TU Delft

Delft University of Technology

From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions

Fenicia, FF; Kavetski, D; Savenije, HHG; Pfister, L

DOI
10.1002/2015WR017398

Publication date
2016

Published in
Water Resources Research

Citation (APA)
Fenicia, FF., Kavetski, D., Savenije, HHG., & Pfister, L. (2016). From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. *Water Resources Research*, *52*(2), 954-989. https://doi.org/10.1002/2015WR017398

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2015WR017398

Key Points:
- Distributed hydrological model development using a systematic hypothesis testing approach
- Use of distinct models structures in different HRUs
- Modeling insights into geology and topography as hydrological controls

Correspondence to:
F. Fenicia,
fabrizio.fenicia@eawag.ch

Citation:
Fenicia, F., D. Kavetski, H. H. G. Savenije, and L. Pfister (2016), From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions, *Water Resour. Res.*, *52*, 954–989, doi:10.1002/2015WR017398.

Received 14 APR 2015
Accepted 16 JAN 2016
Accepted article online 20 JAN 2016
Published online 12 FEB 2016
Corrected 14 MAR 2016

This article was corrected on 14 MAR 2016. See the end of the full text for details.

© 2016. American Geophysical Union. All Rights Reserved.

From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions

Fabrizio Fenicia[1], Dmitri Kavetski[2], Hubert H. G. Savenije[3], and Laurent Pfister[4]

[1]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland, [2]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia, [3]Water Resources Section, Delft University of Technology, Delft, Netherlands, [4]LIST, Luxembourg Institute of Science and Technology, Belvaux, Luxembourg

Abstract This paper explores the development and application of distributed hydrological models, focusing on the key decisions of how to discretize the landscape, which model structures to use in each landscape element, and how to link model parameters across multiple landscape elements. The case study considers the Attert catchment in Luxembourg—a 300 km^2 mesoscale catchment with 10 nested subcatchments that exhibit clearly different streamflow dynamics. The research questions are investigated using conceptual models applied at hydrologic response unit (HRU) scales (1–4 HRUs) on 6 hourly time steps. Multiple model structures are hypothesized and implemented using the SUPERFLEX framework. Following calibration, space/time model transferability is tested using a split-sample approach, with evaluation criteria including streamflow prediction error metrics and hydrological signatures. Our results suggest that: (1) models using geology-based HRUs are more robust and capture the spatial variability of streamflow time series and signatures better than models using topography-based HRUs; this finding supports the hypothesis that, in the Attert, geology exerts a stronger control than topography on streamflow generation, (2) streamflow dynamics of different HRUs can be represented using distinct and remarkably simple model structures, which can be interpreted in terms of the perceived dominant hydrologic processes in each geology type, and (3) the same maximum root zone storage can be used across the three dominant geological units with no loss in model transferability; this finding suggests that the partitioning of water between streamflow and evaporation in the study area is largely independent of geology and can be used to improve model parsimony. The modeling methodology introduced in this study is general and can be used to advance our broader understanding and prediction of hydrological behavior, including the landscape characteristics that control hydrologic response, the dominant processes associated with different landscape types, and the spatial relations of catchment processes.

1. Introduction

1.1. Distributed Catchment Models

Distributed catchment models are attractive to hydrologists because they hold the promise of representing the spatial heterogeneity of hydrological processes. Spatially distributed predictions are required for many applications, such as streamflow forecasting at multiple points within the catchment [e.g., *Thielen et al.*, 2009], and the identification of critical source areas for contaminant or sediment transport [e.g., *Djodjic and Villa*, 2015]. Explicit representation of distributed catchment characteristics is also necessary for impact studies, such as those concerning land use change and the construction of reservoirs [e.g., *Li et al.*, 2009].

One of the earliest and most influential blueprints for developing distributed models dates to *Freeze and Harlan* [1969]. In this blueprint, models are developed based on the current understanding of small-scale environmental physics, obtained through in situ and laboratory experiments under relatively well-defined conditions [*Köhne et al.*, 2009]. This approach has inspired models such as MIKE-SHE [*Refsgaard et al.*, 1995], InHM [*VanderKwaak and Loague*, 2001], and tRIBS [*Ivanov et al.*, 2004]. The development of these models typically follows a "bottom-up" philosophy, where large-scale system behavior is characterized through an explicit and detailed representation of the system's constituent processes and their interactions [*Sivapalan et al.*, 2003]. Models developed following this blueprint are commonly referred to as "physical," in the sense that, at least in principle, their parameters and states correspond to measurable quantities [*Abbott*, 1996].

Models based on small-scale physics, however, are not readily applied at the scale of catchments and beyond. Their data requirements—especially for representing the subsurface—are difficult to meet even using modern measurement technologies, and their computational costs are often infeasible for many practical purposes [*Reggiani and Schellekens*, 2006; *Kampf and Burges*, 2007].

In addition to application issues, models based on small-scale physics may not necessarily represent the best conceptual tools for the research questions that arise at the catchment scale [*Dooge*, 1988]. These models "emphasize the explicit mapping of the heterogeneities of landscape properties and the resulting process complexities" [*McDonnell et al*., 2007], which may be unnecessary—or even counterproductive—if the modelers' interest is in the emergent dynamics of the entire catchment system [*Beven*, 2002; *Sivapalan*, 2003]. In this respect, the Representative Elementary Watershed (REW) approach [*Reggiani and Rientjes*, 2005] provides a theoretical framework to upscale the balance equations from point to catchment scale. Physically rigorous derivation of appropriate averaging volumes and closure relationships at the landscape element scale remain elusive challenges in the application of this approach [e.g., *Beven*, 2006].

An alternative blueprint for developing distributed models is to estimate the equations describing the hydrological processes directly at the spatiotemporal scales of interest—or at least at scales closer to the scales of interest. These models typically follow a "top-down" approach to model development, in the sense of seeking a parsimonious description of the emergent properties of the entire system [*Sivapalan et al*., 2003]. Examples of such models include TOPMODEL [*Beven and Kirkby*, 1979], VIC [*Wood et al*., 1992], TAC-D [*Uhlenbrook et al*., 2004], and MHM [*Samaniego et al*., 2010]. These models are commonly referred to as "conceptual," in the sense of providing an abstract representation of catchment mechanisms using state variables and parameters that may not correspond directly to observable quantities [*Abbott*, 1996]. Their parameters, therefore, need to be estimated via calibration (inverse modeling) techniques.

Depending on model complexity and data availability, the calibration of hydrological models can be prone to poor identifiability. For example, typical calibration to flows at watershed outlets only allows for the identification of a handful of model parameters [e.g., *Jakeman and Hornberger*, 1993], making it difficult to undertake model calibration in distributed contexts [*Obled et al*., 1994].

Although model classifications (e.g., "top-down" versus "bottom-up," "conceptual" versus "physical") are generally helpful, in practice, the distinction is blurred rather than clear-cut. Most practical models cannot be unambiguously assigned to a unique class, as they invariably combine multiple philosophies and approaches when representing different system components [*Butts et al*., 2004]. Moreover, the level of physical rigor clearly depends on how a model is applied [*Grayson et al*., 1992]. As such, hydrological models form a broad continuum ranging from some of the purest physically based hydrological models—pore-scale models—to what might be considered their true antithesis, pure data-mining techniques. Conceptual models fit somewhere in the middle of this range, arguably closer to the "physical" end when motivated by achieving a fidelious process representation [e.g., *Beven et al*., 1984; *Wood et al*., 1992; *Clark et al*., 2008; *Clark and Kavetski*, 2010; *Fenicia et al*., 2014] rather than purely by maximizing streamflow predictive ability. This study will utilize conceptual modeling techniques in a distributed application and will attempt to relate modeling results to available experimental understanding.

1.2. Decisions in Distributed Hydrological Models

The process of constructing and applying any process-motivated hydrological model—whether conceptual or physical, lumped or distributed—is best pursued systematically, through a series of decisions specifying and testing the distinct aspects of a complete model representation [e.g., *Clark et al*., 2011a].

The decisions involved in distributed hydrological modeling can be grouped into three general categories: (1) specification of the spatial discretization approach, (2) definition of the structure and connectivity of the resulting spatial elements, and (3) specification of parameter and state constraints to improve model parsimony. These decisions represent hypotheses of the underlying system behavior, in particular, hypotheses reflecting the landscape characteristics that control hydrologic response, the dominant processes associated with each landscape element and their spatial connectivity, and the relative magnitude of various catchment processes. As outlined next, distributed models differ significantly in terms of these decisions.

Spatial discretization approaches include regular grids [e.g., *Refsgaard et al*., 1995; *Samaniego et al*., 2010], irregular networks [*VanderKwaak and Loague*, 2001; *Ivanov et al*., 2004], hillslopes [e.g., *Zehe et al*., 2001;

Smith et al., 2013], subcatchments [*Lindstrom et al*., 1997; *Reggiani et al*., 1998], and "hydrological response units" (HRUs) [*Leavesley et al*., 1983]. HRUs can lead to different spatial discretization depending on how they are defined. HRUs may be defined based on just a few characteristics, such as topography alone, leading to relatively coarse spatial discretization [e.g., *Gao et al*., 2014a], or may be defined based on many more characteristics, such as vegetation, soil type, presence of macropores, hydrophobicity, permeability, etc., leading to much finer spatial resolution [e.g., *Scherrer and Naef*, 2003].

With respect to the model structure used within individual spatial elements, many models apply the same structure across the entire domain, e.g., the grid-based MHM model [*Samaniego et al*., 2010] and the HRU-based SWAT model [*Arnold et al*., 2012]. However, in some cases, the presence of certain processes can already be ruled out prior to model development, e.g., subsurface flow processes in areas where most streamflow originates from surface runoff [*Kampf and Burges*, 2007]. For this reason, some distributed models tailor the model structures at the spatial model element scale based on the perceived dominant processes, e.g., the grid-based TAC-D model [*Uhlenbrook et al*., 2004] and the HRU-based FLEX-Topo model [*Savenije*, 2010]. The spatial model elements may be fully connected by lateral flows (e.g., MIKE-SHE), or connected through the groundwater (e.g., FLEX-Topo) or discharge into a common stream (e.g., MHM).

The problem of overparameterization typically arises whenever a distributed model is calibrated to limited forcing-response data. The use of "regularization" procedures, which employ additional information to constrain the model parameters, has become increasingly common. For example, the process of "parameter regionalization" attempts to relate hydrological model parameters to each other and to observable watershed characteristics [*Pokhrel et al*., 2008; *Samaniego et al*., 2010; *Pokhrel et al*., 2012; *Kumar et al*., 2013]. The calibration then proceeds in terms of the parameters of the regional relationships instead of the parameters of the individual hydrological model elements. More generally, relativity and magnitude constraints on model parameters, state variables, and internal fluxes can be used [*Gupta and Nearing*, 2014]. For example, *Gharari et al*. [2014] constrained the magnitude of FLEX-Topo model parameters and fluxes according to a priori judgments on the behavior of different landscape types. Finally, a Bayesian application of regularization using priors on individual model parameters was demonstrated by *Frey et al*. [2011].

1.3. Model Decisions as Multiple Working Hypotheses

Model decisions are seldom unique, and as such represent multiple working hypotheses that require testing against observed data [*Chamberlin*, 1965; *Clark et al*., 2011a]. The potential for more incisive hypothesis testing in hydrology has been demonstrated in many recent applications [e.g., *Krueger et al*., 2010; *Buytaert and Beven*, 2011; *Clark et al*., 2011b; *Westerberg et al*., 2011; *Fenicia et al*., 2014; *Hrachowitz et al*., 2014; *Hublart et al*., 2015, and others]. For example, *Buytaert and Beven* [2011] found that the addition of a parallel store into the original TOPMODEL structure improved streamflow predictions in an Andean catchment, and that this store was hydrologically interpretable as disconnected wetlands and depressions that produce very slow response.

The method of multiple working hypotheses is particularly powerful when pursued using flexible modeling frameworks, such as the FUSE [*Clark et al*., 2008], SUPERFLEX [*Fenicia et al*., 2011], and SUMMA [*Clark et al*., 2015]. These frameworks implement the model building process as a combination of multiple customizable elements. From a hypothesis testing perspective, these frameworks enable controlled model comparisons and thus overcome the difficulties encountered when comparing models that differ in a multitude of uncontrolled aspects [*Kampf and Burges*, 2007; *Clark et al*., 2011a]. For example, *Kavetski and Fenicia* [2011] and *van Esse et al*. [2013] explored the advantages and limitations of flexible models for representing streamflow dynamics of diverse catchments; *Fenicia et al*. [2014] used the fine-level modularity of the SUPERFLEX framework to examine the relationships between catchment characteristics and conceptual model structure at a number of experimental locations in Luxembourg.

Most existing studies have considered the method of multiple working hypotheses in the context of lumped hydrological models [e.g., *Clark et al*., 2011b; *Fenicia et al*., 2014; *Hrachowitz et al*., 2014]. In this paper, we apply the method of multiple working hypotheses to distributed models—which involves a broader spectrum of model decisions and a broader range of hydrological information.

1.4. Aims

The aim of this study is to review and systematically assess the hydrological hypotheses associated with the following key decisions in distributed model development:

A1.1: how to discretize the landscape into regions of hydrologic similarity (HRUs),

A1.2: which model structures to assign to the HRUs in order to reflect their dominant processes, and

A1.3: how to relate model parameters across space in a way that reflects the spatial distributions of the associated catchment processes.

In addition to the specific aims A1.1–A1.3, this study also pursues the broader aim:

A2: contribute toward the systematic development and application of distributed hydrological models using the method of multiple working hypotheses, and explore the advantages and limitations of this method.

We focus our analysis on the Attert catchment in Luxembourg, where extensive experimental data are available to support the development and analysis of distributed hydrological models. The case study site is of particular interest because it comprises 10 gauging stations that exhibit markedly different streamflow dynamics. Previous experimental and modeling studies focused mainly on three headwater subcatchments, and led to the hypothesis that geology controls the dominant runoff generating processes [e.g., *Juilleret et al*., 2012; *Fenicia et al*., 2014]. As the three subcatchments differ in many other aspects, including area, topography, soils, and land use, this hypothesis requires further investigation. Moreover, it is unclear whether this hypothesis also holds for larger subcatchments of the Attert, where process heterogeneity and complex stream networks make it more difficult to interpret the catchment response to precipitation. It is hence of interest to consider the alternative hypothesis that the dominant runoff processes are controlled by topography [*Gao et al*., 2014a; *Gharari et al*., 2014], as well as the "null" hypothesis that neither geology nor topography are active controls.

The hypothesis testing process centers on the assessment of model transferability in space and time. Analysis of spatial transferability represents a particularly stringent test of model performance [*Andreassian et al*., 2009]. Model transferability is assessed using multiple streamflow-based criteria, including traditional "sum-of-squared-errors" prediction metrics as well as "hydrological signatures" such as the base flow and flashiness indices [e.g., *Baker et al*., 2004; *Eckhardt*, 2008]. Note that, in this study, signatures are used purely as independent posterior diagnostics [e.g., *Blazkova and Beven*, 2009; *Kavetski and Fenicia*, 2011], rather than as constraints during parameter inference [e.g., *Zhang et al*., 2008; *Westerberg et al*., 2011].

This paper is structured as follows. Section 2 describes the study area and data. Section 3 describes the case study methodology, including the model setup and the hypothesis testing approach. Section 4 describes the iterative model improvement process, which is undertaken and reported in multiple stages associated with aims A1.1–A1.3. Section 5 discusses the case study results, with an emphasis on interpreting the model hypotheses suggested by the empirical analysis. Section 6 summarizes the main conclusions of the study.

2. Study Area and Data

2.1. Physiography and Climate

The Attert catchment in Luxembourg has been the focus of extensive experimental and modeling work. The majority of this work has concentrated on three headwater subcatchments, namely the Huewelerbach, Weierbach, and Wollefsbach [e.g., *Pfister et al*., 2009, 2010; *Juilleret et al*., 2012; *Fenicia et al*., 2014]. These subcatchments are also part of the present study.

The Attert catchment has an area of 296 km^2 and drains through the outlet in Bissen (see Figure 1). Average precipitation is 860 mm/yr (in the period 2001–2012). Precipitation increases from east to west, resulting in a spatial variation of about 100 mm/yr across the entire Attert catchment. Precipitation does not vary significantly within the year. December, January, and February are the wettest months, with about 100 mm/month, whereas April, August, and September are the driest months, with about 70 mm/month. Potential evaporation has a clear seasonal pattern, ranging from about 15 mm/month in winter to 80 mm/month in summer. The average runoff coefficient is about 40%, with seasonal variations clearly related to the evaporation pattern (see also *Pfister et al*. [2005] for a description of the climate in this area).

The catchment lies on three main geological formations (Figure 1a). The north-western part of the catchment is located on shallow schist bedrock, which is considered to be relatively impermeable though it tends to disaggregate toward the soil surface [*Wrede et al*., 2015]. The central part of the catchment is

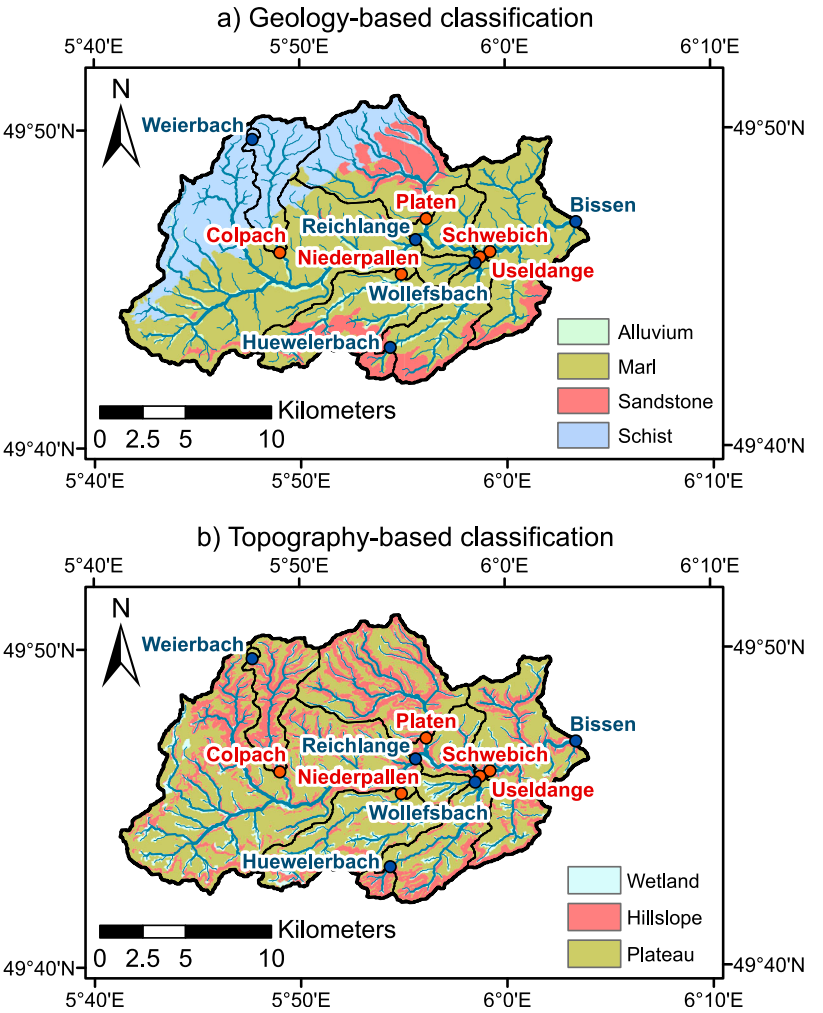

Figure 1. Attert catchment, subcatchments, and stream gauging stations. Landscape classifications based on (a) geology and (b) topography are shown. Stream gauging stations of Group A are marked in blue; stations of Group B are marked in red.

dominated by marl, which is also generally impermeable. The southern part of the catchment lies on sandstone (which is also present in the northern part). The sandstone formation is highly permeable and capable of storing large amounts of water. Alluvial deposits are typically present in the near-stream areas. Further details on the geology of the Attert area are provided by *Fenicia et al*. [2014] and *Wrede et al*. [2015].

The topography of the catchment is illustrated in Figure 1b, following a landscape classification approach based on slope and Height Above Nearest Drainage (HAND) [*Renno et al*., 2008]. Areas with HAND values below a specified HAND threshold (here, 5 m) are classified as "wetlands." The remaining part of the landscape is classified as either "plateau" or "hillslope," depending on whether the slope is below or above a specified slope threshold (here, 11%). The HAND and slope threshold values used in this study are based on the work of *Gharari et al*. [2011] on the Wark catchment in Luxembourg.

Visual inspection of Figures 1a and 1b suggests that a landscape classification based on topography (HAND and slope) would be different from a landscape classification based on geology. Nevertheless, there is some correlation between geological and topographical attributes:

1. Alluvium formations appear to correspond to wetland areas;
2. Marl formations are characterized by gently sloping terrains, whereas sandstone and schist formations typically exhibit, respectively, V-shaped and U-shaped valleys near the river;
3. Marl formations are characterized by extensive plateau areas;

Table 1. Physiography of the Attert Subcatchments (Listed From Smallest to Largest) and Associated Stream Gauging Stations[a]

Gauging Station	Subcatchments Contained	Total Drained Area (km^2)	Partial Drained Area (km^2)	Topographic Composition			Geological Composition			
				Wetland	Hillslope	Plateau	Alluvium	Marl	Sandstone	Schist
Weierbach		0.43	0.43	5%	16%	79%	5%			95%
Huewelerbach		1.6	1.6	9%	65%	26%	10%	22%	68%	
Wollefsbach		4.5	4.5	26%	4%	70%	6%	94%		
Colpach	1	18.5	18.5	8%	49%	43%	7%	14%		79%
Schwebich	2,3	29.1	23.1	14%	28%	58%	14%	62%	24%	
Niederpallen		33.3	33.3	19%	14%	67%	15%	64%	21%	
Platen		43.6	43.6	10%	44%	46%	8%	26%	31%	35%
Reichlange	1,4,6	166.9	114.6	14%	25%	61%	11%	59%	2%	28%
Useldange	1–8	251.3	11.7	17%	20%	63%	14%	86%		
Bissen	1–9	296.3	45.0	14%	25%	61%	9%	81%	10%	

[a]The topographic and geological compositions are calculated for the partial drained areas.

4. Sandstone and schist formations have, on average, higher proportions of hillslope areas than marl formations;

Table 1 reports the total areas drained through the Attert stream gauging stations ("total" subcatchments), the partial areas resulting from subdivision into subcatchments ("partial" subcatchments), and the proportions of geological and topographic landscape types in each partial subcatchment. Table 1 shows that regions with similar geological classes may have different proportions of topographical classes, and vice versa.

The landscape information was derived using: (1) a digital elevation model (50 m resolution), (2) the observed river network, and (3) a geological map of the area. The topography and geology-based classifications, subdivision into subcatchments, and other landscape characteristics were obtained using routine GIS operations. The geology types shown in Figure 1a were obtained by regrouping more detailed information.

2.2. Hydrometeorological Time Series Data

The total observation period used in the present study ranges from 1 September 2003 to 1 September 2012. The observed variables used for the present study include precipitation, streamflow, temperature, and other meteorological variables used to estimate potential evaporation.

Streamflow data are available at subhourly resolution from 10 gauging stations within the catchment. The subcatchments are defined as the areas draining through these gauges (Figure 1). The streamflow data availability is shown schematically in Figure 2; although available streamflow measurements span most of the analysis period, some stations (e.g., Colpach and Bissen) have extensive periods with missing data.

Figure 3a demonstrates the diverse streamflow dynamics of the Attert subcatchments, with different times-to-peak, base flow characteristics, etc. The streamflow diversity can be quantified by the matrix of Pearson cross-correlations R^2 between streamflow time series at multiple locations, shown in Figure 3b. The cross

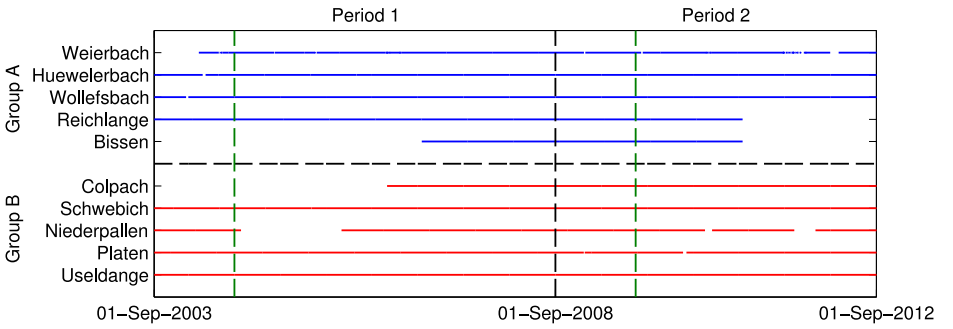

Figure 2. Data availability (indicated by continuous colored lines) and data partitioning for transferability analysis (indicated by dashed black lines). The division into time Periods 1 and 2 is used for split-sample testing of time transferability; the division into subcatchment Groups A and B is used for split-sample testing of spatial transferability (see Table 2). The first year of each period, delimited by green dashed lines, is used for model warmup.

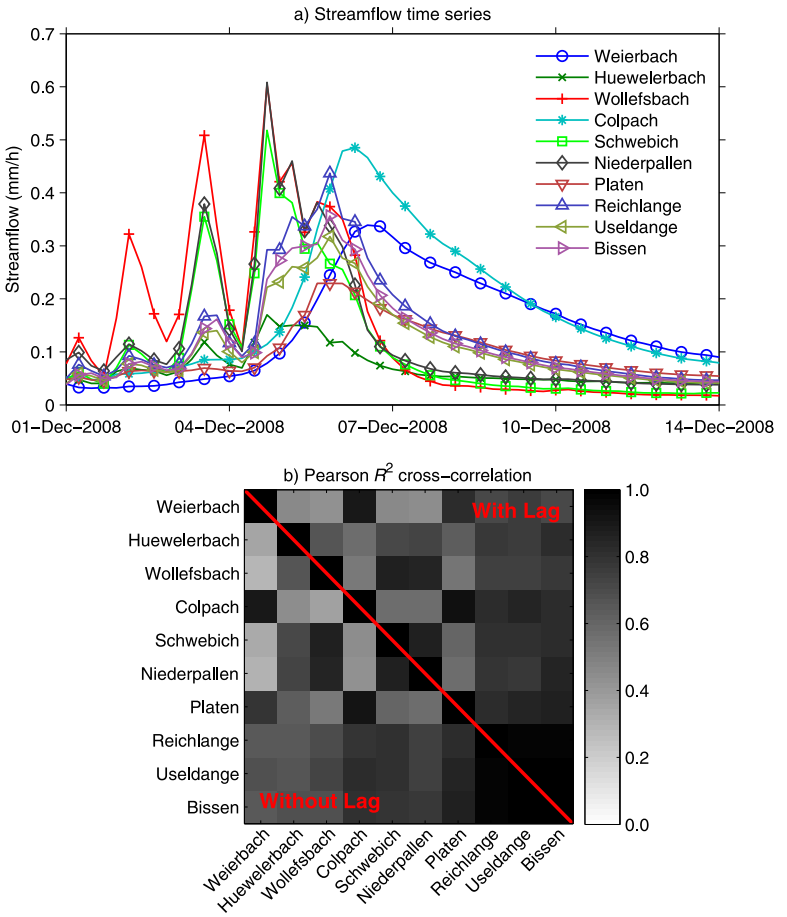

Figure 3. Streamflow diversity in the Attert catchment. (a) observed hydrographs over a 2 week period with a large storm event; (b) matrix of Pearson cross-correlation coefficients between the hydrographs of all subcatchments in the period 2000–2012. The lower triangle shows the R^2 between streamflows at the same time step (i.e., no lag), whereas the upper triangle shows the maximum R^2 computed over lags of 0 to $\pm$4 days. The subcatchments are sorted from smallest to largest by area.

correlation between streamflows at the same time step is shown in the lower triangle, whereas the maximum cross correlation allowing for lags of up to $\pm$4 days is shown in the upper triangle. Figure 3b shows that the smaller subcatchments exhibit particularly diverse streamflow dynamics. For example, the R^2 of the Weierbach versus Wollefsbach hydrographs is 0.29 (unlagged) and 0.42 (allowing for lags). The near symmetry of the lagged/unlagged R^2 matrix suggests that the streamflow variability is due to genuinely different dynamics, rather than merely due to different routing times. Conversely, larger subcatchments show substantially less streamflow variability, due to the nested setup where the larger subcatchments include the smaller ones (Table 1). For example, the hydrographs at Reichlange, Useldange, and Bissen have an $R^2 > 0.97$, with the maximum cross correlation occurring at lag 0.

Precipitation data are available at subhourly resolution from 13 stations inside and outside the catchment. For each subcatchment, spatially averaged precipitation time series were constructed using a nearest neighbor (Thiessen) approach. Note that although continuous time series were constructed for each subcatchment, it was not possible to obtain homogeneous time series because not all rain gauges were operational over the entire analysis period. Periods during which some of the rain gauges were nonoperational were identified using fieldwork inspections and analysis of double mass curves (cumulated streamflow versus cumulated precipitation). Gauge data from these periods were excluded from the analysis. Nonheating rain gauges were excluded from the precipitation analysis over time periods where the temperature dropped below 3°C. The network of available raingauges therefore varied with time.

Daily potential evaporation was calculated as "open water" evaporation using the Penman-Monteith equation [*Monteith*, 1981] and measured climatic data, namely minimum and maximum day temperatures, wind

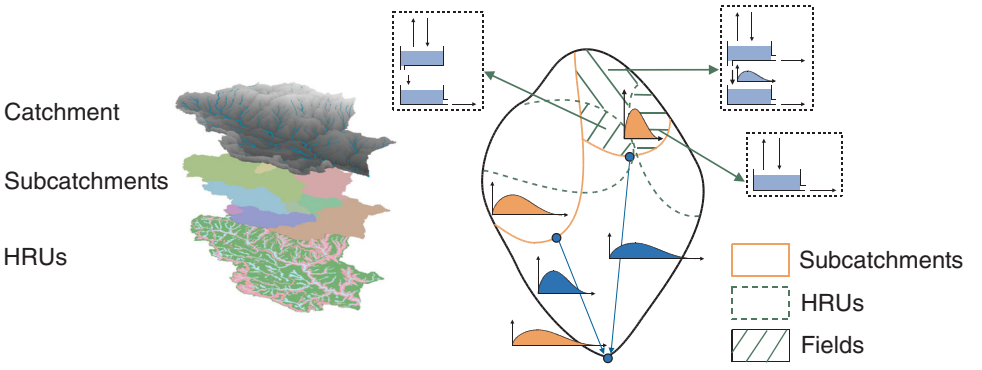

Figure 4. Spatial organization of the catchment model structures. The basic landscape elements are referred to as "fields." Fields are defined based on the spatial discretization of the catchment into HRUs and the spatial discretization of the forcing data. Fields within the same HRU are assigned the same HRU model structure and parameter values; fields within the same subcatchment receive the same forcing data. The fields are modeled independently, hence fields within the same HRU and within the same subcatchment are modeled as a single element. The effect of river routing is represented by convolution functions. We distinguish the routing of the contribution from internal fields (marked in red) versus that from upstream inlets (marked in blue). See section 3.1.1 for further details.

speed, relative humidity, pressure, and sunshine duration. The daily-average estimates were temporally disaggregated assuming a sinusoidal pattern during daylight hours and zero evaporation during the night. Given the 6 h time step used in the model application (see below), this approach usually results in 2 time steps per day with nonzero potential evaporation. The impact of the sinusoidal pattern assumption is hence likely to be minor.

The hydrological modeling was carried out on a 6 h time step, and all model data were averaged to this resolution. The somewhat unusual choice of model time scale was motivated by daily time steps being considered too coarse to characterize the quick response of some of the subcatchments [e.g., see *Kavetski et al*., 2011], whereas using hourly data would have resulted in exceedingly large computational demands.

3. Material and Methods

This section describes the general methodology used for model development and hypothesis testing. The hypotheses specific to the individual stages of iterative model improvement process are reported in section 4, as they are guided by the model analysis results obtained in earlier stages.

3.1. Hydrological Model Setup

3.1.1. Spatial Discretization and Flow Network Structure

This section describes the approach for discretizing the landscape and connecting multiple model elements. The approach is similar to the methods used in other HRU-based distributed models, such as PRMS [*Leavesley et al*., 1983], SWAT [*Arnold et al*., 2012], and in the more general, SUMMA framework [*Clark et al*., 2015].

The flow network structure developed for the present application is shown in Figure 4. The basic spatial elements are referred to as *fields*. Each field is characterized by its own set of state variables, which in general are tracked separately from the state variables of other fields. Each field is assigned a model structure and an associated set of parameter values, and receives a set of forcing data.

The spatial discretization of a catchment into fields requires two distinct steps:

i. Spatial discretization of catchment properties, in order to assign model structures and parameters to each field. In this work, catchment properties are discretized according to HRUs, which in turn are defined based on either topography or geology. For example, the four geology classes of Figure 1a represent 4 HRUs. All fields within the same HRU are assigned the same model structure and parameter values, though not necessarily the same forcing (see below). In this paper, we use the term "landscape classification" to refer to the process of subdividing the catchment landscape into HRUs. Note that landscape subdivision into HRUs is distinct from catchment subdivision into subcatchments; in particular, the same HRU may span multiple subcatchments.

ii. Spatial discretization of the forcing data, in order to assign forcing data to each field. In this work, the forcing is assumed to be uniform over each subcatchment. As a consequence of this choice, fields belonging to the same HRU and located in the same subcatchment will share the same forcing, whereas fields belonging to the same HRU, but located in different subcatchments, will generally receive different forcing data.

The specification of spatial connectivity between the fields also requires two distinct steps:

i. Specification of lateral connectivity between the fields. In this work, we follow the principles of "semidistributed modeling" and assume that the fields are laterally disconnected from each other, so that their state variables are uncoupled and evolve independently. As a consequence of this choice, fields belonging to the same HRU and located in the same subcatchment not only share the same forcing data, but will have identical values of their state variables. Such fields can be aggregated and modeled as a single field, so that each subcatchment will have at most as many fields as there are HRUs. This approach further reduces the number of fields and hence the computational cost of a model simulation. Conversely, fields belonging to the same HRU, but located in different subcatchments, will generally receive different forcing data—the state variables of such fields will therefore have different values and require separate computation.
ii. Specification of a routing network to represent the river network transfer of streamflow from the individual fields to the subcatchment outlets. Within a partial subcatchment, we distinguish the routing of streamflow from its internal fields versus the routing of streamflow from upstream inlets. The routing of streamflow from the internal fields (marked in red in Figure 4) represents the effect of a ramified stream network. This flow represents the contribution of the partial subcatchment to the streamflow of the corresponding total subcatchment. In order to calculate the total streamflow at the subcatchment outlet, it is necessary to include the contribution of upstream inlets. For example, the partial subcatchment at Reichlange receives inflows from the total subcatchments at Colpach and Niederpallen (see Table 1 and Figure 1). Each inflow is routed through a lag function (marked in blue in Figure 4) that represents the effect of a single river stretch. The total streamflow from the (total) subcatchment is then calculated by adding the streamflow from the internal fields to the streamflow routed from the inlets of upstream subcatchments.

In this study, all lag functions associated with the river network have a symmetric triangular shape, as used in the HBV model [*Lindstrom et al*., 1997]. Each lag function is therefore characterized by a single time base parameter (see section 3.1.4).

In view of the nesting of model structures from the HRU level to the catchment level, it is necessary to distinguish between two types of model structures:

HRU Model Structure. Model structure that characterizes a particular HRU (e.g., schist, hillslope, etc.);

Catchment Model Structure. Complete model structure that characterizes the entire catchment (in this study, the Attert catchment).

Note that a catchment model structure comprises the HRU model structures assigned to each field, the (potential) connections between them, and the routing network.

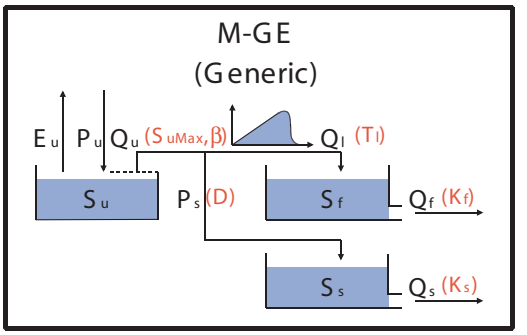

Figure 5. "Generic" HRU model structure, intended to represent a broad range of processes [*van Esse et al*., 2013] and used in catchment model structures M-Uni, M-Top, and M-Geo-1.

3.1.2. HRU Model Structures

The HRU-scale hydrological processes are represented using lumped model structures. These structures are refined as part of an iterative model development process (section 4). Initially, we use the same structure in all HRUs. This model structure is referred to as M-GE ("generic") and is shown in Figure 5. It is similar to models such as HyMod [*Boyle*, 2001] and HBV, and is intended to accommodate a diverse range of streamflow dynamics. In particular, M-GE corresponds to model structure SF11 in the study by *van Esse et al*. [2013], where it provided, on average, good predictive performance over 237 French catchments.

M-GE consists of three reservoirs and a lag function. An unsaturated root zone reservoir (UR) partitions inflow using a saturation excess mechanism, where the runoff coefficient is a nonlinear (power) function of reservoir storage. This reservoir can be emptied only through evaporation. The outflow from UR is partitioned into quick and slow flow components. The quick flow component is first routed through a lag function, and then through a "fast reacting" (FR) linear reservoir. The slow flow component is routed through a "slow reacting" (SR) linear reservoir. The lag function is a smoothed right-angled triangle intended to reproduce the rising limb of the hydrograph.

One of the major aims of this work is to investigate the use of distinct model structures in individual HRUs. Whereas model structure M-GE is sufficiently complex to accommodate a relatively broad range of processes, we consider the "tailoring" of model structures to the specific processes perceived to be dominant in the individual HRUs. The selection of these more parsimonious tailored model structures is detailed in section 4.2.1 and is motivated by the earlier work of *Fenicia et al*. [2014] and *Wrede et al*. [2015].

Snow accumulation and melt are accommodated by augmenting all HRU model structures with a snow reservoir based on a smoothed version of the degree-day model [*Kavetski and Kuczera*, 2007]. The physically distinct processes of soil evaporation, interception, and transpiration are modeled within all HRU model structures as a single "evaporation" component.

The total number of model HRUs, fields, states, and parameters are given in Table 2. The number of fields is given by the total number of HRUs in all subcatchments (by coincidence, both the topography and geology-based discretizations lead to a total of 30 fields). The total number of states is calculated as the total number of distinct states in each field (2–5 states including snow, depending on the HRU model structure, see Table A1) plus the total number of lag functions in the routing network (19 in all cases). Note that states with identical values across multiple fields are counted as a single state. For example, the snow reservoir is common to all fields within the same subcatchment (section 3.1.4), and hence counts as a single state per subcatchment.

Full details of the HRU model structures, including water balance equations, constitutive functions and associated parameters are given in Appendix A. The model equations are solved using the robust implicit Euler time stepping scheme with fixed 6 hourly steps (given by the chosen resolution of forcing data, see section 2.2). The model initialization and numerical solution are detailed in Appendix A.

3.1.3. Model Inputs and Outputs

Model inputs consist of precipitation, potential evaporation, and temperature time series. These inputs are distributed per subcatchment, i.e., each partial subcatchment receives a different set of spatially lumped inputs. The model outputs consist of streamflow time series at all 10 subcatchment outlets. The availability of distributed streamflow data allows joint calibration of model parameters describing different regions of the catchment, as well as independent validation of model predictions at multiple internal catchment locations.

3.1.4. Model Parameters

Within a catchment model structure, model parameters are treated as either: (i) "local" parameters, which are specific to individual HRUs, or (ii) "global" parameters, which are shared (i.e., set to the same value) across the entire catchment. The designation of a parameter as local or global represents a trade-off of parsimony versus spatial representativeness, and is an important model decision considered in our case study.

The sets of parameters in these categories depend on the model decisions, and therefore vary as the model structures are refined during the iterative model development process (section 4). However, the parameters listed below are treated as global in all model development stages:

Table 2. Properties of Catchment Model Structures: Number of HRUs, Parameters, Fields, and States

Catchment Model Structure	Number of HRUs	Number of Global Parameters	Number of Local (HRU-Specific) Parameters	Total Number of Parameters	Total Number of Fields	Total Number of States
M-Uni	1	3	6	9	10	69
M-Top	3	3	18 (6 × 3)	21	30	149
M-Geo-1	4	3	24 (6 × 4)	27	30	149
M-Geo-2	4	3	10 (4 + 3 + 2 + 1)	13	30	83
M-Geo-3	4	4	7 (3 + 2 + 1 + 1)	11	30	73

Snow Module. The "degree-day" snowmelt parameter is calibrated under the assumption that it is uniform over the entire catchment (i.e., the same parameter value was used in all HRUs). The parameters describing the melt temperature of snow in the air and the melt temperature of snow on the ground are kept fixed at 0°C. All numerical smoothing parameters are also fixed. See *Kavetski and Kuczera* [2007] for a description of the snow module parameters.

Evaporation. Parameter C_e, which represents a correction factor to the potential evaporation, is present in all HRU model structures. This parameter is calibrated under the assumption that it is uniform over the catchment. In principle, this evaporation parameter could be linked to vegetation cover, but this information was not considered in the present study. This choice is partly justified because vegetation is relatively uniform at the discretization scales considered.

River routing. The base of the (triangular) lag functions is assumed to be proportional to either: (i) the square root of the area of the subcatchments (convolutions marked in red in Figure 4), or (ii) the length of the stream connecting the nodes (convolutions marked in blue in Figure 4). The same proportionality parameter is used in all lag functions and is calibrated jointly with other model parameters.

The calibration of model parameters is undertaken using a Bayesian Weighted Least Squares approach [*Evin et al.*, 2014], with optimal parameter values estimated using a multistart quasi-Newton optimization algorithm and parameter uncertainty estimated using a Markov Chain Monte Carlo (MCMC) algorithm (Appendix B).

3.2. Hydrological Hypothesis Testing

3.2.1. Generation of Multiple Working Hypotheses

Multiple alternative model structures can be hypothesized using at least two distinct approaches. In the first approach, models are generated a priori and aim to comprehensibly cover the entire space of plausible hypotheses [e.g., *Buytaert and Beven*, 2011; *Fenicia et al.*, 2014; *Hublart et al.*, 2015]. In the second approach, models are generated, trialed, and modified iteratively; in this case, the set of hypotheses at a given model development stage is guided by the outcome of previous stages. For example, the iterative approach was employed by *Fenicia et al.* [2008], where the model structure and parameters were updated as additional sources of information were incorporated into the analysis, and by *Reichert and Mieleitner* [2009], where model diagnostic approaches were used to identify model deficiencies and suggest structural improvements.

In distributed model development, the number of model decisions is generally large, especially compared to the case of lumped models. In these circumstances, a priori model generation becomes impractical. Therefore, this study applies the method of multiple working hypotheses in an iterative way, as elaborated next.

3.2.2. Decision Sets and Stages of Model Development

The hypothesis development and testing framework is designed to address aims A1.1–A1.3 listed in section 1.4. These aims are explored sequentially, by constructing and evaluating model hypotheses associated with each set of model development decisions. The model development process is schematized in Figure 6:

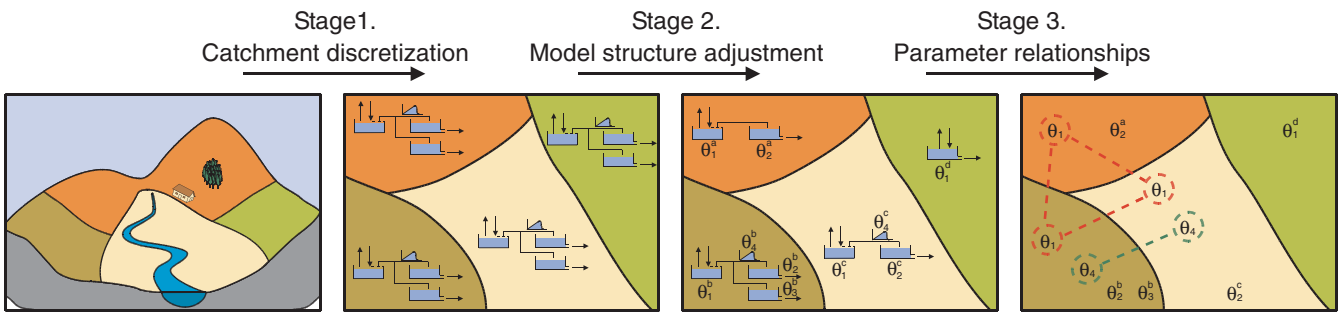

Figure 6. Distributed model development approach. Stage 1 specifies the catchment discretization into HRUs, Stage 2 tailors the model structures to the HRUs, and Stage 3 links model parameters across HRUs.

1. Stage 1 considers multiple landscape classification approaches to discretize the catchment into HRUs;
2. Stage 2 explores the tailoring of model structure to individual HRUs;
3. Stage 3 considers parameter constraints across multiple HRUs.

As we progress through the analysis stages, some of the model decisions are kept fixed at their "optimal" values estimated in the preceding stages. The model hypotheses are compared based on their streamflow predictive ability in split-sample validation setup (section 3.2.3), using multiple metrics (Appendix C).

The model hypotheses are implemented using the SUPERFLEX framework [*Fenicia et al*., 2011]. SUPERFLEX provides a conceptual representation of hydrological processes using a set of generic "hydrological" elements—which can be broadly classified as reservoirs, lag functions, and connections—and implements them using robust numerical schemes. Different model structures can be obtained by combining multiple elements configured to reflect specific processes hypothesized to be dominant by the modeler. The main advantage of the SUPERFLEX framework is that it offers comprehensive fine-grain control over the decisions within the model building process, thus facilitating thorough and systematic model comparisons.

3.2.3. Appraisal of Space-Time Model Transferability

The model hypotheses are evaluated based on their space and time transferability (see Figure 2):

a. Time transferability is tested by subdividing the total observation period into two periods, referred to as Period 1 (1 September 2003 to 31 August 2008, 5 years) and Period 2 (1 September 2008 to 1 September 2012, 4 years);
b. Space transferability is tested by grouping the subcatchments into two groups, referred to as Group A and Group B (see below for the grouping rationale).

The following split-sample analysis procedure is used:

1. Calibrate the model using the Group A stream gauging stations in Period 1;
2. Validate the model *in time*, by evaluating the model on Group A stations in Period 2;
3. Validate the model *in space*, by evaluating the models on Group B stations in Period 1;
4. Validate the model *in space-time*, by evaluating the models on Group B stations in Period 2;
5. Repeat Steps 1–4 for every combination of Group A/B stations and Periods 1/2.

The notation Val-X*m* (Cal-Y*n*) is used to indicate the calibration/validation scenarios. For example, "Val-B2 (Cal-A1)" refers to the scenario where the models are calibrated to Group A stations in Period 1 and validated on Group B stations in Period 2.

The hydrological models are assessed in terms of streamflow predictive ability and in terms of parameter identifiability given available data (sections 4.1.2, 4.2.2, and 4.3.2). The streamflow predictive ability of the calibrated models is quantified using a combination of statistical error metrics and hydrological signatures, selected to provide a more complete picture of the relative merits of individual models and model components [e.g., *Criss and Winston*, 2008; *Gupta et al*., 2008; *Kavetski et al*., 2011]. Details of the predictive performance criteria are provided in Appendix C. Results are then linked to the process understanding available from experimental fieldwork in the Attert catchment (section 5).

3.2.4. Construction of the Space-Time Validation Setup (Groups A and B)

The grouping of subcatchments into Groups A and B is motivated by the following hydrological rationale:

i. We hypothesize that the spatial streamflow variability across the Attert catchment [*Fenicia et al*., 2014] (see Figure 2) is due to differences in the HRU "composition" of its subcatchments.
ii. Based on proposition (i), we hypothesize that the streamflow time series of the headwater subcatchments carry more information about individual HRUs than the streamflow time series of larger subcatchments, due to the headwater subcatchments being smaller and more homogeneous in their HRU composition;
iii. Based on proposition (ii), an "informative" group of subcatchments can be obtained by grouping together the headwater subcatchments;
iv. The informativeness of the group can be enhanced by adding a few of the larger subcatchments, because the inclusion of streamflow time series that combine contributions from multiple HRUs provides an implicit constraint on the calibration of models to the homogeneous headwater subcatchments alone.

Based on this rationale, Group A is constructed to contain all three headwater subcatchments (i.e., Huewelerbach, Weierbach, and Wollefsbach), in combination with the largest subcatchment (Bissen) and a medium sized subcatchment (Reichlange). Group B is constructed to contain medium sized subcatchments (Colpach, Schwebich, Niederpallen, Platen, and Useldange). If our reasoning in steps (i)–(iv) above is valid, calibration of models to Group A subcatchments will be easier than calibration to Group B subcatchments.

The partitioning of the available time series into Periods 1 and 2 is governed primarily by data availability. We opt for simple split-sample validation instead of "leave-*k*-out" cross validation, in order to test the models' ability to extrapolate over longer time periods and to reduce the computational cost/complexity of the analysis. On the other hand, leave-*k*-out cross validation can be more forgiving in the presence of data inconsistencies and nonstationarities; these issues tend to be more frequent in longer time series due to potential major changes in catchment condition and/or changes in the data acquisition approaches. Note that Period 1 is slightly longer than Period 2 (5 and 4 years) to provide a balanced amount of continuous data.

4. Iterative Model Analysis and Improvement

This section describes the hypotheses and results associated with the three model development stages: spatial discretization of the landscape (Stage 1, section 4.1), tailoring the model structure to HRUs (Stage 2, section 4.2), and analysis of spatial relations between model parameters across HRUs (Stage 3, section 4.3).

4.1. Stage 1: Spatial Discretization of the Landscape (Models M-Uni, M-Top, and M-Geo-1)

4.1.1. Hydrological Hypotheses

In Stage 1, the following landscape classification approaches are compared:

M-Uni. Uniform landscape classification. In this classification, the entire catchment is treated as a single HRU, with a single set of model structure parameters. Note that, as precipitation is distributed per subcatchment (section 3.1.3), each subcatchment has its own storage state variables. This hypothesis leads to a model with nine parameters (Table 2, see also paragraph below);

M-Top. Topography-based landscape classification. This classification is described in section 2.1, and results in 3 HRUs: wetland, hillslope, and plateau (Figure 1b). This hypothesis leads to a total of 21 parameters (Table 2).

M-Geo-1. Geology-based landscape classification. This classification is described in section 2.1 and results in 4 HRUs: alluvium, marl, sandstone, and schist (Figure 1a). It leads to 27 parameters (Table 2).

In Stage 1, the same model structure, M-GE, is used across all HRUs. This methodology allows attributing differences in model predictive performance to differences in spatial discretization. In turn, these differences are interpreted as an indication of the degree to which the landscape types of topography and geology control the actual hydrological response in the case study catchment. M-Uni serves as a "null" hypothesis that landscape type does not materially affect hydrological response.

The "generic" HRU model structure M-GE is shown in Figure 5. This model structure is used in Stage 1 because it can accommodate a broad range of processes (section 3.1.2) and is hence less likely to bias the results in favor of any particular landscape classification approach. M-GE leads to six parameters per HRU (Table A1).

As landscape classifications M-Uni, M-Top, and M-Geo-1 lead to a different number of HRUs (see list above), the corresponding catchment models have a different total number of local (HRU-specific) model parameters. In addition, all catchment models have three global parameters (section 3.1.4).

4.1.2. Results

4.1.2.1. Performance in Terms of Maximum Likelihood

Figure 7 shows the (scaled) maximum log likelihood values ($\overline{\Phi}$) achieved by the catchment model structures on the multiple combinations of subcatchments and time periods. At this stage, only the first three catchment model structures of Figure 7 are considered, namely M-Uni, M-Top, and M-Geo-1. The likelihood is the objective function used for model calibration, and it is therefore the logical metric to rank the models in terms of their predictive performance. The scaled log likelihood values reported below represent the logarithm of the average probability density of a streamflow data point within the predictive distribution

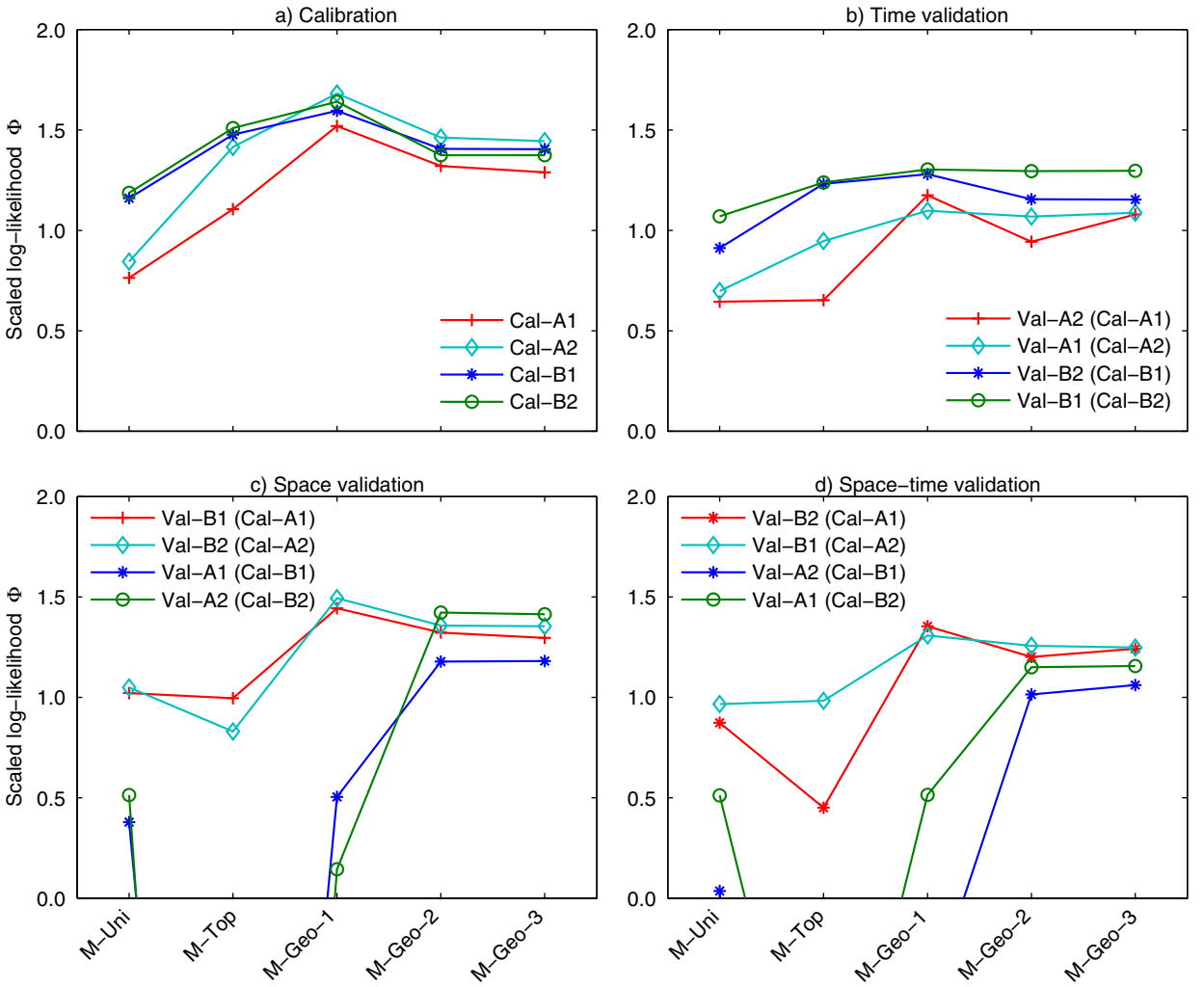

Figure 7. Performance of catchment model structures in terms of scaled log likelihood values (see section 3.2.3 for the definition of the Cal/Val notation). In Figure 7a, the likelihood values are obtained through optimization, in Figures 7b–7d, the likelihood values are obtained by applying the likelihood function with previously calibrated parameters to independent "validation" data sets. Calibration performance is higher for more complex models (M-Geo-1 in Figure 7a). Validation performance favors the simpler geology-based models (M-Geo-2 and M-Geo-3, especially in Figures 7c and 7d). It can also be seen that spatial transferability is generally harder to achieve than time transferability (Figure 7b versus Figures 7c and 7d).

generated by the models. We report the absolute values and comment on the differences. As likelihood values have a statistical rather than hydrological interpretation, the results in this section should be viewed in conjunction with the comparison of other metrics and visual hydrograph comparisons (see remainder of section 4.1.2).

When comparing the performance of M-Uni, M-Top, and M-Geo-1 in Figure 7, the following results can be noted:

1. *Calibration Period* (Figure 7a). M-Geo-1 achieves the highest $\overline{\Phi}$; followed by M-Top and M-Uni. All four curves (representing different periods and groups of catchments) result in the same ranking of models. On average (all four curves), the differences in average log likelihood values, $\overline{\Phi}(\text{M-Uni}) = 0.99$, $\overline{\Phi}(\text{M-Top}) = 1.38$, and $\overline{\Phi}(\text{M-Geo-1}) = 1.61$ correspond to a range of about a factor of 2 in average probability density per observed data point (e.g., a factor of 1.3 difference between M-Geo-1 and M-Top). Note that M-Geo-1 is also the most complex model considered in this study (Table 2).
2. *Time Validation* (Figure 7b). M-Geo-1 continues to be overall better (higher $\overline{\Phi}$), although the differences are even less pronounced than in the calibration period: $\overline{\Phi}(\text{M-Uni}) = 0.83$, $\overline{\Phi}(\text{M-Top}) = 1.02$, and $\overline{\Phi}(\text{M-Geo-1}) = 1.21$ (averaged over all four curves). All four curves result in the same model ranking.

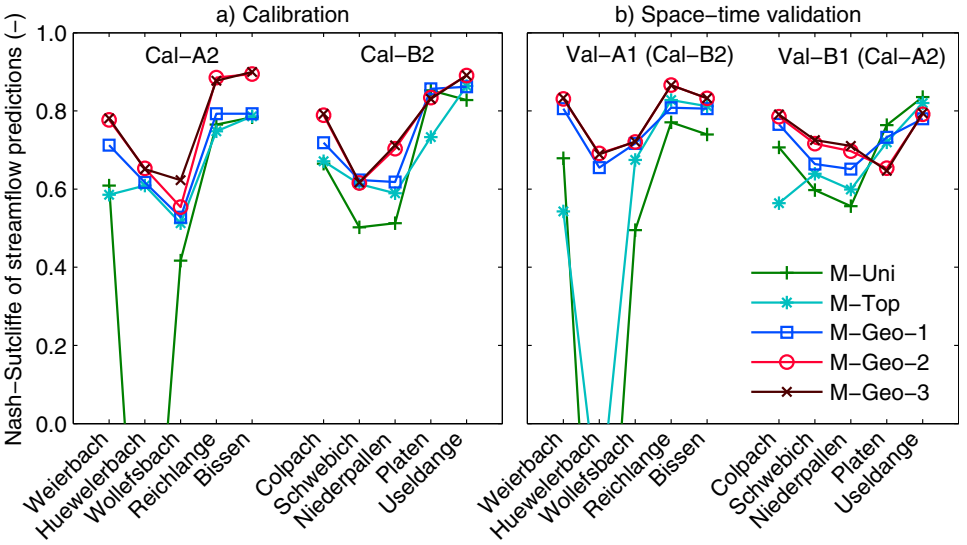

Figure 8. Performance of catchment model structures in terms of Nash-Sutcliffe efficiencies (NSE). Note that the NSE efficiency differs from the WLS objective function used in model calibration and hence represents a (relatively) independent verification metric. (b) Geology-based models exhibit better performance in space-time validation than the topography-based models. The drop in performance in the Huewelerbach subcatchment is discussed in section 5.1.

3. *Space Validation* (Figure 7c). When originally calibrated on Group A subcatchments, M-Geo-1 performs better than the other models. On average over scenarios Val-B*m* (Cal-A*m*), $\overline{\Phi}(\text{M-Uni}) = 1.04$, $\overline{\Phi}(\text{M-Top}) = 0.91$, and $\overline{\Phi}(\text{M-Geo-1}) = 1.47$. However, the spatial transposability of all models deteriorates substantially when originally calibrated on Group B subcatchments, and it is M-Uni that performs best. On average over scenarios Val-A*m* (Cal-B*m*), $\overline{\Phi}(\text{M-Uni}) = 0.45$, $\overline{\Phi}(\text{M-Top}) = -5.91$, and $\overline{\Phi}(\text{M-Geo-1}) = 0.32$.
4. *Space-Time Validation* (Figure 7d). Similarly to space validation, models originally calibrated on Group A subcatchments provide a notably more robust performance than the same models originally calibrated on Group B subcatchments. Averaging over scenarios Val-B*n* (Cal-A*m*) with $m \neq n$, M-Geo-1 achieves the highest likelihood: $\overline{\Phi}(\text{M-Uni}) = 0.92$, $\overline{\Phi}(\text{M-Top}) = 0.72$ and $\overline{\Phi}(\text{M-Geo-1}) = 1.33$. This corresponds to a factor of about 1.8 difference in average probability density per observed data point between models M-Geo-1 and M-Top. In contrast, when averaging over scenarios Val-A*n* (Cal-B*m*), it is M-Uni that achieves the highest likelihood: $\overline{\Phi}(\text{M-Uni}) = 0.27$, $\overline{\Phi}(\text{M-Top}) = -3.64$, and $\overline{\Phi}(\text{M-Geo-1}) = -0.01$.

Comparing time versus space validation results, it can be seen that swapping the calibration-validation time periods has a much smaller effect than swapping the calibration-validation subcatchments. For this reason, the remainder of the paper focuses on the analyses Cal-*X*2 and Val-*X*1(Cal-*Y*2), i.e., on calibration on Period 2 and space-time validation on Period 1.

4.1.2.2. Performance in Terms of Nash-Sutcliffe Efficiency

Figure 8 shows the Nash-Sutcliffe efficiency associated with all three model development stages. The focus in this section is still on M-Uni, M-Top, and M-Geo-1. The left plot displays the calibration performance, whereas the right plot indicates the space-time validation performance.

In Figure 8, catchment model structures M-Uni, M-Top, and M-Geo-1 compare as follows:

1. *Calibration* (Figure 8a). The "null" hypothesis M-Uni has the worst performance of all hypotheses considered in Stage 1, with an average Nash-Sutcliffe index Υ of 0.52 (ranging from negative to 0.85). Note the poor performance of M-Uni in the Huewelerbach ($\Upsilon < 0$), even when this subcatchment is included in the calibration. M-Top offers a better predictive performance than M-Uni, with an average $\Upsilon = 0.67$ (ranging between 0.51 and 0.86); M-Uni achieves $\Upsilon = 0.61$ in the Huewelerbach subcatchment. M-Geo-1 provides the best overall performance, with an average $\Upsilon = 0.71$ (ranging between 0.53 and 0.86). However, it offers little improvement over M-Top in the Huewelerbach subcatchment, $\Upsilon = 0.62$.
2. *Space-Time Validation* (Figure 8b). M-Geo-1 has a distinctly better performance than M-Uni and M-Top. The quality of M-Geo-1 predictions is (relatively) consistently high across the 10 subcatchments, with an

average Υ of 0.74 (ranging between 0.65 and 0.81). The performance of M-Uni and M-Top is lower, with average Υ of 0.53 (ranging from negative to 0.84) and 0.60 (ranging from negative to 0.83), respectively.

The quality of M-Uni and M-Top is highly variable: these models perform well on some subcatchments but poorly on others. For example, in space-time validation, M-Uni fits the Useldange streamflow quite well ($\Upsilon > 0.80$), but its performance drops in the Wollefsbach ($\Upsilon = 0.49$) and particularly in the Huewelerbach ($\Upsilon < 0$). Similarly, M-Top fits Reichlange, Bissen, and Useldange streamflows well ($\Upsilon > 0.80$), but its performance drops in the Weierbach ($\Upsilon = 0.54$) and once again in the Huewelerbach ($\Upsilon < 0$).

It can also be seen that M-Uni and M-Top do not behave consistently in calibration and space-time validation. For example, M-Top provides relatively accurate simulation of streamflow in the Colpach subcatchment in calibration: $\Upsilon = 0.67$, which is essentially similar to the results for M-Uni, $\Upsilon = 0.66$ and for M-Geo-1, $\Upsilon = 0.72$ (Figure 8a). However, in the space-time validation, M-Top predictions only achieve $\Upsilon = 0.56$ whereas M-Uni and M-Geo-1 achieve clearly higher accuracy: $\Upsilon = 0.71$ and 0.77, respectively (Figure 8b).

The Huewelerbach subcatchment clearly represents a challenging catchment in space-time validation. In this catchment, M-Uni is poor in both calibration and validation, $\Upsilon < 0$. The performance of M-Top drops from $\Upsilon = 0.61$ in calibration to $\Upsilon < 0$ in space-time validation. M-Geo-1 has a consistent and relatively high performance in calibration and validation, $\Upsilon = 0.62$ and 0.66, respectively. The Huewelerbach is a groundwater-dominated catchment; its streamflow is dominated by base flow, as will be seen below in the analysis of the base flow index. The base flow, apart from being relatively high, is also characterized by very slow, difficult to predict, dynamics. As a result, the streamflow does not vary as much as in the other catchments (see Figure 3a). Both because the base flow is difficult to predict, and because the total streamflow has little variability, it is harder for the models to achieve a high Nash-Sutcliffe efficiency in this catchment.

The relative model performances in terms of the Nash-Sutcliffe can be compared to the relative performances in terms of the maximum log likelihood values. With respect to the Nash-Sutcliffe efficiency, in the four scenarios represented in Figure 8—Cal-A2, Cal-B2, Val-A1 (Cal-B2), and Val-B1 (Cal-A2)—it can be seen that $\Upsilon(\text{M-Uni}) < \Upsilon(\text{M-Top}) < \Upsilon(\text{M-Geo-1})$. With respect to the maximum log likelihood values (Figure 7), the same ranking is respected, except for the scenario Val-A1 (Cal-B2), where $\overline{\Phi}(\text{M-Top}) < \overline{\Phi}(\text{M-Uni}) \approx \overline{\Phi}(\text{M-Geo-1})$ (Figure 7d). Hence, although the performance of M-Geo-1 in Val-A1 (Cal-B2) is similar to M-Uni in terms of $\overline{\Phi}$ (Figure 7d), M-Geo-1 has a much better performance in terms of Υ: on average, in Val-A1 (Cal-B2), $\Upsilon(\text{M-Geo-1}) = 0.76$ versus $\Upsilon(\text{M-Uni}) = 0.37$. The contrasting results in terms of the log likelihood versus Nash-Sutcliffe efficiency values may be indicative of the different emphases of these metrics: the log likelihood provides a balanced weighting of high and low flows, whereas the (unweighted) Nash-Sutcliffe efficiency tends to emphasize the fitting of high flows only (equations (C1) and (C2) in Appendix C). A more detailed analysis of streamflow time series would be necessary gain a better understanding of these differences in the present study.

4.1.2.3. Performance in Terms of Hydrological Signatures

Figure 9 appraises the model hypotheses in terms of the base flow index I_b and the flashiness index I_f of their streamflow predictions. Note that the signatures of both the observed and simulated streamflows are presented for the space-time validation period Val-*X*1 (Cal-*Y*2).

The signatures of observed hydrographs further illustrate the diversity of streamflow dynamics in the case study subcatchments. I_b ranges from 0.52 to 0.88, and I_f ranges from 0.07 to 0.36. Note that the indices are not strongly correlated. For example, the Huewelerbach has the highest base flow index ($I_b = 0.88$), but ranks average with respect to the flashiness index ($I_f = 0.10$). The Weierbach has the lowest flashiness index ($I_f = 0.07$), but ranks average with respect to the base flow index ($I_b = 0.74$).

M-Uni, with distributed storage accounting but uniform parameters, is unable to capture the streamflow variability across the subcatchments. For example, while the base flow indices of observed streamflows vary from 0.52 to 0.88 depending on the subcatchment, the base flow index of the M-Uni streamflow stays constant at about 0.77 for all subcatchments. A similar finding holds for the flashiness index: it varies from 0.07 to 0.36 when computed using observed streamflows, yet varies only between 0.07 and 0.08 when computed using the streamflows predicted using model M-Uni. This finding suggests that spatial variability in rainfall alone is insufficient—by a wide margin—to explain the spatial variability in the streamflow characteristics.

M-Top provides better predictions than M-Uni, for example, by correctly predicting that Huewelerbach and Platen have relatively high I_b and relatively low I_f, respectively. However, M-Top still exhibits notable

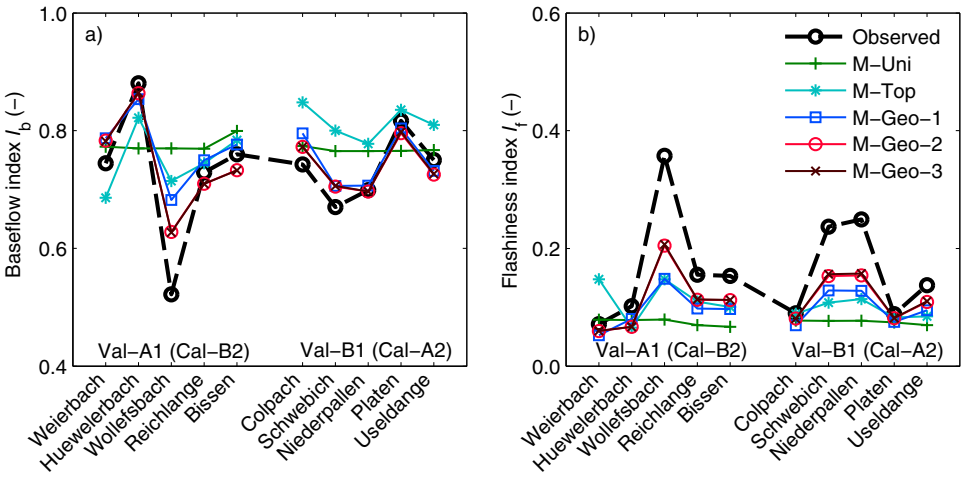

Figure 9. Comparison of hydrological signatures of simulated streamflows versus hydrological signatures of observed streamflow. All signatures are calculated over the space-time validation period. M-Uni, where the same model structure is used in all HRUs, completely misses major differences in subcatchment behavior. The observed variability in hydrological signatures across the subcatchments is best captured by models M-Geo-2 and M-Geo-3; M-Geo-1 and M-Top also follow the observed signature trends, but to lesser degrees.

limitations. For example, M-Top predicts higher I_b for the Wollefsbach than for the Weierbach (0.72 and 0.69, respectively), and predicts higher I_f for the Weierbach than for the Huewelerbach (0.15 and 0.07, respectively), whereas observations indicate opposite rankings in both cases (0.52 and 0.74 and 0.07 and 0.10, respectively).

M-Geo-1 captures the spatial pattern of observed signatures better than M-Top and M-Uni. In particular, it is the only model that correctly predicts the Huewelerbach and Wollefsbach subcatchments to have the highest and lowest base flow indices, respectively (Figure 9a). M-Geo-1 is also the only model that correctly predicts the Weierbach and Wollefsbach to have the lowest and highest flashiness indices respectively (Figure 9b).

We now consider the quantitative goodness of fit of modeled and observed signatures. In terms of the base flow index I_b, M-Geo-1 outperforms M-Uni and M-Top, producing simulated signatures that generally match the observed signature values. Some exceptions can be noted, e.g., in the Wollefsbach catchment, the simulated and observed I_b are 0.68 and 0.52, respectively. M-Uni clearly overestimates I_b for the Group B catchments (by about 0.1 on average). M-Uni is generally poor on all catchments. In terms of the flashiness index I_f, M-Geo-1 once again produces the closest match to the observed values, followed by M-Top and M-Uni. However, all models, including M-Geo-1, tend to underestimate I_f in all catchments. The possible reasons for this underestimation will be discussed in section 5.1.

4.1.2.4. Parameter Identifiability

M-Geo-1, although seemingly the "best" performing model identified thus far, is also the most complex of the models used in this study (Table 2). There are some indications that this model might be too complex: (1) multiplicity of local optima (see below), (2) large parameter uncertainties.

Figure 10 shows the distribution of optimal values scaled log likelihood values obtained using independent quasi-Newton searches initiated from 100 random locations in the feasible parameter space (see Appendix B). The optimal values are sorted and normalized between 0 and 1. For a given model, the extent of the top horizontal segment, where the normalized value of the scaled log likelihood, $\bar{\bar{\Phi}}$ reaches its highest value of 1, indicates the proportion of local searches that converged to the (presumably) global optimum. The extents of other horizontal segments indicate the proportion of searches that converge to corresponding local optima. A single well-defined step in this "distribution of optima" indicates a single well-identifiable optimum, whereas "continuous" increases are indicative of a multitude of poorly distinguishable local optima.

The four "steps" in the modal distribution plot of M-Uni indicate four distinct local optima, with about 25% of the searches leading to the global optimum. The continuous increases in the curves describing the optima of models M-Geo-1 and M-Top suggest that these models do not have well-identifiable optima, but rather exhibit a multitude of local optima. The abundance of multiple optima in these models suggests

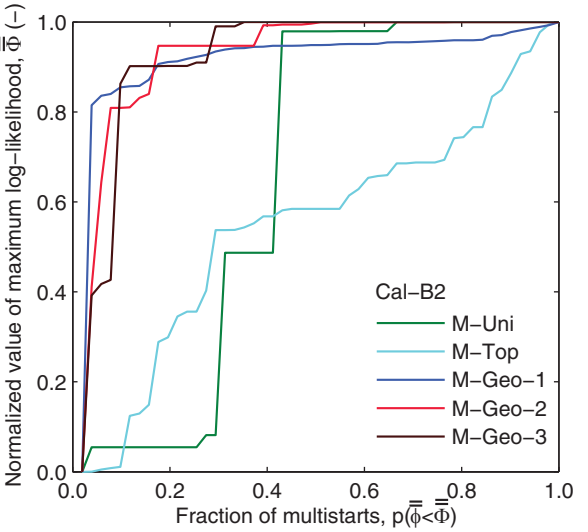

Figure 10. Distribution of maximized scaled log likelihood values (normalized between 0 and 1) obtained from independent local quasi-Newton searches initiated from 100 random locations in the feasible parameter space. See section 4.1.2 for the interpretation of this figure. The distinct "steps" in the distributions of optima of the simpler catchment models (M-Uni, M-Geo-2, and M-Geo-3) suggest that these models have fewer local optima (and hence better identifiability) than the more complex models (M-Top and M-Geo-1).

poor identifiability, possibly due to excessive model complexity. Further model refinement should hence focus on improving parsimony. This leads to Stage 2 and is described next.

4.1.2.5. Visual Hydrograph Comparison

Figure 11 shows the fitted hydrographs generated using catchment models M-Uni, M-Top and M-Geo-1 for three headwater subcatchments, Huewelerbach, Weierbach, and Wollefsbach, in space-time validation. The left plots show hydrographs from a representative time period, whereas the right plots show scatter plots of observed versus predicted streamflow for the entire space-time validation period (Period 1).

Figures 11a and 11b show the performance of M-Uni. This model produces very similar hydrographs in all three subcatchments, which is in stark contrast to the variability exhibited by the observed hydrographs. The poor representation of spatial variability is easy to explain by considering that M-Uni assumes spatially uniform model structure and parameter values. In fact, since M-Uni characterizes all subcatchments using the same HRU model structure and the same parameter set, simulated hydrographs in different subcatchments can differ only due to: (1) rainfall variability over the three subcatchments, and (2) differences in stream length and hence in river routing times. These differences in forcing data and stream length parameters do lead to some differences in the model states and streamflow predictions for different subcatchments. For example, the hydrograph of the Wollefsbach subcatchment is slightly lower than that of the Huewelerbach and Wollefsbach subcatchments (Figure 11a). However, these differences are too minor to explain the actual spatial variability of observed streamflow time series. For example, in the observation window shown in Figure 11a, the observed peak flows of the Huewelerbach, Weierbach, and Wollefsbach subcatchments vary by a factor of 3 (0.5, 1.0, and 1.5 mm/h, respectively), but peak flows predicted by M-Uni for these catchments range merely between 0.3 and 0.5 mm/h.

Figures 11c and 11d show that M-Top is able to produce different streamflows for the three subcatchments. For example, M-Top correctly predicts the Huewelerbach subcatchment to have lower peaks and higher base flow than the Weierbach and Wollefsbach subcatchments. However, the Huewelerbach predictions produced by M-Top are very poor, with Nash-Sutcliffe efficiency below zero. Moreover, M-Top is not able to capture some of the notable spatial patterns across the Attert catchment, such as the flashier response of the Wollefsbach versus the more damped and delayed second-peak response of the Weierbach. The hydrographs predicted by M-Top for these subcatchments are in fact very similar. The predictive deficiencies of M-Top, evident from the visual analysis, suggest that the topography-based spatial discretization has not captured some major sources of spatial streamflow variability.

Figures 11e and 11f show that M-Geo-1 captures to a significant extent the differences in the hydrograph dynamics of the three subcatchments, including the delayed response of the Weierbach, the flashier response of the Wollefsbach and the smoother response of the Huewelerbach. For example, in contrast to the other models, M-Geo-3 correctly predicts the Wollefsbach and Huewelerbach to have the highest and lowest peak flow respectively, and it correctly captures the time to peak (1–2 days) of the Weierbach catchment. That said, M-Geo-1 is far from a perfect model and consistently underestimates the peaks in the Huewelerbach catchment. For example, during the storm event of 17 January 2007, M-Geo-1 predicts a peak streamflow of 0.8 mm/h, whereas the observed streamflow is 1.5 mm/h.

The performance of models M-Uni, M-Top, and M-Geo-1 calibrated to the Huewelerbach subcatchment is further illustrated in Figures 12a–12c, where the 95% prediction limits are shown. It can be seen that these

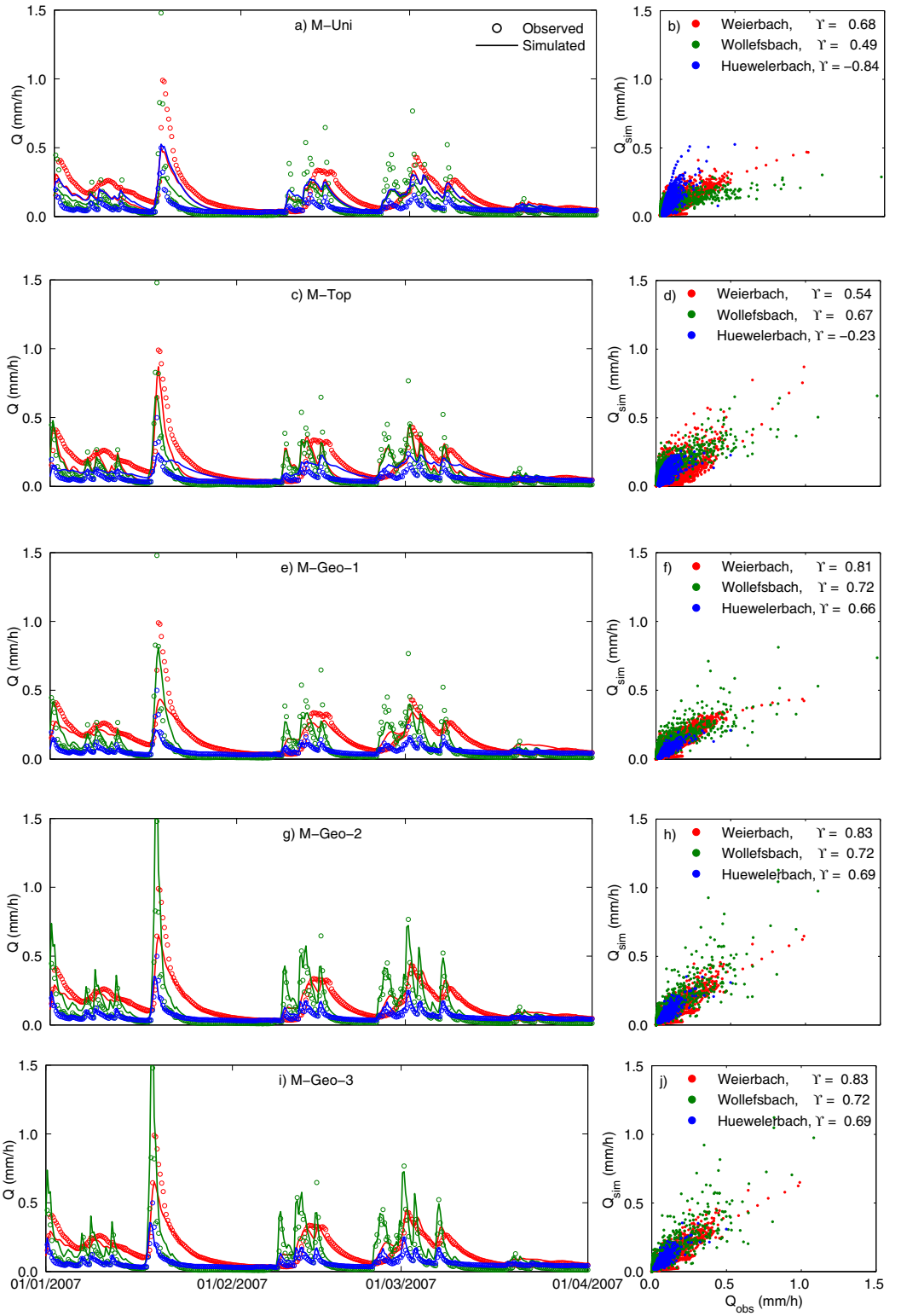

Figure 11. Hydrographs of the headwater subcatchments Weierbach, Wollefsbach, and Huewelerbach, selected because they exhibit clearly different streamflow dynamics. Left plots show the streamflow time series for a subset of the Val-A1 (Cal-B2) analysis. Right plots show scatterplots of observed versus predicted streamflows and the Nash-Sutcliffe efficiency ϒ calculated over Period 1 (excluding warmup). It can be seen that catchment model structures M-Geo-2 and M-Geo-3 (which produce nearly identical predictions) capture best the differences in streamflow dynamics of the three subcatchments.

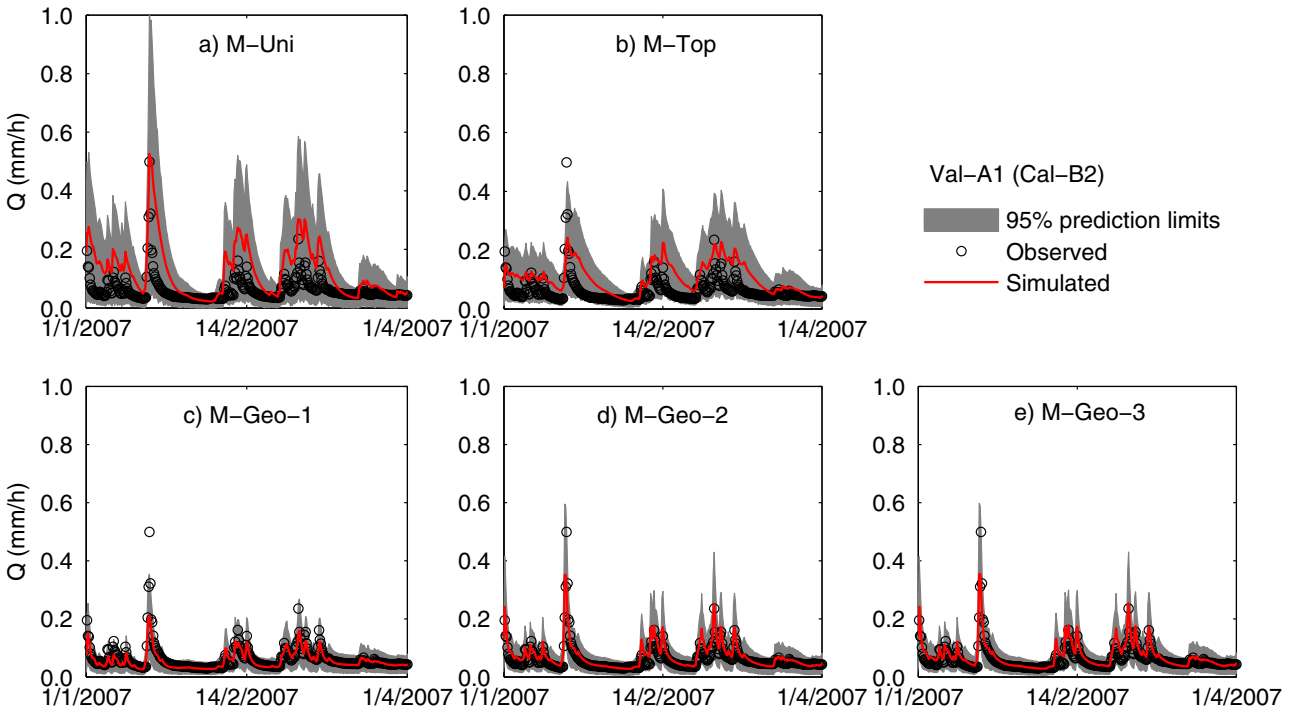

Figure 12. Streamflow prediction limits computed over a portion of the validation period in the Huewelerbach subcatchment. The catchment models M-Uni, M-Top, and M-Geo-1 provide a progressively better match to the observations, with the 95% prediction limits become tighter while still capturing the observed time series. Models M-Geo-2 and M-Geo-3, despite being simpler than M-Geo-1, provide slightly more reliable predictive distributions (e.g., see the storm around 17 January 2007).

catchment model structures provide a progressively better match to the observations. As a result of the better fit, the 95% prediction limits become notably tighter while still capturing the observed time series.

4.2. Stage 2: Tailoring Model Structure to HRUs (Model M-Geo-2)

4.2.1. Hydrological Hypotheses

Stage 1 results indicate that model hypothesis M-Geo-1 provides better predictive performance than the alternative model hypotheses M-Uni and M-Top. This finding is in turn indicative of the hydrological dynamics of the study area being controlled primarily by geology. Stage 2 investigates the tailoring of the HRU model structure to specific HRUs—instead of using the same model structure in all HRUs as was done in Stage 1. All subsequent analyses are based on M-Geo-1, i.e., no further analysis of M-Uni and M-Top was undertaken.

Stage 2 takes advantage of our previous work on the Attert catchment [*Fenicia et al*., 2014], where we trialed multiple lumped model structures on three headwater subcatchments, namely the Huewelerbach, Weierbach, and Wollefsbach. *Fenicia et al*. [2014] suggested that these three subcatchments are best represented by distinct model structures. These distinct structures were interpreted in terms of the distinct dominant processes in the three subcatchments, such as surface runoff, lateral flow, deep percolation, etc.

The three headwater subcatchments considered by *Fenicia et al*. [2014] lie on mixed geologies (see Table 1). We hypothesize that the lumped models that best represented the behavior of these subcatchments included the combined effect of the different geologies. In order to attempt to isolate the effect of the different geologies, we revisit our previous results [*Fenicia et al*., 2014]. The HRU model structures selected to represent the different geologies are detailed below (all model equations are reported in Appendix A):

1. **Schist** is represented using the model structure M-SC shown in Figure 13a. This structure comprises a threshold reservoir, a lag function, and a nonlinear reservoir all connected in series. The threshold reservoir is intended to represent the threshold-like streamflow response of schist, whereas the lag function is intended to represent the pronounced delay in streamflow response of schist during wet conditions. The threshold behavior of the schist-dominated Weierbach subcatchment can be seen in Figure 3a, whereas the threshold effects are discussed in *Wrede et al*. [2015]. Note that model structure M-SC is simpler than model structure M-GE—it has one reservoir less—because schist is considered impermeable and unlikely

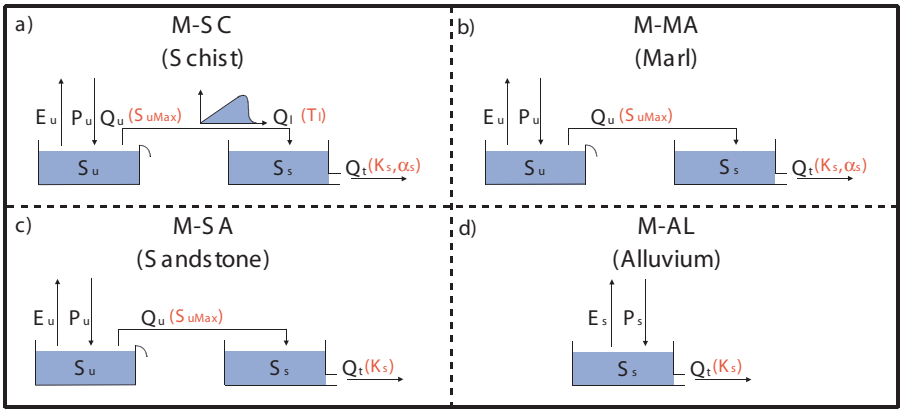

Figure 13. HRU model structures used in models M-Geo-2 and M-Geo-3. These HRU model structures are simpler than the model structure M-GE (Figure 6), as they are intended to simulate a narrower range of processes.

to sustain significant groundwater processes (section 2.1). The nonlinear reservoir in model structure M-SC is intended to represent lateral flows through schist.

2. **Marl** is represented using model structure M-MA shown in Figure 13b. This structure is essentially the same as M-SC, but excludes the lag function. Model structure M-MA is similar to model structure M-SC, because, similar to schist, marl is considered unlikely to sustain groundwater flows yet is perceived to support considerable lateral flows (section 2.1) [see also *Wrede et al*., 2015]. The lag function is excluded from M-MA because marl tends to respond quickly to rainfall, as observed in the marl-dominated Wollefsbach subcatchment (Figure 3a).
3. **Sandstone** is represented using model structure M-SA shown in Figure 13c. This structure is similar to M-MA, but the SR reservoir is linear rather than nonlinear. The linear reservoir is intended to represent groundwater flow in sandstone formations, such as those in the sandstone-dominated Huewelerbach subcatchment (Figure 3a) [see also *Wrede et al*., 2015]. As groundwater dynamics are often satisfactorily represented using linear reservoirs [e.g., *Tallaksen*, 1995; *Fenicia et al*., 2006], the inclusion of nonlinearity was not considered necessary in this reservoir.
4. **Alluvium** is represented using model structure M-AL shown in Figure 13d. This model structure comprises a linear reservoir with evaporation. Whereas model structures M-SC, M-MA, and M-SA all include an unsaturated root zone reservoir with a threshold (UR), such a reservoir was not considered necessary in model structure because alluvium formations in the Attert catchment are typically saturated and permanently connected to the stream. The simple linear reservoir is deemed sufficient to reproduce the quick response of the alluvial or riparian zone to precipitation. For example, this quick response manifests itself as the first peak of the Weierbach [*Kavetski et al*., 2011; *Wrede et al*., 2015].

The HRU model structures M-SC, M-MA, M-SA, and M-AL represent the building blocks that, when combined based on the geology of the subcatchments, closely resemble the best performing lumped models M03, M07 and M09-M12 identified by *Fenicia et al*. [2014]. For example, model M07, which provided the best fit to the double streamflow peak of the Weierbach (schist and alluvium), is similar to a combination of models M-SC and M-AL. Model M03, which worked well on the Wollefsbach (marl), is essentially M-MA. The "parallel" structures M09-M12 found to be suited to the Huewelerbach (sandstone and marl) represent combinations of models M-SA and M-MA.

The catchment model structure designed in Stage 2 is referred to as **M-Geo-2**. This model is simpler than M-Geo-1. In particular, the total number of states is reduced from 149 to 83 because M-MA, M-SA, and M-AL have fewer reservoirs than M-GE, and the total number of parameters is reduced from 27 to 13 because the exclusion of reservoirs also reduces the number of HRU-specific parameters.

4.2.2. Results

4.2.2.1. Performance in Terms of Maximum Likelihood

Figure 7 shows that M-Geo-2 has a lower calibration performance than M-Geo-1 (Figure 7a), likely as a result of reduced model complexity. On average (all 4 curves), $\overline{\Phi}(\text{M-Geo-2}) = 1.39$ whereas $\overline{\Phi}(\text{M-Geo-1}) = 1.61$.

In time validation (Figure 7b), M-Geo-2 is either slightly worse (Cal-*X*1) or approximately the same (Cal-*X*2). On average (all four curves), $\overline{\Phi}(\text{M-Geo-2}) = 1.12$, whereas $\overline{\Phi}(\text{M-Geo-1}) = 1.21$.

In space validation (Figure 7c), M-Geo-2 offers slightly worse performance than M-Geo-1 when originally calibrated on Group A subcatchments. On average over scenarios Val-B*m* (Cal-A*m*), $\overline{\Phi}(\text{M-Geo-2}) = 1.34$ whereas $\overline{\Phi}(\text{M-Geo-1}) = 1.47$. However, M-Geo-2 is drastically more robust than M-Geo-1 when originally calibrated to Group B subcatchments. On average over scenarios Val-A*m* (Cal-B*m*), $\overline{\Phi}(\text{M-Geo-2}) = 1.30$—which is similar to the values in Val-B*m* (Cal-A*m*), whereas $\overline{\Phi}(\text{M-Geo-1}) = 0.32$—which is clearly lower.

Space-time validation (Figure 7d) leads to similar qualitative results as space validation. When originally calibrated to Group A subcatchments, M-Geo-2 has a slightly lower likelihood than M-Geo-1: $\overline{\Phi}(\text{M-Geo-2}) = 1.23$ whereas $\overline{\Phi}(\text{M-Geo-1}) = 1.33$. In contrast, when originally calibrated to Group B subcatchments, M-Geo-2 is clearly preferable to M-Geo-1: $\overline{\Phi}(\text{M-Geo-2}) = 1.08$ whereas $\overline{\Phi}(\text{M-Geo-1}) = -0.01$. In terms of average probability density per observed data point, M-Geo-2 is ahead of M-Geo-1 by almost a factor of 3.

These results indicate that the use of HRU-specific model structures in M-Geo-2 has improved the space transferability of the M-Geo model, with its temporal transferability remaining stable. The minor reduction in performance incurred in scenarios where the models are originally calibrated to Group A subcatchments is more than compensated by a major improvement in scenarios where the models are originally calibrated to Group B subcatchments. The minor reduction in performance mentioned above may in fact represent the effects of overfitting—although this requires further investigation to be established more conclusively.

4.2.2.2. Performance in Terms of Nash-Sutcliffe Efficiency

Figure 8 shows that, in terms of Nash-Sutcliffe efficiency, M-Geo-2 performs slightly better than M-Geo-1—both in calibration and in space-time validation. In calibration, on average (all subcatchments), $\Upsilon(\text{M-Geo-2}) = 0.76$ (ranging between 0.55 and 0.89) whereas $\Upsilon(\text{M-Geo-1}) = 0.71$ (ranging between 0.53 and 0.86). In space-time validation, on average (all subcatchments), $\Upsilon(\text{M-Geo-2}) = 0.76$ (ranging between 0.65 and 0.87), whereas $\Upsilon(\text{M-Geo-1}) = 0.74$ (ranging between 0.65 and 0.81). In fact, when considering individual subcatchments, $\Upsilon(\text{M-Geo-2}) \geq \Upsilon(\text{M-Geo-1})$ in all subcatchments except Platen, both in calibration and validation.

The relative model performances in terms of the Nash-Sutcliffe efficiency can be compared to the relative performances in terms of the maximum log likelihood values. In terms of the Nash-Sutcliffe efficiency, in the four scenarios represented in Figure 8, on average, $\Upsilon(\text{M-Geo-2}) > \Upsilon(\text{M-Geo-1})$. In terms of the maximum log likelihood values, the same ranking, $\overline{\Phi}(\text{M-Geo-2}) > \overline{\Phi}(\text{M-Geo-1})$, is respected in space-time validation (Figure 7d, Val-B1; Val-A1). However, in calibration, the ranking is reversed and $\overline{\Phi}(\text{M-Geo-1}) > \overline{\Phi}(\text{M-Geo-2})$ (Figure 7a, Cal-A2; Cal-B2). Once again, this could be symptomatic of overfitting by the more complex M-Geo-1 model.

4.2.2.3. Performance in Terms of Hydrological Signatures

Figure 9 shows that, in terms of hydrological signatures, M-Geo-2 produces a slightly wider range of streamflow dynamics than M-Geo-1. Whereas the base flow index I_b varies between 0.68 and 0.85 for M-Geo-1, it varies between 0.63 and 0.86 for M-Geo-2, and whereas the flashiness index I_f varies between 0.05 and 0.15 for M-Geo-1, it varies between 0.06 and 0.20 for M-Geo-2.

This finding provides further insights into the differences between the behavior of M-Geo-1 and M-Geo-2, indicating that M-Geo-2 produces more diverse hydrograph responses than M-Geo-1. For example, the hydrograph responses simulated by M-Geo-1 for the Weierbach, Huewelerbach, and Wollefsbach catchments are more similar to each other than those simulated by M-Geo-2. In fact, on the Wollefsbach catchment, M-Geo-2 produces a lower I_b (0.63 versus 0.68) and higher I_f (0.20 versus 0.15) than M-Geo-1, whereas I_b and I_f in the Weierbach and Huewelerbach catchments stay at similar values.

4.2.2.4. Parameter Identifiability

Figure 10 shows that the distribution of optima in model M-Geo-2 has two well-defined "steps," with about 50% of the searches leading to the estimated global optimum ($\bar{\bar{\Phi}} = 1$). The presence of well-defined steps in the distribution of optima of M-Geo-2 can be contrasted with the "continuous" increase in the distribution of optima in model M-Geo-1. These differences suggest that model M-Geo-2 has fewer local optima than model M-Geo-1—in other words, it has better identifiable parameters.

Figure 14 shows the posterior distributions of model parameter S_{uMax} in M-Geo-2. This parameter represents the threshold value of the UR reservoir (note that S_{uMax} is absent in the model of the alluvial HRU). Figure 14

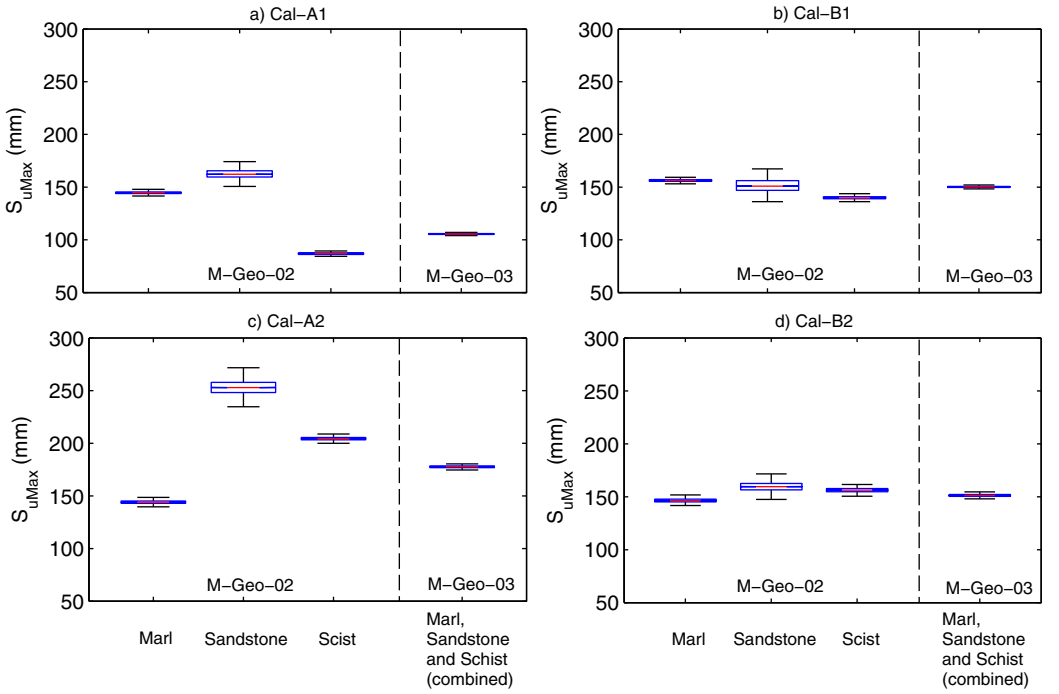

Figure 14. Box plots of the posterior distribution of parameter S_{uMax} in catchment models M-Geo-2 and M-Geo-3. The edges of the box indicate the 25th and 75th percentiles, the whiskers extend to the min/max values. In M-Geo-2, parameter S_{uMax} is calibrated independently in the marl, sandstone, and schist HRU models (note that the alluvium HRU model does not have the S_{uMax} parameter). In M-Geo-3, the same S_{uMax} value is used in the three HRUs models listed above. It can be seen that: (i) S_{uMax} values across the four calibration scenarios are more consistent in model M-Geo-3 than in M-Geo-2, (ii) in Cal-B*m* and M-Geo-2, S_{uMax} takes similar values in the three HRUs, and (iii) in M-Geo-3, using a single S_{uMax} parameter reduces posterior uncertainty.

shows that when calibration is carried out on Group B subcatchments (Figures 14b and 14d), S_{uMax} has similar values in the HRU models representing marl, sandstone, and schist (e.g., in Cal-B2: 146.4 $\pm$ 1.9, 159.5 $\pm$ 4.1, and 156.2 $\pm$ 2.1 mm, respectively). Moreover, the calibrated values and uncertainty ranges are generally consistent in the two calibration periods, B1 and B2.

In contrast, when calibrated on Group A subcatchments (Figures 14a and 14c), S_{uMax} has different values in the 3 HRU models and the values are less consistent across the two calibration periods, A1 and A2. For example, in scenario Cal-A1, S_{uMax} of Marl is larger than S_{uMax} of Schist, and the opposite occurs in scenario Cal-A2. These considerations are important for Stage 3 of model development, which is described next.

4.2.2.5. Visual Hydrograph Comparison

Figures 11g and 11h show that model M-Geo-2, although simpler than M-Geo-1, is still able to capture the difference in streamflow dynamics of the Weierbach, Wollefsbach and Huewelerbach catchments. In the Huewelerbach catchment, M-Geo-2 shows much better performance than M-Geo-1, particularly in the ability to match the peaks. In the Weierbach catchment, M-Geo-2 performs slightly better than M-Geo-2, as it produces slightly higher peaks. In the Wollefsbach catchment, the performances of M-Geo-1 and M-Geo-2 are very similar.

Figure 12d examines in greater detail the performance of M-Geo-2 in the Huewelerbach catchment. The 95% prediction limits show that M-Geo-2 provides a better match to the observations than M-Geo-1, particularly in the peaks.

4.3. Stage 3: Spatial Relations Between Model Parameters Across HRUs (M-Geo-3)

4.3.1. Hydrological Hypotheses

In Stage 2, catchment model M-Geo-2 was found to have better space transferability than catchment model M-Geo-1, despite using simpler HRU model structures. In Stage 3, we take M-Geo-2 as the basis of further improvements and investigate opportunities for further reduction of its complexity, by using the same model parameter values across all HRUs.

A candidate choice of model parameter to link across HRUs is S_{uMax}. This choice is motivated by the following insights from Figure 14: (1) the similarity of S_{uMax} values across multiple HRUs in Cal-B*m*, and (2) the inconsistency of parameter estimates in Cal-A*m* (ideally, parameter estimates would not be strongly dependent on the calibration period). We therefore investigate the treatment of S_{uMax} as a "global" parameter. The resulting catchment model structure is referred to as model **M-Geo-3**.

Compared to M-Geo-2, M-Geo-3 reduces model complexity as follows:

1. Total number of model parameters reduced by 2 from 13 to 11 (Table 2), as three distinct S_{uMax} parameters in the HRU model structures for marl, sandstone, and schist are replaced by a single common S_{uMax} parameter;
2. Total number of model states reduced by 10 from 83 to 73 (Table 2). This simplification occurs because S_{uMax} is the only parameter of the UR reservoir, and giving it the same value across the three HRU model structures results in identical UR equations and parameters. In turn, all UR reservoirs within a single subcatchment receive the same forcing data and hence will have identical state values. Hence, similarly to the snow component, only a single UR state value per subcatchment needs to be tracked.

A complete list of all model parameters and states of M-Geo-3 is given in Table 2.

4.3.2. Results

4.3.2.1. Performance in Terms of Maximum Likelihood

Figure 7a shows that making S_{uMax} a "global" parameter—despite reducing the number of parameters from 13 to 11 and the number of states from 83 to 73—makes virtually no difference on calibration performance. On average (all four curves), $\overline{\Phi}(\text{M-Geo-3}) = 1.38$ whereas $\overline{\Phi}(\text{M-Geo-2}) = 1.39$.

In validation, M-Geo-3 offers essentially the same performance as M-Geo-2. On average (all 4 curves), in time validation $\overline{\Phi}(\text{M-Geo-3}) = 1.15$ versus $\overline{\Phi}(\text{M-Geo-2}) = 1.12$; in space validation $\overline{\Phi}(\text{M-Geo-3}) = 1.31$ versus $\overline{\Phi}(\text{M-Geo-2}) = 1.32$; and in space-time validation $\overline{\Phi}(\text{M-Geo-3}) = 1.18$ versus $\overline{\Phi}(\text{M-Geo-2}) = 1.16$.

4.3.2.2. Performance in Terms of Nash-Sutcliffe Efficiency

Figure 8 shows little difference in performance between M-Geo-2 and M-Geo-3. In calibration, on average (all subcatchments), $\Upsilon(\text{M-Geo-3}) = 0.77$ whereas $\Upsilon(\text{M-Geo-2}) = 0.76$. In space-time validation, on average (all subcatchments), $\Upsilon(\text{M-Geo-3}) = \Upsilon(\text{M-Geo-2}) = 0.76$. This result is consistent with the similarity of the likelihood values listed earlier in this section, and indicates that the model simplifications carried out in M-Geo-3 have not reduced the streamflow predictive ability of the M-Geo-2 model used as a basis for M-Geo-3.

4.3.2.3. Performance in Terms of Hydrological Signatures

Figure 9 shows that, in terms of hydrological signatures, the performance of M-Geo-3 is virtually identical to the performance of M-Geo-2, for all subcatchments. This result indicates that the model simplifications carried out in M-Geo-3 have not reduced the ability of the model to reproduce spatial streamflow patterns.

4.3.2.4. Parameter Identifiability

Figure 10 shows the effect of reducing the dimension of the parameter space on the convergence of the parameter optimization. M-Geo-3 appears characterized by a well-identifiable global optimum, which attracted about 65% of the local quasi-Newton searches. In comparison, M-Geo-2 had a slightly less defined global optimum, which attracted about 50% of the local searches.

Figure 14 compares the distribution of parameter S_{uMax} in M-Geo-3 versus M-Geo-2. The following results can be observed:

1. In all cases, the single S_{uMax} parameter in M-Geo-3 has lower uncertainty than any of the three HRU-specific S_{uMax} parameters of M-Geo-2;
2. In Cal-B*m* (Figures 14b and 14d), the single S_{uMax} parameter in M-Geo-3 has similar values (around 150 mm) to the three S_{uMax} parameters of M-Geo-2;
3. In Cal-A*m*, the single S_{uMax} parameter in M-Geo-3 varies between 110 mm (Figure 14c) and 180 mm (Figure 14a). These values are closer to each other than the estimates of S_{uMax} in M-Geo-2, where, across the two calibrated periods, S_{uMax} varied between 150 and 270 mm for sandstone, and between 80 and 200 mm for schist. Moreover, in M-Geo-3, the average S_{uMax} value in the calibration periods B1 and B2 is about 135 mm, which is close to the average values of about 150 mm obtained when calibrating on the catchments of Group B.

These findings lend weight to the hypothesis underpinning Stage 3, namely that S_{uMax} values do not vary significantly between HRUs. In turns, this finding provides direct support for constraining parameter S_{uMax} to have the same value in the marl, sandstone, and schist HRUs.

4.3.2.5. Visual Hydrograph Comparison

Figures 11i and 11j show that M-Geo-3 produces nearly identical hydrograph to M-Geo-2 on the Huewelerbach, Weierbach, and Wollefsbach catchments. Hence, the reduction in model complexity represented by M-Geo-3 does not penalize model performance in terms of streamflow predictions.

Figure 12e shows the 95% prediction limits of M-Geo-3 on the Huewelerbach catchments. It can be noted that the prediction limits of M-Geo-3 are virtually indistinguishable from those of M-Geo-2.

5. Discussion

5.1. Selecting a Landscape Classification Approach (Stage 1)

Landscape discretization is a key decision in distributed hydrological model applications [*Beven*, 1996; *Uhlenbrook et al*., 2007]. In general, spatial discretization should reflect the landscape properties that dominate hydrological response. Previous studies have shown that a wide range of landscape properties can dominate hydrological response in different applications. Topography [e.g., *Hjerdt et al*., 2004], geology [e.g., *Tague and Grant*, 2004], vegetation [e.g., *Gao et al*., 2014b], soils [e.g., *Soulsby et al*., 2006], earthworm burrows [e.g., *Wienhofer and Zehe*, 2014], etc., have all been suggested as hydrological controls in the cited studies. Although the relative importance of these controls may vary depending on the type of environment and spatiotemporal scales, most model applications specify the dominant landscape controls—and the resulting landscape classifications—a priori, either guided by previous work, or purely as an assumption.

Topography is assumed to be a dominant control of hydrological processes in many models, including TopModel [*Beven and Kirkby*, 1979], the Catchment Connectivity Model [*Smith et al*., 2013] FLEX-topo [*Savenije*, 2010], and others. However, as argued by *Grayson and Western* [2001], the combination of features that allows topography to be a dominant control on hydrologic processes occurs less often than believed. For example, *Devito et al*. [2005] contend that topography, on its own, would provide misleading predictions of the flow pathways of the Mink Lake catchment (Alberta, Canada), which are largely controlled by geology. Based on previous fieldwork, geology is also believed to be an important control in the Attert catchment in Luxembourg [e.g., *Juilleret et al*., 2012; *Fenicia et al*., 2014; *Wrede et al*., 2015], which is the focus of the present study.

In this study, we contrasted a geology-based landscape classification M-Geo-1 that reflects the perception that geology controls streamflow response in the Attert catchment, a topography-based classification M-Top that corresponds to the assumptions made by many conceptual models that topography controls catchment response, and a uniform classification M-Uni that serves as the null hypothesis that neither geology or topography control streamflow response.

The comparison of model hypotheses M-Uni, M-Top, and M-Geo-1 vividly illustrates the impact of the landscape classification approach on model predictions. M-Geo-1 provided a more robust predictive performance than M-Uni or M-Top, in the sense of providing a comparatively high performance in calibration and generally maintaining performance in spatiotemporal validation (see section 4.1.2). This finding lends more weight to the hypothesis that, in the Attert catchment, geology exerts more control than other landscape features on the streamflow generation processes.

An exception in the performance of M-Geo-1 can be seen in Figures 7c and 7d, where M-Geo-1 has lower values of the likelihood in space and space-time validation than M-Uni. Stage 2 analysis, where the simpler M-Geo-2 is seen to provide an improvement over M-Geo-1 in terms of spatial validation, suggests that lack of parsimony/overparameterization is the likely reason for the poor performance of M-Geo-1 in Figures 7c and 7d.

M-Geo-1 is the model that best reproduces the observed ranking of catchments with respect to the hydrological signatures (Figure 9). In terms of the model abilities to match the observed signature values, it is apparent that the flashiness index I_f is generally underestimated by all models. Considering that I_f is defined as the average difference between streamflow values at successive time steps, the underestimation of I_f indicates that on average, simulated streamflows are "smoother" than observed streamflows. This difference in "smoothness" could be (speculatively) attributed to fine time scale processes that are not captured by the models, and/or by sampling and measurement errors in the calibration data.

The Huewelerbach subcatchment represents an important test for model performance. Both M-Uni and M-Top effectively "fail" in this catchment, given their low performance in terms of Nash-Sutcliffe efficiency (Figure 8). The Huewelerbach is dominated by sandstone, and its response is characterized by a pronounced and relatively stable groundwater component (e.g., the base flow index of this subcatchment is 0.9, see Figure 9). A geology-based classification is clearly needed to represent the spatial variability of streamflow dynamics arising from differences in the hydraulic properties of sandstone (Huewelerbach) versus schist and marls (other catchments) [see also *Wrede et al*., 2015]. In terms of topography, the Huewelerbach is dominated by hillslope. In contrast to hillslopes in the schist or marl formations, which promote lateral flow contributing to quickflow, the hillslopes in the sandstone formations generate little lateral flow. Instead, sandstone hillslopes tend to promote vertical flow, which eventually contributes to base flow. The topography-based classification, which assumes all hillslopes exhibit the same behavior, cannot capture these differences in hillslope behavior. Although the Huewelerbach subcatchment is where the contribution of sandstone to base flow is most visible, similar behavior is also apparent elsewhere. For example, the Platen subcatchment has a relatively high base flow index (about 0.8) (Figure 9), and it comprises large sandstone areas (Figure 1 and Table 1).

Previous work in the Attert area has already highlighted the role of geology in controlling the hydrological behavior in this region [e.g., *Juilleret et al*., 2012; *Fenicia et al*., 2014]. It is hence not entirely surprising that our modeling results indicate geology to play a more important role than topography in controlling the Attert streamflow dynamics. However, the earlier work focused mainly on the headwater subcatchments, which, in addition to differing in geology, differed in many other aspects, including topography, land use, area, etc. Moreover, the earlier studies did not investigate the dominant controls in larger and more heterogeneous subcatchments.

Stage 1 contributes to the existing knowledge of the Attert catchment by showing that: (1) geological information can be effectively regionalized by means of HRUs, and (2) geology is the main control on streamflow generation processes – not only in smaller subcatchments that are relatively uniform in terms of landscape properties, but also in larger subcatchments characterized by mixed geologies. In terms of broader impact, Stage 1 shows that (1) the selection of the landscape classification approach has a strong impact on model results, and it therefore deserves careful analysis, and (2) the assumption considered by many models that topography controls hydrologic response does not hold in general. Future model development work will focus on enhancing the geology discretization with a topography discretization [e.g., *Winter*, 2001], taking advantage of the insights reported in the application of FLEX-Topo [*Gao et al*., 2014a; *Gharari et al*., 2014].

Finally, in the present study, the basic landscape units (fields) are assumed to behave independently, and their contributions are linearly superimposed on each other. This approximation substantially simplifies model implementation and reduces computational costs. It appears to hold in the present case, given the relatively good predictive performance of models such as M-Geo-3. However, the general validity of this approach is likely to depend on the scale of application and may be location specific. The modeling of smaller spatial units is likely to require approaches that explicitly represent their spatial connectivity. For example, *Meerveld and Weiler* [2008] found that the representation of the trenchflow response in the Panola hillslope required the explicit representation of bedrock topography and variable soil depths.

5.2. Defining Model Structures for Different HRUs (Stage 2)

Stage 1 used a "generic" HRU model structure (M-GE). As shown in previous work, this structure can reproduce well the behavior of a wide range of catchments [*van Esse et al*., 2013] and is similar to well-known models such as HBV. The use of a "generic" model structure in all landscape elements is common in many distributed models, e.g., SWAT and MHM. However, as noted by *Kampf and Burges* [2007], it may be unnecessary to represent processes that are known (with a sufficient degree of confidence!) to be absent in certain environments—e.g., subsurface flow in areas where all the streamflow is known to originate from surface runoff, or Hortonian runoff where most of the streamflow is fed by groundwater.

Stage 2 illustrates the benefit of "distributing" the model structure, i.e., tailoring the model structure to individual HRUs. We stress that the tailoring was limited strictly to reducing model complexity, as M-Geo-2 is a special case of M-Geo-1, with some model components removed but no new components added. Stage 2 resulted in (1) a significant simplification of the model structures, and (2) more robust predictive performance. The reduction of model complexity is a result of ruling out processes that are not dominant in

individual geology types. For example, "fast" processes, such as surface runoff or lateral flow, are "ruled out" in sandstone, in the sense of being assumed to play a secondary role. For analogous reasons, "slow" processes, such as groundwater flow, are ruled out in marl and schist.

The improvement of M-Geo-2 over M-Geo-1 is clearly seen in terms of Nash-Sutcliffe efficiency (Figure 8), where calibration and space-time validation performance is equaled or improved in 9 of the 10 subcatchments. In terms of the max likelihood values, the improvement of M-Geo-2 over M-Geo-1 is particularly evident in space and space-time validation of the Group B subcatchments (Figures 7c and 7d, Cal-B*m*). Group B contains subcatchments with mixed geology, whereas Group A contains the three headwaters, which have relatively homogeneous geologies (section 3.2.4). It is therefore easier to infer the behavior of the individual geologies when calibrating on Group A then on Group B. The improvements of M-Geo-2 over M-Geo-1 are also evident in its better ability to capture the spatial trends in the observed hydrological signatures (Figure 9).

Given that M-Geo-2 is a special case of M-Geo-1, the improved validation performance of M-Geo-2 appears due purely to improved parsimony of M-Geo-2: M-Geo-2 is less prone than M-Geo-1 to overfit the calibration data. The tendency of complex hydrological models to overfit calibration data and provide poorer predictions in validation periods is well known in hydrology [for example, see *Schoups et al*., 2008].

The robustness of streamflow predictions generated using M-Geo-2 suggests that large-scale hydrological response of the different geologies could be characterized by remarkably simple HRU model structures, with only one to four parameters each (Figure 13). All three dominant geologies (marl sandstone and schist) have in common a threshold-like reservoir, which represents a "saturation excess" partitioning of incoming flow between water retained in the reservoir (which eventually evaporates) and streamflow generation. This threshold reservoir is linked in series with model elements specific to the individual geologies: a lag function (schist), a linear reservoir (sandstone), or nonlinear reservoir (marl and schist). The alluvium appears well represented by an even simpler HRU model structure, with just a single reservoir.

It is of interest to compare the HRU model structures used to represent the three dominant geologies in this study to the single storage model inferred by *Kirchner* [2009] in the Plynlimon catchment in mid-Wales. As seen in the model schematics in Figure 13, the HRU model structures M-SC, M-MA, and M-SA are more complex than the single nonlinear reservoir model used by *Kirchner* [2009]. According to our analysis, the proposed HRU model structures are likely to represent a lower bound on model complexity for the subcatchments of the Attert catchment, as their structures are broadly related to the characteristics of the observed streamflow time series and further reduction in model complexity resulted in a clear deterioration in predictive capabilities. As shown in *Fenicia et al*. [2014], the threshold reservoir (UR) was necessary to reproduce the seasonality in streamflow response. The lag function in the schist HRU model structure M-SC was needed to reproduce the large delay observed in the schist formations. As shown in our previous work [*Fenicia et al*., 2014], the single reservoir model performed relatively poorly on the three headwaters. Only the alluvium formation, which represents a relatively minor portion of the Attert subcatchments, is well represented by a single storage unit. Noting that the alluvium formations are usually in a wet state, at least in the region considered, our hypothesis is that the "class" of catchments that are well represented by a single storage model, such as the Plynlimon catchment [*Kirchner*, 2009] and the Maimai catchment [*Kavetski and Fenicia*, 2011], are generally characterized by relatively wet conditions. Drier climates would eventually reveal thresholds in catchment response, which may require more complex model representations.

In relation to our previous work on the three headwater subcatchments [*Fenicia et al*., 2014], the HRU model structures associated with the different geologies could be interpreted as parts of the lumped models that were found to best represent these headwaters (see section 4.2.1 in this work and *Fenicia et al*. [2014, "Discussion" section]). The fact that the HRUs are defined according to observable characteristics (geology) strengthens the physical motivation for the model structures developed in this work.

5.3. Linking Model Parameters Across HRUs (Stage 3)

Model hypotheses M-Geo-2 and M-Geo-3 differ in complexity (number of parameters and number of states, see Table 2), yet demonstrate almost equal predictive performance. The likelihood values, Nash-Sutcliffe efficiencies, and signature values are very similar (Figures 7–9). The Nash-Sutcliffe efficiency value for M-Geo-3 is relatively high (on average 0.76 in space time validation), and this model closely matches the

spatial patterns of observed signatures (Figure 9). Moreover, M-Geo-3 has better identifiable model parameters than M-Geo-2 (Figures 10 and 14), which can be attributed to the reduction in model complexity.

The sole difference between M-Geo-2 and M-Geo-3 is the use of a single value of parameter S_{uMax} across the main geology types (marl, sandstone, and schist). This apparently minor model difference is in fact quite important, when considering the similar predictive performance of M-Geo-2 and M-Geo-3 and similar values of S_{uMax} in these two models. This similarity implies that the HRU model structures that represent the 3 dominant geologies (marl, sandstone, and schist) use the same UR model component, with identical reservoir structure and parameter values. Therefore, UR cannot be responsible for the variability in *simulated* responses across multiple geologies (e.g., Figure 11i). Instead, these response differences must be caused by differences in other model components in the geological HRU structures, namely the lag function (LF) and the reservoir (SR), via differences in the associated constitutive functions and parameter values.

Based on the hydrologic processes that the model components are intended to represent, it is possible to distinguish processes that are controlled primarily by geology from processes that are not. In particular, UR is the only reservoir from which evaporation takes place, and it is emptied solely through evaporation. Therefore, UR can be seen to partition the incoming flow between storage available for evaporation versus streamflow generation. The finding that, following calibration, S_{uMax} has similar values for different geologies suggests that, at least according to models calibrated in the Attert catchment, evaporation does not depend strongly on geology. We note that this modeling finding requires independent experimental analysis and verification.

Factors other than geology may control the partitioning of water between evaporation and flow generation. For example, vegetation and climate effects were not investigated in this study. The omission of vegetation information is partly justified because land use is fairly uniform over the relatively large HRUs considered in this work. The climate is also relatively uniform over the study region, due to the small area of the latter. The regional or global scale is more appropriate for the analysis of these effects.

At the global scale, it is well known that vegetation behavior is largely controlled by climate [e.g., *Budyko*, 1974; *Milly*, 1994]. Recently, *Gao et al*. [2014b] related the values of S_{uMax}—interpreted as the root zone storage capacity—to climatic and ecosystem characteristics, over a large number of catchments on a global scale. *Gao et al*. [2014b] reported that climate and land cover are largely responsible for the variability in root zone storage capacity. In the present case study, considering that geology may condition soil depth and consequently rooting depth, it could have been hypothesized that geology also conditions the accessibility of plants to water. Moreover, it may have been reasonable to hypothesize that the markedly different dynamics of the Attert catchment subcatchments are attributable to differences in the vegetation and its transpiration dynamics. However, the results of this study do not appear to support this hypothesis, as the root zone storage capacity of vegetation was estimated to have similar values in different geologies.

In terms of complexity, the catchment model structure M-Geo-3 obtained using the iterative process was surprisingly simple. It is well known that, when calibrating hydrological models on individual hydrographs, only a "handful" of parameters are often identifiable [e.g., *Jakeman and Hornberger*, 1993]. Less clear is what would be an appropriate level of model complexity for a distributed hydrological model. In our previous work, we found that five to eight parameters are sufficient to reliably capture the hydrographs of some of the case study catchments [*Fenicia et al*., 2014]. If these numbers scaled up linearly, a tenfold amount of model parameters would be necessary to simulate 10 hydrographs using a distributed model. In this study, where we attempted to simulate 10 seemingly different hydrographs, two "handfuls" (11 parameters) were found to be sufficient. This result may give a broad indication of the complexity of a distributed model in similar conditions of application, yet care must be taken because this result is derived solely based on case study location data. The likely reason for the relatively low complexity of the resulting model M-Geo-3 is that the differences in the observed hydrographs appear to be well explained by the permutation of a few HRUs, which in turn appear to be well represented by parsimonious conceptual model structures with several shared parameters.

In terms of generalizing our findings to other regions, we stress that our hydrological results are largely empirical. Our specific conclusions are therefore currently limited to streamflow generation, and to the Attert region considered in this work. It is possible that alternative types of data may suggest controls other than geology on the streamflow responses of the Attert catchment. Furthermore, it is certainly possible—indeed, likely—that topography may play a stronger role in controlling the streamflow responses of larger catchments with more heterogeneous topography than the headwater subcatchments of the Attert catchment.

6. Conclusions

This study presented an iterative methodology for developing distributed hydrological models that parsimoniously capture the spatial variability of catchment properties and streamflow behavior. The methodology considers, sequentially, the following three sets of model development decisions: (1) landscape classification into hydrological response units (HRUs), (2) selection of model structures to be associated with the individual HRUs, and (3) linking of model parameters across the multiple HRUs. Each set of model decisions is investigated using multiple alternative hypotheses, and appraised using multiple diagnostics.

The following conclusions were obtained based on an application of the proposed model development methodology to the Attert catchment in Luxembourg:

1. The observed differences in streamflow response across the Attert catchment appear related to the geological variability of the area, as suggested by the robust performance of models based on geological HRUs. This result contrasts with many conceptual hydrological modeling applications, where it is more commonly assumed that topography rather than geology controls hydrologic response. A landscape classification based on topography, although having stronger explanatory power than a uniform landscape classification, could not fully explain the observed spatial variation in streamflow dynamics. We stress that this finding was established through the analysis of model transferability in space and time, and would not have been detected from the analysis of calibration performance alone;
2. The hydrological behavior of the different geological units can be represented by distinct and remarkably simple conceptual model structures, with one to four parameters per structure. These model structures could be tailored to, and interpreted in terms of, the dominant processes perceived to characterize distinct HRUs. In the study catchment, we found that 11 parameters were sufficient to reproduce the observed differences in streamflow response observed at 10 stream gauging stations. This is arguably a surprisingly low number of parameters for a distributed hydrological model, and is comparable to the typical range of complexity of lumped model structures. The relative model simplicity appears due to observed streamflow diversity of the Attert catchment being explained by the contribution of streamflow from just a few geology-based HRUs, which were used in the model discretization.
3. In the Attert catchment, the partitioning of precipitation into evaporation and streamflow appears independent from geology, and is well represented by the inclusion of a saturation excess threshold component into the HRU-specific model structures. If corroborated by independent experimental work, these modeling results can advance our understanding of dominant runoff generation processes and their conceptualization, in particular in the Attert catchment in Luxembourg.
4. When applying a model jointly to multiple subcatchments, calibration to subcatchments with relatively homogeneous internal HRU compositions leads to better validation performance than calibration to subcatchments that comprise multiple HRUs. This improvement is likely due to "homogeneous" subcatchments offering more easily accessible information about the behavior of individual HRUs. We stress that both groups of subcatchments had multiple geologies present, and it was the *distribution* of geologies in the *individual* catchments that led to differences in the robustness of the calibrated catchment model.

More broadly, this study highlights important benefits of approaching hydrological model development as a systematic testing of multiple working hypotheses: (1) improved predictive performance of the models, (2) improved model parsimony, and (3) quantitative and qualitative modeling insights that complement experimental process understanding. The application of the method of multiple working hypotheses was facilitated by the use of the flexible framework SUPERFLEX, which provided a detailed "fine grain" control over each individual modeling decision.

We stress that our particular hydrological findings, in particular the indication that geology trumps topography as a dominant control on streamflow dynamics in the Attert catchment, should be treated as specific to the case study area. Further modeling and experimental research is needed to investigate the extent to which similar conclusions hold in catchments with different climatic and physical properties.

On the other hand, the modeling and hypothesis testing methodology introduced in this study is general. It can be used to advance our broader understanding and prediction of hydrological behavior, including the landscape characteristics that control hydrologic response, the dominant processes associated with different landscape types, and the spatial relations of catchment processes.

Appendix A: Hydrological Model Details

A1. Model Equations

The overall model organization and the HRU model structures are shown schematically in Figures 4, 5, and 13. The parameters of the HRU model structures are listed in Table A1. The water balance equations are listed in Table A2. The constitutive equations are listed in Table A3, and the functions used in the constitutive relations are listed in Table A4.

All catchment model structures incorporate a degree-day snow reservoir (section 3.1.2). The snow module receives as inputs the measured total precipitation and air temperature, and outputs the estimated rainfall plus snowmelt. The snow module outputs are in turn fed as inputs into the HRU model structures:

1. HRU model structures M-GE, M-SC, M-MA, and M-SA: rainfall and snowmelt fed into UR;
2. HRU model structure M-AL: rainfall and snowmelt fed into SR.

As the precipitation input is distributed per subcatchment, the snow reservoirs associated with the HRU model structures within the same subcatchment receive the same input. Hence, considering that all snow model parameters are global, any given subcatchment can be characterized by a single snow storage state. In contrast, snow reservoirs in different subcatchments receive different inputs and hence, despite having the same parameter values, are characterized by individual snow storage states.

A2. Initial Conditions

The model reservoirs are initialized to the same values in all model hypotheses:

1. Snow storage and fast reservoir (FR) storage: assumed to be initially empty;
2. Unsaturated root zone (UR) storage: initialized to 20% of its capacity;
3. Slow reservoir (SR) storage: as this reservoir can have a long characteristic time scale, it is initialized to its long-term steady state value, calculated using a constant forcing equal to the average of the observed forcing data. Note that the steady state storage value depends on the model parameter values.

In order to reduce the effect of the initial conditions, a 1 year warm-up period is used. All simulations (calibration and validation) start in September, which is the beginning of the hydrological year.

Table A1. Components and Parameters of HRU Model Structures[a]

HRU Model Structure			Components				Parameters							
Name	N_s	N_θ	UR	FR	SR	LF	C_e (0.1–3.0)	S_{uMax} (mm) (0.1–500)	β (10^{-3}–10)	T_l (day) (0–10)	D (0–1)	K_f (1/day) (10^{-6}–10^{-1})	K_s ($mm^{1-\alpha}$/day) (10^{-6}–10^{-1})	α (0.1–5.0)
M-GE	4	7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-
M-SC	3	5	✓	-	✓	✓	✓	✓	-	✓	-	-	✓	✓
M-MA	2	4	✓	-	✓	-	✓	✓	-	-	-	-	✓	✓
M-SA	2	3	✓	-	✓	-	✓	✓	-	-	-	-	✓	-
M-AL	1	2	-	-	✓	-	✓	-	-	-	-	-	✓	-

[a] N_θ is the number of parameters and N_s is the number of states. UR, FR, and SR denote different model reservoirs, and LF denotes the lag function (✓ and "-" indicate, respectively, presence or absence).

Table A2. Water Balance Equations of the Models Used in the Experiments[a]

Water Balance Equations	M-GE	M-SC	M-MA	M-SA	M-AL
$\frac{dS_u}{dt} = P_u - Q_u - E_u$	✓	✓	✓	✓	-
$\frac{dS_f}{dt} = P_f - Q_f$	✓	-	-	-	-
$\frac{dS_s}{dt} = P_s - Q_s$	✓	✓	✓	✓	-
$\frac{dS_s}{dt} = P_s - Q_s - E_s$	-	-	-	-	✓
$Q_u = P_l + P_s$	✓	-	-	-	-
$Q_t = Q_f + Q_s$	✓	-	-	-	-

[a] ✓ and "-" indicate, respectively, presence or absence. S, P, Q, and E refer to storage, inflow, discharge and evaporation, respectively. The subscripts u, f, s, and l refer to UR, FR, SR, and LF, respectively.

Table A3. Constitutive Functions of the Models Used in the Experiments[a]

Constitutive Functions	M-GE	M-SC	M-MA	M-SA	M-AL
$\bar{S}_u = S_u / S_{uMax}$	✓	✓	✓	✓	-
$Q_u = P_u f_p(\bar{S}_u \mid \beta)$	✓	✓	✓	✓	-
$E_u = C_e E_p f_m(\bar{S}_u \mid m_1)$	✓	✓	✓	✓	-
$E_s = C_e E_p f_e(S_s \mid m_2)$	-	-	-	-	✓
$Q_l = (P_l * h_l)(t)$	✓	✓	-	-	-
$h_l = \begin{cases} 2t/T_l^2, & t < T_l \\ 0, & t > T_l \end{cases}$	✓	✓	-	-	-
$P_s = DQ_u$	✓	-	-	-	-
$Q_f = k_f S_f$	✓	-	-	-	-
$Q_s = k_s S_s$	✓	-	-	✓	✓
$Q_s = k_s f_p(S_s \mid \alpha)$	-	✓	✓	-	-

[a] ✓ and "-" indicate, respectively, presence or absence. The functions *f* are defined in Table A4. Parameters, fluxes, states are defined in Tables A1 and A2. The operator * in the equation for Q_l denotes the convolution operator (lag function smoothed using the method in *Kavetski and Kuczera* [2007]). The parameters m_1 and m_2 are threshold smoothing parameters [*Kavetski and Kuczera*, 2007] and are fixed to a value of 10^{-2} mm. In all models except M-GE, parameter β is treated as a threshold smoothing parameter and is fixed to a value of 10.

Table A4. Constitutive Functions

Functions	Name
$f_p(x \mid \theta) = x^\theta$	Power function
$f_m(x \mid \theta) = \frac{x(1+\theta)}{x+\theta}$	Monod-type kinetics, adjusted so that $f_m(1 \mid \theta) = 1$
$f_e(x \mid \theta) = 1 - \exp(-x/\theta)$	Tessier function; note that $f_e(x \mid \theta) \to 1$ as $x \to \infty$

A3. Numerical Solution

The model differential equations are solved using the implicit Euler approximation with fixed daily time steps. Our previous research indicated that, in addition to numerical robustness, this approach offers a good practical compromise between numerical accuracy and computational efficiency [*Kavetski and Clark*, 2010; *Kavetski et al.*, 2011]. An appealing practical feature of the fixed-step implicit Euler approximation is that, provided the model's constitutive functions are smooth, the model state variables and outputs will also be smooth with respect to the model parameters. This behavior is computationally highly beneficial—it avoids the tyrannical difficulties with optimization and general analysis that bedevil models with discontinuous and/or geometrically irregular response surfaces [*Kavetski et al.*, 2006a, 2006b; *Kavetski and Clark*, 2010].

Appendix B: Model Calibration

B1. Bayesian WLS Inference of Model Parameters

The hydrological model parameters are calibrated to observed data using a Bayesian inference approach,

$$p(\boldsymbol{\theta}, \boldsymbol{\Xi} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Q}}, M) = p(\tilde{\boldsymbol{Q}} \mid \tilde{\boldsymbol{X}}, \boldsymbol{\theta}, \boldsymbol{\Xi}, M) p(\boldsymbol{\theta}, \boldsymbol{\Xi} \mid M) \tag{B1}$$

where $p(\boldsymbol{\theta}, \boldsymbol{\Xi} \mid \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Q}}, M)$ is the posterior of the parameters θ of hydrological model M and the parameters $\boldsymbol{\Xi}$ of the residual error model, $p(\boldsymbol{\theta}, \boldsymbol{\Xi} \mid M)$ is the prior, and $p(\tilde{\boldsymbol{Q}} \mid \tilde{\boldsymbol{X}}, \boldsymbol{\theta}, \boldsymbol{\Xi}, M)$ is the likelihood.

The posterior distribution is conditioned on the observed streamflow data $\tilde{\boldsymbol{Q}}$ and on the model forcing data $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{P}}, \tilde{\boldsymbol{E}}, \tilde{\boldsymbol{T}})$, where $\tilde{\boldsymbol{P}}$ is the observed precipitation, $\tilde{\boldsymbol{E}}$ is the estimated potential evaporation and $\tilde{\boldsymbol{T}}$ is the measured air temperature. In the absence of additional knowledge, we used noninformative "flat" priors for θ and $\boldsymbol{\Xi}$, $p(\boldsymbol{\theta}, \boldsymbol{\Xi} \mid M) \propto const$, over the range of feasible parameter values [*Box and Tiao*, 1992].

The likelihood function used in this study is based on a Weighted Least Squares (WLS) error model [*Evin et al.*, 2014]. We assume that the residuals errors $\varepsilon_{c,t}$ in the model predictions $\hat{Q}_{c,t}$ of observed streamflow $\tilde{Q}_{c,t}$ in subcatchment c and time step t are statistically independent and can be approximated by a zero-mean Gaussian distribution,

$$\varepsilon_{c,t} \equiv \tilde{Q}_{c,t} - \hat{Q}_{c,t} \sim N\left(0, \sigma^2_{\varepsilon(c,t)}\right), \quad t = 1, 2, \ldots, N_{c,T}, \quad c = 1, 2, \ldots, N_C \tag{B2}$$

where $N_{c,T}$ is the length of streamflow time series (excluding missing data) in subcatchment c, and N_C is the total number of subcatchments.

The standard deviation $\sigma_{\varepsilon(c,t)}$ of residual errors is assumed to depend linearly on the simulated streamflow,

$$\sigma_{\varepsilon(c,t)} = a + b\hat{Q}_{c,t} \tag{B3}$$

where $\boldsymbol{\Xi} = (a, b)$ are error model parameters, assumed to be the same across all subcatchments.

The assumptions in equations (B2) and (B3) lead to the following likelihood function:

$$p(\tilde{\boldsymbol{Q}}|\tilde{\boldsymbol{X}}, \boldsymbol{\theta}, \boldsymbol{\Xi}, M) = \prod_{c=1}^{N_C} \prod_{t=1}^{N_{c,T}} f_N\left(\varepsilon_{c,t}[\boldsymbol{\theta}; \tilde{\boldsymbol{X}}, M] | 0, \sigma^2_{\varepsilon(c,t)}[\boldsymbol{\theta}, \tilde{\boldsymbol{X}}, M, \boldsymbol{\Xi}]\right) \tag{B4}$$

where $f_N(x|m, s^2)$ is the Gaussian probability density function with mean m and variance s^2. Note that equation (B4) explicitly lists all dependencies needed to evaluate $\varepsilon_{c,t}$ and $\sigma^2_{\varepsilon(c,t)}$.

B2. Multistart Quasi-Newton Parameter Optimization

Parameter calibration consisted of maximizing the posterior distribution in equation (B1) jointly with respect to the hydrological model parameters $\boldsymbol{\theta}$ and the error model parameters $\boldsymbol{\Xi}$. The optimization was carried out using a multistart quasi-Newton method [e.g., *Kavetski and Clark*, 2010]. Here we initiated 100 independent quasi-Newton searches from random initial seeds in the feasible parameter space. The quasi-Newton algorithm used trust-region safeguards to stabilize convergence to the nearest optimum.

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

B3. Representative CPU Runtimes

The CPU runtime of a model application depends on the number of calibrated parameters and the number of model states. In general, the number of parameters affects the number of model runs required for model optimization and MCMC analysis, whereas the number of model states affects the runtime of an individual model evaluation. However, these properties are not independent—a model with more states will generally have more calibration parameters (unless more parameter constraints are employed).

In this study, the simplest catchment model (M-Uni) required about 100,000 iterations and 12 h of runtime on a single standard desktop CPU (i.e., without using any multicore parallel computation within a single model calibration). The most complex catchment model (M-Geo-1) required 1,000,000 iterations and about 10 days of runtime. The runtime for a single model run ranged from 0.4 to 0.9 s. All models were implemented in Fortran-95.

Appendix C: Predictive Performance Criteria

This appendix lists the evaluation criteria used to assess model transferability in space and time.

Criteria (i)–(ii) are based on statistical error metrics:

i. The value of the likelihood function (equation (B4)), which is used as the objective function when optimizing the model parameters (Appendix B). The likelihood function in equation (B4) is highest for simulations that minimize the prediction errors standardized by the estimated standard deviation of residual errors. To facilitate the comparison of likelihood values for evaluation periods that differ in time series length, we report a scaled (log) likelihood function, defined as the arithmetic average of log likelihood terms corresponding to each observed streamflow data point:

$$\bar{\Phi}\left[\hat{\mathbf{Q}}_{1:N_{c,T},1:N_C}, \tilde{\mathbf{Q}}_{1:N_{c,T},1:N_C}\right] = \frac{1}{\sum_{c=1}^{N_C} N_{c,T}} \sum_{c=1}^{N_C} \sum_{t=1}^{N_{c,T}} \log f_N\left(\tilde{Q}_{c,t} | \hat{Q}_{c,t}, \sigma^2_{\varepsilon(c,t)}\right) \tag{C1}$$

Equation (C1) is applied separately to time periods 1 and 2, and separately to subcatchment groups A and B.

The scaled log likelihood $\bar{\Phi}$ represents the natural logarithm of the average probability density of an observed streamflow observation within the predictive distribution of a model. These values are best interpreted in terms of their differences. For example, the difference between $\bar{\Phi} = 0.5$ and $\bar{\Phi} = 1.5$ corresponds to a factor of 2.7 in probability space. We stress that, unlike in the case of the Nash-Sutcliffe efficiency metric listed next, values $\bar{\Phi} = 0$ and $\bar{\Phi} = 1$ have no special significance.

ii. The Nash-Sutcliffe efficiency of streamflow time series predictions [*Nash and Sutcliffe*, 1970]:

$$\Upsilon\left[\hat{\mathbf{Q}}_{1:N_T}, \tilde{\mathbf{Q}}_{1:N_T}\right] = 1 - \frac{\sum_{t=1}^{N_T} \left(\tilde{Q}_t - \hat{Q}_t\right)^2}{\sum_{t=1}^{N_T} \left(\tilde{Q}_t - \text{ave}[\tilde{\mathbf{Q}}_{1:N_T}]\right)^2} \tag{C2}$$

where $\text{ave}[\tilde{\mathbf{Q}}_{1:N_T}]$ denotes the average (mean) of the observed data.

In addition to the statistical streamflow prediction metrics above, two hydrological signatures are used:

iii. The base flow index I_b [e.g., *Eckhardt*, 2008],

$$I_b[\mathbf{Q}_{1:N_T}] = \frac{\sum_{t=1}^{N_T} Q_t^{(b)}}{\sum_{t=1}^{N_T} Q_t} \tag{C3}$$

which is defined as the fraction of base flow $Q^{(b)}$ within total streamflow.

The base flow was estimated by applying a low-pass filter to the streamflow time series. A Gaussian moving-average window of 30 days was used, with centered mean and standard deviation of 15 days [e.g., *Kavetski et al*., 2011]. The base flow was calculated as the minimum of the streamflow and the low-pass filter output. Note that the type of filter and filter parameter values are likely to affect the base flow index [e.g., *Eckhardt*, 2008], however these aspects were not investigated in this work.

iv. The flashiness index I_f [*Baker et al*., 2004],

$$\frac{I_f[\mathbf{Q}_{1:N_T}] = \sum_{t=1}^{N_T} |Q_t - Q_{t-1}|}{\sum_{t=1}^{N_T} Q_t} \tag{C4}$$

which describes the "responsiveness" of a catchment in terms of the step-by-step variability of its streamflow.

When applied to simulated streamflows, the predictive performance criteria (ii)–(iv) are calculated for the "optimal" model simulations, i.e., using the parameter sets that maximize Φ. Moreover, they are calculated for the optimal model simulations alone, without accounting for the effect of the error model.

The evaluation of model performance in terms of the signatures serves the two main purposes: (1) identify whether there is a consistent ranking of catchments with respect to the observed and simulated signature values, and (2) aid the interpretation of the relative differences between the models, i.e., compare subcatchment responses within each catchment model, and between catchment models.

Acknowledgments
We thank the Administration des Services Techniques de l'Agriculture (ASTA), Administration de la Gestion de l'Eau (AGE) of Luxembourg, and the Luxembourg Institute of Science and Technology (LIST) for making available the data used in this study. Part of this work was funded through the DFG-FNR research project CAOS (INTER/DFG/11/01). The second author acknowledges the support by the Australian Research Council's Discovery Projects funding scheme (project number DP110105408). The corresponding Author can make the data and software available for the sole purpose of replicating the results of this study. Data requests for other purposes should be directed to ASTA, AGE, and LIST. We thank Jerôme Juilleret for providing expert advice on Luxembourgish geology.

References

Abbott, M. B. (1996), *Distributed Hydrological Modelling*, vol. IX, 321 S. pp., Kluwer Acad., Dordrecht, Netherlands.
Andreassian, V., C. Perrin, L. Berthet, N. Le Moine, J. Lerat, C. Loumagne, L. Oudin, T. Mathevet, M. H. Ramos, and A. Valery (2009), HESS Opinions 'Crash tests for a standardized evaluation of hydrological models', *Hydrol. Earth Syst. Sci.*, *13*(10), 1757–1764.
Arnold, J. G., et al. (2012), Swat: Model use, calibration, and validation, *Trans. ASABE*, *55*(4), 1491–1508.
Baker, D. B., R. P. Richards, T. T. Loftus, and J. W. Kramer (2004), A new flashiness index: Characteristics and applications to midwestern rivers and streams, *J. Am. Water Resour. Assoc.*, *40*(2), 503–522, doi:10.1111/j.1752-1688.2004.tb01046.x.

Beven, K. (1996), The limits of splitting: Hydrology, *Sci. Total Environ.*, *183*(1-2), 89–97.
Beven, K. (2002), Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system, *Hydrol. Processes*, *16*(2), 189–206.
Beven, K. (2006), Searching for the Holy Grail of scientific hydrology: Q(t) = H((S), (R), Delta t)A as closure, *Hydrol. Earth Syst. Sci.*, *10*(5), 609–618.
Beven, K., and M. J. Kirkby (1979), A physically based, variable contributing area model of basin hydrology, *Hydrol. Sci. Bull.*, *24*(1), 43–69.
Beven, K. J., M. J. Kirkby, N. Schofield, and A. F. Tagg (1984), Testing a physically-based flood forecasting-model (Topmodel) for 3 UK catchments, *J. Hydrol.*, *69*(1-4), 119–143.
Blazkova, S., and K. Beven (2009), A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, *Water Resour. Res.*, *45*, W00B16, doi:10.1029/2007WR006726.
Box, G. E. P., and G. C. Tiao (1992), *Bayesian Inference in Statistical Analysis*, 588 pp., John Wiley, N. Y.
Boyle, D. P. (2001), *Multicriteria Calibration of Hydrological Models*, Dep. of Hydrol. and Water Resour., Univ. of Ariz., Tucson.
Budyko, M. I. (1974), *Climate and Life*, 508 pp., Academic, N. Y.
Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, *298*(1-4), 242–266, doi:10.1016/j.jhydrol.2004.03.042.
Buytaert, W., and K. Beven (2011), Models as multiple working hypotheses: Hydrological simulation of tropical alpine wetlands, *Hydrol. Processes*, *25*(11), 1784–1799, doi:10.1002/hyp.7936.
Chamberlin, T. C. (1965), The method of multiple working hypotheses, *Science*, *148*(3671), 754–759.
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., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.
Clark, M. P., D. Kavetski, and F. Fenicia (2011a), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.
Clark, M. P., H. K. McMillan, D. B. G. Collins, D. Kavetski, and R. A. Woods (2011b), Hydrological field data from a modeller's perspective: Part 2: Process-based evaluation of model hypotheses, *Hydrol. Processes*, *25*(4), 523–543, doi:10.1002/hyp.7902.
Clark, M. P., et al. (2015), A unified approach for process-based hydrologic modeling: 1. Modeling concept, *Water Resour. Res.*, *51*, 2498–2514, doi:10.1002/2015WR017198.
Criss, R. E., and W. E. Winston (2008), Do Nash values have value? Discussion and alternate proposals, *Hydrol. Processes*, *22*(14), 2723–2725, doi:10.1002/hyp.7072.
Devito, K., I. Creed, T. Gan, C. Mendoza, R. Petrone, U. Silins, and B. Smerdon (2005), A framework for broad-scale classification of hydrologic response units on the Boreal Plain: Is topography the last thing to consider?, *Hydrol. Processes*, *19*(8), 1705–1714, doi:10.1002/hyp.5881.
Djodjic, F., and A. Villa (2015), Distributed, high-resolution modelling of critical source areas for erosion and phosphorus losses, *Ambio*, *44*, suppl. 2, 241–251, doi:10.1007/s13280-014-0618-4.
Dooge, J. C. I. (1988), Hydrology in perspective, *Hydrol. Sci. J.*, *33*(1), 61–85.
Eckhardt, K. (2008), A comparison of baseflow indices, which were calculated with seven different baseflow separation methods, *J. Hydrol.*, *352*(1–2), 168–173, doi:10.1016/j.jhydrol.2008.01.005.
Evin, G., M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera (2014), Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resour. Res.*, *50*, 2350–2375, doi:10.1002/2013WR014185.
Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2006), Is the groundwater reservoir linear? Learning from data in hydrological modelling, *Hydrol. Earth Syst. Sci.*, *10*(1), 139–150.
Fenicia, F., J. J. McDonnell, and H. H. G. Savenije (2008), Learning from model improvement: On the contribution of complementary data to process understanding, *Water Resour. Res.*, *44*, W06419, doi:10.1029/2007WR006386.
Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, *Water Resour. Res.*, *47*, W11510, doi:10.1029/2010WR010174.
Fenicia, F., D. Kavetski, H. H. G. Savenije, M. P. Clark, G. Schoups, L. Pfister, and J. Freer (2014), Catchment properties, function, and conceptual model representation: Is there a correspondence?, *Hydrol. Processes*, *28*(4), 2451–2467, doi:10.1002/hyp.9726.
Freeze, R. A., and R. L. Harlan (1969), Blueprint for a physically-based, digitally-simulated hydrologic response model, *J. Hydrol.*, *9*(3), 237–258, doi:10.1016/0022-1694(69)90020-1.
Frey, M. P., C. Stamm, M. K. Schneider, and P. Reichert (2011), Using discharge data to reduce structural deficits in a hydrological model with a Bayesian inference approach and the implications for the prediction of critical source areas, *Water Resour. Res.*, *47*, W12529, doi:10.1029/2010WR009993.
Gao, H., M. Hrachowitz, F. Fenicia, S. Gharari, and H. H. G. Savenije (2014a), Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China, *Hydrol. Earth Syst. Sci.*, *18*(5), 1895–1915, doi:10.5194/hess-18-1895-2014.
Gao, H., M. Hrachowitz, S. J. Schymanski, F. Fenicia, N. Sriwongsitanon, and H. H. G. Savenije (2014b), Climate controls how ecosystems size the root zone storage capacity at catchment scale, *Geophys. Res. Lett.*, *41*, 7916–7923, doi:10.1002/2014GL061668.
Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004), *Bayesian Data Analysis*, vol. XXV, 2nd ed., 668 pp., Chapman and Hall, Boca Raton, Fla.
Gharari, S., M. Hrachowitz, F. Fenicia, and H. H. G. Savenije (2011), Hydrological landscape classification: Investigating the performance of HAND based landscape classifications in a central European meso-scale catchment, *Hydrol. Earth Syst. Sci.*, *15*(11), 3275–3291, doi:10.5194/hess-15-3275-2011.
Gharari, S., M. Hrachowitz, F. Fenicia, H. Gao, and H. H. G. Savenije (2014), Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration, *Hydrol. Earth Syst. Sci.*, *18*(12), 4839–4859, doi:10.5194/hess-18-4839-2014.
Grayson, R., and A. Western (2001), Terrain and the distribution of soil moisture, *Hydrol. Processes*, *15*(13), 2689–2690, doi:10.1002/hyp.479.
Grayson, R. B., I. D. Moore, and T. A. Mcmahon (1992), Physically based hydrologic modeling: 2. Is the concept realistic, *Water Resour. Res.*, *28*(10), 2659–2666.
Gupta, H. V., and G. S. Nearing (2014), Debates—The future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science, *Water Resour. Res.*, *50*, 5351–5359, doi:10.1002/2013WR015096.

Gupta, H. V., T. Wagener, and Y. Q. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*(18), 3802–3813, doi:10.1002/hyp.6989.
Hjerdt, K. N., J. J. McDonnell, J. Seibert, and A. Rodhe (2004), A new topographic index to quantify downslope controls on local drainage, *Water Resour. Res.*, *40*, W05602, doi:10.1029/2004WR003130.
Hrachowitz, M., O. Fovet, L. Ruiz, T. Euser, S. Gharari, R. Nijzink, J. Freer, H. H. G. Savenije, and C. Gascuel-Odoux (2014), Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, *Water Resour. Res.*, *50*, 7445–7469, doi:10.1002/2014WR015484.
Hublart, P., D. Ruelland, A. Dezetter, and H. Jourde (2015), Reducing structural uncertainty in conceptual hydrological modelling in the semi-arid Andes, *Hydrol. Earth Syst. Sci.*, *19*(5), 2295–2314, doi:10.5194/hess-19-2295-2015.
Ivanov, V. Y., E. R. Vivoni, R. L. Bras, and D. Entekhabi (2004), Preserving high-resolution surface and rainfall data in operational-scale basin hydrology: A fully-distributed physically-based approach, *J. Hydrol.*, *298*(1–4), 80–111, doi:10.1016/j.jhydrol.2004.03.041.
Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model, *Water Resour. Res.*, *29*(8), 2637–2649.
Juilleret, J., J. F. Iffly, L. Hoffmann, and C. Hissler (2012), The potential of soil survey as a tool for surface geological mapping: A case study in a hydrological experimental catchment (Huewelerbach, Grand-Duchy of Luxembourg), *Geol. Belgica*, *15*(1-2), 36–41.
Kampf, S. K., and S. J. Burges (2007), A framework for classifying and comparing distributed hillslope and catchment hydrologic models, *Water Resour. Res.*, *43*, W05423, doi:10.1029/2006WR005370.
Kavetski, D., and M. P. Clark (2010), Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction, *Water Resour. Res.*, *46*, W10511, doi:10.1029/2009WR008896.
Kavetski, D., and F. Fenicia (2011), Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, *Water Resour. Res.*, *47*, W11511, doi:10.1029/2011WR010748.
Kavetski, D., and G. Kuczera (2007), Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration, *Water Resour. Res.*, *43*, W03411, doi:10.1029/2006WR005195.
Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, *J. Hydrol.*, *320*(1-2), 173–186, doi:10.1016/j.jhydrol.2005.07.012.
Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis, *J. Hydrol.*, *320*(1-2), 187–201, doi:10.1016/j.jhydrol.2005.07.013.
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.
Kirchner, J. W. (2009), Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, *Water Resour. Res.*, *45*, W02429, doi:10.1029/2008WR006912.
Köhne, J. M., S. Köhne, and J. Šimůnek (2009), A review of model applications for structured soils: a) Water flow and tracer transport, *J. Contam. Hydrol.*, *104*(1–4), 4–35, doi:10.1016/j.jconhyd.2008.10.002.
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.
Kumar, R., B. Livneh, and L. Samaniego (2013), Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme, *Water Resour. Res.*, *49*, 5700–5714, doi:10.1002/wrcr.20431.
Leavesley, G. H., R. W. Lichty, B. M. Troutman, and L. G. Saindon (1983), Precipitation-runoff modeling system: User's manual, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, *83-4238*, 214 pp.
Li, Z., W.-Z. Liu, X.-C. Zhang, and F.-L. Zheng (2009), Impacts of land use change and climate variability on hydrology in an agricultural catchment on the Loess Plateau of China, *J. Hydrol.*, *377*(1–2), 35–42, doi:10.1016/j.jhydrol.2009.08.007.
Lindstrom, G., B. Johansson, M. Persson, M. Gardelin, and S. Bergstrom (1997), Development and test of the distributed HBV-96 hydrological model, *J. Hydrol.*, *201*(1-4), 272–288.
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.
Meerveld, I. T.-V., and M. Weiler (2008), Hillslope dynamics modeled with increasing complexity, *J. Hydrol.*, *361*(1–2), 24–40, doi:10.1016/j.jhydrol.2008.07.019.
Milly, P. C. D. (1994), Climate, soil-water storage, and the average annual water-balance, *Water Resour. Res.*, *30*(7), 2143–2156, doi:10.1029/94WR00586.
Monteith, J. L. (1981), Evaporation and surface-temperature, *Q. J. R. Meteorol. Soc.*, *107*(451), 1–27, doi:10.1256/smsqj.45101.
Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.
Obled, C., J. Wendling, and K. Beven (1994), The sensitivity of hydrological models to spatial rainfall patterns—An evaluation using observed data, *J. Hydrol.*, *159*(1-4), 305–333.
Pfister, L., G. Drogue, C. Poirier, and L. Hoffmann (2005), Spatial variability of trends in hydrological extremes induced by orographically enhanced rainfall events due to westerly atmospheric circulations, *Water Sci. Technol.*, *51*(5), 15–21.
Pfister, L., J. J. McDonnell, S. Wrede, D. Hlubikova, P. Matgen, F. Fenicia, L. Ector, and L. Hoffmann (2009), The rivers are alive: On the potential for diatoms as a tracer of water source and hydrological connectivity, *Hydrol. Processes*, *23*(19), 2841–2845, doi:10.1002/hyp.7426.
Pfister, L., J. J. McDonnell, C. Hissler, and L. Hoffmann (2010), Ground-based thermal imagery as a simple, practical tool for mapping saturated area connectivity and dynamics, *Hydrol. Processes*, *24*(21), 3123–3132, doi:10.1002/hyp.7840.
Pokhrel, P., H. V. Gupta, and T. Wagener (2008), A spatial regularization approach to parameter estimation for a distributed watershed model, *Water Resour. Res.*, *44*, W12419, doi:10.1029/2007WR006615.
Pokhrel, P., K. K. Yilmaz, and H. V. Gupta (2012), Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures, *J. Hydrol.*, *418*, 49–60, doi:10.1016/j.jhydrol.2008.12.004.
Refsgaard, J. C., B. Storm, and A. Refsgaard (1995), Recent developments of the Systeme Hydrologique Europeen (SHE) towards the MIKE SHE, *Modell. Manage. Sustainable Basin Scale Water Resour. Syst.*, *231*, 427–434.
Reggiani, P., and T. H. M. Rientjes (2005), Flux parameterization in the representative elementary watershed approach: Application to a natural basin, *Water Resour. Res.*, *41*, W04013, doi:10.1029/2004WR003693.
Reggiani, P., and J. Schellekens (2006), Rainfall-runoff modeling: Distributed models, in *Encyclopedia of Hydrological Sciences*, edited by M. G. Anderson, pp. 1–16, John Wiley, Hoboken, N. J., doi:10.1002/0470848944.hsa135.
Reggiani, P., M. Sivapalan, and S. M. Hassanizadeh (1998), A unifying framework for watershed thermodynamics: Balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics, *Adv. Water Resour.*, *22*(4), 367–398.

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.
Renno, C. D., A. D. Nobre, L. A. Cuartas, J. V. Soares, M. G. Hodnett, J. Tomasella, and M. J. Waterloo (2008), HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia, *Remote Sens. Environ.*, *112*(9), 3469–3481.
Samaniego, L., R. Kumar, and S. Attinger (2010), Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, *Water Resour. Res.*, *46*, W05523, doi:10.1029/2008WR007327.
Savenije, H. H. G. (2010), HESS opinions "Topography driven conceptual modelling (FLEX-Topo)", *Hydrol. Earth Syst. Sci.*, *14*(12), 2681–2692, doi:10.5194/hess-14-2681-2010.
Scherrer, S., and F. Naef (2003), A decision scheme to indicate dominant hydrological flow processes on temperate grassland, *Hydrol. Processes*, *17*(2), 391–401, doi:10.1002/hyp.1131.
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.
Sivapalan, M. (2003), Process complexity at hillslope scale, process simplicity at the watershed scale: Is there a connection?, *Hydrol. Processes*, *17*(5), 1037–1041, doi:10.1002/hyp.5109.
Sivapalan, M., G. Bloschl, L. Zhang, and R. Vertessy (2003), Downward approach to hydrological prediction, *Hydrol. Processes*, *17*(11), 2101–2111.
Smith, T., L. Marshall, B. McGlynn, and K. Jencso (2013), Using field data to inform and evaluate a new model of catchment hydrologic connectivity, *Water Resour. Res.*, *49*, 6834–6846, doi:10.1002/wrcr.20546.
Soulsby, C., D. Tetzlaff, P. Rodgers, S. Dunn, and S. Waldron (2006), Runoff processes, stream water residence times and controlling landscape characteristics in a mesoscale catchment: An initial evaluation, *J. Hydrol.*, *325*(1-4), 197–221, doi:10.1016/j.jhydrol.2005.10.024.
Tague, C., and G. E. Grant (2004), A geological framework for interpreting the low-flow regimes of Cascade streams, Willamette River Basin, Oregon, *Water Resour. Res.*, *40*, W04303, doi:10.1029/2003WR002629.
Tallaksen, L. M. (1995), A review of baseflow recession analysis, *J. Hydrol.*, *165*(1-4), 349–370.
Thielen, J., J. Bartholmes, M. H. Ramos, and A. de Roo (2009), The European Flood Alert System—Part 1: Concept and development, *Hydrol. Earth Syst. Sci.*, *13*(2), 125–140, doi:10.5194/hess-13-125-2009.
Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00b14, doi:10.1029/2008WR006825.
Uhlenbrook, S., S. Roser, and N. Tilch (2004), Hydrological process representation at the meso-scale: The potential of a distributed, conceptual catchment model, *J. Hydrol.*, *291*(3-4), 278–296, doi:10.1016/j.jhydrol.2003.12.038.
Uhlenbrook, S., J. Didszun, N. Tilch, J. McDonnell, and K. McGuire (2007), Breaking up is always difficult—Landscape discretization as a process-transfer approach for prediction in ungauged basins, in *Proceedings of the PUB Kick-Off Meeting*, *IAHS Publ.*, *309*, 102–109.
VanderKwaak, J. E., and K. Loague (2001), Hydrologic-response simulations for the R-5 catchment with a comprehensive physics-based model, *Water Resour. Res.*, *37*(4), 999–1013, doi:10.1029/2000WR900272.
van Esse, W. R., C. Perrin, M. J. Booij, D. C. M. Augustijn, F. Fenicia, D. Kavetski, and F. Lobligeois (2013), The influence of conceptual model structure on model performance: A comparative study for 237 French catchments, *Hydrol. Earth Syst. Sci.*, *17*(10), 4227–4239, doi:10.5194/hess-17-4227-2013.
Westerberg, I. K., J. L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, J. E. Freer, and C. Y. Xu (2011), Calibration of hydrological models using flow-duration curves, *Hydrol. Earth Syst. Sci.*, *15*(7), 2205–2227, doi:10.5194/hess-15-2205-2011.
Wienhofer, J., and E. Zehe (2014), Predicting subsurface stormflow response of a forested hillslope—The role of connected flow paths, *Hydrol. Earth Syst. Sci.*, *18*(1), 121–138, doi:10.5194/hess-18-121-2014.
Winter, T. C. (2001), The concept of hydrologic landscapes, *J. Am. Water Resour. Assoc.*, *37*(2), 335–349, doi:10.1111/j.1752-1688.2001.tb00973.x.
Wood, E. F., D. P. Lettenmaier, and V. G. Zartarian (1992), A land-surface hydrology parameterization with subgrid variability for general-circulation models, *J. Geophys. Res.*, *97*(D3), 2717–2728.
Wrede, S., F. Fenicia, N. Martínez-Carreras, J. Juilleret, C. Hissler, A. Krein, H. H. G. Savenije, S. Uhlenbrook, D. Kavetski, and L. Pfister (2015), Towards more systematic perceptual model development: A case study using 3 Luxembourgish catchments, *Hydrol. Processes*, *29*(12), 2731–2750, doi:10.1002/hyp.10393.
Zehe, E., T. Maurer, J. Ihringer, and E. Plate (2001), Modeling water flow and mass transport in a loess catchment, *Phys. Chem. Earth, Part B*, *26*(7–8), 487–507, doi:10.1016/S1464-1909(01)00041-7.
Zhang, Z. X., T. Wagener, P. Reed, and R. Bhushan (2008), Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization, *Water Resour. Res.*, *44*, W00B04, doi:10.1029/2008WR006833.

Erratum

The order of Figure 5 and Figure 6 as well as the corresponding text has been changed to reflect the correct numerical order for both figures and text. This version may be considered the authoritative version of record.