during storm events, while log transformation of responses will emphasize low flows during hydrograph recessions (see *Schaefli and Gupta* [2007] for a more general discussion of benchmarking against specific solution features). As coarse data averaging smears high-frequency features of the rainfall-runoff dynamics, it will affect parameter inference and structure identification.

7.3. A Short Theoretical Exploration of Time Scale Dependencies

[99] While the main focus of this paper is on experimental data analysis, its findings can be put on a more rigorous theoretical footing, which is briefly outlined next.

7.3.1. Trends Arising From Numerical Discretization Errors

[100] Numerical discretization errors arising when approximating ODEs such as (1) are quite well understood [e.g., *Butcher*, 2008]. In general, the numerical error e of an ODE approximation XX is a Taylor power series in Δt,

$$e_n^{(\mathrm{XX})} = S(t_n) - S_n^{(\mathrm{XX})} = \sum_{k=1}^{\infty} c_k \Delta t^k , \qquad (8)$$

where $S(t_n)$ is the exact solution of the ODE. In equation (8), the coefficients c depend on the nonlinearity of the solution, and a jth-order method has $c_{j-1} = 0$. For example, the global error of the explicit Euler approximation (2) is

$$e_n^{(\mathrm{EE})} = S(t_n) - S_n^{(\mathrm{EE})} \propto \Delta t \frac{\mathrm{d}^2 S}{\mathrm{d}t^2} + O\left(\Delta t^2 \frac{\mathrm{d}^3 S}{\mathrm{d}t^3}\right) + \ldots , \qquad (9)$$

where the derivative terms represent averages over the entire approximation period.

[101] The influence of numerical stability is also critical: if we define an error amplification factor $\xi = e_{n+1}^{(\mathrm{XX})}/e_n^{(\mathrm{XX})}$, an unstable scheme has $\xi > 1$, and the errors accumulate uncontrollably.

[102] The spurious numerical trends in parameter estimates reported in this paper are a consequence of such time discretization errors. Several observations can be made on the basis of error analysis (8) and (9).

[103] 1. The direct dependence of numerical errors on the time step size Δt necessarily implies a dependence of model predictions, and hence calibrated parameter estimates, on Δt [*Kavetski et al.*, 2003].

[104] 2. While for first-order approximations (such as the EE and IE schemes) the numerical errors are asymptotically linear with respect to Δt, the behavior for larger step sizes can be highly nonlinear because of higher-order terms in equation (8). In the absence of error-controlled substepping, it is difficult to guarantee that the approximation is operating in its asymptotic region, and this may explain the variety of linear and nonlinear trends reported in the hydrological modeling literature [e.g., *Littlewood and Croke*, 2008; *Wang et al.*, 2009].

[105] 3. While a fixed step, unconditionally stable approximation such as the IE scheme cannot be guaranteed to be free of numerical errors, the behavior of a fixed step, conditionally stable method such as the EE scheme is particularly fragile: even episodic numerical instabilities will significantly aggravate any time scale dependencies.

7.3.2. Trends Arising From Temporal and Spatial Averaging of the Data

[106] Although exact solutions and numerically accurate adaptive approximations (which reduce Δt until the error $e_n^{(\mathrm{XX})}$ is below a user-prescribed tolerance τ) will be free of spurious numerical time scale trends, they still remain subject to genuine data-averaging time scale effects (and to data sampling and measurement errors).

[107] In particular, the very act of averaging the rainfall forcing over a time step Δt introduces "smearing" errors into the forcing data, which necessarily translates into Δt-dependent errors in the model predictions and calibrated parameters. In calibration, similar smearing effects will arise with respect to the observed streamflow that is being fitted. The smearing affects both rainfall and runoff data and can be seen in Figure 5 [see also *Ostrowski et al.*, 2010]. Smearing effects will impact particularly strongly on the identification of fast processes. This can be seen for quick-flow parameters T_f and K_r in Figure 3: the step-size-dependent smearing of fast-scale forcing and response features creates time scale dependencies that are unrelated to numerical solution errors and hence cannot be removed even using exact analytical solutions. Conversely, slow (recession) processes will be less affected by such averaging. These observations emphasize the interplay between data resolution, identifiability, and the time scale at which parameters and associated processes operate.

[108] More generally, data-averaging and sparse-sampling time scale effects are inherent to any model forced with (and/or calibrated to) averaged data, whether the model is based on conceptual or physical principles [e.g., *Clark et al.*, 2008], or transfer functions [e.g., *Young and Garnier*, 2006], whether it is formulated in discrete or continuous time and regardless of the inference method. In particular, although previous publications have suggested that continuous time models have time-scale-invariant parameters [e.g., *Young and Garnier*, 2006; *Littlewood and Croke*, 2008], this holds only in the theoretical context of calibrating to continuous time forcing response data. As soon as averaging errors are present in the data, dynamics below this averaging scale are smeared and no longer identifiable. Unless an accurate subsampling scale interpolation is used for both the forcing and response time series (which, essentially, requires prior knowledge of the very dynamics the modeler is trying to explore), the inferred model parameters and structures can become time scale dependent.

[109] Analogous arguments apply with respect to using spatially averaged data. Perhaps more generally, since spatiotemporal averaging errors are tied to the very way a lumped model is constructed and applied, they could also be viewed as a type of model structural error (e.g., see the discussion by *Kuczera et al.* [2006]).

7.4. Implications for Calibration and Model Analysis

7.4.1. A Comment on Empirical Strategies for Relating Parameters to Time Step Size

[110] As part of an analysis of time scale dependencies of conceptual hydrological model parameters, *Littlewood and Croke* [2008] and *Wang et al.* [2009] derived empirical

relations between parameter values and time step size. Such relationships could be useful, for example, when regionalizing parameter values or when determining the limiting parameter values corresponding to a continuous time version of the original discrete time model.

[111] We note that these relationships are specific to the model structure for which they are obtained (and hence may be affected by model modifications, such as including additional components and process representations). They are also conditioned on the forcing response data and the likelihood function used in their estimation and will also include numerical approximation effects. Hence, while the overall time scaling may, indeed, be linear or otherwise correctly determined from empirical analysis, its specific quantification will generally be problem dependent. Therefore, the use of such relationships in extrapolating the model to different hydrological data resolution, especially in different catchments, must be approached with caution. They should not be viewed as a panacea from spurious numerical and statistical artifacts and are likely to be meaningful only when combined with properly formulated and inferred governing equations solved using robust numerical techniques.

7.4.2. Adopting the "Right" Numerics

[112] The time scale analyses in this paper lend particular urgency to the problem of robust numerical model design in conceptual hydrological modeling [e.g., *Kavetski et al.*, 2003; *Clark and Kavetski*, 2010]. The fixed step explicit methods, which remain frequently used in hydrological research and operations, significantly exacerbate any time scale dependencies of the governing equations. In many cases, given the propensity for instabilities and erratic behavior of fixed step explicit approximations, the time scale "dependencies" of models using these fragile time-stepping schemes are nothing but spurious numerical artifacts (Figures 2 and 3).

[113] Note that Figure 4 also shows cases where the truncation error of the explicit Euler scheme fortuitously improves the model performance for certain combinations of model complexity and data resolution. While superficially beneficial, we emphatically recommend against such "good results for the wrong reasons" (echoing analogous concerns of *Kirchner* [2006] in the context of confounding model component interaction in process representation). In particular, virtually regardless of the data resolution, a modeler using the fixed step EE approximation will be unable to confidently distinguish between a genuinely good performance of the model equations versus a fortuitous and fragile compensation of model errors by numerical artifacts. Similarly, poor and/or inconsistent performance, at virtually any time scale, could be attributed either to conceptualization errors and poorly defined constitutive functions or to numerical errors. Using uncontrollable numerical errors to compensate for structural errors of the model conceptualization is imprudent, resulting, as this study shows, in spurious time step trends, loss of predictive performance, and other problems (see *Kavetski and Clark* [2010, 2011] for a further showcasing of Pandora's box of numerical artifacts in conceptual hydrology).

[114] Instead, these problems should be addressed using adaptive substepping with error control or, at a minimum, using unconditionally stable methods. In particular, the lack of spurious time scale trends in the parameter distributions of models implemented using the IE scheme and its enviably stable performance in validation are further evidence of its robustness for conceptual hydrologic modeling [*Kavetski and Clark*, 2010].

7.4.3. Toward "Better" Objective Functions and Diagnostic Measures

[115] This study also suggests that the objective function not only defines the optimized parameters and their entire distributions but, importantly, also controls their time scale dependencies. In Bayesian analysis, the posterior distribution (objective function) is a consequence of the hypothesized error models. Given their demonstrable impact on inference and prediction, these hypotheses can, and should, be carefully checked a posteriori [e.g., *Box and Tiao*, 1992; see also *Laio and Tamea*, 2007; *Thyer et al.*, 2009; *Renard et al.*, 2010]. This study shows that objective (likelihood) functions that do not adequately describe the error structure are not only poorly suited for estimating posterior parameter distributions (e.g., as discussed by *Stedinger et al.* [2008]) but can also obscure and/or alter the time dependence characteristics of the model, hampering the interpretation of model parameters and structure and providing potentially erroneous predictions for streamflow and other quantities of interest. In conjunction with insufficient data resolution, deficiencies in the likelihood function also affect the types of processes that are captured in the inferred model (e.g., Figure 6).

[116] A key challenge in hydrological model identification is that adequate error models are difficult to construct, and until recently, there were no conceptually robust and computationally tractable frameworks for actually utilizing them. In cases where structural errors dominate data uncertainties, the error models underlying the likelihood (objective) function as well as any associated (Bayesian) priors must reflect not only data uncertainty but, more importantly, structural errors. This requires the likelihood function to provide a probabilistic description of (epistemic) uncertainties arising from the limited knowledge of the system and is clearly one of the harder challenges facing any probabilistic inference method [e.g., *Beven*, 2008; *Bulygina and Gupta*, 2009; *Renard et al.*, 2010]. A particular challenge is to adequately reflect the autocorrelated nature of these errors [e.g., see *Sorooshian and Dracup*, 1980; *Kavetski et al.*, 2002b; *Beven*, 2008; *Reichert and Mieleitner*, 2009] (also see the short exposé on structural errors by *Doherty and Welter* [2010]), which is expected to be particularly strong when working with high-resolution data. Yet we are also optimistic regarding progress in several distinct directions [e.g., *Kavetski et al.*, 2006b; *Kuczera et al.*, 2006; *Vrugt et al.*, 2008; *Reichert and Mieleitner*, 2009; *Bulygina and Gupta*, 2009; *Kirchner*, 2009; *Renard et al.*, 2010; *Schoups and Vrugt*, 2010].

[117] Useful insights into the physical fidelity of conceptual models can also be obtained from diagnostic approaches [e.g., *Gupta et al.*, 1998, 2008] and from sensitivity analysis [e.g., *van Werkhoven et al.*, 2008; *Spear and Hornberger*, 1980], including the popular GLUE methodology [*Beven and Binley*, 1992]. Although, strictly speaking, these methods do not support quantitative uncertainty estimation (e.g., see the robust critique of attempting to give subjectively defined goodness-of-fit measures quantitative probabilistic meaning by *Stedinger et al.* [2008, paragraphs 5 and 6]), they can provide useful qualitative guidance for

model selection [e.g., *Yilmaz et al.*, 2008; *Rupp et al.*, 2008; *Krueger et al.*, 2010]. What then becomes important is that the selected performance criteria provide strong discriminatory power (e.g., rainfall-runoff cross-correlation diagnostics can be much more incisive than the Nash-Sutcliffe index; see the broader critique by *Schaefli and Gupta* [2007]).

[118] This study also highlights that the resolution of the data governs what signatures are preserved and hence could be represented in the models. For example, analysis of the observed at daily time scale shows only the slow-scale lagged response, with the signature of the faster-scale first peak erased by data averaging. Hence, depending on the resolution of the observed data, a timing-oriented diagnostic can be sharp and incisive (hourly data, Figures 7a–7d) or blunt and noninformative (daily data, Figures 7e–7h).

7.5. Broader Implications for Hypothesis Testing in Hydrological Sciences

[119] Meaningful hypothesis testing hinges on the information content present in and extracted from the available data [e.g., *Jakeman and Hornberger*, 1993; *Gupta et al.*, 1998; *Schoups et al.*, 2008; *Gupta et al.*, 2008] and must also rigorously account for data uncertainties if structural hypothesis errors are to be "isolated" and analyzed, at least in a distributional sense [e.g., *Renard et al.*, 2010]. The influence of data resolution on hypothesis testing is therefore a topic of evident significance. Our findings illustrate not only the ability of higher-resolution data to reveal increasingly finer-scale hydrological processes but also the feasibility of their mathematical representation within comparatively simple conceptual models (necessarily subject to the key caveats in the next paragraph and the limitations in section 7.6). More generally, the findings regarding the consistency of parameter estimation and model identification across different time scales also highlight the importance of appropriate selection of the resolution of the supporting data in the context of pursuing and scrutinizing "unified" theories and models of hydrology at the catchment scale [e.g., *Sivapalan*, 2005; *McDonnell et al.*, 2007; *Troch et al.*, 2009].

[120] However, because of the apparent numerical fragility of the current implementation of many conceptual hydrological models (e.g., see the critique by *Clark and Kavetski* [2010], and references therein), harmful and confounding interactions between data resolution, objective functions, and numerical approximation errors can no longer continue to be largely ignored by the conceptual hydrological community. They pose a clear practical concern in the context of meaningful hypothesis development and testing and require proper attention [*Kavetski and Clark*, 2010, 2011].

[121] For example, the consistency of the calibration and validation performances is often used when evaluating supported model complexity, with inconsistent performance in predictive ("validation") applications generally viewed as a symptom of overfitting [e.g., *Jakeman and Hornberger*, 1993; *Kingston et al.*, 2008; *Schoups et al.*, 2008]. Yet these inconsistencies could well be numerical (fixed step explicit time stepping) and/or statistical (inadequate objective function) artifacts. For example, Figure 5 shows artifacts dependent on the time scale of the system, while *Kavetski and Clark* [2010] show artifacts dependent on the calibration period and catchment. Hence, reliance on simplistic model implementation and analysis techniques can easily result in spurious mathematical artifacts dictating conclusions of physically oriented hypothesis testing and complexity assessment.

[122] Since this study utilized least squares regression, similar behavior is expected for calibration methods based on modifications of similar objective functions. For example, subjectively rescaling Nash-Sutcliffe or other sums-of-squares criteria (e.g., as attempted in GLUE applications to heuristically mimic the effects of unaccounted data and structural errors using inflated parameter uncertainty [e.g., *Franks et al.*, 1998; *Blasone et al.*, 2008]) may not remove inconsistent behavior with respect to the data resolution (and length), though it could simply shroud it in additional parameter uncertainty and introduce new errors by omitting residual terms. Indeed, even a likelihood function correctly describing data uncertainty would lead to a poor inference, unless it also included a representation of numerical errors as if they were structural errors of the model hypothesis!

[123] Our findings reinforce the critical need to use more robust and accurate numerical and statistical techniques in hydrology and, logically, to use stringent diagnostic tests to identify and improve deficient components, be they model structures, numerical approximations, or likelihood functions. Despite the challenges of pursuing this rigorously, we are optimistic regarding progress in the distinct directions of robust model formulation [e.g., *Kavetski et al.*, 2003; *Clark et al.*, 2008; *Clark and Kavetski*, 2010; *Schoups et al.*, 2010], the development and application of increasingly realistic likelihood functions [e.g., *Kavetski et al.*, 2006b; *Stedinger et al.*, 2008; *Reichert and Mieleitner*, 2009; *Renard et al.*, 2010; *Schoups and Vrugt*, 2010], stringent process-oriented diagnostics [e.g., *Gupta et al.*, 2008; *Yilmaz et al.*, 2008; *Clark et al.*, 2011], and experimental insights [e.g., *Seibert and McDonnell*, 2002; *Weiler and McDonnell*, 2004; *Western et al.*, 2004; *Vaché and McDonnell*, 2006]. When exploited as part of systematic multimodel analyses and hypothesis testing [e.g., *Clark et al.*, 2008], we believe that pursuing these advances offers promising prospects for developing more scientifically defensible, and operationally reliable, hydrological models.

7.6. Current Limitations and a View to the Future

[124] Several methodological compromises were made in the empirical case study, chiefly because of currently severe data analysis and computational limitations when modeling at short (down to subhourly) time scales, but also to maintain a clearer focus on methods commonly used by hydrological scientists and practitioners.

7.6.1. Simplistic Specification of the Likelihood Function

[125] The case studies reported here do not distinguish between data (rainfall and streamflow) and structural uncertainties [e.g., *Kavetski et al.*, 2002b; *Vrugt et al.*, 2008; *Reichert and Mieleitner*, 2009]. At best, these are imperfectly lumped into residual errors, which in SLS and WLS represent a crude mixture of all sources of uncertainty. Hence, we did not attempt to a priori constrain the parameters of the residual error model, e.g., using streamflow rating curve analysis [*Thyer et al.*, 2009]. Poor treatment of

data and structural errors reduces the robustness of the inference, e.g., resulting in (probabilistic) streamflow predictions and parameter estimates that are unduly dependent on the specific realization of data errors in the calibration period (e.g., as illustrated by *Thyer et al.* [2009]; see also *Renard et al.* [2010]). In the context of parameter estimation, failing to reliably account for input errors and residual autocorrelations (usually) results in underestimated posterior parameter uncertainties [e.g., *Beven and Young*, 2003; *Thyer et al.*, 2009; *Schoups and Vrugt*, 2010]. Moreover, the distributional adequacy of the streamflow predictions will generally be poor [e.g., *Thyer et al.*, 2009; *Renard et al.*, 2010]. However, note that predictive uncertainty will not necessarily be "uniformly" underestimated. For example, a homoscedastic residual error model (SLS) will tend to underestimate the uncertainty during high flows and overestimate it during low flows (Figure 5). A poorly specified heteroscedastic error model (e.g., WLS with equation (6b)) will also fail to produce reliable prediction limits (e.g., as shown by *Thyer et al.* [2009]).

[126] The limitations of traditional regression are well known and represent a consequence of a poor specification of the likelihood function in equations (5) and (6), rather than any purported deficiency of the Bayesian methodology itself (a point stressed by many authors [e.g., *Kavetski et al.*, 2002b; *Reichert and Mieleitner*, 2009; *Stedinger et al.*, 2008]). Indeed, the additional spurious time scale dependencies suggested in the empirical studies of this paper could themselves be viewed as a manifestation of the misspecified likelihoods. It is difficult to see how could such deficiencies be overcome merely by subjective, unverified alterations of the objective function and/or heuristic sampling procedures, let alone those that violate basic axioms of probability theory (as seemingly advocated, for example, by *Beven et al.* [2008] and the GLUE references therein).

[127] Instead, even with an inexact model of the system dynamics, the Bayesian method can produce statistically reliable inference and prediction, but only provided statistically reliable characterizations of the observational and structural errors are specified [e.g., *Renard et al.*, 2010]. By no means an easy challenge, estimating data error models requires a careful analysis of the observational system, yet progress in this direction has been apparent for nearly a decade (e.g., see the work of *Willems* [2001] and *Villarini et al.* [2008] on rainfall sampling errors). More interestingly, the results of *Renard et al.* [2010] indicate that to the extent that the data error models are reliable and precise, Bayesian inference can proceed even with vague priors on the structural errors. In other words, given a set of data error models, a Bayesian scheme can achieve "closure" of total errors by inferring the structural uncertainty [*Renard et al.*, 2009; see also *Kuczera et al.*, 2010b]. A key restriction is that the distributional reliability of this closure depends on the distributional reliability of the data error models and on the flexibility of the (possibly very vague) structural error model. With respect to the latter challenge, exploration of structural errors using flexible model frameworks and experimental insights is of interest [e.g., *Clark et al.*, 2008; *Fenicia et al.*, 2008; *Clark et al.*, 2011]. We also note useful advances in Bayesian (probabilistic) characterization of the (epistemic) uncertainties in the system structure (e.g., using stochastic parameters [*Kuczera et al.*, 2006; *Reichert and Mieleitner*, 2009] or, more generally, using nonparametric techniques [*Bulygina and Gupta*, 2009]). Other promising methods include quantile regression [e.g., *Koenker*, 2005], more flexible residual error models [e.g., *Schoups and Vrugt*, 2010], and inductive methods [e.g., *Young and Ratto*, 2009]. We anticipate that direct, robust, and conceptually appealing ways to confront the apparent impasse currently surrounding hydrological calibration require not just "changing the question" [*Sivapalan*, 2009] but exploiting a fusion of these technical advances with experimental basin insights [e.g., *Seibert and McDonnell*, 2002; *Tromp-van Meerveld and McDonnell*, 2006b; *van den Bos et al.*, 2006a] and a stringent posterior scrutiny that combines statistical [e.g., *Laio and Tamea*, 2007; *Thyer et al.*, 2009] and process-oriented [e.g., *Gupta et al.*, 2008; *Yilmaz et al.*, 2008; *Clark et al.*, 2011] diagnostics.

[128] A final current limitation is computational: while input- and/or structural-error-sensitive Bayesian technologies can increasingly (and still imperfectly) be applied in specific case studies [e.g., *Kavetski et al.*, 2006b; *Vrugt et al.*, 2008; *Thyer et al.*, 2009; *Reichert and Mieleitner*, 2009], they remain expensive for large-scale experiments such as those undertaken in this study. Perhaps optimistically, we view this as a lesser challenge than the error model specification tasks: more efficient optimization and sampling algorithms, including multimethod approaches [*Vrugt and Robinson*, 2007] and limited memory MCMC strategies [*Kuczera et al.*, 2010a], are being rapidly developed and refined and are increasingly benefitting from parallel computing facilities.

7.6.2. Relatively Basic Numerical Time-Stepping Techniques

[129] We note the absence of sufficient experiments with high-order numerical approximations and near-exact solutions in our empirical evaluations. This was partially due to limiting the scope of this paper to time-stepping schemes commonly used in conceptual hydrological modeling (see *Clark and Kavetski* [2010] for a recent review) but was also motivated by the nontrivial practical issue of juggling numerical accuracy and solution smoothness (section 3.3), especially in the context of maintaining methodological consistency across the multitude of runs and experiments in this study. Hence, for pragmatic reasons, we relied upon the unconditional stability of the implicit Euler approximation and exploited its smoothness to speed up calibration without imposing a truncation error tolerance (see *Kavetski et al.* [2003] and *Kavetski and Clark* [2010] for further discussion). Since computationally demanding practical numerical work does require carefully considered accuracy-cost tradeoffs [e.g., *Gill et al.*, 1981], a separate study should focus on state-of-the-art variable order–variable step algorithms [e.g., *Butcher*, 2008] vis-à-vis fixed step implicit approximations, in the specific context of conceptual hydrological modeling with uncertain averaged data.

7.6.3. Focus on Lumped Conceptual Models and a Single Catchment

[130] Finally, largely for reasons of space and to maintain a (relatively) compact presentation, we limited the scope of the empirical evaluations in this study to four lumped conceptual models and a single experimental catchment. Of particular interest is the extension of the analysis to multiple catchments [*Merz et al.*, 2009] and spatially

distributed hydrological models [*Ivanov et al.*, 2004; *Immerzeel and Droogers*, 2008], whose promise of better representation of catchment and forcing heterogeneities has often been hampered by overparameterizations and, consequently, poor identifiability if given only exceedingly aggregated data [e.g., *Beven*, 1993] and potentially weak identification techniques. Such analysis may highlight even more prominently the nontrivial interplay between physical realism and statistical identifiability, forcing and structural errors, inference data requirements and practical predictive ability under a variety of temporal and spatial scales, and, finally, practical tradeoffs between numerical accuracy and computational complexity.

8. Conclusions

[131] This study investigated the interplay between data resolution versus parameter inference and model structure identification in conceptual hydrological modeling, with a dual focus on (1) practical computation using simple numerical approximations of conceptualized catchment-scale dynamics, calibrated using common likelihood functions, and (2) the ability of calibrated models to represent important qualitative and quantitative aspects of hydrological behavior that may not be captured in simplistic statistical calibration approaches. Recent progress notwithstanding, these issues remain quite poorly understood, both empirically and mathematically, especially in the context of probabilistic model estimation, uncertainty analysis, and predictive application.

[132] The empirical analysis was carried out using four conceptual rainfall-runoff models of varying complexity, applied to the experimental Weierbach catchment (Luxembourg) over time scales ranging from 30 min to 3 days. Its schist geology makes it an interesting, and challenging, case study, as higher data resolution reveals finer-scale dynamics, such as double-peaked hydrographs representing delayed catchment response, and forces the models to reproduce these features. The availability of experimental insights in this catchment leads to a hydrological process-oriented perspective that complements the statistical identification analyses reported in this paper.

[133] The empirical analysis highlighted the subtle influence of seemingly unrelated aspects of model implementation and calibration on model identification, interpretation, and predictive use. The following have been shown:

[134] 1. Conditionally stable and otherwise erratic time-stepping schemes introduce spurious time scale trends into the inferred hydrological model parameters and model structures. Although still seemingly underestimated by sections of the hydrological community, this finding should not be surprising given the veritable numerical daemons unearthed in recent work [e.g., *Clark and Kavetski*, 2010; *Kavetski and Clark*, 2010] and also given a flock of earlier harbingers [e.g., *Michel et al.*, 2003; *Kavetski et al.*, 2003].

[135] 2. Simplistic likelihood functions can also introduce spurious time scale artifacts into the inference. This ties well with previous work on parameter consistency in the context of statistical identification of conceptual hydrological models [e.g., *Kavetski et al.*, 2002b; *Clark and Vrugt*, 2006; *Stedinger et al.*, 2008; *Thyer et al.*, 2009; *Schoups and Vrugt*, 2010].

[136] From a more process-oriented perspective, the following conclusions can be drawn.

[137] 1. When robust numerical implementations and heteroscedastic (albeit still exceedingly crude) residual error models are used, the inferred parameter distributions stabilized considerably. Parameters describing slow dynamics remained largely invariant over the time scales of consideration, whereas parameters describing fast dynamics converged toward increasingly precise and stable estimates as the data resolution was refined and approached the characteristic time scale of these processes. The improved model and parameter behavior is expected to help process-oriented interpretation [e.g., *Seibert and McDonnell*, 2002; *Birkel et al.*, 2010] and regionalization [e.g., *Kling and Gupta*, 2009; *Bai et al.*, 2009] of model parameters and structures.

[138] 2. While using temporally and spatially averaged forcing response data necessarily introduces (nonnumerical) process time scale dependencies, provided robust numerical and statistical techniques are employed, the unveiling of finer-scale features of the forcing response data by higher-resolution sampling supported the identification of a larger number of parameters and components, in particular, quick-flow dynamics, within increasingly complex model structures. Since this can be interpreted as a more stable and realistic behavior of inferred model parameters supporting more physical meaningful modeling, this is a strong argument in favor of collecting high-resolution hydrological data for the purpose of hypothesis testing and model development.

[139] We stress that the numerical maladies illustrated in this work are naturally avoided by using robust solutions of the hypothesized model equations. In hydrology, this is generally already the case with most physically based models [e.g., *Ivanov et al.*, 2004]. As eloquently affirmed by a colleague, "facing the numerical conundrum squarely in the face has been a hallmark of subsurface flow and transport modeling since the very advent of the computer era" (S. P. Neuman, personal communication, 2010). Although numerical robustness in itself guarantees neither scientific nor operational adequacy, it is an essential precondition for meaningfully pursuing these objectives. We hence reiterate an earlier call (outlined by *Kavetski and Clark* [2011]) to the conceptual hydrological community (and any other field of environmental modeling or science that is tempted to disregard robust mathematics, perhaps on the pretext of "the models are simplistic anyway") that a laissez-faire attitude toward numericostatistical errors is scientifically indefensible, and that this has been recognized, or is being recognized, in most other branches of hydrology and science (as can be seen by consulting the vast field of both popular and specialized books and journals on mathematical modeling, of which our reference list is but a small sample).

[140] From a broader perspective of the hypothesis-based approach to hydrological science [e.g., *Kuczera and Franks*, 2002], this study provides insights into the information content of data and, through robust numerical and statistical techniques, furthers the utilization of high-resolution data from experimental catchments to advance our understanding of catchment-scale dynamics. When selected on the basis of experimental insights, implemented using robust numerical techniques, and tested using a spectrum of statistical and process-oriented diagnostics, the more complex

model structures supported by the higher-resolution data provided better overall performance both in terms of aggregate measures of model performance (here goodness of fit of streamflow predictions) and in terms of reproducing important qualitative signatures in the data. Importantly, consistent results were obtained both across multiple time scales and across calibration-validation periods. While empirical evidence is still limited, theoretical arguments support the generality of these encouraging findings.

[141] **Acknowledgments.** We are grateful to our colleagues, including Hoshin Gupta, George Kuczera, Shlomo Neuman, Nuria Martinez-Carreras, Laurent Pfister, Hubert Savenije, Mark Thyer, and Sebastian Wrede, for valuable discussions on the topic of this manuscript. We thank Manfred Ostrowski for bringing his research report to our attention. We also thank the Editors and the reviewers for their valuable comments, criticisms, and suggestions.

References

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

Bai, Y., T. Wagener, and P. Reed (2009), A top-down framework for watershed model evaluation and selection under uncertainty, *Environ. Modell. Software*, *24*, 901–916.

Baker, D. L. (1995), Applying higher order DIRK time steps to the “modified Picard” method, *Ground Water*, *33*(2), 259–263.

Bárdossy, A., and S. K. Singh (2008), Robust estimation of hydrological model parameters, *Hydrol. Earth Syst. Sci.*, *12*, 1273–1283.

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

Beven, K. (1993), Prophecy, reality and uncertainty in distributed hydrological modeling, *Adv. Water Resour.*, *16*, 41–51.

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

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

Beven, K. J., and A. M. Binley (1992), The future of distributed hydrological models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*, 279–298.

Beven, K., and P. Young (2003), Comment on “Bayesian recursive parameter estimation for hydrologic models” by M. Thiemann, M. Trosset, H. Gupta, and S. Sorooshian, *Water Resour. Res.*, *39*(5), 1116, doi:10.1029/2001WR001183.

Beven, K. J., P. J. Smith, and J. E. Freer (2008), So just why would a modeller choose to be incoherent?, *J. Hydrol.*, *354*, 15–32.

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

Blasone, R. S., J. A. Vrugt, H. Madsen, D. Rosbjerg, B. A. Robinson, and G. A. Zyvoloski (2008), Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling, *Adv. Water Resour.*, *31*, 630–648.

Blöschl, G., and M. Sivapalan (1995), Scale issues in hydrological modeling—A review, *Hydrol. Processes*, *9*, 251–290.

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

Brath, A., A. Montanari, and E. Toth (2004), Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model, *J. Hydrol.*, *291*, 232–253.

Brillinger, D. R. (1989), Consistent detection of a monotonic trend superposed on a stationary time series, *Biometrika*, *76*, 23–30.

Bulygina, N., and H. Gupta (2009), Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation, *Water Resour. Res.*, *45*, W00B13, doi:10.1029/2007WR006749.

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., Littleton, Colo.

Butcher, J. (2008), *Numerical Methods for Ordinary Differential Equations*, 2nd ed., John Wiley, Chichester, U. K.

Cho, J., S. Mostaghimi, M. S. Kang, and J. A. Chun (2009), Sensitivity to grid and time resolution of hydrology components of Dansat, *Trans. ASABE*, *52*(4), 1121–1128.

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

Clark, M. P., and J. A. Vrugt (2006), Unraveling uncertainties in hydrologic model calibration: Addressing the problem of compensatory parameters, *Geophys. Res. Lett.*, *33*, L06406, doi:10.1029/2005GL025604.

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. E. Rupp, R. A. Woods, H. J. Tromp-van Meerveld, N. E. Peters, and J. E. Freer (2009), Consistency between hydrological models and field observations: Linking processes at the hillslope scale to hydrological responses at the watershed scale, *Hydrol. Processes*, *23*, 311–319.

Clark, M. P., H. K. McMillan, D. B. G. Collins, D. Kavetski, and R. A. Woods (2011), Hydrological field data from a modeller’s perspective. Part 2: Process-based evaluation of model hypotheses, *Hydrol. Processes*, *25*(4), 523–543, doi:10.1002/hyp.7902.

Cressie, N., C. A. Calder, J. S. Clark, J. M. ver Hoef, and C. K. Wikle (2009), Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling, *Ecol. Appl.*, *19*(3), 553–570.

Dawdy, D. R. (1983), A review of rainfall-runoff modeling, in *Experience in the Development and Application of Mathematical Models in Hydrology and Water Resources in Latin America (Proceedings of the Tegucigalpa Hydromath Symposium)*, *IAHS Publ.*, *152*, 97–113.

Dingman, S. L. (1994), *Physical Hydrology*, Prentice-Hall, Englewood Cliffs, N. J.

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

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

Farmer, D., M. Sivapalan, and C. Jothityangkoon (2003), Climate, soil, and vegetation controls upon the variability of water balance in temperate and semiarid landscapes: Downward approach to water balance analysis, *Water Resour. Res.*, *39*(2), 1035, doi:10.1029/2001WR000328.

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

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

Finnerty, B. D., M. B. Smith, D. J. Seo, V. Koren, and G. E. Moglen (1997), Space-time scale sensitivity of the Sacramento model to radar-gage precipitation inputs, *J. Hydrol.*, *203*, 21–38.

Franks, S., P. Gineste, K. J. Beven, and P. Merot (1998), On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas into the calibration process, *Water Resour. Res.*, *34*, 787–797, doi:10.1029/97WR03041.

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

Gill, P. E., W. Murray, and M. H. Wright (1981), *Practical Optimization*, Academic, London.

Götzinger, J., and A. Bárdossy (2008), Generic error model for calibration and uncertainty estimation of hydrological models, *Water Resour. Res.*, *44*, W00B07, doi:10.1029/2007WR006691.

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*, 751–763, doi:10.1029/97WR03495.

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

Hamed, K. H., and A. R. Rao (1998), A modified Mann-Kendall trend test for autocorrelated data, *J. Hydrol.*, *204*, 182–196.

Hamon, W. R., and G. H. Belt (1973), Energy balance-resistance model for computing evapotranspiration, *Eos Trans. AGU*, *54*(4), 270.

Hopp, L., and J. J. McDonnell (2009), Connectivity at the hillslope scale: Identifying interactions between storm size, bedrock permeability, slope angle and soil depth, *J. Hydrol.*, *376*, 378–391.

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

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

Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model?, *Water Resour. Res.*, *29*, 2637–2649, doi:10.1029/93WR00877.

Kahaner, D., C. Moler, and S. Nash (1989), *Numerical Methods and Software*, Prentice-Hall, Englewood Cliffs, N. J.

Kandel, D. D., A. W. Western, and R. B. Grayson (2005), Scaling from process timescales to daily steps: A distribution function approach, *Water Resour. Res.*, *41*, W02003, doi:10.1029/2004WR003380.

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

Kavetski, D., and M. P. Clark (2011), Numerical troubles in conceptual hydrology: Approximations, absurdities and impact on hypothesis-testing, *Hydrol. Processes*, *25*(4), 661–670, doi:10.1002/hyp.7899.

Kavetski, D., P. Binning, and S. W. Sloan (2002a), Noniterative time stepping schemes with adaptive truncation error control for the solution of Richards equation, *Water Resour. Res.*, *38*(10), 1211, doi:10.1029/2001WR000720.

Kavetski, D., S. Franks, and G. Kuczera (2002b), Confronting input uncertainty in environmental modelling in calibration of watershed models, in *Calibration of Watershed Models, Water Sci. Appl. Ser.*, vol. 6, edited by Q. Y. Duan et al., pp. 49–68, AGU, Washington, D. C.

Kavetski, D., G. Kuczera, and S. W. Franks (2003), Semidistributed hydrological modeling: A "saturation path" perspective on TOPMODEL and VIC, *Water Resour. Res.*, *39*(9), 1246, doi:10.1029/2003WR002122.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis, *J. Hydrol.*, *320*, 187–201.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.

Kingston, G. B., H. R. Maier, and M. F. Lambert (2008), Bayesian model selection applied to artificial neural networks used for water resources modeling, *Water Resour. Res.*, *44*, W04419, doi:10.1029/2007WR006155.

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.

Kirchner, J. W. (2009), Catchments as simple dynamic systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, *Water Resour. Res.*, *45*, W02429, doi:10.1029/2008WR006912.

Kirchner, J. W., X. H. Feng, C. Neal, and A. J. Robson (2004), The fine structure of water-quality dynamics: The (high-frequency) wave of the future, *Hydrol. Processes*, *18*, 1353–1359.

Kling, H., and H. Gupta (2009), On the development of regionalization relationships for lumped watershed models: The impact of ignoring sub-basin scale variability, *J. Hydrol.*, *373*, 337–351.

Koenker, R. (2005), *Quantile Regression*, Cambridge Univ. Press, New York.

Krueger, T., J. Freer, J. N. Quinton, C. J. A. Macleod, G. S. Bilotta, R. E. Brazier, P. Butler, and P. M. Haygarth (2010), Ensemble evaluation of hydrological model hypotheses, *Water Resour. Res.*, *46*, W07516, doi:10.1029/2009WR007845.

Kuczera, G., and S. Franks (2002), Testing hydrologic models: Fortification or falsification?, in *Mathematical Modelling of Large Watershed Hydrology*, edited by V. P. Singh and D. K. Frevert, pp. 141–186, Water Resour. Publ., Littleton, Colo.

Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, *J. Hydrol.*, *211*, 69–85.

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

Kuczera, G., D. Kavetski, B. Renard, and M. Thyer (2010a), A limited-memory acceleration strategy for MCMC sampling in hierarchical Bayesian calibration of hydrological models, *Water Resour. Res.*, *46*, W07602, doi:10.1029/2009WR008985.

Kuczera, G., B. Renard, M. Thyer, and D. Kavetski (2010b), There are no hydrological monsters, just models and observations with large uncertainties!, *Hydrol. Sci. J.*, *55*(6), 980–991.

Laio, F., and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, *11*, 1267–1277.

Linsley, R. K., and M. A. Kohler (1958), *Hydrology for Engineers*, McGraw-Hill, London.

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

Littlewood, I. G., and B. F. W. Croke (2008), Data time-step dependency of conceptual rainfall-streamflow model parameters: An empirical study with implications for regionalization, *Hydrol. Sci. J.*, *53*(4), 685–695.

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

Maniak, U. (1997), *Hydrologie und Wasserwirtschaft: Eine Einführung für Ingenieure*, 4th ed., Springer, Berlin.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *J. Hydrol.*, *330*, 368–381.

Martinez-Carreras, N., T. Udelhoven, A. Krein, F. Gallart, J. F. Iffly, J. Ziebel, L. Hoffmann, L. Pfister, and D. E. Walling (2010), The use of sediment colour measured by diffuse reflectance spectrometry to determine sediment sources: Application to the Attert River catchment (Luxembourg), *J. Hydrol.*, *382*, 49–63.

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

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

Michel, C., C. Perrin, and V. Andreassian (2003), The exponential store: A correct formulation for rainfall-runoff modelling, *Hydrol. Sci. J.*, *48*(1), 109–124.

Miller, C. T., G. A. Williams, C. T. Kelley, and M. D. Tocci (1998), Robust solution of Richards' equation for nonuniform porous media, *Water Resour. Res.*, *34*, 2599–2610, doi:10.1029/98WR01673.

Mitchell, T. (1997), *Machine Learning*, McGraw-Hill, New York.

Ostrowski, M., M. Bach, S. V. DeSimone, and V. Gamerith (2010), Analysis of the time-step dependency of parameters in conceptual hydrologic models, report, urn:nbn:de:tuda-tuprints-20996, Tech. Univ. Darmstadt, Darmstadt, Germany.

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

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

Pfister, L., et al. (2006), Study of the water cycle components in the Attert River Basin (CYCLEAU), report, Fonds Natl. de la Rech., Luxembourg.

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

Renard, B., D. Kavetski, and G. Kuczera (2009), Comment on "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction" by Newsha K. Ajami et al., *Water Resour. Res.*, *45*, W03603, doi:10.1029/2007WR006538.

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

Rupp, D. E., J. Schmidt, R. A. Woods, and V. J. Bidwell (2008), Analytical assessment and parameter estimation of a low-dimensional groundwater model, *J. Hydrol.*, *377*, 143–154.

Schaake, J. C., V. I. Koren, Q. Y. Duan, K. Mitchell, and F. Chen (1996), Simple water balance model for estimating runoff at different spatial and temporal scales, *J. Geophys. Res.*, *101*(D3), 7461–7475, doi:10.1029/95JD02892.

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

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

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

Schoups, G., J. A. Vrugt, F. Fenicia, and N. C. van de Giesen (2010), Corruption of accuracy and efficiency of Markov chain Monte Carlo simulation by inaccurate numerical implementation of conceptual hydrologic models, *Water Resour. Res.*, *46*, W10530, doi:10.1029/2009WR008648.

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.

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

Sivapalan, M. (2009), The secret to "doing better hydrological science": Change the question!, *Hydrol. Processes*, *23*, 1391–1396.

Sivapalan, M., G. Blöschl, L. Zhang, and R. Vertessy (2003a), Downward approach to hydrological prediction, *Hydrol. Processes*, *17*, 2101–2111.

Sivapalan, M., et al. (2003b), IAHS decade on predictions in ungauged basins (PUB), *Hydrol. Sci. J.*, *48*(6), 857–880.

Sorooshian, S. (1981), Parameter estimation of rainflow-runoff models with heteroscedastic streamflow errors—The noninformative data case, *J. Hydrol.*, *52*, 127–138, doi:10.1016/0022-1694(81)90099-8.

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

Spear, R. C., and G. M. Hornberger (1980), Eutrophication in Peel Inlet—II. Identification of critical uncertainties via generalized sensitivity analysis, *Water Res.*, *14*, 43–49.

Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resour. Res.*, *44*, W00B06, doi:10.1029/2008WR006822.

Tang, Y., P. Reed, T. Wagener, and K. van Werkhoven (2007), Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation, *Hydrol. Earth Syst. Sci.*, *11*, 793–817.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modelling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.

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

Tromp-van Meerveld, H. J., and J. J. McDonnell (2006a), Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope, *Water Resour. Res.*, *42*, W02410, doi:10.1029/2004WR003778.

Tromp-van Meerveld, H. J., and J. J. McDonnell (2006b), Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis, *Water Resour. Res.*, *42*, W02411, doi:10.1029/2004WR003800.

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

van den Bos, R., L. Hoffmann, J. Juilleret, P. Matgen, and L. Pfister (2006a), Conceptual modelling of individual HRU's as a trade-off between bottom-up and top-down modelling: A case study, paper presented at 3rd Biennial Meeting of International Environmental Modelling and Software Society, Burlington, Vt., 9–13 Jul.

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

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2008), Characterization of watershed model behavior across a hydroclimatic gradient, *Water Resour. Res.*, *44*, W01429, doi:10.1029/2007WR006271.

Villarini, G., P. Mandapaka, W. F. Krajewski, and R. Moore (2008), Rainfall and sampling uncertainties: A rain gauge perspective, *J. Geophys. Res.*, *113*, D11102, doi:10.1029/2007JD009214.

Vrugt, J. A., and B. A. Robinson (2007), Improved evolutionary optimization from genetically adaptive multimethod search, *Proc. Natl. Acad. Sci. U. S. A.*, *104*(3), 708–711.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.

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

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

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, *Hydrol. Processes*, *17*, 455–476.

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

Wang, Y., B. He, and K. Takase (2009), Effects of temporal resolution on hydrological model parameters and its impact on prediction of river discharge, *Hydrol. Sci. J.*, *54*(5), 886–898.

Weiler, M., and J. J. McDonnell (2004), Virtual experiments: A new approach for improving process conceptualization in hillslope hydrology, *J. Hydrol.*, *285*, 3–18.

Western, A. W., S. L. Zhou, R. B. Grayson, T. A. McMahon, G. Blöschl, and D. J. Wilson (2004), Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes, *J. Hydrol.*, *286*, 113–134.

Western, A. W., S. L. Zhou, R. B. Grayson, T. A. McMahon, G. Blöschl, and D. J. Wilson (2005), Reply to comment by Tromp van Meerveld and McDonnell on Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes, *J. Hydrol.*, *303*, 313–315.

Willems, P. (2001), Stochastic description of the rainfall input errors in lumped hydrological models, *Stochastic Environ. Res. Risk Assess.*, *15*(2), 132–152.

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.

Young, P. (1998), Data-based mechanistic modelling of environmental, ecological, economic and engineering systems, *Environ. Modell. Software*, *13*, 105–122.

Young, P. (2003), Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale, *Hydrol. Processes*, *17*, 2195–2217.

Young, P. C., and H. Garnier (2006), Identification and estimation of continuous-time, data-based mechanistic (DBM) models for environmental systems, *Environ. Modell. Software*, *21*, 1055–1072.

Young, P. C., and M. Ratto (2009), A unified approach to environmental systems modeling, *Stochastic Environ. Res. Risk Assess.*, *23*(7), 1037–1057.

M. P. Clark, National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307-3000, USA.

F. Fenicia, Department of Environment and Agro-Biotechnologies, Centre de Recherche Public–Gabriel Lippmann, L-4422 Belvaux, Luxembourg.

D. Kavetski, Environmental Engineering, University of Newcastle, Callaghan, NSW 2308, Australia. (dmitri.kavetski@newcastle.edu.au)