Pursuing the method of multiple working hypotheses for hydrological modeling

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

Received 29 July 2010; revised 6 August 2011; accepted 15 August 2011; published 28 September 2011.

[1] Ambiguities in the representation of environmental processes have manifested themselves in a plethora of hydrological models, differing in almost every aspect of their conceptualization and implementation. The current overabundance of models is symptomatic of an insufficient scientific understanding of environmental dynamics at the catchment scale, which can be attributed to difficulties in measuring and representing the heterogeneity encountered in natural systems. This commentary advocates using the method of multiple working hypotheses for systematic and stringent testing of model alternatives in hydrology. We discuss how the multiple-hypothesis approach provides the flexibility to formulate alternative representations (hypotheses) describing both individual processes and the overall system. When combined with incisive diagnostics to scrutinize multiple model representations against observed data, this provides hydrologists with a powerful and systematic approach for model development and improvement. Multiple-hypothesis frameworks also support a broader coverage of the model hypothesis space and hence improve the quantification of predictive uncertainty arising from system and component nonidentifiabilities. As part of discussing the advantages and limitations of multiple-hypothesis frameworks, we critically review major contemporary challenges in hydrological hypothesis-testing, including exploiting different types of data to investigate the fidelity of alternative process representations, accounting for model structure ambiguities arising from major uncertainties in environmental data, quantifying regional differences in dominant hydrological processes, and the grander challenge of understanding the self-organization and optimality principles that may functionally explain and describe the heterogeneities evident in most environmental systems. We assess recent progress in these research directions, and how new advances are possible using multiple-hypothesis methodologies.

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

1. Introduction

1.1. Ambiguities in the Choice of Model Structure

[2] Building an environmental model requires making a series of decisions regarding the appropriate representation of natural processes. Some of these decisions can already be based on well-established physical understanding. For example, the snow component of a model can be designed to explicitly simulate all energy and mass fluxes at the snow–atmosphere interface (as opposed to "just" representing snowmelt as an empirical function of temperature). However, gaps in our current understanding of environmental dynamics, combined with incomplete knowledge of the properties and boundary conditions of most environmental systems, make many important modeling decisions far more ambiguous. For example, how should saturation-excess runoff be represented? What about macropore flow: is it significant or even dominant, and, if so, how should it be represented? What is the best way to quantify the impact of (unknown) bedrock topography/permeability on subsurface water retention? Other modeling decisions may be driven by pragmatic considerations, such as the modeler's background, computer budget, and study objectives. How can we represent the spatial variability in snow depth across a hierarchy of scales? Is an application of Beer's law to a single canopy layer sufficient to simulate the transmission of shortwave radiation through the forest canopy, or are more sophisticated methods required? Finally, some decisions are more "holistic" in nature. How do we represent the heterogeneity of flow paths through a catchment, and hydrological controls such as topography and soil properties? From a higher vantage point, can we account for the geomorphologic and biological drivers that may have shaped the current traits of a landscape of interest and its vegetation, or indeed, entire classes of environmental systems? The point is that there is currently little agreement regarding what a "correct" model structure is, especially at relatively larger spatial scales such as catchments and beyond. In current practice,

[1]Research Applications Laboratory, National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA.
[2]Environmental Engineering, University of Newcastle, Callaghan, Australia.
[3]Department of Environment and Agro-Biotechnologies, Centre de Recherche Public–Gabriel Lippmann, Grand-Duchy of Luxembourg.
[4]Water Resources Section, Delft University of Technology, Netherlands.

0043-1397/11/2010WR009827

faced with such a range of decisions, different modelers will generally make different modeling decisions, often in an ad hoc manner, on the basis of their balancing of process understanding, the data available to evaluate the model, the purpose of the modeling exercise, and other considerations.

[3] Whether directly or indirectly, ambiguities in the choice of model structure have led to a plethora of hydrological models (e.g., see the collections of models described by *Singh and Frevert* [2002a, 2002b], and the classification of models proposed by *Kampf and Burges* [2007]). Faced with such remarkable overabundance, the community has struggled to identify the "most appropriate" models even in the relatively simple terms of "best empirical performance," let alone in terms of their scientific validity. In several model intercomparison experiments to date, participating models often produced similar levels of accuracy with respect to streamflow predictions, despite employing quite different conceptualizations of natural processes [e.g., *Reed et al.*, 2004; *Duan et al.*, 2006].

1.2. Are We Failing?

[4] In our opinion, the apparent superfluity of catchment-scale hydrological models is symptomatic of a current failure for the discipline of hydrology in its efforts to develop a general model structure that is "physically realistic" (in the sense of agreeing with experimental insights), operationally adequate, and applicable in different environments and climatic conditions. This failure is evident in the differences among the "general" models described by *Burnash et al.* [1973], *Liang et al.* [1994], *Reggiani et al.* [1998], *VanderKwaak and Loague* [2001], *Maxwell and Miller* [2005], and *Qu and Duffy* [2007] and many other authors.

[5] The debates regarding appropriate methods to represent natural processes are particularly symptomatic of an insufficient scientific understanding of environmental dynamics at the catchment scale. This lack of understanding can be attributed to our current inability to adequately quantify the impact of subcatchment heterogeneities on the catchment's hydrological response [*McDonnell et al.*, 2007; *Kumar*, 2011], or formulated more generally as a failure to resolve the catchment-scale "closure" problem [*Reggiani et al.*, 1998, 1999; *Beven*, 2006b]. Several commentators have also noted the lack of a unified theory of hydrology at the catchment scale, and suggested that such a theory should reflect the self-organization and optimality principles that may functionally explain and describe the heterogeneities evident in most environmental systems [e.g., *Sivapalan*, 2005; *McDonnell et al.*, 2007; *Troch et al.*, 2009].

[6] These challenges raise several important questions, including: Does a single "correct" catchment-scale hydrological model exist at all, or is the current glut of models a consequence of poor identifiability from currently available hydrological data? Do differences in hydrological functioning across diverse hydrological landscapes require region-specific model structures? Do some (or all) catchment systems inherently require spatially distributed representations? Or are we on the right track and "just haven't made it there yet"? In this commentary, we contend not only that these questions are currently poorly understood, but, more concerningly for the discipline of hydrology, that we lack, or sometimes are unwilling (or unable) to develop or apply, the scientific tools to answer these questions.

1.3. Aims and Scope of Commentary

[7] This commentary advocates systematically adopting the method of multiple working hypotheses [*Chamberlin*, 1890] for hydrological model development and evaluation. We make two specific contributions. First, we frame major contemporary challenges in catchment-scale hydrological modeling (and their potential solutions) from a more unified perspective of hypothesis testing in hydrology. Second, we propose a tractable research strategy that combines the capabilities of modular (flexible) frameworks to isolate individual model hypotheses, with the diagnostic approach to model evaluation that exploits different types of observed data and data signatures to scrutinize both individual model components *and* their connectivity within the complete system model. Our broader objective is to contribute a powerful and systematic "multiple-hypothesis" approach for improving model representations of hydrological processes, and for handling uncertainties arising from model structural ambiguities and data errors. In addition to tying together themes raised as key challenges in previous commentaries [e.g., *Beven*, 2006a, 2008; *Blöschl*, 2006; *Dunn et al.*, 2008; *Gupta et al.*, 2008; *Kirchner*, 2006; *McDonnell et al.*, 2007; *Savenije*, 2001, 2009; *Sivapalan*, 2009; *Soulsby et al.*, 2008; *Tetzlaff et al.*, 2008; *Troch et al.*, 2009], it is our express intent to promote several practical solutions, which, while still in their nascence, are already yielding encouraging progress and insights.

[8] The commentary is structured as follows. We begin by posing model development decisions, both on overall model structure as well as on individual components, as testable hypotheses. We then review several widely used strategies for hydrological model development and evaluation, ranging from ad hoc considerations during model development and refinement, to "top-down" and rejectionist frameworks, to more systematic model intercomparison experiments. These approaches are critiqued in the context of hypothesis-testing, including their stringency in exposing both entire models and individual internal components to scrutiny, their breadth of coverage of the feasible model hypothesis space, and, finally, their ability to support controlled model evaluation and analysis. Multiple-hypothesis frameworks are then proposed as a more systematic method for model development that can address the shortcomings of current techniques. We discuss the major functional aspects of multiple-hypothesis approaches, highlighting both similarities and contrasts to existing multimodel frameworks. As part of our discussion of the advantages and limitations of multiple-hypothesis frameworks, we tie together several important challenges in hydrological hypothesis-testing, including exploiting data from both experimental watersheds and operational networks to carry out incisive diagnostic tests of the model hypotheses, accounting for model structure ambiguities arising from the limited availability and large uncertainty in environmental data, the challenge of understanding regional differences in dominant hydrological processes, and the challenge of representing heterogeneity at different spatial scales within the model domain. While the presentation and example focuses primarily on catchment-scale hydrological modeling, we

contend that the flexible model approach has a great potential when applied for hypothesis testing in other fields of environmental modeling, whether pursued using physically based methods or using more conceptual perspectives.

2. Hydrological Models as Hypotheses of Catchment Function and Behavior

2.1. Defining Models as Hypotheses

[9] We define an environmental model as the entire set of coupled variables (states) and functional relationships used to represent a catchment or environmental system. In this definition, the functional relationships can be based on theory (or, more generally, on process understanding) or derived from data. This perspective provides a clear focus on representing the system dynamics, patterns, and functionality [e.g., *McDonnell et al.*, 2007; *Sivapalan*, 2009; *Kumar*, 2011]. This definition applies to all models, regardless of their form (e.g., deterministic versus stochastic) and regardless of the model complexity.

[10] In the context of scientific hydrology, a model is a hypothesis of catchment function: it encompasses a description of dominant hydrological processes and predicts how those processes combine to produce the catchment's response to external forcing. The general characteristics of a model, and hence the types of hypotheses and assumptions embodied in it, may depend on the broad perspectives taken during model development. For example, the "bottom-up" perspective focuses on hydrological theory at the small scale, and aggregates the output to the scale of interest, such as the entire catchment. This is the approach underlying the current generation of physics-based distributed models [e.g., *VanderKwaak and Loague*, 2001; *Ivanov et al.*, 2004; *Maxwell and Miller*, 2005; *Qu and Duffy*, 2007]. Conversely, the "top-down" perspective attempts to describe the system directly at the scale of interest [e.g., *Klemeš*, 1983; *Dooge*, 1986; *Sivapalan et al.*, 2003]. This can be pursued in several ways, such as by analytically integrating the small-scale equations [e.g., *Reggiani et al.*, 1999], by imposing optimality constraints [*Schymanski et al.*, 2007, 2009], by employing system identification techniques [e.g., *Young*, 1998, 2003], or by conceptualizing the overall system dynamics on the basis of experimental and other perceptions [e.g., *Beven and Kirkby*, 1979; *Lamb and Beven*, 1997; *Uhlenbrook et al.*, 2004; *Vaché and McDonnell*, 2006; *Birkel et al.*, 2010]. It is not our intention here to debate the relative merits of each approach nor are they necessarily mutually exclusive within a given model application [e.g., see *Butts et al.*, 2004]. Rather, our intention is to emphasize that all environmental models represent simplified hypotheses of the real world, and that these hypotheses require rigorous construction, implementation, and testing. We then show how flexible frameworks can help as hypothesis-testing tools within the contexts of both the "bottom-up" and "top-down" model development and application strategies.

2.2. Toward Testable Hypotheses

[11] Although hydrological models are often recognized as hypotheses of catchment behavior [e.g., *Beven*, 2001; *Kuczera and Franks*, 2002; *Andréassian et al.*, 2009], a model is more akin to an *assemblage of coupled hypotheses*. The constituent hypotheses may include, for example, descriptions of subsurface flows through the soil matrix, hypotheses of surface runoff and base flow generation, and, more generally, hypotheses regarding dominant processes and the scale/resolution of their representation within the overall system architecture. The hypothesis-testing process must therefore seek to scrutinize (and perhaps reject) model components, as well as models in their entirety. However, the constituent hypotheses may not be testable at the integrated system level because individual model subcomponents interact in complex ways that are not distinguishable using aggregate measures of model performance.

[12] Hypothesis-testing in catchment-scale hydrology therefore requires both isolating *and* linking a myriad of model decisions (i.e., hypotheses) of different types and at different levels of conceptualization. Important model structure decisions include (1) delineating the system of interest, including its initial and boundary conditions and its forcing and response variables, (2) selecting the processes and state variables to include in the model (e.g., explicitly modeling nitrogen fluxes, canopy storage, and snowpack temperature), and (3) selecting among alternative representations of a particular process (e.g., different model representations of canopy interception). Decisions on model structure also include the critical choice of the appropriate model *architecture*, which ties together the individual elements of a model. Model architecture may include process separation (e.g., base flow versus interflow in a "fully lumped" model), spatial discretization into grid cells, subbasins, or land cover types (in a spatially distributed model) and/or vertical discretization into layers representing the soil, vegetation canopy and snowpack (in a vertically resolved model), and the model representation of subgrid heterogeneity. There are usually strong interdependencies and implications between modeling decisions at different levels. For example, representing infiltration using Richards' equation implies using (at least) a vertically resolved model.

[13] When approached from this perspective, hypothesis-testing requires that hydrological models be decomposed into a set of testable components (constituent hypotheses). Each such hypothesis can then be subjected to *independent* scrutiny, minimizing, wherever possible, the confounding interactions with other model hypotheses. For such hypothesis testing to be scientifically meaningful, the decomposition of a (hydrological) model into its constituent hypotheses must be carried out in a systematic manner. This requires explicitly identifying the (usually interrelated) individual decisions regarding system and process conceptualization, selection, and representation made during model development.

3. Scrutiny of Model Hypotheses: Are Current Approaches Adequate?

[14] In view of the Scientific Method of Hypothesis Testing [*Popper*, 1959], it has been recognized that the hydrologist often fails to rigorously scrutinize the constituent hypotheses within their models [e.g., *Mroczkowski et al.*, 1997; *Uhlenbrook et al.*, 1999; *Kuczera and Franks*, 2002; *Beven*, 2005; *Vaché and McDonnell*, 2006; *Sivakumar*, 2008]. Model development and refinement is often ad hoc, insufficiently documented, and, as noted by *McDonnell et al.* [2007], is still failing to adequately reflect field-based knowledge. Models are too often evaluated using subjective

and highly aggregated performance metrics, such as discrepancies between simulated and observed streamflows expressed solely using the Nash-Sutcliffe criterion, which do not directly test any individual hypothesis within the overall model [e.g., *Uhlenbrook et al.*, 1999; *Gupta et al.*, 2008]. Finally, in the absence of rigorous quantitative accounting for the uncertainty in the observed data, meaningful evaluation of a model is impossible: its failure can then at best be merely attributed to a nebulous mix of data and structural errors [e.g., as discussed by *Renard et al.*, 2010]. In our opinion, the absence of rigorous hypothesis testing is impeding scientific progress and preventing operational improvements in many areas of hydrology. It also necessarily reduces the confidence in the predictive abilities of current models.

3.1. Hypothesis Selection as Part of Model Development and Refinement

[15] The first opportunity to evaluate a model's constituent hypotheses is during the initial stages of model development. Ideally, a discerning model developer will carefully scrutinize each modeling decision and thoughtfully evaluate modeling alternatives (for example, on the basis of the literature and previous experience). However, although multiple alternatives may be considered when a model is developed, it is typical that only one approach is implemented and tested. For example, *Ivanov et al.* [2004] construct a physically based model of catchment hydrology that combines existing representations for the processes of rainfall interception, evapotranspiration, infiltration, groundwater flow, and runoff routing, yet only a single approach was selected for each model component, and experiments with alternative process representations were not reported. Omitting the analysis of alternative model representations is arguably quite common in hydrology; indeed, some of our own studies did not assess alternative process representations [e.g., *Clark et al.*, 2008b]. However, the ubiquity of a problem is no defense for neglect: lack of experimentation with alternative process representations can result in the model building and evaluation process being dominated by the individual, potentially biased, perspective of the modeler.

[16] Opportunities for model evaluation also arise during subsequent stages of model development and refinement, and may involve the inclusion of missing processes and/or refinement of existing representations. For example, when incorporating missing groundwater processes into existing land surface models, *Liang et al.* [2003] and *Niu et al.* [2007] adopted different groundwater representations, yet neither study evaluated possible alternatives and their impacts on the predicted land-atmosphere interactions. It is therefore difficult to judge whether the selected model enhancements are the most appropriate. Similarly, *Livneh et al.* [2010] address negative biases in land surface model simulations of snow by modifying the model's albedo formulation and snowpack temperature estimation, and by including a provision for the refreeze of liquid water in the snowpack. However, as is too common in practice, experiments with alternative representations of the processes they modified were not reported, again making it difficult to judge whether the model refinements are most appropriate.

[17] Even in cases where the developer experiments extensively with different model representations to address model deficiencies, the insights gained in these experiments often remain hidden because model failures are rarely fully reported in the peer-reviewed literature (with some notable exceptions, such as the Hydrological Monsters workshop described by *Andréassian et al.* [2010]). As a consequence, we (as a community) have acquired only limited knowledge of the comparative performance (and hence suitability) of different system representations.

3.2. Model Evaluation Along the Axis of Complexity

[18] Model development can also be based on complexity considerations [e.g., *Desborough*, 1999; *Atkinson et al.*, 2002]. For example, most applications of the "top-down" strategy for model improvement in hydrology [*Klemeš*, 1983; *Dooge*, 1986; *Sivapalan et al.*, 2003] progressively increased model complexity to improve model performance [e.g., *Jothityangkoon et al.*, 2001; *Atkinson et al.*, 2002; *Eder et al.*, 2003; *Farmer et al.*, 2003; *Bai et al.*, 2009]. This approach can produce parsimonious models that provide useful insights into catchment behavior. For example, *Bai et al.* [2009] used eight models of increasing complexity to evaluate the significance of subsurface flow routing in drier basins. Another important recent development is the process-based application of the top-down approach by *Son and Sivapalan* [2007], where an initial model was evolved along the axis of complexity using streamflow data alone, followed by testing and refinement of its internal structure using auxiliary data including observed groundwater levels and deuterium concentrations in streamflow. However, most practical applications of the top-down approach tend to consider a limited number of alternatives, and seldom consider competing process representations of equivalent complexity [e.g., *Jothityangkoon et al.*, 2001; *Atkinson et al.*, 2002; *Eder et al.*, 2003; *Farmer et al.*, 2003; *Bai et al.*, 2009]. This effectively restricts the investigation to the axis of complexity along a single branch of the model development tree, which limits the coverage of the model hypothesis space and may result in overlooking more plausible alternative model structures. In addition, by relying on the information content of catchment response data alone, the models developed with this approach are often perceived as unrealistic, in particular, over simplistic, by experimentalists [e.g., *Kirchner*, 2006].

3.3. Model Intercomparison Experiments

[19] Perhaps the closest we currently get to systematic hypothesis-testing in hydrology is during model inter-comparison experiments [e.g., *Wood et al.*, 1998; *Slater et al.*, 2001; *Reed et al.*, 2004; *Duan et al.*, 2006; *Breuer et al.*, 2009]. Surely they provide an extensive evaluation of modeling alternatives, since one of their intended aims is to understand inter-model differences [*Pitman and Henderson-Sellers*, 1998]? Yet, in our opinion, multimodel experiments to-date have been largely thwarted from fulfilling this objective for two main reasons. First, from a purely logistic perspective, when comparing an ad hoc collection of participating models, as is typical in current intercomparison studies, there are simply too many structural and implementation differences to meaningfully attribute the performance differences between any two models to specific individual components and hypotheses [*Koster and Milly*, 1997]. Second, the output of multicomponent models conveys only

limited information on the internal system states and fluxes. Hence, in studies where models are evaluated solely on the basis of aggregated output performance (e.g., goodness-of-fit of streamflow time series alone), the individual constituent hypotheses remain hidden from comparison and scrutiny [*Kuczera and Franks*, 2002]. We therefore conclude that model comparison studies are still a long way from reliably elucidating the appropriateness of different model representations.

3.4. Rejectionist Frameworks: Generalized Likelihood Uncertainty Estimation (GLUE)

[20] In recognition of the Popperian principle of falsification of testable hypotheses [*Popper*, 1959], a number of "rejectionist" frameworks have been proposed in hydrological sciences [e.g., *Beven*, 2002; *Vaché and McDonnell*, 2006]. Of these, the GLUE methodology [*Beven and Binley*, 1992; *Beven and Freer*, 2001], a broader philosophy that includes model inference, evaluation, and application, has been widely adopted in environmental studies (e.g., *Beven* [2006a] and references therein; see also *Stedinger et al.* [2008] for a critique). GLUE involves classifying a model as "behavioral" or "nonbehavioral" on the basis of summary statistics of model performance (e.g., behavioral models are those for which the sum of squared errors between simulated and observed streamflow is below a specified threshold). This is a form of hypothesis testing because it facilitates the rejection of model hypotheses that perform inadequately with respect to the evaluation criteria employed. For example, *Franks et al.* [1998] employed GLUE to scrutinize models using streamflow data and remotely sensed estimates of saturated areas; *Blazkova et al.* [2002] and *Freer et al.* [2004] used streamflow data and distributed water table measurements. These studies separated parameter sets into behavioral and nonbehavioral groups based on streamflow data alone, and then additional data on saturated areas and distributed water tables were used to further reject some of the "behavioral" parameter sets. Though in most applications of GLUE the different models corresponded to different parameter sets within a single model structure (e.g., uniformly sampled from the feasible parameter space), multiple model structures could also be evaluated using essentially the same approach [*Krueger et al.*, 2010].

[21] However, from our perspective, GLUE provides a superficial approach to model evaluation. In our opinion, a rejectionist framework is useful only inasmuch as it is based on reasonable measures to separate "good" from "bad" hypotheses. Yet GLUE offers little new insights into the key question of how to separate behavioral from nonbehavioral models: it rejects models based on a subjectively defined threshold in a subjectively defined pseudo-likelihood function. Furthermore, typical applications of GLUE do not even attempt to specify or infer any distinction between data and model errors. Instead, they lump a multitude of model and data problems into an inflated parametric uncertainty [*Beven*, 2006a], and, at least in applications to date, appear exempt from posterior scrutiny (e.g., on grounds of difficulties in deriving adequate rainfall uncertainty models [e.g., *Beven et al.*, 2008], and/or difficulties in representing epistemic uncertainties in the hydrological model structure using probability theory [e.g., *Beven*, 2008]). In doing so, rather than exposing model hypotheses to scrutiny, GLUE effectively justifies leaving them hidden behind a cloak of unresolved data and model errors.

[22] The original motivation for GLUE included the need for better hypothesis testing [*Beven*, 2001], and recent work has aimed to extend the GLUE methodology in this respect [*Beven*, 2006a]. For example, *Krueger et al.* [2010] consider a number of potentially testable hypotheses within the GLUE inference and, importantly, consider data errors when discriminating among competing hydrological models: Their trapezoidal penalty function for the residuals represents an assumed error model that combines streamflow and structural errors, while their rainfall input ensemble represents an assumed rainfall error model (albeit quite simplistic, using a six-member ensemble to characterize the uncertainty in 6 months of hourly data). These GLUE enhancements largely mimic the standard specification of error models within theoretically based statistical inference methods, and reflect the general trend in the broader environmental sciences toward a more careful treatment of uncertainties in environmental data since the original GLUE method was proposed 20 years ago. However, the GLUE extensions themselves still do not address the key issue of isolating constituent model hypotheses and subjecting them to independent scrutiny: any new techniques for model decomposition, evaluation, and improvement must be developed separately from GLUE and, as such, could be applied in other, more theoretically grounded, inference frameworks. Moreover, to the extent that the GLUE likelihood function components and rejection thresholds are not subjected to scrutiny and improvement (e.g., as required within a more formal application of Bayesian principles), it is our opinion that the GLUE approach does not adequately address the quest for rigorous evaluation of hydrological hypotheses.

4. Construction and Use of Multiple-hypothesis Approaches

4.1. From Single Models to Multiple-hypothesis Frameworks

[23] We introduce the term "multiple-hypothesis framework" to describe any modeling framework that facilitates experimenting with different ways to represent the behavior of a system. Note that this includes model frameworks developed primarily to evaluate model representations of increasing complexity [e.g., *Desborough*, 1999; *Fenicia et al.*, 2008b; *Bai et al.*, 2009; *Krueger et al.*, 2010; *Buytaert and Beven*, 2011], as well as model frameworks used to evaluate competing hypotheses of comparable complexity [*Moore and Clarke*, 1981; *Wagener et al.*, 2002; *Clark et al.*, 2008a; *Smith and Marshall*, 2010]. More broadly, we consider the multiple-hypothesis framework as an umbrella category, which includes "multiphysics" models used in the numerical weather prediction and land-surface modeling communities [*Jankov et al.*, 2005; *Niu et al.*, 2011], as well as frameworks designed to integrate model components or entire models [*Leavesley et al.*, 2002; *Kumar et al.*, 2006; *Pomeroy et al.*, 2007]. Provided multiple options are available for individual components and/or for the connectivity of the components, all of these modeling frameworks can be used for particular hypothesis testing

applications. Sections 4.2 through 4.5 elaborate on the requirements of multiple-hypothesis framework and on their practical limitations.

4.2. Key Requirements of a Multiple-hypothesis Framework

[24] The hypothesis-decomposition requirements outlined in section 2.2 can be accommodated within model frameworks that are flexible (and extensible) in their selection and representation of hydrological processes, including their overall connectivity within the model architecture. The following key aspects are of significance:

[25] 1. Support multiple alternative decisions regarding process selection and representation, e.g., based on a literature review, established theory, discussion with experimentalists, or on other prior perceptions. Consider the case of a physics-based snow model developed for streamflow forecasting applications. Relevant modeling decisions include, but are not limited to: (1) What stability function is used to compute turbulent heat fluxes?; (2) What method is used to represent snow albedo (function of time since last snowfall, or explicitly simulating grain growth)?; and (3) What method is used to represent interception and unloading of snow from the forest canopy? Allowing for empirical components, snow melt may also be represented as an empirical function of temperature. More generally, the number and complexity of decisions will depend on the size and complexity of the modeled system, and, in engineering contexts, also on the purpose of the model.

[26] 2. Accommodate different options for the model architecture, representing the connectivity between different model components. A key architectural consideration is the representation of heterogeneities, which can affect many different hydrological processes. For example, subgrid heterogeneity in land-surface models is commonly represented using a "mosaic" approach, by disaggregating a grid cell into a number of different vegetation types [e.g., *Koster and Suarez*, 1992; *Liang et al.*, 1994]. The mosaic approach accounts for the impact of subgrid heterogeneities in vegetation on fluxes such as canopy throughfall/drip, canopy evaporation, and transpiration, but does not typically consider horizontal fluxes of water within a grid cell, which may be important in order to account for the connections between vegetation type and water availability [e.g., *Tromp-van Meerveld and McDonnell*, 2006a]. Allowing for multiple architectural options enables a comparison of the mosaic approach with other model architectures, such as those where multiple vegetation types can coexist in a one-dimensional column [*Oleson et al.*, 2010], and those that allow for subgrid variability in soil moisture [e.g., *Famiglietti and Wood*, 1994].

[27] 3. The ability to separate the hypothesized model equations from their solutions, especially if the latter require numerical approximations. For example, for a continuous-simulation rainfall-runoff model, the governing equations should be formulated and reported in continuous-time state-space form, and only then approximated in discrete time [e.g., *Kavetski et al.*, 2003; *Young and Garnier*, 2006; *Clark and Kavetski*, 2010, and references therein]. Distinguishing between a hypothesis of hydrological behavior and its practical implementation (which may require additional approximations, linearizations, or smoothing) allows for a clearer physically oriented analysis unobscured by mathematical solution aspects. It also facilitates upgrading the model as more accurate and/or efficient solution techniques become available. In our opinion, the distinction between posing a hypothesis and implementing that hypothesis in a model appears all too often confused. This can seriously, and unnecessarily, compromise model development, analysis, and application [*Kavetski and Clark*, 2010, 2011].

[28] The modular approach to model development is consistent with the philosophy employed by *Beven and Kirkby* [1979] in TOPMODEL, which is presented as a set of concepts, rather than a fixed structure, based on the hypothesis that topographic indices control saturated areas and base flow. For example, *Ambroise et al.* [1996] explored linear versus power forms of the topographic index function suggested by a priori recession analysis. It also parallels the "multiphysics" options available in numerical weather prediction models and land-surface models [e.g., *Jankov et al.*, 2005; *Niu et al.*, 2011]. A key issue is the granularity of the decision tree. For example, as discussed next, a single model component could be used to represent the entire "soil zone" or, alternatively, distinct components could be used to characterize associated subprocesses, e.g., saturation-excess runoff or vertical drainage. Though some personal- and context-specific subjectivity in the granularity of model components is unavoidable, ideally we should strive to isolate as many modeling decisions as possible.

4.3. Relationship to General Modular Frameworks

[29] The multiple-hypothesis approach can be implemented within a modular framework (and, eventually, software) that provides the ability to mix and match different model components, and to select multiple options for each modeling decision. The implementation must also support the interdependencies that will often be encountered between different modeling decisions (e.g., as illustrated in section 5.4).

[30] Noting the key requirement of isolating individual model hypotheses, it is important to distinguish the aim of providing multiple representations of different physical processes from other possible aims of modular frameworks, such as the integration of existing models to simulate larger-scale systems. In a general modular modeling system, the different models or model components (e.g., components that represent hydrological processes in the soil zone, or components that represent the below-canopy snowpack) may have widely different philosophies with respect to key aspects such as degree of process conceptualization, spatial/temporal discretization, numerical approximation, and software implementation [e.g., *Leavesley et al.*, 2002; *Kumar et al.*, 2006; *Pomeroy et al.*, 2007]. The multiple and often significant differences between different models (and between model components) make it difficult to test hypotheses in a controlled fashion. From this perspective, the individual components used in many modular modeling systems are often too coarse: e.g., the inclusion of a submodel representing the entire soil zone lumps together several important modeling decisions. Hence, in order to maximize their utility for hypothesis testing, modular modeling frameworks should be designed with much finer granularity, and allow for multiple options for different modeling decisions.

4.4. Use of Multiple-hypothesis Frameworks for Systematic Scrutiny of Model Decisions

[31] There are several recent examples in the literature where multiple fine-grain hypotheses were compared. In a land-surface modeling study, *Niu et al.* [2011] used a multiple-hypothesis framework to compare different physically based options for representing turbulent heat transfer, soil moisture stress, and snow processes. The ability to isolate and compare individual modeling options allowed *Niu et al.* [2011] to attribute differences in overall model performance to specific modeling decisions. In a catchment hydrology study, *Clark et al.* [2011] evaluated alternative model representations of evapotranspiration, vertical drainage, surface runoff, and base flow in an experimental basin where several data sources where available. *Clark et al.* [2011] only used five model structures, with these structures carefully selected to isolate differences in model components. Also consider in this context the study of *Kavetski et al.* [2011], where several rainfall-runoff models were analyzed across a range of time scales: the analysis considered model hypotheses of different structure and complexity, implemented using different numerical methods, and calibrated using different inference schemes. The analysis was conducted in a controlled way, so as to disentangle the effects of model complexity, data resolution, and numerical time stepping schemes on the models' ability to reproduce hydrological signatures of interest. These examples provide an initial demonstration of the effectiveness of multiple-hypothesis frameworks for model evaluation, and we anticipate future studies along these lines will further advance our understanding of various modeling options.

[32] More generally, the flexibility in the selection of model architecture and components can be exploited to design various strategies for a controlled and thorough exploration of the hypothesis space. For example, multiple options for a single model component could be evaluated against observed data using multifaceted model diagnostics while keeping the rest of the model fixed. Next, multiple options for a single model component could be trialed within different model architectures. Interactions between multiple model decisions can be examined next, under the same overall approach. This multistage model evaluation strategy is analogous to the Sobol strategy for parameter sensitivity analysis [*Saltelli*, 2002] applied using multiple sensitivity metrics. Restricting the variations to one (or few) components at a time provides a more controlled model evaluation, whereas evaluating variants of one model component within multiple parent model structures provides a more comprehensive evaluation of model alternatives. In many cases, the range of model alternatives can be narrowed down based on process understanding, whereas in other cases a wider range of options must be considered.

4.5. Practical Limitations

[33] Especially in the earlier stages, practical implementations of multiple-hypothesis frameworks necessarily requires trade-offs between model flexibility, complexity, comprehensiveness, and computational cost. For example, current applications of the FUSE approach remain quite limited in scope because they include only comparatively simple representations of the soil zone and do not directly resolve spatial variability [*Clark et al.*, 2008a]. Such limitations often arise for pragmatic reasons: resource constraints inevitably reduce the number and complexity of modeling decisions (hypotheses) that are included and/or considered within a particular application. At least initially, it may also be necessary to nest the new multiple-hypothesis configurations within one or more parent models.

[34] Computational costs impose another important constraint on the exploration of different model representations of environmental behavior, requiring explicit or implicit trade-offs between the range of modeling decisions considered and the spatial resolution and size of the model domain. For example, given a practical set of resources and computational budgets, should a modeler reduce the number of competing hypotheses under investigation in order to use higher spatial resolution? Or use coarser discretization in order to simulate a larger domain size?

[35] While practical constraints necessarily affect all applied sciences, the multiple-hypothesis approach is entirely general and offers solid prospects for rigorous hypothesis testing in increasingly realistic environmental modeling contexts [e.g., *Niu et al.*, 2011]. Though still in its nascence, this progress gives us confidence that the ability of multiple-hypothesis frameworks to isolate individual model hypotheses, when combined with increased availability and more intelligent use of data, will lead to more scientifically defensible environmental models.

5. Requirements for Meaningful Hypothesis Testing

[36] At its core, hypothesis testing aims to scrutinize the consistency of the predictions of a hypothesis against empirical observations [*Popper*, 1959]. In disciplines such as physics, where the experimental conditions can be carefully controlled, it is often possible to rigorously apply concepts of statistical significance [e.g., *Lehmann and Romano*, 2005]. In hydrology and many other environmental disciplines, events of interest may be infrequent or nonrepeatable, and the uncertainty in the observations is seldom fully characterized. In these common cases, it may be preferable to use Bayesian approaches, where model hypotheses are evaluated more subjectively in light of both the plausibility of the model hypothesis (which could be expressed on the basis of prior process understanding) and of the data used in a particular evaluation. This section outlines more specific requirements for carrying out informative hypothesis testing, and surveys some of the recent developments and applications in this direction.

5.1. The Need for Stringent Model Diagnostics and Clever Use of Data

[37] Crafting a model as a set of testable hypotheses must proceed hand-in-hand with the development and application of stringent model diagnostics that challenge both individual constituent hypotheses and the overall model architecture. As discussed by *Kuczera and Franks* [2002], a major challenge "is to expose internal variables to scrutiny. This is not a trivial challenge, but one that must be vigorously pursued if conceptual catchment modeling is to avoid degenerating into a sterile curve-fitting exercise." Similarly, *Gupta et al.* [2008] outline the limitations of

traditional aggregate metrics (such as the sum of squared differences between model simulations and observations of the overall system response), and highlight the need for incisive model diagnostics that can scrutinize different subcomponents of a model. This lesson has been repeatedly highlighted by experimentalists, who show that scrutiny of multivariate data collected in experimental basins not only exposes major model deficiencies, but may be indispensable for improving model realism [e.g., *Seibert and McDonnell*, 2002; *Uhlenbrook et al.*, 2004; *Weiler and McDonnell*, 2004; *Vaché and McDonnell*, 2006; *McGuire et al.*, 2007; *Sayama and McDonnell*, 2009; *Birkel et al.*, 2010]. Two main categories of model diagnostics (hypothesis-testing tools) can be employed:

[38] 1. The improved use of traditional data, in a way that focuses hypothesis testing on reproducing hydrological behavior rather than merely matching data with model simulations. For example, streamflow measurements can support a richer set of diagnostics than traditional time series analysis alone. Hydrographs can be separated into recession periods versus periods that are actively "driven" by rainfall [e.g., *Boyle et al.*, 2000], or used to generate a set of indices (diagnostic signatures) that have explanatory power for different processes within a model [*Gupta et al.*, 2008]. For example, *Yilmaz et al.* [2008] use the slope of the flow duration curve to evaluate the model representation of vertical drainage. Note that the application of some model diagnostic measures may require additional hypotheses.

[39] 2. Using new types of data. In many locales, hydrological investigations remain thwarted by a dearth of data. In particular, much of the data needed for better model development is only available in a small set of experimental catchments, and, even then, scrutinizing certain model hypotheses may require data for which measurement technologies are still unreliable, or not yet available. One of the major challenges in catchment-scale hydrology is therefore to design effective methods to measure fluxes and storages at the relevant spatial scales [e.g., *Beven*, 2006b].

[40] Yet, at least in experimental catchments, notable progress is already apparent in collecting and utilizing independent data on internal hydrological processes. For example, *Western et al.* [2004] explored the spatial variability of soil moisture to understand the dominant processes controlling soil moisture patterns, *Tromp van-Meerveld and McDonnell* [2006b, 2006c] measured the spatial variability of water table depth to diagnose the controls of bedrock topography on hydrological connectivity and storm runoff, and *Vaché and McDonnell* [2006] used isotopic estimates of residence time to diagnose the heterogeneity of flow paths within a catchment. In land-surface modeling, it is common to evaluate models using eddy-covariance measurements of sensible and latent heat [e.g., *Abramowitz et al.*, 2008]. Advances are also evident in the development of techniques for measuring hydrological processes over larger spatial scales, including the use of cosmic ray sensors for soil moisture estimation [*Zreda et al.*, 2008], the use of GPS for soil moisture and snow monitoring [*Larson et al.*, 2009, 2010], the use of terrestrial scanning lidar for snow monitoring [*Prokop*, 2008; *Hood and Hayashi*, 2010], and the use of fiber optics for distributed temperature sensing [e.g., *Selker et al.*, 2006; *Tyler et al.*, 2009]. Limitations notwithstanding, new sources of information are clearly beneficial, if not critical, for model development and testing. Experimental endeavors must therefore be vociferously encouraged and generously funded.

5.2. Accounting for Both Prior Information and Data Uncertainty Within Hypothesis Testing

[41] The issue of exploiting traditional and new types of data for model analysis brings us to the thorny issue of balancing prior expectations with the uncertainty in the data. Practical issues such as data availability and data quality necessarily affect the insights that can be gained in a particular hydrological system. Put simply, *data uncertainty constrains our ability to discriminate among competing hydrological hypotheses*. If there is a strong prior confidence in a particular model hypothesis (e.g., from theory, or from data collected earlier, or from data collected in other basins), strong new evidence will be needed to reject that model hypothesis. Addressing this critical issue requires a careful analysis of the sampling and measurement errors of observational systems and carefully reflecting this observational uncertainty in model inference, analysis, and prediction. For example, *Rodriguez-Iturbe and Meija* [1974], *Morrissey et al.* [1995], *Krajewski et al.* [2003], and *Villarini et al.* [2008] present approaches to rain gage error analysis; *Di Baldassarre and Montanari* [2009] and *McMillan et al.* [2010] discuss rating curve errors; and *Kavetski et al.* [2006] and *Renard et al.* [2010] formulate methods for incorporating data error models into model analysis.

5.3. Avoiding Simplistic Descriptions of Complex Systems

[42] When contemplating the identifiable complexity of a system given limited quantitative data [e.g., *Jakeman and Hornberger*, 1993; *Perrin et al.*, 2001, 2003; *Schoups et al.*, 2008; *Pande et al.*, 2009; *van Dijk*, 2010], it is customary to refer to Occam's razor to shear unwarranted complexity from a model [e.g., *Young et al.*, 1996; *Perrin et al.*, 2001]. However, this must be done thoughtfully, without imposing simplistic solutions on complex problems. Indeed, developing a model that "works for the right reasons" [*Kirchner*, 2006] is likely to require the philosophy purportedly attributed to Albert Einstein that "everything should be as simple as possible, *but not one bit simpler*" (italics added). For example, the apparent structural simplicity suggested in the absence of accurate quantitative ("hard") data in a particular application should be judged against independent knowledge available from general hydrological theory and/or any qualitative ("soft") fieldwork evidence. In other words, model evaluation requires a mix of qualitative and quantitative insights [e.g., *Seibert and McDonnell*, 2002; *Young and Ratto*, 2009; *McMillan et al.*, 2011; *Clark et al.*, 2011; *Kavetski et al.*, 2011].

[43] While a modeler will generally strive to construct a model that is a "complete" representation of the system of interest, evaluation of this model is limited by inevitable practicalities such as data limitations (see section 5.2). We may therefore question if inference drawn solely from limited and highly uncertain hydrological data justifies the scientific acceptance of models that appear simplistic when judged against fieldwork evidence (e.g., *Jakeman and Hornberger* [1993] indicate that rainfall-runoff data alone

supports the inference of models with 4–5 parameters at most, in the case studies they examined). Or we may question if high resolution data from a densely gaged experimental catchment can support the inference of a more complex model [*Kavetski et al.*, 2011]. Hence, iterative model improvement, which aims to take advantage of new information, generally leads to new hypotheses being proposed in response to new data and/or other new independent insights and theories [e.g., *Son and Sivapalan*, 2007; *Fenicia et al.*, 2008]. These and other questions can be approached using a multiple-hypothesis framework, and are relevant if we are to overcome the fragmented status quo and move toward more unified catchment-scale modeling theories [*Sivapalan*, 2009].

5.4. Recognizing Interactions Between Process Descriptions and Model Architecture

[44] Hypothesizing individual process descriptions requires considering not only the choice of representation method, but also how the method is implemented within the overall model architecture. Understanding how different model development decisions interact is therefore important in order to both isolate key modeling decisions and to design experiments and diagnostics that evaluate their impact on the overall system response. For example, at the integrated model level we may question if the spatial variability in transpiration matches observations, e.g., does the model adequately represent the dominance of transpiration in riparian areas? We may then question if the model captures the relative controls of catchment-average transpiration by soil moisture and depth to the water table, and if the model suitably represents hillslope–riparian interactions.

[45] As another example, consider alternative model representations of surface runoff at the catchment scale. TOPMODEL and VIC surface runoff representations are analogous in terms of the functional dependence of contributing areas on storage [e.g., *Sivapalan et al.*, 1987; *Kavetski et al.*, 2003], but can behave very differently depending on how they are incorporated into a multicomponent model [*Clark et al.*, 2008a]. For example, runoff-generating areas in TOPMODEL are conceptualized as dependent on saturated zone storage (i.e., the depth to the water table), whereas in VIC they are formulated as dependent on unsaturated zone storage. These differences affect the expansion and contraction of runoff-generating areas and hence the streamflow response dynamics of the catchment [*Clark et al.*, 2008a]. Therefore, the hypothesized overall model architecture within which individual process representations are embedded must also be subjected to scrutiny and evaluation.

[46] Note that the fidelity of model simulations depends both on the appropriate choice of the model equations as well as on the choice of model parameter values. In practice, the distinction between the model "structure" (the model equations) and the model "parameters" (the adjustable coefficients in the model equations) is often rather subjective and imprecise. As a contrived example, the "different" structural equations $q = kx$ and $q = kx^{\alpha}$ are identical when $\alpha = 1$. But more generally, "different" algebraic expressions, such as $\cos(kx)$, x^k, $\exp(kx)$, "different" systems of differential equations, and other mathematical constructs, can behave functionally very similarly depending on the range of application and parameter values. The appropriate choice of model equations and the appropriate values of model parameters are therefore both hypotheses that should be subject to careful scrutiny.

6. Benefits of Multiple-hypothesis Methods

6.1. Guidance for Model Selection and Improvement

[47] Multiple-hypothesis frameworks facilitate controlled and comprehensive model comparisons, which provide guidance for model improvement. In the standard approach to model development (in which only one approach is implemented and tested) the model developer may *believe* their approach is suitable, but they often have little information and capabilities to investigate if alternative approaches are more suitable for their intended purpose. Multiple-hypothesis approaches provide a systematic framework for generating and comparing competing hypotheses, and hence significantly facilitate model improvement, both in general and site-specific contexts (for example, where independent evidence may favor a particular modeling decision on the basis of additional data or previous investigations). Moreover, by comparing model representations at the level of model subcomponents it becomes possible to select the best component hypotheses from different models, thereby avoiding the need to reject entire models (this makes better use of insights gained during model development).

[48] Multiple-hypothesis frameworks also facilitate examining trade-offs between complexity and practicality more systematically, in particular, with respect to the computational costs associated with mathematical representations of specific processes. For example, modeling unsaturated flows using Richards' equation at the scale of its constitutive functions may require resolution beyond the current data and computational resources. Hence, one-dimensional models of infiltration over depths of several meters are typically implemented using only 5–10 soil layers [e.g., *Boone and Wetzel*, 1996]. In a practical context, where it may be reasonable to assign computational budgets to model components depending on their relative importance, multiple-hypothesis frameworks can also be used to isolate individual decisions (such as the number of soil layers) and facilitate more informed pragmatic trade-offs between model complexity and computational expense.

[49] Multiple-hypothesis model approaches may also reduce biases in model selection arising from the understandable subjectivity in human judgment. A century earlier, *Chamberlin* [1890] suggested that scientists develop "parental affection" for their theories, and advocated the method of multiple working hypotheses where "the effort is to bring up into view every rational explanation of new phenomena . . . the investigator then becomes parent of a family of hypotheses, and, by his parental relation to all, he is forbidden to fasten his affections unduly upon any one." Employing multiple-hypothesis approaches in a rigorous and quantitative manner can reduce undue favoritism arising from individual perspectives [see also *Holländer et al.*, 2009].

6.2. Confronting Ambiguities in the Apparent/identifiable System Structure

[50] The term "equifinality" is often used in hydrology to describe situations where multiple parameter sets and/or entire model structures appear equally plausible. For example,

different models may generate near-identical predictions, or have near-identical indices of model performance [e.g., see *Beven and Freer*, 2001; *Beven*, 2006a]. This is a manifestation of "nonidentifiability," a key statistical inference limitation encountered in many scientific fields, especially in data-scarce contexts [e.g., *Tarantola*, 2005; *Renard et al.*, 2010].

[51] In the majority of current hydrological analyses, model nonidentifiability should not be surprising. Given the potential interactions between the different components of a catchment model, the behavior of many different configurations and parameter values may be indistinguishable when evaluated against a single (and uncertain) response time series "*unless the detailed characteristics of these components can be specified independently*" (italics added) [*Beven and Freer*, 2001], or unless additional types of data are available [e.g., *Freer et al.*, 2004; *Fenicia et al.*, 2008a]. The key word from *Beven and Freer* [2001] is "unless"; the challenge for the community is to identify ways to independently estimate internal storages and fluxes of water in a catchment (and importantly, associated uncertainties), and exploit these estimates either as qualitative or quantitative diagnostic tools [e.g., *Gupta et al.*, 2008; *Kollet and Maxwell*, 2008a; *Lawrence et al.*, 2011], or directly build them into the inference [e.g., *Seibert*, 2000; *Fenicia et al.*, 2008a]. Recent progress notwithstanding, certain model identification ambiguities are likely to persist in the foreseeable future, especially in describing operational, let alone ungaged, catchments.

[52] Multiple-hypothesis frameworks may help quantify the predictive uncertainty stemming from system nonidentifiability by generating ensembles of competing model representations, both of equal and varying complexity. For example, available data can be used to hypothesize a set of reasonable model architectures and components for a given catchment, and the corresponding ensemble of models then used to represent structural uncertainty because of system nonidentifiability. A major unresolved challenge for the ensemble method to work is to ensure that the ensemble includes at least one hypothesis that approximates "reality" within the range of data uncertainty or, more leniently, within the design requirements of an application. It is also necessary to avoid cases where all model representations are wrong for the same reasons (e.g., in view of the arguments of section 5.4, making an artificial distinction between "structures" and "parameters" can result in degenerate model ensembles). By formulating the multiple working hypotheses at levels ranging from system architecture down to process subcomponents, a multiple-hypothesis framework can offer a much broader coverage of the model space than current multimodel approaches [e.g., *Marshall et al.*, 2007; *Hsu et al.*, 2009; *Bohn et al.*, 2010], in which a small number of individual models of varying complexity are included, often on ad hoc considerations (see also section 3.2 on the axis-of-complexity).

6.3. Understanding Regional Differences in Catchment Behavior

[53] Another important challenge is understanding "uniqueness of place" [*Beven*, 2000]. In a recent opinion paper, *Andréassian et al.* [2009] follow up on the seminal work by *Klemeš* [1986] and suggest that, in addition to transposability in time, hydrological models should also be transposable in space, in particular, under "very different climatic conditions" and, presumably, in basins with (very) different geology. However, efforts to identify a "universal" catchment-scale model have arguably been generally unsuccessful to date, as suggested by the results of calibrating single models across hundreds of catchments [*Le Moine et al.*, 2007], and by attempts to estimate model parameters a priori from spatial data on soils and vegetation [*Reed et al.*, 2004; *Duan et al.*, 2006]. While a spatially transposable model may yet be achieved (indeed, *Andréassian et al.* urge pursuing this quest more vigorously) it may be that a unique set of equations for catchment-scale dynamics applicable over the entire range of environmental systems simply does not exist. For example, spatially distributed models developed on the basis of Darcy's law for porous media may not be appropriate to simulate vertical water movement in mountainous scree slopes, or to simulate groundwater flow in Karst catchments. Alternatively, it may be that a unique set of equations for catchment-scale dynamics is not identifiable from the kind of data currently used in model development and evaluation [e.g., *Reed et al.*, 2004; *Duan et al.*, 2006]. Answering these questions hinges critically on addressing the challenges listed earlier.

[54] While there are clearly several distinct perspectives for exploring "uniqueness of place" (e.g., identifying different model structures to describe different dominant processes in different hydrological landscapes, or continuing the pursuit of the hitherto elusive universal model at the catchment scale), a common consideration is to ensure that the model architecture reflects the connectivity between small-scale processes and the system scale response. In particular, it is important that models represent the influence of the landscape on the partitioning, storage and release of water at the catchment scale [*Wagener et al.*, 2007; *Kumar*, 2011]. Multiple-hypothesis frameworks may hence be used to test hypotheses on intercatchment differences in hydrological behavior arising from differences in climate, vegetation, topography and soils, as well as differences in the evolutionary history of the landscape and human activities [e.g., *Savenije*, 2009; *Sivapalan*, 2009]. Catchment-scale signatures, such as flow duration curves, provide insights into catchment-scale function [e.g., *Farmer et al.*, 2003; *Wagener et al.*, 2007; *Kavetski et al.*, 2011], and these signatures are indispensible in evaluating the mapping between model architecture and landscape architecture. Experimenting with different model configurations in multiple catchments is therefore an informative learning exercise that helps us understand the hydrological functioning at the catchment scale across different hydrological landscapes.

6.4. Combining A Priori and Data-Based Hypotheses of Model Structure

[55] So far, we have focused on methods for essentially a priori formulation of model hypotheses (e.g., from theory, or using perceptual insights, or prior fieldwork evidence), followed by posterior diagnostics and improvement strategies. Yet a number of important techniques approach model development from a different perspective. They make fewer a priori assumptions regarding model structure and instead try to let the data generate and/or refine the constituent hypotheses describing system behavior. In

hydrology, these include recession curve analysis [*Lamb and Beven*, 1997; *Clark et al.*, 2009; *Kirchner*, 2009], data-based mechanistic (DBM) modeling [*Young*, 1998, 2003], and, more recently, a Bayesian approach [*Bulygina and Gupta*, 2009]. As we argue next, though not without their own set of limitations, these methods are not only well suited to hypothesis testing in hydrology [e.g., *Young*, 2003; *Young et al.*, 1996], but can be readily exploited within multiple-hypothesis frameworks such as those we propose here.

[56] For example, consider recession analysis, which focuses on empirical identification of the base flow function from data periods where quick flow processes are assumed dormant [*Lamb and Beven*, 1997; *Kirchner*, 2009]. In its simplest form, it uses a single-state variable to describe base flow processes, which may overlook base flow generation from different storage units and landscape types. The combination of insights from recession analysis with more traditional modeling approaches can be powerful. For example, it can be used to specify the form of the base flow function of a more complex multicomponent model [*Atkinson et al.*, 2003; *Fenicia et al.*, 2006]. Yet care must be taken: e.g., seemingly nonlinear reservoir behavior could instead be a manifestation of a linear reservoir with inflow [*Fenicia et al.*, 2006] or several linear reservoirs [*Clark et al.*, 2009; *Harman et al.*, 2009].

[57] The data-based mechanistic (DBM) modeling framework [*Young*, 1998, 2003] is another, more formal, technique that aims to let the data, rather than prior hypotheses, dictate the mathematical structure of the model. The model structure in the DBM scheme is formulated using transfer functions. The nonlinearity of hydrological systems is then approximated either using a nonlinear transformation of streamflow to obtain "effective rainfall" (and assuming the remaining routing system is linear), or using time- and state-dependent parameters in the transfer function model [*Young*, 2003]. While this can be restrictive in general modeling contexts [*Reichert and Mieleitner*, 2009], the DBM method has been useful not only in hydrology, but also across broader environmental sciences [*Young*, 1998]. Furthermore, there is scope to use the DBM method to combine "data-based" techniques with "reductionist" approaches on the basis of the prior hypotheses of the system [e.g., *Young and Ratto*, 2009].

[58] More recently, *Bulygina and Gupta* [2009, 2010] proposed a nonparametric Bayesian approach to more directly explore the probabilistic mapping between rainfall and runoff using mixtures of Gaussian distributions, conditioning the inference on (1) some prior analysis of the overall system structure (e.g., the number of state variables which can be estimated using the false neighbor method [*Bulygina and Gupta*, 2009], or taken from an existing model [*Bulygina and Gupta*, 2010]) and (2) on assumed data error models (a current limitation that could be remedied using observational network analysis e.g., *Willems* [2001]; see *Renard et al.* [2010], for further discussion).

[59] In our opinion, the Bayesian paradigm is particularly attractive for hypothesis testing in environmental sciences, offering the capability to directly describe system nonlinearities and data uncertainties [e.g., *Kavetski et al.*, 2002; *Vrugt et al.*, 2008; *Cressie et al.*, 2009; *Hsu et al.*, 2009; *Renard et al.*, 2010; and many others], while exploiting independent process understanding as prior knowledge [e.g., *Bulygina and Gupta*, 2009]. The use of nonparametric probabilistic techniques to approximate epistemic structural uncertainties [e.g., as shown by *Bulygina and Gupta*, 2010] is an encouraging advance: it facilitates the development of models that are more consistent with the functional view of catchment dynamics advocated in section 2.1, with fewer constraints arising from the mathematically convenient yet often restrictive forms of particular parametric relationships. There is also scope to exploit control-theory identification techniques such as dominant mode analysis [e.g., *Young and Ratto*, 2009] to help hypothesize the overall model architecture. Outstanding challenges in the development of Bayesian structural inference include its extension to more complex multistate models (including spatially distributed contexts), the independent derivation of reliable and precise data uncertainty models (e.g., using observational network analysis), as well as the use of more probing, physically oriented hypothesis-testing methods and independent information to cope with remaining nonidentifiabilities and ambiguities in the model inference and interpretation.

6.5. Looking for "New" Laws and Addressing the Closure Problem

[60] Hypothesis testing using multiple-hypothesis frameworks can advance not only process representations in the current suite of hydrological models, but be used to rigorously implement and evaluate new hydrological theories. In particular, our current inability to adequately quantify the impact of subcatchment heterogeneities on the catchment's hydrological response can be related to the problem of "closure" (here, quantifying the relationships between water, energy, and momentum fluxes) at the catchment scale [*Reggiani et al.*, 1998, 1999]. Catchment-scale closure has remained an elusive challenge, recently referred to as "The Holy Grail" of catchment-scale hydrology [*Beven*, 2006b]. A key difficulty has been ensuring that flux calculations based on space-time averaged properties of a medium (the catchment) are sufficiently representative of aggregated behavior over potentially highly heterogeneous smaller scales. For example, using Richards' equation to represent the unsaturated zone within a hydrological model with spatial resolution on the scale of hundreds of meters necessarily assumes that the governing equations, identified at the "lab" scale, remain valid even when applied using "effective" parameters that implicitly represent the heterogeneity of the subsurface at the scale of the model discretization (e.g., see *Nordbotten et al.* [2007] for an illustration using the Darcy equation).

[61] Changing the scale at which a process is described requires alternative modeling methodologies. "Physically based" distributed models derived a priori from smaller-scale physical laws could, at least in principle, close the scale gap by directly aggregating small-scale heterogeneous behavior to larger scales. This can be done numerically, e.g., *Kollet et al.* [2010] use supercomputing resources to demonstrate a proof-of-concept variably saturated groundwater model configured at hydrological resolution, with billions of grid cells. Alternatively, changing the scale at which a process is described may require changing the form of the governing equations and constitutive relations

(e.g., see the analytical work by *Reggiani et al.* [1998, 1999]; and the discussion of principles by *Kirchner* [2006]). It is also possible that upscaling hydrological dynamics to entire catchments and beyond will require abandoning the mathematical convenience of deterministic models and mandate a more fundamental shift toward stochastic descriptions. For example, subscale variability in catchment properties and forcings may inherently prevent a deterministic prediction of the system response given coarsely aggregated forcing data (e.g., *Kuczera et al.* [2006]; see also the use of random fields to parameterize subsurface hydraulic properties, e.g., *Kollet and Maxwell* [2008b]). Multiple-hypothesis methodologies can help understand the differences among different modeling paradigms.

7. Final Perspectives

[62] The hypothesis-based method is entirely general and can be applied to search for and evaluate "new" hydrological laws [*Dooge*, 1986], identify the dominance of different hydrological processes [*Sivakumar*, 2008], explore the impact of hydrological connectivity on catchment response [*Western et al.*, 2004], and other endeavors. New modeling approaches are motivated by the lure of novel theories of hydrological functioning at the catchment scale. For example, *McDonnell et al.* [2007] suggest moving beyond a mere description of heterogeneities to a broader analysis of the self-organizing and optimality principles that may be responsible for the emergence and maintenance of hydrological and larger environmental systems. Disciplines such as geomorphology, soil science, biogeochemistry, and ecology can provide useful insights into how hydrological systems have evolved and why certain patterns and functions, such as the Budyko curve, emerge over increasing space and time scales [e.g., *Sivapalan*, 2005; *Schymanski et al.*, 2007, 2009; *Troch et al.*, 2009]. Looking at the problem through a different lens, such as optimality, provides additional metrics that can be used to falsify model hypotheses [*Schymanski et al.*, 2007, 2008, 2009; *Schaefli et al.*, 2011]. A multiple-hypothesis methodology, where competing hypotheses can be systematically constructed and evaluated within a single robust numerical framework, holds as much promise for testing new hydrological theories as for testing competing model representations within a "traditional" catchment model. The novel perspectives and new sources of information being uncovered through interdisciplinary collaboration introduce exciting opportunities to advance hydrological science.

[63] In conclusion, we argue that the ongoing quest for physically realistic catchment-scale models, including more appropriate representations of heterogeneous hydrological processes, needs to be embedded in a hypothesis-testing framework that rigorously scrutinizes hypotheses against observed data. It is our proposition that this is best achieved using multiple-hypothesis frameworks, where different process representations and overall system hypotheses can be evaluated in a controlled and relatively independent way. We are optimistic that, when model hypotheses are stringently tested using available data from both experimental watersheds and operational observing networks, multiple-hypothesis approaches can become useful learning tools and lead to considerably more scientifically defensible and operationally reliable hydrological models.

[64] **Acknowledgments.** We are grateful to Andrew Barrett, David Gochis, Hoshin Gupta, Ethan Gutmann, Thomas Hopson, Roy Rasmussen, and David Rupp for valuable comments on an earlier version of the manuscript. We also thank editor Praveen Kumar and the reviewers, including Kellie Vaché, Murugesu Sivapalan, Valeriy Ivanov, and an anonymous reviewer, for their constructive and insightful criticisms, comments, and suggestions, which have greatly improved this paper. This research was partially funded by the NOAA Climate Program Office under grant R4310142.

References

Abramowitz, G., R. Leuning, M. P. Clark, and A. Pitman (2008), Evaluating the Performance of Land Surface Models, *J. Clim.*, *21*, 5468–5481, doi:10.1175/2008JCLI2378.1.

Ambroise, B., K. Beven, and J. Freer (1996), Toward a generalization of the TOPMODEL concepts: Topographic indices of hydrological similarity, *Water Resour. Res.*, *32*(7), 2135–2145, doi:10.1029/95WR03716.

Andréassian, V., C. Perrin, L. Berthet, N. Le Moine, J. Lerat, C. Loumagne, L. Oudin, T. Mathevet, M.-H. Ramos, and A. Valery (2009), Crash tests for standardized evaluation of hydrological models, *Hydrol. Earth Syst. Sci.*, *13*, 1757–1764, doi:10.5194/hess-13-1757-2009.

Andréassian, V., C. Perrin, E. Parent, and A. Bardossy (2010), The Court of Miracles of Hydrology: Can failure stories contribute to hydrological science?, *Hydrol. Sci. J.*, *55*, 849–856.

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

Atkinson, S. E., M. Sivapalan, R. A. Woods, and N. R. Viney (2003), Dominant physical controls on hourly flow predictions and the role of spatial variability: Mahurangi catchment, New Zealand, *Adv. Water Resour.*, *26*, 219–235, doi:10.1016/S0309-1708(02)00183-5.

Bai, Y., T. Wagener, and P. Reed (2009), A top-down framework for watershed model evaluation and selection under uncertainty, *Environ. Modell. Software*, *24*, 901–916, doi:10.1016/j.envsoft.2008.12.012.

Beven, K. J. (2000), Uniqueness of place and process representations in hydrological modelling, *Hydrol. Earth Syst. Sci.*, *4*(2), 203–213, doi:10.5194/hess-4-203-2000.

Beven, K. J. (2001), On hypothesis testing in hydrology, *Hydrol. Process.*, *15*, 1655–1657, doi:10.1002/hyp.436.

Beven, K. J. (2002), Towards an alternative blueprint for a physically based digitally simulated hydrological response modelling system, *Hydrol. Process.*, *16*, 189–206, doi:10.1002/hyp.343.

Beven, K. J. (2005), On the concept of model structural error, *Water Sci. Technol.*, *52*, 167–175.

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

Beven, K. J. (2006b), Searching for the Holy Grail of scientific hydrology: Qt = H(S, R, dt) A as closure, *Hydrol. Earth Syst. Sci.*, *10*, 609–618.

Beven, K. J. (2008), On doing better hydrological science, *Hydrol. Process.*, *22*, 3549–3553, doi:10.1002/hyp.7108.

Beven, K. J., and M. J. Kirkby (1979), A physically based, variable contributing model of basin hydrology, *Hydrol. Sci. Bulletin*, *24*, 43–69.

Beven, K. J., and A. M. Binley (1992), The future of distributed hydrological models: model calibration and uncertainty prediction, *Hydrol. Process.*, *6*, 279–298, doi:10.1002/hyp.3360060305.

Beven, K. J., and J. E. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology, *J. Hydrol.*, *249*, 11–29, doi:10.1016/S0022-1694(01)00421-8.

Beven, K. J., P. J. Smith, and J. E. Freer (2008), So just why would a modeller choose to be incoherent?, *J. Hydrol.*, *354*(1–4), 15–32.

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

Blazkova, S., K. Beven, P. Tacheci, and A. Kulasova (2002), Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): The death of TOPMODEL?, *Water Resour. Res.*, *38*(11), 1257, doi:10.1029/2001WR000912.

Blöschl, G. (2006), Hydrologic synthesis: Across processes, places, and scales, *Water Resour. Res.*, *42*, W03S02, doi:10.1029/2005WR004319.

Bohn, T. J., M. Y. Sonessa, and D. P. Lettenmaier (2010), Seasonal hydrologic forecasting: Do multi-model ensemble averages always yield improvements in forecast skill?, *J. Hydrometeor.*, *11*, 1358–1372, doi:10.1175/2010JHM1267.1.

Boone, A., and P. J. Wetzel (1996), Issues related to low resolution modeling of soil moisture: Experience with the PLACE model, *Global and Planetary Change*, *13*, 161–181.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, *36*, 3663–3674, doi:10.1029/2000WR900207.

Breuer, L., et al. (2009), Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM) I: model intercomparison of current land use, *Adv. Water Resour.*, *32*, 129–146, doi:10.1016/j.advwatres.2008.10.003.

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

Bulygina, N., and H. Gupta (2010), How Bayesian data assimilation can be used to estimate the mathematical structure of a model, *Stochastic Environmental Research and Risk Assessment*, *24*, 925–937, doi:10.1007/s00477-010-0387-y.

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

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire (1973), *A generalized streamflow simulation system: Conceptual modeling for digital computers*, Technical Report, U.S. National Weather Service, Sacramento, California.

Buytaert, W., and K. Beven (2011), Models as multiple working hypotheses: Hydrological simulation of tropical alpine wetlands, *Hydrol. Processes*, doi:10.1002/hyp.7936.

Chamberlin, T. C. (1890), The method of multiple working hypotheses, *Science* (*old series*), *15*, 92.

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

Clark, M. P., D. E. Rupp, R. A. Woods, X. Zheng, R. P. Ibbitt, A. G. Slater, J. Schmidt, and M. Uddstrom (2008b), Hydrological data assimilation with the Ensemble Kalman Filter: Use of streamflow data to update the states in a distributed hydrological model, *Adv. Water Resour.*, *31*, 1309–1324.

Clark, M. P., D. E. Rupp, R. A. Woods, H. J. Tromp-van Meerveld, N. E. Peters, and J. E. Freer (2009), Consistency between hydrological models and field observations: Linking processes at the hillslope scale to hydrological responses at the watershed scale, *Hydrol. Process.*, *23*(2), 311–319, doi:10.1002/hyp.7154.

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

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

Cressie, N., C. A. Calder, J. S. Clark, J. M. ver Hoef, and C. K. Wikle (2009), Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling, *Ecol. Appl.*, *19*(3), 553–570, doi:10.1890/07-0744.1.

Desborough, C. E. (1999), Surface energy balance complexity in GCM land surface models, *Climate Dynamics*, *15*, 389–403, 10.1007/s0038 20050289.

Di Baldassarre, G., and A. Montanari (2009), Uncertainty in river discharge observations: A quantitative analysis, *Hydrol. Earth Syst. Sci.*, *13*(6), 913–921.

Dooge, J. C. I. (1986), Looking for hydrologic laws, *Water Resour. Res.*, *22*(9S), 46S–58S, doi:10.1029/WR022i09Sp0046S.

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

Dunn, S. M., J. Freer, M. Weiler, M. J. Kirkby, J. Seibert, P. F. Quinn, G. Lischeid, D. Tetzlaff, and C. Soulsby (2008), Conceptualization in catchment modelling: Simply learning?, *Hydrol. Process.*, *22*(13), 2389–2393.

Eder, G., M. Sivapalan, and H. P. Nachtenbel (2003), Modelling of water balances in an Alpine catchment through exploitation of emerging properties over changing time scales, *Hydrol. Process.*, *17*, 2125–2149.

Famiglietti, J. S., and E. F. Wood (1994), Multi-scale modeling of spatially variable water and energy balance Processes, *Water Resour. Res.*, *30*(11), 3061–3078, doi:10.1029/94WR01498.

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

Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2006), Is the groundwater reservoir linear? Learning from data in hydrological modelling, *Hydrol. Earth Syst. Sci.*, *10*, 139–150.

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

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

Franks, S., P. Gineste, K. J. Beven, and P. Merot (1998), On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas into the calibration process, *Water Resour. Res.*, *34*(4), 787–797, doi:10.1029/97WR03041.

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

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

Harman, C. J., M. Sivapalan, and P. Kumar (2009), Power law catchment-scale recessions arising from heterogeneous linear 638 small-scale dynamics, *Water Resour. Res.*, *45*, W09404, doi:10.1029/2008WR007392.

Holländer, H. M., et al. (2009), Comparative predictions of discharge from an artificial catchment (Chicken Creek) using sparse data, *Hydrol. Earth Syst. Sci.*, *13*, 2069–2094.

Hood, J. L., and M. Hayashi (2010), Assessing the application of a laser rangefinder for determining snow depth in inaccessible alpine terrain, *Hydrol. Earth Syst. Sci.*, *14*, 901–910, doi:910.5194/hess-5114-5901-2010.

Hsu, K. L., H. Moradkhani, and S. Sorooshian (2009), A sequential Bayesian approach for hydrologic model selection and prediction, *Water Resour. Res.*, *45*, W00B12, doi:10.1029/2008WR006824.

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

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, doi:10.1029/93WR00877.

Jankov, I., W. A. Gallus, M. Segal, B. Shaw, and S. E. Koch (2005), The impact of different WRF model physical parameterizations and their interactions on warm season MCS rainfall, *Weather and Forecasting*, *20*, 1048–1060.

Jothityangkoon, C., M. Sivapalan, and D. Farmer (2001), Process controls on water balance variability in a large semi-arid catchment: Downward approach to hydrological model development, *J. Hydrol.*, *254*, 174–198.

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., S. Franks, and G. Kuczera (2002), Confronting Input Uncertainty in Environmental Modelling, in *Calibration of Watershed Models*, edited by Q. Y. Duan, et al., pp. 49–68, American Geophysical Union, Washington DC.

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

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

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

Kavetski, D., G. Kuczera, and S. W. Franks (2006), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*(3), W03407, doi:10.1029/2005WR004368.

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

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

Klemeš, V. (1983), Conceptualization and scale in hydrology, *J. Hydrol.*, *65*, 1–23, doi:10.1016/0022-1694(83)90208-1.

Klemeš, V. (1986), Operational testing of hydrologic simulation models, *Hydrol. Sci. J.*, *31*, 13–24, doi:10.1080/02626668609491024.

Kollet, S. J., and R. Maxwell (2008a), Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model, *Water Resour. Res.*, *44*, W02402, doi:10.1029/2007WR006004.

Kollet, S. J., and R. Maxwell (2008b), Quantifying the effects of three-dimensional subsurface heterogeneity on Hortonian runoff processes using a coupled numerical, stochastic approach, *Adv. Water Resour.*, *31*, 807–817, doi:10.1016/j.advwatres.2008.01.020.

Kollet, S. J., R. M. Maxwell, C. S. Woodward, S. Smith, J. Vanderborght, H. Vereecken, and C. Simmer (2010), Proof of concept of regional scale hydrologic simulations at hydrologic resolution utilizing massively parallel computer resources, *Water Resour. Res.*, *46*, W04201, doi:10.1029/2009WR008730.

Koster, R. D., and P. C. D. Milly (1997), The interplay between transpiration and runoff formulations in land surface schemes used with atmospheric models, *J. Clim.*, *10*, 1578–1591.

Koster, R. D., and M. J. Suarez (1992), Modeling the land surface boundary in climate models as a composite of independent vegetation stands, *J. Geophys. Res.*, *97*, 2697–2715.

Krajewski, W. F., G. J. Ciach, and E. Habib (2003), An analysis of small scale rainfall variability in different climatic regimes, *Hydrol. Sci. J.*, *48*, 151–162.

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

Kuczera, G., and S. Franks (2002), Testing hydrologic models: Fortification or falsification?, in *Mathematical Modelling of Large Watershed Hydrology*, edited by V. P. Singh and D. K. Frevert, Water Resources Pub., Littleton, CO.

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

Kumar, P. (2011), Typology of hydrologic predictability, *Water Resour. Res.*, *47*, W00H05, doi:10.1029/2010WR009769.

Kumar, S. V., et al. (2006), Land information system: An interoperable framework for high resolution land surface modeling, *Environ. Modell. Software*, *21*(10), 1402–1415, doi:10.1016/j.envsoft.2005.07.004.

Lamb, R., and K. Beven (1997), Using interactive recession curve analysis to specify a general catchment storage model, *Hydrol. Earth Syst. Sci.*, *1*, 101–113, doi:10.5194/hess-1-101-1997.

Larson, K. M., E. Gutmann, V. Zavorotny, J. Braun, M. Williams, and F. Nievinski (2009), Can we measure snow depth with GPS receivers?, *Geophys. Res. Lett.*, *36*, L17502, doi:10.1029/2009GL039430.

Larson, K. M., J. Braun, E. E. Small, V. Zavorotny, E. Gutmann, and A. Bilich (2010), GPS multipath and its relation to near-surface soil moisture content, *IEEE J-STARS*, *3*, 91–99, doi:10.1109/JSTARS.2009.2033612.

Lawrence, D. M., et al. (2011), Parameterization improvements and functional and structural advances in version 4 of the Community Land Model, *J. Adv. Model. Earth Syst.*, *3*, M03001, 27 pp., doi:10.1029/2011MS000045.

Le Moine, N., V. Andrassian, C. Perrin, and C. Michel (2007), How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments, *Water Resour. Res.*, *43*, W06428, doi:10.1029/2006WR005608.

Leavesley, G. H., S. L. Markstrom, P. J. Restrepo, and R. J. Viger (2002), A modular approach to addressing model design, scale, and parameter estimation issues in distributed hydrological modelling, *Hydrol. Process.*, *16*(2), 173–187.

Lehmann, E. L., and J. P. Romano (2005), *Testing statistical hypotheses*, 3rd ed., 786 pp., Springer, NY.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges (1994), A simple hydrologically based model of land-surface water and energy fluxes for general-circulation models, *J. Geophys. Res.*, *99*(D7), 14415–14428, doi:10.1029/94JD00483.

Liang, X., Z. H. Xie, and M. Y. Huang (2003), A new parameterization for surface and groundwater interactions and its impact on water budgets with the variable infiltration capacity (VIC) land surface model, *J. Geophys. Res.*, *108*(D16), 8613, doi:10.1029/2002JD003090.

Livneh, B., Y. Xia, K. E. Mitchell, M. B. Ek, and D. P. Lettenmaier (2010), Noah LSM snow model diagnostics and enhancements, *J. Hydrometeorol.*, *11*, 721–738, doi:10.1175/2009JHM1174.1.

Marshall, L., D. Nott, and A. Sharma (2007), Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework, *Hydrol. Process.*, *21*(7), 847–861, doi:10.1002/hyp.6294.

Maxwell, R. M., and N. L. Miller (2005), Development of a coupled land surface and groundwater model, *J. Hydrometeorol.*, *6*, 233–247, doi:10.1175/JHM422.1.

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.

McGuire, K. J., M. Weiler, and J. J. McDonnell (2007), Integrating tracer experiments with modeling to assess runoff processes and water transit times, *Adv. Water Resour.*, *30*, 824–837, doi:10.1016/j.advwatres.2006.07.004.

McMillan, H., J. Freer, F. Pappenberger, T. Krueger, and M. P. Clark (2010), Impact of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, *Hydrol. Process.*, *24*, 1270–1284.

McMillan, H. K., M. P. Clark, W. B. Bowden, M. Duncan, and R. A. Woods (2011), Hydrological field data from a modeler's perspective. Part 1: Diagnostic tests for model structure, *Hydrol. Process.*, *25*, 511–522, doi:10.1002/hyp.7841.

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

Morrissey, M. L., J. A. Maliekal, J. S. Greene, and J. M. Wang (1995), The uncertainty of simple spatial averages using rain-gauge networks, *Water Resour. Res.*, *31*, 2011–2017, doi:10.1029/95WR01232.

Mroczkowski, M., R. G. Paul, and G. Kuczera (1997), The quest for more powerful validation of conceptual catchment models, *Water Resour. Res.*, *33*, 2325–2335, doi:10.1029/97WR01922.

Niu, G. Y., Z. L. Yang, R. E. Dickinson, L. E. Gulden, and H. Su (2007), Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data, *J. Geophys. Res.*, *112*(D7), D07103, doi:10.1029/2006JD007522.

Niu, G.-Y., et al. (2011), The Community Noah Land Surface Model with Multi-Parameterization Options (NOAH-MP): 1. Model description and evaluation with local-scale measurements, *J. Geophys. Res.*, *116*, D12109, doi:10.1029/2010JD015139.

Nordbotten, J. M., M. A. Celia, H. K. Dahle, and S. M. Hassanizadeh (2007), Interpretation of macroscale variables in Darcy's law, *Water Resour. Res.*, *43*(8), W08430, doi:10.1029/2006WR005018.

Oleson, K. W., et al. (2010), *Technical Description of version 4.0 of the Community Land Model (CLM)*, NCAR Technical Note NCAR/TN-478+STR, National Center for Atmospheric Research, Boulder, CO, 257 pp.

Pande, S., M. McKee, and L. A. Bastidas (2009), Complexity-based robust hydrologic prediction, *Water Resour. Res.*, *45*, W10406, doi:10.1029/2008WR007524.

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

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

Pitman, A. J., and A. Henderson-Sellers (1998), Recent progress and results from the project for the intercomparison of land surface parameterization schemes, *J. Hydrol.*, *213*, 128–135, doi:10.1016/S0022-1694(98)00206-6.

Pomeroy, J. W., D. M. Gray, T. Brown, N. R. Hedstrom, W. L. Quinton, R. J. Granger, and S. K. Carey (2007), The cold regions hydrological model: A platform for basing process representation and model structure on physical evidence, *Hydrol. Process*, *21*(19), 2650–2667, doi:10.1002/hyp.6787.

Popper, K. (1959), *The Logic of Scientific Discovery*, Hutchinson, London, 480 pp.

Prokop, A. (2008), Assessing the applicability of terrestrial laser scanning for spatial snow depth measurements, *Cold Regions Science and Technology*, *54*(3), 155–163, doi:10.1016/j.coldregions.2008.07.002.

Qu, Y. Z., and C. J. Duffy (2007), A semidiscrete finite volume formulation for multiprocess watershed simulation, *Water Resour. Res.*, *43*, W08419, doi:10.1029/2006WR005752.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, and D. J. Seo (2004), Overall distributed model intercomparison project results, *J. Hydrol.*, *298*, 27–60, doi:10.1016/j.jhydrol.2004.03.031.

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, doi:10.1016/S0309-1708(98)00012-8.

Reggiani, P., S. M. Hassanizadeh, M. Sivapalan, and W. G. Gray (1999), A unifying framework for watershed thermodynamics: constitutive relationships, *Adv. Water Resour.*, *23*, 15–39, doi:10.1016/S0309-1708(99)00005-6.

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

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

Rodriguez-Iturbe, I., and J. M. Meija (1974), Design of rainfall networks in time and space, *Water Resour. Res.*, *10*, 713–728, doi:10.1029/WR010i004p00713.

Saltelli, A. (2002), Making best use of model evaluations to compute sensitivity indices, *Comput. Phys. Commun.*, *145*(2), 280–297, doi:10.1016/S0010-4655(02)00280-1.

Savenije, H. H. G. (2001), Equifinality, a blessing in disguise, *Hydrol. Earth Syst. Sci.*, *15*, 2835–2838.

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

Sayama, T., and J. J. McDonnell (2009), A new time-space accounting scheme to predict stream water residence time and hydrograph source components at the watershed scale, *Water Resour. Res.*, *45*, W07401, doi:10.1029/2008WR007549.

Schaefli, B., C. J. Harman, M. Sivapalan, and S. J. Schymanski (2011), HESS Opinions: Hydrologic predictions in a changing environment: Behavioral modeling, *Hydrol. Earth Syst. Sci.*, *15*, 635–646.

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.

Schymanski, S., M. Roderick, M. Sivapalan, L. Hutley, and J. Beringer (2007), A test of the optimality approach to modeling canopy properties and CO2 uptake by natural vegetation, *Plant Cell Environ.*, *30*, 1586–1598, doi:10.1111/j.1365-3040.2007.01728.x.

Schymanski, S. J., M. Sivapalan, M. L. Roderick, J. Beringer, and L. B. Hutley (2008), An optimality-based model of the coupled soil moisture and root dynamics, *Hydrol. Earth Syst. Sci.*, *12*, 913–932, doi:10.5194/hess-12-913-2008.

Schymanski, S. J., M. Sivapalan, M. L. Roderick, L. B. Hutley, and J. Beringer (2009), An optimality-based model of the dynamic feedbacks between natural vegetation and the water balance, *Water Resour. Res.*, *45*, W01412, doi:10.1029/2008WR006841.

Seibert, J. (2000), Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, *Hydrol. Earth Syst. Sci.*, *4*(2), 215–224.

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

Selker, J., N. van de Giesen, M. Westhoff, W. Luxemburg, and M. B. Parlange (2006), Fiber optics opens window on stream dynamics, *Geophys. Res. Lett.*, *33*, L24401, doi:10.1029/2006GL027979.

Singh, V. P., and D. K. Frevert (2002a), *Mathematical Models of Small Watershed Hydrology and Applications*, 972 pp., Water Resources Pub., Highland Ranch, CO.

Singh, V. P., and D. K. Frevert (2002b), *Mathematical Models of Large Watershed Hydrology*, 914 pp., Water Resources Pub., Highland Ranch, CO.

Sivakumar, B. (2008), Dominant processes concept, model simplification and classification framework in catchment hydrology, *Stochastic Environmental Research and Risk Assessment*, *22*, 737–748, doi:10.1007/s00477-007-0183-5.

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

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

Sivapalan, M., K. Beven, and E. F. Wood (1987), On hydrologic similarity: 2. A scaled model of storm runoff production, *Water Resour. Res.*, *23*, 2266–2278, doi:10.1029/WR023i012p02266.

Sivapalan, M., G. Blöschl, L. Zhang, and R. Vertessy (2003), Downward approach to hydrological prediction, *Hydrol. Process.*, *17*(11), 2101–2111.

Slater, A. G., et al. (2001), The representation of snow in land surface schemes: Results from PILPS 2(d), *J. Hydrometeor.*, *2*, 7–25.

Smith, T. J., and L. A. Marshall (2010), Exploring uncertainty and model predictive performance concepts via a modular snowmelt-runoff modeling framework, *Environ. Modell. Software*, *25*(6), 691–701.

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

Soulsby, C., C. Neal, H. Laudon, D. A. Burns, P. Merot, M. Bonell, S. M. Dunn, and D. Tetzlaff (2008), Catchment data for process conceptualization: simply not enough?, *Hydrol. Process.*, *22*, 2057–2061.

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

Tarantola, A. (2005), *Inverse Problem Theory and Methods for Model Parameter Estimation*, 342 pp., Society for Industrial and Applied Mathematics, Philadelphia, PA.

Tetzlaff, D., J. J. McDonnell, S. Uhlenbrook, K. J. McGuire, P. W. Bogaart, F. Naef, A. J. Baird, S. M. Dunn, and C. Soulsby (2008), Conceptualizing catchment processes: Simply too complex?, *Hydrol. Process.*, *22*, 1727–1730, doi:10.1002/hyp.7069.

Troch, P. A., G. A. Carrilo, I. Heidbuchel, S. Rajagopal, M. Switanek, T. H. M. Volkman, and M. Yaeger (2009), Dealing with landscape heterogeneity in watershed hydrology: A review of recent progress toward new hydrological theory, *Geography Compass*, *3*, 375–392.

Tromp-van Meerveld, H. J., and J. J. McDonnell (2006a), On the interrelations between topography, soil depth, soil moisture, transpiration rates and species distribution at the hillslope scale, *Adv. Water Resour.*, *29*, 293–310, doi:10.1016/j.advwatres.2005.02.016.

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

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

Tyler, S. W., J. S. Selker, M. B. Hausner, C. E. Hatch, T. Torgersen, C. E. Thodal, and S. G. Schladow (2009), Environmental temperature sensing using Raman spectra DTS fiber-optic methods, *Water Resour. Res.*, *45*, W00D23, doi:10.1029/2008WR007052.

Uhlenbrook, S., J. Seibert, C. Leibundgut, and A. Rodhe (1999), Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure, *Hydrol. Sci. J.*, *44*(5), 779–797.

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

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

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*, 999–1013, doi:10.1029/2000WR900272.

van Dijk, A. I. J. M. (2010), Selection of an appropriately simple storm runoff model, *Hydrol. Earth Syst. Sci.*, *14*, 447–458.

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

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

Wagener, T., M. J. Lees, and H. S. Wheater (2002), A toolkit for the development and application of parsimonious hydrological models, in *Mathematical Models of Small Watershed Hydrology*, Volume 2, edited by V. P. Singh, D. K. Frevert, and D. Meyer, Water Resources Publications, Colorado, USA.

Wagener, T., M. Sivapalan, P. Troch, and R. Woods (2007), Catchment classification and hydrologic similarity, *Geography Compass*, *1*, doi:10.1111/j.1749-8198.2007.00039.

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

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

Willems, P. (2001), Stochastic description of the rainfall input errors in lumped hydrological models, *Stochastic Environmental Research and Risk Assessment*, *15*(2), 132–152. 10.1007/s004770000063.

Wood, E. F., et al. (1998), The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) phase 2(c) Red-Arkansas River basin experiment: 1. Experiment description and summary intercomparisons, *Global and Planetary Change*, *19*, 115–135, doi:10.1016/S0921-8181(98)00044-7.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, *44*, W02501, doi:10.1029/2008WR007347.

Young, P. (1998), Data-based mechanistic modelling of environmental, ecological, economic and engineering systems, *Environ. Modell. Software*, *13*(2), 105–122.

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

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

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

Young, P., S. Parkinson, and M. Lees (1996), Simplicity out of complexity in environmental modelling: Occam's razor revisited, *J. Appl. Statistics*, *23*(2–3), 165–210.

Zreda, M., D. Desilets, T. P. A. Ferre, and R. L. Scott (2008), Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons, *Geophys. Res. Lett.*, *35*, L21402, doi:10.1029/2008GL035655.

M. P. Clark, Research Applications Laboratory, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307–3000, USA. (mclark@ucar.edu)

F. Fenicia, Department of Environment and Agro-Biotechnologies, Centre de Recherche Public, Gabriel Lippmann, 41 Rue du Brill, L-4422 Belvaux, Grand-Duchy of Luxembourg.

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