Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-36, 2017
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 24 January 2017

Hydrology and Earth System Sciences
Discussions

HESS Opinions: The complementary merits of top-down and bottom-up modelling philosophies in hydrology

Markus Hrachowitz[1], Martyn Clark[2]

[1] Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN Delft, Netherlands
[2] National Center for Atmospheric Research, Boulder CO, 80301, USA

Correspondence to: Markus Hrachowitz (m.hrachowitz@tudelft.nl)

Abstract. In hydrology, the two somewhat competing modelling philosophies of bottom-up and top-down approaches are the basis of most process-based models. Differing mostly (1) in their respective degree of detail in resolving the modelling domain and (2) in their respective degree of explicitly treating conservation laws, these two philosophies suffer from similar limitations. Nevertheless, a better understanding of their respective basis (i.e. micro-scale vs. macro-scale) as well as their respective short comings bears the potential of identifying the complementary value of the two philosophies for improving our models. In this manuscript we analyse several frequently communicated beliefs and assumptions to identify, discuss and emphasize the functional similarity of the two modelling philosophies. We argue that deficiencies in model applications largely do not depend on the modelling philosophy but rather on the way a model is implemented. Based on the premises that top-down models can be implemented at any desired degree of detail and that any type of model remains to some degree conceptual we argue that a convergence of the two modelling strategies may hold some value for progressing the development of hydrological models.

1 What is the issue?

Hydrological models are used to predict floods, droughts, groundwater recharge and land-atmosphere exchange, and are of critical importance as tools to develop strategies for water resources planning and management. This is in particular true in the light of the increasing effects of climate and land-use change on the terrestrial water cycle. Yet, in spite of their central importance, these models frequently fail to reproduce the hydrological response in periods they have not been calibrated for, thereby providing unreliable predictions.

As models aim to encapsulate our understanding of the system, their weakness for predictions suggests that, besides the impact of observational uncertainties (e.g. Beven and Westerberg, 2011), at least some of the underlying processes that control how water and energy are stored in, transferred through and released from different parts of a catchment are not sufficiently well represented, in terms of both, parameters and parameterizations, in state-of-the-art current generation models.

The hydrologic modeling community sets out to design process-based system descriptions explicitly based on our understanding of the actual mechanisms involved using a range of different philosophies and approaches. The one end of the spectrum are detailed, high resolution descriptions of small-scale processes that are numerically integrated to larger scales (e.g. catchments). These models are commonly referred to as bottom-up or "physically based". The other end of the spectrum are less detailed, often spatially lumped representations of the system at the catchment-scale. These models, consisting of suites of storage elements ("buckets") that are linked by fluxes, are typically referred to as top-down or "conceptual" models. More generally, the "bottom-up" and "top-down" distinction reflects two end points along the continuum of complexity and existing hydrologic models are typically classified into one of these two groups depending on their *process complexity*, i.e., the extent to which models explicitly represent specific processes; and their *spatial complexity*, i.e., the extent to which models explicitly represent details of the landscape and the lateral flow of water across model elements. In addition, formulations of bottom-up models are expressed through a detailed and explicit treatment of conservation of mass, energy and momentum as well as parametrizations that are directly based on observations of fluxes on the small scale (i.e. the closure relations; Beven, 2006a). Current-generation top-down models are less detailed in that aspect.

Over the past four decades innumerable studies illustrated the value but also the limitations of these two competing modeling philosophies, i.e. top-down vs. bottom-up. Irrespective of the model type, it is often observed that, for training periods and with respect to the calibration objective(s), models exhibit considerable skill to reproduce the response patterns of a given system. However, it is likewise observed that many models cannot adequately reproduce aspects of the observed system response other than the calibration objectives, and which may include descriptors of emergent patterns, i.e. catchment signatures, such as flow duration curves (e.g. Jothityangkoon et al. 2001; Eder et al., 2003; Yadav et al., 2007; Martinez and Gupta, 2011; Sawicz et al., 2011; Euser et al., 2013; Willems et al., 2014; Westerberg and McMillan, 2015) but also variables the model may not have been calibrated to, such as ground- or soil water fluctuations. This failure to mimic system internal dynamics and patterns in a meaningful way indicates that models do a good curve-fitting job, but do not represent the dominant processes of the system in a meaningful way, thereby providing the right answers for the wrong reasons (cf. Kirchner, 2006). Together with the largely inevitable errors introduced by data uncertainty (e.g. Beven and Westerberg, 2011; Beven et al., 2011; Renard et al., 2011; Beven, 2013; McMillan et al., 2012; Kauffeldt et al., 2013; McMillan and Westerberg, 2015; Coxon et al., 2015) and insufficient model evaluation and testing (cf. Klemes, 1986; Wagener 2003; Clark et al., 2008; Gupta et al., 2008, 2012; Andreassian et al., 2009), models then often experience substantial performance decreases when used to predict the hydrological response for time periods they were not calibrated for .

Notwithstanding the similar skills of their models (e.g. Reed et al. 2004; Breuer et al., 2009; Smith et al., 2012; Lobligeois et al., 2013, Vansteenkiste et al., 2014), there is surprisingly little fruitful exchange between the bottom-up and top-down modelling communities. Top-down models are criticized for lacking a robust physical or theoretical basis (e.g. Paniconi and Putti, 2015; Fatichi et al., 2016), whereas bottom-up models are often viewed as having inferior representations of sub-grid variability (e.g. Beven and Cloke, 2012) and are not sufficiently agile to represent the dominant processes in different environments (e.g. Mendoza et al., 2015). Even more, instead of joining forces and integrating the respective efforts, communication between the communities is often limited to mutually highlighting the deficiencies of and dismissing the respective modelling strategies.

We think that to achieve progress in the discipline of scientific hydrology and to develop models for more reliable predictions, it is necessary for the different hydrologic modelling communities to take a step back. Reflecting on failures and successes can not only help to design better models but also to appreciate the *complementary* nature and value of microscale process understanding on the one hand and the quest for general laws at the catchment scale (Klemes, 1983; Dooge, 1986; Sivapalan, 2005) on the other hand. This commentary is based on detailed and, at times, refreshingly heated discussions during and after the *1st Workshop on Improving the Theoretical Underpinnings of Hydrologic Models* (Bertinoro, April 2016). Our aim is to identify, discuss and clarify common misunderstandings and misinterpretations of competing modeling approaches. By emphasizing the importance of zooming out to the macroscale we intend to scrutinize the value of top-down models with respect to bottom-up models. More generally, we provide a perspective of how to take advantage of different modelling philosophies in order to improve our predictions.

2 Modelling philosophies

2.1 The basis of bottom-up models

Bottom-up models provide a description of the flow system that is physically consistent with our understanding of the forces acting on and controlling the release of water from the control volumes under consideration. The individual control volumes in a model domain are typically designed as regular grids at spatial resolutions between centimetres and hundreds of meters, each representing the lateral and vertical flow processes of water (and energy) through the porous and heterogeneous soil column and at the land-vegetation-atmosphere interface. As basic building blocks of these models, output from the individual control volumes, i.e. the boundary fluxes (Beven, 2006a), is then aggregated along their respective surface and subsurface flow directions through the adjacent control volumes to the channel and eventually routed to the catchment outlet.

The bottom-up strategy therefore has two main features: It explicitly accounts for spatial heterogeneity and its influence on the hydrological response; and it provides a rigorous and physically consistent way to encapsulate and formalize our theoretical knowledge of the dominant processes that are known to be active in terrestrial hydrology. This allows for clearly separable descriptions of distinct processes, which in turn provides hypotheses that can in principle be individually scrutinized and tested against observations (Gupta et al., 2008; Clark et al., 2015). Firmly based on our understanding of the underlying, small-scale physics, the development of bottom-up models is directed towards a meaningful representation of natural feedback between individual parts of the system. In a non-linear system, characterized by spatial heterogeneity, threshold processes and boundary conditions that are variable over a wide range of scales, newest-generation implementations of this inductive approach to hydrology were shown to have the potential to reproduce emergent pattern, such as the effect of preferential flow paths (e.g. Zehe and Blöschl, 2004; Kollet and Maxwell, 2006; Zehe et al., 2006; Sudicky et al., 2008).

2.2 The basis of top-down models

Top-down models are in the first instance based on the information in the available data. For most catchments worldwide this is, at best, limited to (ideally areal) estimates of precipitation and potential evaporation as well as stream flow. Thus, the starting point are not process descriptions that result from *anecdotal* observations and theoretical considerations at the *small scale* but rather the system integrated, emergent response pattern as characterized by the available observations at the *catchment scale*.

The difference between the input (e.g. precipitation) and output signals (e.g. stream flow) integrates the effects of natural heterogeneity, internal organization and process feedback and describes the low-pass filter properties of a catchment. As a low-pass filter a catchment buffers the high-frequency components of a random input signal (i.e. precipitation) in storage components and eventually releases the input signal with system specific time lags as stream flow or evaporation. In other words, the input is approximated by a dirac delta function for each model time step. This input signal then undergoes, at all scales of relevance, spatial and temporal dispersion in its most general sense before it reaches the catchment outlet. Such a domain-integrated dispersion pattern results from a suite of individual, functionally distinct dispersive effects. On the one hand, a spatially uniform input signal experiences dispersion due the distribution of different flow path lengths in a catchment: water that enters the system close to the catchment divide in the headwaters will have to travel longer distances to the outlet than water entering the system just next to the outlet. Water travelling longer distances is therefore likely to arrive later at the outlet ("geomorphologic dispersion"; e.g. Rinaldo et al., 1991; Snell and Sivapalan, 1994). On the other hand, a water volume entering at one specific location in a catchment can be partitioned to follow different flow paths such as deep groundwater or overland flow. Controlled by the respective flow resistances and gradients along these flow paths, this results in different flow velocities, thereby introducing a second source of dispersion ("kinematic dispersion"; e.g. Botter and Rinaldo, 2003), which is also reflected in the dichotomy of hillslope vs. channel dispersion (e.g. Robinson et al., 1995).

Top-down models, interpreted as simple filters, aim to reproduce these catchment-scale integrated dispersion patterns, which reflect catchment-internal organization and feedback. Such approaches which do not explicitly invoke process interpretations have a long history and include the instantaneous unit hydrograph (IUH) and extended concepts (e.g. Rodriguez-Iturbe and Valdes, 1979; Rinaldo et al., 2006). The main challenge for top-down models is to identify and represent the dominant individual flow paths with typically distinct dispersion properties as well as their temporally variable hydrological connectivity and interactions that link the various subcomponents of the heterogeneous flow domain. In doing so, top-down models obtain a suitable description of the non-linear effective input into these flow generating subcomponents from a buffer component whose storage volume is controlled by the energy input and thus by the evaporative demand.

Many top-down modelling studies use only a simple distinction between low and medium pass filter properties, i.e. two storage elements operating at different timescales, which are fed by effective input obtained from a rough approximation of a non-linear storage buffer, which describes the dynamic partitioning into evaporative fluxes and drainage that reflect the underlying spatial organization of connectivity. These simple flux parameterizations can in many cases produce good first order estimates of hydrological response dynamics.

Thus, top-down models are based on observed input-output relationships, without further assumptions on the actual flow system; in contrast to bottom-up models, this approach does not, in a first instance, explicitly define and describe actually detailed underlying processes. For example, storage elements with the shorter process timescales can represent different mechanisms. Depending on the catchment under consideration they can be interpreted as representation of preferential flow, transmissivity feedback, overland flow or other fast responding processes. In its fundamentals rooted in purely data-driven thinking, top-down models therefore provide a means to reproduce dynamics that emerge at the scale of interest (e.g. a catchment) and that can actually be observed with very limited need for additional assumptions on other processes, which albeit active cannot be discriminated by the available data. Zooming out to the actual scale of the model application, strictly maintaining mass balance, and implementing a parsimonious representation of the energy balance, top-down models have proven to effectively represent the emergent patterns of partitioning of water fluxes through a few dominant flow paths with different dispersive properties, integrating catchment organization and internal feedback of the entire model domain, in spite of largely disregarding catchment-internal process complexity.

3 Modelling myths – or not?

There is a wide range of frequently communicated beliefs and assumptions on alternative approaches to modelling. They reflect different perceptions of modelling limitations. In the following sections we will scrutinize common modelling critiques (*C1*-*C3*) and discuss the extent to which we believe they are justified.

3.1 (*C1*) *"Top-down models have a poor physical and theoretical basis."*

Since top-down models originate from empirical approaches to mimic the hydrological response only based on available observations, without further assumptions on the system internal processes, this statement does certainly have an element of truth. However, evaluating this statement requires considering the effects of scale and emergent properties of a system. Put simply, how do different modeling philosophies represent large-scale fluxes?

To illustrate the physical basis of large-scale models, consider a gas volume with given boundary conditions and energy inputs (e.g. Savenije, 2001; Blöschl and Zehe, 2005). Under conservation of mass and energy, a functional relationship between energy input (i.e. pressure) and temperature can be established, characterizing the emergent properties of the system at the actual scale of the gas volume, i.e. the Gay-Lussac law. To understand the response of the system to energy inputs, it is in principle possible to describe the exact trajectories and velocities of the individual gas molecules and their interactions within the volume. This approach could, in theory, also form the basis of the Gay-Lussac law, and the same temperature dynamics should emerge. Yet, currently available observational technology, e.g. to define the initial conditions of the systems, at scales of actual interest renders the molecular dynamics approach practically infeasible, untestable, and unnecessary. In spite of disregarding the actual small-scale physics, it is difficult to argue that Gay-Lussac's law and similar observation-based, functional relationships at the macroscale, often considered natural laws, have a poor physical and theoretical basis. Rather, they are valuable descriptors of the system at the macroscale without loss of essential information, which remains encapsulated in the functional relationships by integration of system-internal heterogeneity and complexity. Similarly, water flows in a catchment follow the observable, physical phenomenon of dispersion, controlled by water and energy input, gravity, and flow resistances. The development of top-down models is then the process of identifying functional relationships between system input and the integrated output pattern emerging through organization at the catchment scale, i.e. the testing of competing hypotheses (e.g. Clark et al., 2011; Fenicia et al., 2011), without the need of resorting to small scale physics.

The above arguments to some degree reflect the contrasting approaches of testing hypothesis vs. testing emerging patterns, and therefore the difference between deductive and inductive scientific reasoning (e.g. Salmon, 1967). If a large enough sample of different systems is available, the emerging patterns and the associated functional relationships (i.e. model hypotheses) can facilitate similarity analysis and classification, eventually allowing to "search for general laws at the macroscale" (Sivapalan, 2005). Although the value and need for this way of thinking in hydrology was already emphasized several decades ago (e.g. Klemes, 1983; Dooge, 1986) and echoed in several subsequent publications (e.g. Sivapalan et al., 2003; McDonnell et al., 2007; Blöschl et al., 2013; Sivakumar et al., 2013; Gupta et al., 2014) it only recently gained significant momentum (e.g. Lyon and Troch, 2007; Carillo et al., 2011; Sawicz et al., 2011; Coopersmith et al., 2012; Berghuijs et al., 2014; Fenicia et al., 2014; Li et al., 2014; McMillan et al., 2014).

In spite of not being explicitly based on small scale physics, top-down models may thus nevertheless be considered as physically based, yet on a different scale: on the macroscale. This is furthermore in particular true as on the one hand they do largely satisfy conservation laws at the catchment scale, as (1) they rigorously maintain mass balance, and (2) they provide a reasonable and parsimonious representation of the energy balance, which can be meaningful *if carefully constrained* not only with respect to the hydrograph but also with respect to the observed runoff coefficients on a range of scales (e.g. annual, seasonal and event-based), which defines the partitioning between streamflow and evaporative fluxes (e.g. Budyko, 1974; Donohue et al., 2007; Sivapalan et al., 2011) plus potential deep infiltration losses (e.g. Andreassian and Perrin, 2012). Such large-scale conservation equations require large-scale flux parameterizations (i.e. the closure problem). Such large-scale fluxes are typically estimated as a function of system-average water quantities (storage S), reflecting the integrated gradients in the model domain, and system-average resistances (storage coefficient k), i.e. $Q=f(S,k)$.

The purported physical basis of macroscale laws permits that a physical meaning can and actually should be assigned to all processes in top-down models. Yet, the physical structure of top-down models has in the past frequently only been defined in a casual way by loosely "interpreting" the hydrological function of individual model components with respect to the real-world system under investigation and the model used. For example, for typical 3-box models such as HBV, the fast responding component has been previously interpreted as overland flow, shallow subsurface flow, pipe flow or simply as a vague quick flow component, essentially lumping these processes (e.g. Wood et al., 1992; Jakeman and Hornberger, 1993; Madsen, 2000; Seibert et al., 2003; Koren et al., 2004; Uhlenbrook et al., 2004; Fenicia et al., 2008). However, without detailed testing, such interpretations of their physical basis remain somewhat ambiguous and subjective.

It is therefore desirable and eventually necessary to explore methods to more objectively and rigorously test individual model sub-components against observations (Clark et al., 2011) and/or to *assign* physical meaning to them *a priori* (cf. Bahremand, 2016). A potentially effective starting point for the latter is to use observations at the modelling scale to infer information about the functional shape (i.e. parameterization) and to quantify the actual parameters of individual processes at that scale. Examples include the recession behaviour of catchments in dry periods. If long enough observation records are available, Master Recession Curves (MRC) can be constructed directly from data (Lamb and Beven, 1997). Notwithstanding potential sources of uncertainty, such as (1) in observations (Beven et al., 2011), (2) arising from potentially interacting processes, e.g. evaporative fluxes (Fenicia et al., 2006) or deep infiltration losses (Hrachowitz et al., 2014), as well as (3) the role of spatial storage heterogeneities (Spence et al., 2010) a MRC can hold considerable information. From this both, a meaningful set of possible parameterizations as well as feasible prior parameter distributions for the slow responding model component can if not determined at least be considerably constrained. As it is well established that stream flow in dry periods is most commonly sustained by deep groundwater, assigning this quantitative information to the slow responding model component then objectively defines it as groundwater component at the observation and modelling scale. Similarly, there is strong evidence that the water holding capacity in the unsaturated root zone ($S_{U,max}$), which is the core of many hydrological systems as it controls the partitioning of drainage and evaporative fluxes, can be robustly estimated at the catchment scale based on the long-term water balance, i.e. observations of precipitation and stream flow (or, if available, actual evaporation; Gao et al., 2014; deBoer-Euser et al., 2016; Nijzink et al., 2016a).

These system components quantify actual physical properties present and physical processes active at the observation and modelling scale and therefore provide a clear physical meaning to different parts of a model. Defining the ranges of operation of individual model components in such an evidence-based way and thus assigning some level of physical meaning to the different components, does not only reduce the feasible model space (i.e. parameterizations and parameter values) but it has also the potential to increase a model's hydrological consistency while reducing its predictive uncertainty. Such a reduction of degrees of freedom in a model will in many instances lead to "sub-optimal" model performance with respect to some calibration objectives, as the *a priori* constrained processes will have less possibility to compensate for other model errors and to produce the right answers for the wrong reasons (cf. Kirchner, 2006). Reduced performance is then a clear indication of, besides observational uncertainty, an incomplete or inadequate representation of the remaining dominant partitioning points of the system.

We therefore argue that top-down models do have in principle, if well implemented and tested, a robust physical and theoretical basis and that it is possible to relate the structure of top-down models to stores and fluxes in nature (e.g. Clark et al., 2008; Fenicia et al., 2016), albeit at a different spatial scale and process resolution than bottom-up models. These types of models emphasize the value of zooming out and understanding the system from the point of holistic empiricism. However, this is not to say that there is not considerable room for improvement. This is in particular true for an explicit physically consistent treatment of the energy and momentum balances. Closely linked to that is the question to which level of detail the dominant catchment processes can and have to be resolved to reproduce the observed system response in a meaningful way, which directly leads to statement *C2*.

3.2 (*C2*) *"Top-down models are too simplistic and cannot adequately represent natural heterogeneity."*

Simple lumped top-down models, such as HBV, have a long track record of, at first glance, successful applications in a wide range of catchments world-wide. However, this success is in many cases deceptive and needs to undergo more critical scrutiny. The reason is that these models are often used in a quasi-inductive way with an implicit *a priori* assumption that they are a meaningful representation of the system, thereby not treating the model as a hypothesis and not testing alternative formulations. This is exacerbated by a frequent lack of robust (multivariate) calibration and systematic and exhaustive (multivariate) post-calibration evaluation procedures to ensure that the overall modelled system response, including emerging patterns such as flow duration curves, reproduces the observed response dynamics in a meaningful way.

The importance of adequate representations natural heterogeneity has for a long time been highlighted (e.g. Klemes, 1986; Andreassian et al., 2009; Clark et al., 2011; Gupta et al., 2012). However, typical model calibration is limited to time series of streamflow observations and provides merely insight into a very small number of parameters (Jakeman and Hornberger, 1993). Thus, although any additional model process has the potential to improve the representation of heterogeneity and subsequently the calibrated model performance, the required additional flux parameterizations and calibration parameters increase the feasible model (or parameter) space and the resulting potential for equifinality (Beven, 1993), thereby turning models into the oft-cited "mathematical marionettes" (Kirchner, 2006). In spite of a high skill to reproduce the calibration objective, such a model will in many situations struggle to simultaneously reproduce different additional system internal dynamics (e.g. groundwater fluctuations) and emerging patterns (e.g. flow duration curves), indicating its failure to meaningfully represent dominant processes and their heterogeneity in a catchment, which in turn often results in a poor predictive power of these models. This was in the past demonstrated by many studies (e.g. Jothityangkoon et al., 2001; Atkinson et al., 2002; Fenicia et al., 2008; Euser et al., 2013; Coxon et al., 2014; Fenicia et al., 2014; Hrachowitz et al., 2014; Willems, 2014).

The lack of an adequate model calibration, testing and evaluation culture partly arises both from insufficient exploitation of the information content of the available data, and also the real lack of suitable data to more effectively constrain models (Gupta et al., 2008; Clark et al., 2011). Under these conditions, many models remain ill-posed inverse problems. To limit the associated problems, i.e. equifinality, Occam's razor is commonly invoked to make models "as simple as possible but not simpler" (e.g. Clark et al., 2011). But how simple is "as simple as possible"? Or in other words, how large a model space (i.e. possible parameterizations and prior parameter space) can be constrained with available information to identify reasonably narrow posterior distributions while ensuring a high as possible multivariate model performance?

The fundamental question that needs to be addressed in any model application is which model complexity is supported by the available data, including both *process complexity*, i.e. the detail to which models explicitly represent specific processes; and *spatial complexity*, i.e. the extent to which models explicitly represent details of the landscape and the lateral flow of water across model elements. We will address each of these issues in subsequent sections.

3.2.1 Process complexity

Process complexity in terrestrial hydrological systems is, at its fundamental level, characterized by two major partitioning points that control how water is stored in and released from catchments through upward, downward or lateral fluxes (e.g. Rockström et al. 2009; Clark et al., 2015; Savenije and Hrachowitz, 2016). Near the land surface, precipitation is split into (a) evaporation and sublimation from vegetation and ground surface interception (including snow) as well as from open water bodies, (b) overland flow and (c) infiltration into the root zone. Water entering into the root zone, is further partitioned into (d) soil evaporation, (e) plant transpiration, (f) shallow, lateral subsurface flow through preferential drainage features, such as shallow high permeability soil layers, soil pipe networks or a combination thereof and (g) percolation to the unsaturated zone and the groundwater *below* the root zone.

All fluxes (a-g) are present in essentially any catchment, albeit with different relative importance in different environments, and therefore need to be represented in a model. This can be illustrated with the occurrence of weather events that are uncommon for a specific region. In the Atacama Desert, one of the driest places on earth with little or no vegetation under average conditions, uncommonly high spring precipitation, such as in 2015, can cause episodic appearance of abundant vegetation. This temporally changes the partitioning pattern and thus the hydrological functioning of the region as plant transpiration that is otherwise absent is "activated". Similarly, rare occurrences of snow fall can cause temporal anomalies in the hydrological functioning of otherwise warm regions, such as 2013 in the Middle East. In spite of them being "de-activated" most of the time, such processes are in principle present and need therefore also be conceptually reflected in any hydrological model structure. However, if considered negligible in a specific environment during a modelling period of interest, the modeler can decide to deactivate individual processes by using informed prior parameter distributions. In other words, the respective parameters will be set to suitable fixed values that effectively switch off the process using Dirac delta functions as prior distributions.

Largely independent of modelling strategy (top-down vs. bottom-up) the key decision for the modeler in any given catchment is then to decide to which level of detail the individual processes at the two partitioning points will be resolved and how they can be parametrized (cf. Gupta et al., 2012). The questions to be answered are: How much detail is *necessary* to reproduce observed dynamics and pattern? How much detail is *warranted* by the available data to meaningfully parameterize and test the chosen process representation? Two examples of different processes are provided in the Supplementary Material (S1 and S2) to illustrate the thought process involved.

Specifically, many parts of the system need to remain conceptual simplifications, as with increasing complexity, non-linear systems become increasingly problematic to predict with detailed, small-scale descriptions, due to uncertainties in the necessary observations of boundary conditions, forcing and system states (e.g. Zehe et al., 2007). Conceptualization can then entail either, for a detailed process representation, using a high number of effective parameters, which are mostly unknown (and thus need to be calibrated) or potentially unrepresentative (if inferred from direct observations) for the model domain, to describe the system, or zooming out and exploiting simple functional relationships (or pattern) emerging as a result of organization at the macroscale, thereby significantly reducing the number of required effective model parameters. Importantly, this lumping process does not, as long as the simplification encapsulates the relevant dynamics of the system, necessarily involve a loss of information. Rather, it has the potential to *integrate* the interaction of heterogeneous processes at the microscale over the entire domain of interest and thereby to provide a system description that is consistent with real world observations at the scale of interest without the need for further assumptions and the related uncertainties.

3.2.2 Spatial complexity

Natural systems are commonly characterized by considerable heterogeneity in boundary conditions and system forcing, the “uniqueness of place” as eloquently referred to by Beven (2000). This heterogeneity spans over several orders of magnitude in scale, from the microscale (e.g. soil particles) to the continental scale (e.g. mountain ranges) and it is highly unlikely that observation technology will ever enable a comprehensive and non-invasive description of the heterogeneity in hydrologic systems, especially for large model domains.

Analogous to the process complexity, the degree of spatial complexity that can be incorporated in a model hinges on the detail of available information on the system. More specifically, which types of heterogeneity are present? How do they affect water storage and release in different ways? Which types of heterogeneity can be captured by a single emergent functional relationship and for which types of heterogeneity several individual functional relationships at the macroscale are necessary to meaningfully represent real world pattern?

In summary, the problem of spatial complexity is therefore rather multifaceted and an illustrative example is provided in Supplementary Material S3. It is true that untested and poorly evaluated applications of standard top-down models are often oversimplifications that do not adequately reflect natural heterogeneity and its effects on the hydrological response. However, top-down models can be formulated at any level of process and spatial complexity, limited only by the available information. The actual problem is therefore not the top-down model *per se* but rather the way it is implemented and applied. The decision, which degree of zooming out, i.e. which level of detailed process representation is feasible and which level is necessary, eventually needs to be made by the modeller on basis of the available observations, acknowledging that *all* hydrological models at the catchment scale are to a certain extent conceptualizations. When carefully implemented, spatially distributed formulations with an equilibrated balance between process heterogeneity and information/data availability and tested and evaluated against multivariate observed response dynamics, top-down models have been shown to be versatile enough to identify and represent the dominant hydrological processes and their heterogeneity in a catchment (e.g. Fenicia et al., 2008a,b; Hrachowitz et al., 2014; Nijzink et al., 2016b) within limited uncertainty. Mirroring the statement that top-down models are too simplistic and do not represent heterogeneity, it may however in a similar way be valuable to discuss the question if, in the absence of appropriate observations at the scale and resolution of interest, bottom-up models with high process and spatial complexity are not somewhat deceptive about the accuracy that is implied by their formulation, which directly leads to *C3*.

3.3 (*C3*) *“Top-down models are ad-hoc formulations of untestable hypotheses and always need calibration”*

It is true that many applications of standard top-down models, such as HBV or FLEX, have lacked due diligence in the choice of suitable process and spatial complexity and the constraints/information placed on the prior parameter distributions (to de-/activate processes) of a model in a given catchment. Such a frequently unquestioned use of off-the-shelve models implicitly assumes that these models can adequately represent observed hydrological response dynamics in different catchments. This ignores that any model is an assemblage of hypotheses consisting of individual building blocks and their parametrizations, encapsulating the modeler’s understanding how a specific environment shapes the hydrological system. For example, HBV was developed and tested for applications in cool and humid environments, characterized by high volumes and limited seasonality of precipitation together with limited energy supply for evaporation. In such a situation, many of the processes that introduce non-linearity and control the emergence of hydrologic connectivity are not dominant or even negligible, in contrast to arid environments (see also the example S4 in the Supplementary material). The point is that different environmental conditions dictate the need to test if the prior information on the parameters needs to be changed and/or relaxed so as to activate a process that was deactivated in a model previously used in other environments (or vice-versa) to adjust the model to the prevailing environmental conditions.

A meaningful decision on the use of given prior parameter distributions and their information content for a model application in a specific environment can be made if the model hypothesis is carefully tested. However, it is sometimes argued that entire models are untestable hypotheses, as they represent a range of different processes or parts of the system. Models, therefore, need to be seen as sets of distinct hypotheses that need to be tested independently to avoid the adverse effects of equifinality (Clark et al., 2011). When disaggregating a system, the pattern emerging at each subsequent level of detail are, down to molecular levels (or maybe even beyond), a result of the interactions of heterogeneous processes at yet smaller scales (see section “*process complexity*”). Thus, down to that level, every hypothesis consists of several other, smaller scale hypotheses. The relevant question arising here is, to which level do model components then have to be disaggregated to be constitute testable hypotheses? Thus, of course, treating a model as a single hypothesis does not make the hypothesis *untestable*. Rather, given the system-integrated nature of many observations and the frequently limited number of performance indicators considered to test the model against, it may in many cases remain a relatively *weak* test. In contrast, individually testing sub-components of the system will provide the modeler with more information because its sub-components are necessarily less complex than the overall model. This, in turn, provides less possibilities for compensating misrepresentations of one process by wrongly adjusting other processes. In other words, it will have higher potential to avoid Type I errors (i.e. false positives), therefore resulting in a stricter test. The obvious problem arising here is less of theoretical than of practical nature: observations of system sub-components, including the often cited boundary fluxes (Beven, 2006a), to test the model components against are typically not available at the scale and/or resolution of interest or not available at all, although with the ever improving spatio-temporal resolution and quality of remote sensing products the problem will potentially be somewhat alleviated in the near future. Clearly, from that perspective, weak model tests are in the frequent absence of other options preferable to no tests at all.

The above point is very closely related to the necessity of calibration. If the system could be observed at the scale and resolution of interest (e.g. catchment scale for lumped models, grid scale for distributed models), there would be little additional need for testing as the system would be well constrained and its functioning well understood. Thus, much of the problems discussed above is a direct consequence of the absence of such observations. Whenever no adequate observations are available, a model requires calibration. Any model. This is not a limitation that is specific to top-down models. It equally applies to bottom-up models. The difference being that detailed bottom-up models may often provide a deceptive sense of accuracy if operated with highly informed prior parameters distributions (e.g. fixed parameter values or regularized estimates), based on anecdotal, point or plot scale observations that do not match the scale and resolution of the model domain (e.g. grid cell).

We therefore argue that top-down models are not "*ad-hoc formulations of untestable* hypotheses" but rather often *untested* hypotheses that indeed do require calibration, as any other type of model. The actual problem therefore not being the model type ("top-down"), but the way these models are frequently applied in a careless way.

4 Implications, potential ways forward and concluding remarks

From the above discussion, a few relatively clear and unambiguous points define the basis, functioning and limitations of competing approaches for process-based hydrologic modeling. Condensing these points, it emerges that:

(1) Top-down models, in spite of lacking explicit representation of small scale physics, represent the clear physical and observable phenomenon of dispersion. They strictly obey conservation of mass, and can provide, if well implemented, a parsimonious representation of conservation of energy. At the catchment-scale, the lack of small scale detail in these models is offset by embracing the value of zooming out to the macroscale, which in the realm of organized complexity is frequently characterized by the emergence of relatively simple functional relationships that describe individual processes and that, most importantly, integrate typically unobservable natural heterogeneity over the model domain. This provides a physical basis that is firmly rooted in holistic empiricism, similar to statistical physics (e.g. gas laws).

(2) Top-down models can, in principle, be implemented with any desired detail. The key question is whether additional process complexity can be tested against and is justified by the available data. This is true for both process and spatial complexity, which also highlights that we are really crossing a continuum of complexity, where top-down models converge towards small-scale physics based bottom-up formulations.

(3) Top-down models can reflect our conceptual understanding of the system if all water fluxes associated to the two major terrestrial partitioning points, i.e. the near-surface and the unsaturated root zone, are represented. As all these fluxes can, in principle, be present in any environment, albeit with different relative importance, all top-down models therefore need to have the same fundamental model structure (but not necessarily the same parameterization) to reflect these processes.

(4) Top-down models provide the flexibility for modelers to reduce the feasible parameter space and, if desired, to even deactivate specific processes, simplifying the model and reducing the number of calibration parameters and the associated problem of equifinality. For example, if a process is rarely occurring and/or otherwise not exerting significant influence on the system dynamics and is thus not warranted by data in a specific environment, the modeler can decide to use an informed prior parameter distribution for this process. Thereby the process-specific prior parameter distribution will be constrained or, in the extreme case, fixed to one value (i.e. Dirac delta function).

(5) All hydrological models applied at scales beyond the plot scale require some degree of calibration, as direct observations of effective parameters at these modelling scales and resolutions are typically not available. This is also true for bottom-up models. The common practice in bottom-up model applications of applying parameters from observations that do not match the modelling scale and/or resolution may not provide a sufficient representation of the natural heterogeneity of this parameter can lead to considerable misrepresentations of the system and give a deceptive impression of accuracy.

(6) All hydrological models are to some extent "conceptual" and to some extent "physical", they largely only differ in the degree of detail they resolve the system, which in turn is dictated by the available data. While top-down models approach the problem from a macroscale physical understanding, bottom-up models emphasize the microscale perspective. An ideal model would, almost needless to say, provide an equally good representation of both aspects.

(7) All macroscale hydrological models (i.e. hillslope, catchment), remain, in the absence of sufficient observations at the modelling scale and resolution, hypotheses and thus require rigorous, testing and post-calibration evaluation.

(8) The fundamental problems in catchment modelling do not lie in the type of model used, but rather in the way a model is applied. All too often, models are not understood and treated as hypotheses, and thus insufficiently calibrated and tested for applications in different environments.

Progress in catchment-scale understanding of hydrological functioning and the related development of models for more reliable predictions will not only benefit but does actually hinge on a better understanding of how natural heterogeneities at all scales aggregate to larger scales and how this influences the hydrological response. As already emphasized previously by many authors (e.g. Beven, 1989,2001,2006a; Kirchner, 2006; Zehe et al., 2014), these efforts to approach the closure problem in hydrology need to involve both, ways to reliably determine effective model parameters, i.e. the system boundary conditions, that integrate and reflect the natural heterogeneity within the model domain as well as the development of equations that are physically consistent at the scale of application. The latter can be illustrated with the use of the Darcy-

Richards equation, which requires the assumption of local equilibrium, which does not hold beyond the plot scale (e.g. Or at al., 2015), for defining a meaningful matric potential.

These scale and heterogeneity issues were acknowledged already in the early 1980s to be at the core of many problems for our understanding and modelling of hydrological systems (e.g. Dooge, 1986; Wood et al., 1988; Wood et al., 1990; Blöschl and Sivapalan, 1995). It was, for two decades or so, indeed a very active and fruitful field of research but it has somewhat lost momentum. Ten years after the landmark papers of Beven (2006a) and Kirchner (2006), remarkably little progress was made and many ideas and concepts did not find their way into mainstream hydrology. Nevertheless, it is imperative to understand how processes scale, heterogeneity aggregates and how this controls the emergence of patterns at the large scale. This then has the potential to enhance our understanding of what controls catchment functioning and our ability to develop models (e.g. Vinogradov et al., 2011). A potential way forward towards achieving this, may be the much advocated large sample, comparative hydrology to identify pattern and generally applicable, functional relationships (e.g. Blöschl et al., 2013; Gupta et al., 2014).

A further point that is indispensable if progress in catchment-scale modelling wants to be achieved and, in particular, top-down models want to be used as scientific tools, is the need to establish a mainstream culture of robust model calibration, rigorous testing of alternative model formulations (i.e. hypotheses) as well as the systematic assessment of model uncertainties. We are currently in a position where we, in an exaggerated way, feed wrong models with wrong input data and calibrate them to wrong output data to obtain wrong parameters. In the light of so many unknowns, comprehensive, systematic, end-to-end uncertainty analysis needs finally to become a standard requirement of any scientific paper that involves modelling (e.g. Beven, 2006b; Pappenberger and Beven, 2006). With uncertainties in both parameters and parameterizations (i.e. "hypotheses" and thus model equations), besides observational uncertainties, it may be more coherent to combine both in a joint model uncertainty framework and to consider reporting results in model ensembles, similar to what is common practice, for example, in atmospheric sciences. In any case, systematic uncertainty analysis has the potential to significantly reduce type II errors, i.e. rejecting a good model when it should have been accepted ("false negative"; Beven 2010) and is thus instrumental to avoid giving a false impression of accuracy in our models.

Next to uncertainty analysis, stronger and more meaningful model tests, i.e. model calibration and post-calibration evaluation with respect to multiple variables and model states, including, data permitting, model sub-components (e.g. Willems et al., 2014; Clark et al., 2015), as well as to multiple criteria besides to commonly used time series of these variables needs to become a standard procedure. It was argued and shown in a range of papers that, although models frequently exhibit considerable skill to reproduce the hydrograph during calibration and, albeit to a lesser degree, also during "validation", many of these models struggle to reproduce other system relevant features. This includes, for example groundwater table fluctuations (e.g. Fenicia et al., 2008), long-term average runoff coefficients as a proxy of average actual evaporation (e.g. Gharari et al., 2014; Hrachowitz et al., 2014) and solute dynamics (e.g. Birkel et al., 2010; Fenicia et al., 2010) as well as hydrological signatures of these components other than the time series, e.g. duration curves. While some of these signatures (e.g. runoff coefficient) can often be readily reproduced, it is was observed that others, such as the autocorrelation structure as a metric for the memory or persistence of a system are often less well captured by a model (e.g. Euser et al., 2013; Hrachowitz et al., 2014). Such a comprehensive model testing approach has the potential to identify and reject models (i.e. parameters and parameterizations) that “do not meet minimum requirements” (Vache and McDonnell, 2006), which in turn has the potential to considerably reduce type I errors, i.e. falsely accepting poor models when they should be rejected (“false positive”; Beven, 2010).

Finally, it may also be desirable to drop the somewhat arbitrary dichotomy between top-down and bottom-up models, which has in the past caused considerable confusion. Rather, acknowledging that all models are to some degree conceptual, and that often not the actual models are the problem but the inadequate way they are applied, may open up the view towards the real fundamental questions in catchment-scale modelling: how much detail do we *need* in our models and how much detail is *warranted* by data? To find a balance that allows us to best describe the system based on scientifically robust grounds will therefore benefit from accepting a more rigorous culture of model testing, to adjust process and spatial complexity to the environmental conditions and data availability in specific catchments. This will reduce the risk for oversimplifications and system misrepresentations when approaching the problem from the top-down perspective. In contrast, approaching the issue from the bottom-up perspective could potentially substantially benefit from embracing the value of zooming out and making use of emergent processes, which otherwise would be highly problematic to identify and parametrize. On balance, we believe that modelling of catchments will significantly benefit from and may even require a convergence of top-down and bottom-up strategies, in particular with respect to exploiting the features of organization in these complex systems (Dooge, 1986) in a hierarchical way, as for example suggested by Zehe et al. (2014). In that sense we would like to strongly encourage researchers to not only acknowledge but to actively make use of advantages the respective other modelling strategy has to offer in order to strengthen their very own models.

References

Andréassian, V., & Perrin, C. (2012). On the ambiguous interpretation of the Turc‐Budyko nondimensional graph. Water Resources Research, 48(10).

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.H., & Valéry, A. (2009). Crash tests for a standardized evaluation of hydrological models. Hydrology and Earth System Sciences, (13), 1757-1764.

Atkinson, S. E., R. A. Woods, and M. Sivapalan. "Climate and landscape controls on water balance model complexity over changing timescales." Water Resources Research 38.12 (2002).

Bahremand, A. (2016). HESS Opinions: Advocating process modelling and de-emphasizing parameter estimation. Hydrology and Earth System Sciences, 20, 1433-1445.

Berghuijs, W. R., Sivapalan, M., Woods, R. A., & Savenije, H. H. (2014). Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. Water Resources Research, 50(7), 5638-5661.

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in water resources, 16(1), 41-51.

Beven, K. (2000). Uniqueness of place and process representations in hydrological modelling. Hydrology and Earth System Sciences, 4(2), 203-213.

Beven, K. (2001). How far can we go in distributed hydrological modelling?. Hydrology and Earth System Sciences, 5(1), 1-12.

Beven, K. (2006). Searching for the Holy Grail of scientific hydrology: Q t=(S, R,? t) A as closure. Hydrology and Earth System Sciences, 10(5), 609-618.

Beven, K. (2006). On undermining the science?. Hydrological Processes, 20(14), 3141-3146.

Beven, K. J. (2010). Preferential flows and travel time distributions: defining adequate hypothesis tests for hydrological process models. Hydrological Processes, 24(12), 1537-1547.

Beven, K. (2013). So how much of your error is epistemic? Lessons from Japan and Italy. Hydrological Processes, 27(11), 1677-1680.

Beven, K. J., & Cloke, H. L. (2012). Comment on "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water" by Eric F. Wood et al. Water Resources Research, 48(1).

Beven, K., & Westerberg, I. (2011). On red herrings and real herrings: disinformation and information in hydrological inference. Hydrological Processes, 25(10), 1676-1680.

Beven, K., Smith, P. J., & Wood, A. (2011). On the colour and spin of epistemic error (and what we might do about it). Hydrology and Earth System Sciences, 15(10), 3123-3133.

Birkel, C., Dunn, S. M., Tetzlaff, D., & Soulsby, C. (2010). Assessing the value of high‐ resolution isotope tracer data in the stepwise development of a lumped conceptual rainfall–runoff model. Hydrological Processes, 24(16), 2335-2348.

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: a review. Hydrological processes, 9(3‐ 4), 251-290.

Blöschl, G., & Zehe, E. (2005). On hydrological predictability. Hydrological processes, 19(19), 3923-3929.

Blöschl, G. (Ed.). (2013). Runoff prediction in ungauged basins: synthesis across processes, places and scales. Cambridge University Press.

Botter, G., & Rinaldo, A. (2003). Scale effect on geomorphologic and kinematic dispersion. Water resources research, 39(10).

Breuer, L., Huisman, J.A., Willems, P., Bormann, H., Bronstert, A., Croke, B.F.W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A.J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D.P., Lindström, G., Seibert, J., Sivapalan, M., Viney, N.R. (2009). Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) I: Model intercomparison with current land use. Adv. Water Resour. 32 (2), 129–146.

Budyko, M. I. (1974). Climate and Life, 508 pp., Academic, Orlando

Carrillo, G., Troch, P. A., Sivapalan, M., Wagener, T., Harman, C., & Sawicz, K. (2011). Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient. Hydrology and Earth System Sciences, 15(11), 3411-3430.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., & Hay, L. E. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resources Research, 44(12).

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resources Research, 47(9).

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., & Rasmussen, R. M. (2015). A unified approach for process- based hydrologic modeling: 1. Modeling concept. Water Resources Research, 51(4), 2498-2514.

Coopersmith, E., Yaeger, M. A., Ye, S., Cheng, L., & Sivapalan, M. (2012). Exploring the physical controls of regional patterns of flow duration curves–Part 3: A catchment classification system based on regime curve indicators. Hydrology and Earth System Sciences, 16(11), 4467-4482.

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits- of- acceptability framework for 24 UK catchments. Hydrological Processes, 28(25), 6135-6150.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., & Smith, P. J. (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. Water resources research, 51(7), 5531-5546.

de Boer-Euser, T., McMillan, H. K., Hrachowitz, M., Winsemius, H. C., & Savenije, H. H. (2016). Influence of soil and climate on root zone storage capacity. Water Resources Research, 52, 2009-2024.

Donohue, R. J., Roderick, M. L., & McVicar, T. R. (2008). On the importance of including vegetation dynamics in Budyko's hydrological model. Hydrology and Earth System Sciences, 11, 983-995.

Dooge, J. C. (1986). Looking for hydrologic laws. Water Resources Research, 22(9S).

Eder, G., Sivapalan, M., & Nachtnebel, H. P. (2003). Modelling water balances in an Alpine catchment through exploitation of emergent properties over changing time scales. Hydrological Processes, 17(11), 2125-2149.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. Hydrology and Earth System Sciences, 17 (5), 1893-1912.

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., Downer, C. W., Camporese, M., Davison, J. H., Ebel, B., Jones, N., Kim, J., Mascaro, G., Niswonger, R., Restrepo, P., Rigon, R., Shen, C., Sulis, M., & Tarboton, D. (2016). An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. Journal of Hydrology, 537, 45-60.

Fenicia, F., Savenije, H. H. G., Matgen, P., & Pfister, L. (2006). Is the groundwater reservoir linear? Learning from data in hydrological modelling. Hydrology and Earth System Sciences, 10(1), 139-150.

Fenicia, F., Savenije, H. H., Matgen, P., & Pfister, L. (2008a). Understanding catchment behavior through stepwise model concept improvement. Water Resources Research, 44(1).

Fenicia, F., McDonnell, J. J., & Savenije, H. H. (2008b). Learning from model improvement: On the contribution of complementary data to process understanding. Water Resources Research, 44(6).

Fenicia, F., Wrede, S., Kavetski, D., Pfister, L., Hoffmann, L., Savenije, H. H., & McDonnell, J. J. (2010). Assessing the impact of mixing assumptions on the estimation of streamwater mean residence time. Hydrological Processes, 24(12), 1730-1741.

Fenicia, F., Kavetski, D., & Savenije, H. H. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resources Research, 47(11).

Fenicia, F., Kavetski, D., Savenije, H. H., Clark, M. P., Schoups, G., Pfister, L., & Freer, J. (2014). Catchment properties, function, and conceptual model representation: is there a correspondence?. Hydrological Processes, 28(4), 2451-2467.

Fenicia, F., Kavetski, D., Savenije, H. H., & Pfister, L. (2016). From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. Water Resources Research.

Gao, H., Hrachowitz, M., Schymanski, S. J., Fenicia, F., Sriwongsitanon, N., & Savenije, H. H. G. (2014). Climate controls how ecosystems size the root zone storage capacity at catchment scale. Geophysical Research Letters, 41(22), 7916-7923.

Gharari, S., Hrachowitz, M., Fenecia, F., Gao, H., & Savenije, H. H. G. (2014). Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. Hydrology and Earth System Sciences, 18, 4839-4859.

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrological Processes, 22(18), 3802-3813.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. Water Resources Research, 48(8).

Gupta, H. V., Perrin, C., Bloschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: a need to balance depth with breadth. Hydrology and Earth System Sciences, 18(2), 463-477.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., & Gascuel-Odoux, C. (2014). Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. Water Resources Research, 50(9), 7445-7469.

Jakeman, A. J., & Hornberger, G. M. (1993). How much complexity is warranted in a rainfall-runoff model?. Water resources research, 29(8), 2637-2649.

Jothityangkoon, C., Sivapalan, M., & Farmer, D. L. (2001). Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development. Journal of Hydrology, 254(1), 174-198.

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C. Y., & Westerberg, I. K. (2013). Disinformative data in large-scale hydrological modelling. Hydrology and Earth System Sciences, 17(7), 2845-2857.

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research, 42(3).

Klemeš, V. (1983). Conceptualization and scale in hydrology. Journal of hydrology, 65(1-3), 1-23.

Klemeš, V. (1986). Dilettantism in hydrology: transition or destiny?. Water Resources Research, 22(9S).

Kollet, S. J., & Maxwell, R. M. (2006). Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model. Advances in Water Resources, 29(7), 945-958.

Koren, V., Reed, S., Smith, M., Zhang, Z., & Seo, D. J. (2004). Hydrology laboratory research modeling system (HL-RMS) of the US national weather service. Journal of Hydrology, 291(3), 297-318.

Lamb, R., & Beven, K. (1997). Using interactive recession curve analysis to specify a general catchment storage model. Hydrology and Earth System Sciences Discussions, 1(1), 101-113.

Li, H. Y., Sivapalan, M., Tian, F., & Harman, C. (2014). Functional approach to exploring climatic and landscape controls of runoff generation: 1. Behavioral constraints on runoff volume. Water Resources Research, 50(12), 9300-9322.

Lobligeois, F., Andréassian, V., Perrin, C., Tabary, P., & Loumagne, C. (2014). When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events. Hydrology and Earth System Sciences, 18(2), 575-594.

Lyon, S. W., & Troch, P. A. (2007). Hillslope subsurface flow similarity: Real- world tests of the hillslope Péclet number. Water Resources Research, 43(7).

Madsen, H. (2000). Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. Journal of hydrology, 235(3), 276-288.

Martinez, G. F., & Gupta, H. V. (2011). Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. Water Resources Research, 47(12).

McDonnell, J. J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., & Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. Water Resources Research, 43(7).

McMillan, H. K., & Westerberg, I. K. (2015). Rating curve estimation under epistemic uncertainty. Hydrological Processes, 29(7), 1873-1882.

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. Hydrological Processes, 26(26), 4078-4111.

McMillan, H., Gueguen, M., Grimon, E., Woods, R., Clark, M., & Rupp, D. E. (2014). Spatial variability of hydrological processes and model structure diagnostics in a 50 km2 catchment. Hydrological Processes, 28(18), 4896-4913.

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., & Gupta, H. (2015). Are we unnecessarily constraining the agility of complex process- based models?. Water Resources Research, 51(1), 716-728.

Nijzink, R., Hutton, C., Pechlivanidis, I., Capell, R., Arheimer, B., Freer, J., Han, D., Wagener, T., McGuire, K., Savenije, H., & Hrachowitz, M. (2016a). The evolution of root-zone moisture capacities after deforestation: a step towards hydrological predictions under change?. Hydrology and Earth System Sciences, 20(12), 4775-4799.

Nijzink, R. C., Samaniego, L., Mai, J., Kumar, R., Thober, S., Zink, M., Schäfer, D., Savenije, H. H. G., & Hrachowitz, M. (2016b). The importance of topography-controlled sub-grid process heterogeneity and semi-quantitative prior constraints in distributed hydrological models. Hydrology and Earth System Sciences, 20(3), 1151-1176.

Or, D., Lehmann, P., & Assouline, S. (2015). Natural length scales define the range of applicability of the Richards equation for capillary flows. Water Resources Research, 51(9), 7130-7144.

Paniconi, C., & Putti, M. (2015). Physically based modeling in catchment hydrology at 50: Survey and outlook. Water Resources Research, 51(9), 7090-7129.

Pappenberger, F., & Beven, K. J. (2006). Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water resources research, 42(5).

Reed, S., Koren, V., Smith, M.B., Zhang, Z., Moreda, F., Seo, D., Dmip participants, A. (2004). Overall distributed model intercomparison project results. Journal of Hydrology 298, 27–60.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., & Franks, S. W. (2011). Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. Water Resources Research, 47(11).

Rinaldo, A., Marani, A., & Rigon, R. (1991). Geomorphological dispersion. Water Resources Research, 27(4), 513-525.

Rinaldo, A., Botter, G., Bertuzzo, E., Uccelli, A., Settin, T., & Marani, M. (2006). Transport at basin scales: 1. Theoretical framework. Hydrology and Earth System Sciences Discussions, 10, 19-29.

Robinson, J. S., Sivapalan, M., & Snell, J. D. (1995). On the relative roles of hillslope processes, channel routing, and network geomorphology in the hydrologic response of natural catchments. Water Resources Research, 31(12), 3089-3101.

Rockström, J., Falkenmark, M., Karlberg, L., Hoff, H., Rost, S., & Gerten, D. (2009). Future water availability for global food production: the potential of green water for increasing resilience to global change. Water Resources Research, 45(7).

Rodríguez‐ Iturbe, I., & Valdes, J. B. (1979). The geomorphologic structure of hydrologic response. Water resources research, 15(6), 1409-1420.

Salmon, W. (1967). The foundations of scientific inference (Vol. 28). University of Pittsburgh Pre.

Savenije, H. H. G. (2001). Equifinality, a blessing in disguise?. Hydrological processes, 15(14), 2835-2838.

Savenije H. H. G., & Hrachowitz, M. (2016). Opinion paper: Modelling catchments as living organisms. Hydrology and Earth System Sciences.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carillo, G. (2011). Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrology and Earth System Sciences, 15, 2895-2911.

Seibert, J., Rodhe, A., & Bishop, K. (2003). Simulating interactions between saturated and unsaturated storage in a conceptual runoff model. Hydrological Processes, 17(2), 379-390.

Sivakumar, B., Singh, V. P., Berndtsson, R., & Khan, S. K. (2013). Catchment classification framework in hydrology: challenges and directions. Journal of Hydrologic Engineering, 20(1), A4014002.

Sivapalan, M. (2005). Pattern, process and function: elements of a unified theory of hydrology at the catchment scale. Encyclopedia of hydrological sciences.

Sivapalan, M., Blöschl, G., Zhang, L., & Vertessy, R. (2003). Downward approach to hydrological prediction. Hydrological processes, 17(11), 2101-2111.

Sivapalan, M., Yaeger, M. A., Harman, C. J., Xu, X., & Troch, P. A. (2011). Functional model of water balance variability at the catchment scale: 1. Evidence of hydrologic similarity and space‐ time symmetry. Water Resources Research, 47(2).

Smith, M.B., Koren, V., Zhang, Z., Zhang, Y., Reed, S.M., Cui, Z., Moreda, F., Cosgrove,B.A., Mizukami, N., Anderson, E.A. (2012). Results of the DMIP 2 Oklahoma experiments. Journal of Hydrology, 418–419, 17–48.

Snell, J. D., & Sivapalan, M. (1994). On geomorphological dispersion in natural catchments and the geomorphological unit hydrograph. Water Resources Research, 30(7), 2311-2323.

Spence, C., Guan, X. J., Phillips, R., Hedstrom, N., Granger, R., & Reid, B. (2010). Storage dynamics and streamflow in a catchment with a variable contributing area. Hydrological Processes, 24(16), 2209-2221.

Sudicky, E. A., Jones, J. P., Park, Y. J., Brookfield, A. E., & Colautti, D. (2008). Simulating complex flow and transport dynamics in an integrated surface-subsurface modeling framework. Geosciences Journal, 12(2), 107-122.

Uhlenbrook, S., Roser, S., & Tilch, N. (2004). Hydrological process representation at the meso-scale: the potential of a distributed, conceptual catchment model. Journal of Hydrology, 291(3), 278-296.

Vaché, K. B., & McDonnell, J. J. (2006). A process‐ based rejectionist framework for evaluating catchment runoff model structure. Water Resources Research, 42(2).

Vansteenkiste, T., Tavakoli, M., Van Steenbergen, N., De Smedt, F., Batelaan, O., Pereira, F., & Willems, P. (2014). Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation. Journal of Hydrology, 511, 335-349.

Vinogradov, Y. B., Semenova, O. M., & Vinogradova, T. A. (2011). An approach to the scaling problem in hydrological modelling: the deterministic modelling hydrological system. Hydrological processes, 25(7), 1055-1073.

Wagener, T. (2003). Evaluation of catchment models. Hydrological Processes, 17(16), 3375-3378.

Westerberg, I. K., & McMillan, H. K. (2015). Uncertainty in hydrological signatures. Hydrology and Earth System Sciences, 19(9), 3951-3968.

Willems, P. (2014). Parsimonious rainfall–runoff model construction supported by time series processing and validation of hydrological extremes – Part 1: Step-wise model-structure identification and calibration approach. Journal of Hydrology, 510, 578-590.

Willems, P., Mora, D., Vansteenkiste, T., Taye, M. T., & Van Steenbergen, N. (2014). Parsimonious rainfall-runoff model construction supported by time series processing and validation of hydrological extremes–Part 2: Intercomparison of models and calibration approaches. Journal of Hydrology, 510, 591-609.

Wood, E. F., Sivapalan, M., Beven, K., & Band, L. (1988). Effects of spatial variability and scale with implications to hydrologic modeling. Journal of hydrology, 102(1-4), 29-47.

Wood, E. F., Sivapalan, M., & Beven, K. (1990). Similarity and scale in catchment storm response. Reviews of Geophysics, 28(1), 1-18.

Wood, E. F., Lettenmaier, D. P., & Zartarian, V. G. (1992). A land‐ surface hydrology parameterization with subgrid variability for general circulation models. Journal of Geophysical Research: Atmospheres, 97(D3), 2717-2728.

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. Advances in Water Resources, 30(8), 1756-1774.

Zehe, E., & Blöschl, G. (2004). Predictability of hydrologic response at the plot and catchment scales: Role of initial conditions. Water Resources Research, 40(10).

Zehe, E., Lee, H., & Sivapalan, M. (2006). Dynamical process upscaling for deriving catchment scale state variables and constitutive relations for meso-scale process models. Hydrology and Earth System Sciences, 10(6), 981-996.

Zehe, E., Elsenbeer, H., Lindenmaier, F., Schulz, K., & Blöschl, G. (2007). Patterns of predictability in hydrological threshold systems. Water Resources Research, 43(7).

Zehe, E., Ehret, U., Pfister, L., Blume, T., Schröder, B., Westhoff, M., Jackisch, C., Schymanski, S. J., Weiler, M., Schulz, K., Allroggen, N., Tronicke, J., van Schaik, L., Dietrich, P., Scherer, U., Eccard, J., Wulfmeyer, V., & Kleidon, A. (2014). HESS Opinions: From response units to functional units: a thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments. Hydrology and Earth System Sciences, 18(11), 4635-4655.