Understanding the Information Content in the Hierarchy of Model Development Decisions Learning From Data

Gharari, Shervan ; Gupta, Hoshin V.; Clark, Martyn P.; Hrachowitz, Markus; Fenicia, Fabrizio; Matgen, Patrick; Savenije, Hubert H.G.

DOI
10.1029/2020WR027948

Publication date
2021

Document Version
Final published version

Published in
Water Resources Research

Citation (APA)
Gharari, S., Gupta, H. V., Clark, M. P., Hrachowitz, M., Fenicia, F., Matgen, P., & Savenije, H. H. G. (2021). Understanding the Information Content in the Hierarchy of Model Development Decisions: Learning From Data. *Water Resources Research*, *57*(6), 1-35. Article e2020WR027948. https://doi.org/10.1029/2020WR027948

AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR027948

Understanding the Information Content in the Hierarchy of Model Development Decisions: Learning From Data

Shervan Gharari[1], Hoshin V. Gupta[2], Martyn P. Clark[1], Markus Hrachowitz[3], Fabrizio Fenicia[4], Patrick Matgen[5], and Hubert H. G. Savenije[3]

[1]Centre for Hydrology, University of Saskatchewan, Canmore, Alberta, Canada, [2]Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, Arizona, USA, [3]Faculty of Civil Engineering and Geosciences (CITG), Delft University of Technology, Delft, Netherlands, [4]Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dübendorf, Switzerland, [5]Luxembourg Institute of Science and Technology (LIST), Belval, Luxembourg

Key Points:

- We present a strategy for characterizing and quantifying the information added at each model building step
- Model building steps are interdependent in a hierarchical manner
- We call for the focus of model calibration to shift from "*parameter spaces*" to "*function spaces*"

Correspondence to:
S. Gharari,
shervan.gharari@usask.ca

Citation:
Gharari, S., Gupta, H. V., Clark, M. P., Hrachowitz, M., Fenicia, F., Matgen, P., & Savenije, H. H. G. (2021). Understanding the information content in the hierarchy of model development decisions: Learning from data. *Water Resources Research*, *57*, e2020WR027948. https://doi.org/10.1029/2020WR027948

Received 14 MAY 2020
Accepted 28 APR 2021

Abstract Process-based hydrological models seek to represent the dominant hydrological processes in a catchment. However, due to unavoidable incompleteness of knowledge, the construction of "*fidelius*" process-based models depends largely on expert judgment. We present a systematic approach that treats models as hierarchical assemblages of hypotheses (conservation principles, system architecture, process parameterization equations, and parameter specification), which enables investigating how the hierarchy of model development decisions impacts model fidelity. Each model development step provides information that progressively changes our uncertainty (increases, decreases, or alters) regarding the input-state-output behavior of the system. Following the principle of maximum entropy, we introduce the concept of "*minimally restrictive process parameterization equations—MR-PPEs*," which enables us to enhance the flexibility with which system processes can be represented, and to thereby investigate the important role that the system architectural hypothesis (discretization of the system into subsystem elements) plays in determining model behavior. We illustrate and explore these concepts with synthetic and real-data studies, using models constructed from simple generic buckets as building blocks, thereby paving the way for more-detailed investigations using sophisticated process-based hydrological models. We also discuss how proposed MR-PPEs can bridge the gap between current process-based modeling and machine learning. Finally, we suggest the need for model calibration to evolve from a search over "*parameter spaces*" to a search over "*function spaces*."

Plain Language Summary Modelers make many decisions in their quest to formulate a working model. It is important to evaluate the impact of each modeling decision, and to assess the extent to which different decisions improve the representation of the actual system. Building upon past work, we present a framework that enables an improved assessment of individual modeling decisions. Specifically, we suggest that modelers should pay more attention to the hierarchical structure of model building decisions, and to the impact that each such decision can have on the fidelity of resulting process representation.

1. Introduction

1.1. Model as Assemblages of Hypotheses

Process-based models are abstract simplified representations of the underlying physical processes governing the behavior of natural systems. Such models are used to characterize our knowledge about the spatio-temporal structural and behavioral properties of natural systems, thereby enabling decision-makers to approximate the behavior of a system at a specific space-time location and, ultimately, to infer the impacts of various natural changes or anthropogenic modifications to that system. Process-based models can be as simple as characterizing only water movement through a homogeneous porous medium under a constant head difference based on Darcy's Law (e.g., Fitts, 2012), or be very complex involving large numbers of interacting processes as in terrestrial land [surface] models (Pitman, 2003).

Many factors can shape the process of model building, including the purpose, the data available, the experience and background of the modelers, the available computational power, and the desired predictive

accuracy (Addor & Melsen, 2019). Generally speaking, constructing a process-based model begins with our *perceptions* regarding the nature of the natural system, resulting in a "*perceptual-conceptual*" mental model that identifies the processes (and their interconnections) believed to be important in determining its behavior. The model building process then continues with a formalization of these perceptions into a "*conceptual model*" (that facilitates communication and discussion), and then progressively is translated into a "*symbolic/mathematical*" model (the model equations) and finally a "*computational*" model realized in the form of computer code (Beven, 2011; Blöschl & Sivapalan, 1995; Gupta, Clark, et al., 2012). Finally, the model parameter values must be specified based on expert knowledge, or by manual or automatic calibration, and the resulting model simulations must be evaluated (in each particular context) for consistency and adequacy of performance (Refsgaard & Henriksen, 2004).

Expanding upon this understanding, it is important to also note that the model building practice involves assembling together various hypotheses at each of the aforementioned model building levels (Clark, Kavetski, & Fenicia, 2011). We can characterize this *assemblage of hypotheses* as being at four hierarchical levels as follows:

- ***System Diagram and Conservation Law Hypotheses:*** This establishes the control volumes within the model domain, specifies the major processes that must be represented, identifies the boundary conditions and main external forces (disturbances, forcings) acting upon the system, specifies the major state variables that must be tracked in order to represent the internal system dynamics, and consequently identifies the conservation principles (e.g., mass, energy, momentum) that must be obeyed. The result is a high-level system diagram that formalizes the main aspects of the model.
- ***System Architecture Hypothesis:*** This establishes the manner and level of detail by which the internal structure of the system is to be represented, and can vary from a hyper resolution 3-dimensional spatial discretization using a finite difference/element grid to a low-resolution directed graph characterized by interacting conceptual buckets (as in so-called bucket style watershed models). The result is a finite number of subsystem control volumes and associated state variables to be tracked, and a specification of which such elements are to be linked via interconnecting fluxes. The resulting overall node and link graph specifies the pathways by which mass, energy and momentum are transferred through the system to the outputs (Bancheri et al., 2019). In general, each node and link must obey the conservation law hypothesis specified by the system diagram and conservation law hypothesis.
- ***Process Parameterization Hypothesis:*** This defines the mathematical forms of the *Process Parameterization Equations* (hereafter referred to as PPEs), that relate the state variables to the fluxes, and that consequently control the internal (and therefore overall) dynamics of the system. In general, the PPE's must conform to the laws of thermodynamics (fluxes arise in response to gradients) and other pertinent laws of physics that govern the behaviors of the fluxes represented. Further, suitable choices of the PPEs must be consistent with the selected system architecture hypothesis, and therefore must reflect the system scale and spatiotemporal resolution to be represented.
- ***Parameter Specification Hypothesis:*** This defines the values of the parameters used to specify the behaviors of the PPE's for any given physical location and/or application. Further, being conditional on the PPE specification, suitable choices for these parameter values must reflect the system scale and spatiotemporal resolution to be represented.

Additionally, there are hypotheses that pertain to things outside of the modeled system. These include:

- ***Forcing/Input Uncertainty Hypothesis:*** The information provided by the 4 levels of modeling hypotheses mentioned above are conditional on the nature of the system inputs/forcings and their uncertainties. In fact, it is the forcing/input that conditions and shapes the dynamical response of a system. For example, no matter how well the model of a watershed is formulated, it may be impossible to adequately characterize a particular flood event if information regarding the existence of the causal precipitation event(s) is not properly captured by the rain gauges (Beven & Westerberg, 2011). Extensive efforts have been carried out in recent years on ensemble probabilistic methods to address the uncertainties in input/forcing, for example, precipitation and temperature (Clark & Slater, 2006; Cornes et al., 2018; Newman et al., 2015; Tang et al., 2020).
- ***System Response Uncertainty and Evaluation Metric (Objective Function) Hypotheses:*** Similar to the input uncertainty, the measured response of the system, which is often used to infer internal system

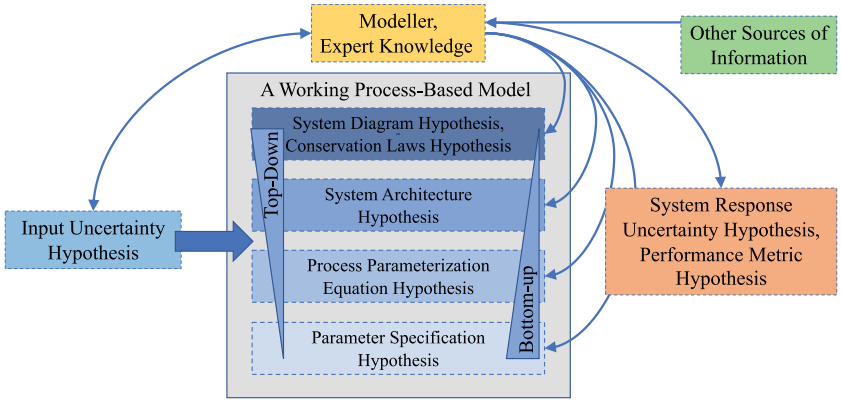

Figure 1. The chart illustrates the general manner in which a modeler interacts with the various levels of hypotheses (see description in the text) for developing a working process-based model.

properties (such as system architecture) and to allocate values for the model parameters, also has its associated uncertainties (Kiang et al., 2018; Westerberg et al., 2011). As an example, a large portion of the effort to relate signatures built on observed system response(s), such as streamflow, to model parameters and processes, falls under the evaluation metric hypothesis (e.g., Bárdossy & Singh, 2008; Euser et al., 2013; McMillan, 2020; Westerberg & McMillan, 2015). Given that inference is typically performed by selecting one or more performance metrics, different performance metrics have the effect of differently regularizing the result of the inferential procedure, thereby dramatically influencing the inferred representation of internal system structure and behavior (Santos et al., 2018).

Figure 1 provides an illustration of how a modeler interacts with the four hierarchical levels of hypothesis specification. Since the model is intended to reflect the behavior of the system, the modeler should give early attention to the system input and response uncertainty hypotheses, which should be based on both the spatial representativeness of the station network and the physical characteristics of the sensors and instruments used to acquire the data (rather than being inferred along with the four hierarchical levels of model building). Similarly, early attention to the regularizing performance metric hypothesis can help to clarify what aspects of system response are important (determined by the modeling goals); this can be as basic as a visual comparison of observed and simulated responses or involve the design of mathematical metrics to quantify the differences between the observed and simulated system behaviors. Note that a modeler may choose to treat the measurement errors as being negligible, and therefore represent the inputs and responses as being deterministic. This can affect the design of the metric used for model performance evaluation.

While the four levels of modeling hypotheses are not entirely separable, each is essential to the construct a working process-based model. Certainly, different emphases can be placed on which level is given primary importance. For example, the downward or top-down modeling approach (Klemeš, 1983; Sivapalan et al., 2003; Young, 2003) tends to place primary emphasis on the levels that specify the conservation principles (the system diagram) and system architecture, leaving PPE forms and parameter values to be inferred at the spatiotemporal scale of interest (e.g., Schulz & Beven, 2003). By contrast, the bottom-up approach tends to emphasize specification of a system architecture that conforms to available small-scale knowledge regarding the spatial distributions of material properties and process behaviors (e.g., 3-D subsurface flow based on Richards' Equation; Condon et al., 2013). In cases where analytical solutions exist, such as when modeling water table deformations around an abstraction well using the Darcy equation, the system architecture can be expressed in terms of elements of infinite spatial extent. In general, the various modeling hypotheses are refined in an iterative fashion as the modeler gains experience with the model and learns more about its strengths and weaknesses. The attempt to formulate an improved working hypothesis can involve drawing upon broader sources of knowledge and involve the use of so-called "*soft*" data (Seibert & McDonnell, 2002).

Each of the aforementioned levels of hypotheses progressively changes (increases, decreases, or somehow alters) our uncertainty regarding the input-state-output behavior of the system (Figure 2). This change in

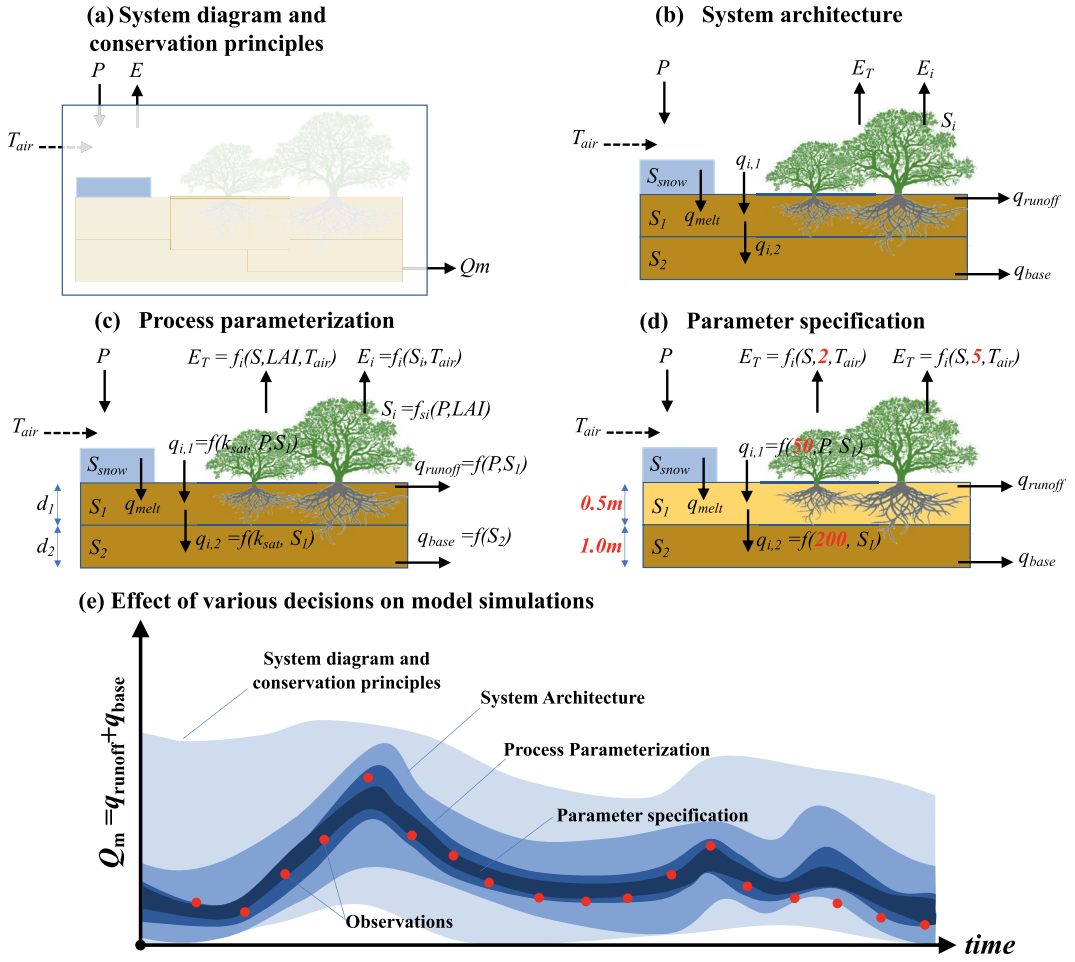

Figure 2. An example of model building hypotheses in a predominantly top-down approach. (a) A System Diagram that identifies the boundary of the system and incoming and outgoing mass and energy fluxes. The mass fluxes are precipitation, transpiration, interception, surface runoff, baseflow, and snow/ice sublimation. The energy fluxes are incoming short-wave radiation and outgoing latent heat fluxes. (b) A System Architecture Diagram that identifies the sub-system components to be modeled as a directed graph of nodes (state variables) and links (processes); here the state variables are canopy water holding capacity, soil moisture, and snow. (c) A Process Parameterization Diagram that specifies the state-flux relationships determining mass and energy fluxes linking the states. (d) A Parameter Specification Diagram, that illustrates the fixing of parameters to appropriate values based on their conceptual and physical properties. (e) A cartoon illustrating how the hierarchical progression of model building hypotheses translates into information (read change of uncertainty) regarding the system output (total streamflow as summation of surface runoff and baseflow; $q_{runoff} + q_{base}$). When the hypotheses are properly specified, the simulated output trajectories should closely track the observations, and bracket them with an acceptable level of uncertainty.

uncertainty occurs due to the fact that "*information*" is added at each level of the modeling process (information in considered to be added when there is a change of uncertainty in a given context). If the added information at each level is "*good*," we can expect the simulated state-output response to progressively converge around the corresponding "*true*" state-output response as approximated by the observed system response and knowledge of the system. In actual practice, there is typically considerable lack of knowledge regarding the true nature of the natural system and its input-state-output behaviors. This means that the modeling building steps of such a system are inherently uncertain. Consequently, poor hypothesis choices at any of these stages can result in uncertainty that must be reflected in the simulated trajectories of the state-output responses, and in systematic deviations of those trajectories from the corresponding observed ones (as characterized by available measurements). Such tendencies to deviate from the observed measurements are usually masked (made more difficult to detect) by the process of (automated or manual) parameter calibration.

Differences in emphasis, regarding how the system architecture and process parameterization (and eventually parameter specification) hypotheses should be specified, have resulted in a variety of modeling approaches. So-called "*physically based*" models generally seek to define the parameters of the process equations as being somehow related (at least in principle) to observable physical characteristics of the system. The question of how best to do this has resulted in a significant body of literature (e.g., Abbott et al., 1986; Antonetti et al., 2017; Beven & Kirkby, 1979; Flügel, 1995; Gao et al., 2019; Knudsen et al., 1986; Loritz, Gupta, et al., 2018; Naef et al., 2002; Reggiani et al., 1998; Uhlenbrook et al., 2004; Winter, 2001; Zehe, Loritz, et al., 2019). However, it is fair to say that the predominant approach to such decisions is mainly to employ expert judgment and tacit knowledge that is often not well documented, so that model building is considered to be largely an "*art*" (Savenije, 2009).

It is, of course, well recognized that model parameters tend to be "*conceptual and empirical*" representations of system properties (Hrachowitz & Clark, 2017), and it was acknowledged decades ago that it is not possible to infer "*true*" parameter values for a given model (Beck, 1983; Johnston & Pilgrim, 1976; among many others). A considerable body of literature has been devoted to parameter value inference and its associated uncertainties. In particular, the issue of parameter uncertainty (and sensitivity) has received significant attention and spawned numerous methods including Global Likelihood Uncertainty Estimation (GLUE; Beven & Binley, 1992), Bayesian Recursive Estimation (BaRE; Gupta, Thiemann, et al., 2003; Misirli et al., 2003; Thiemann et al., 2001), formal Bayesian methods such as DREAM (Vrugt, ter Braak, et al., 2008), BATEA (Kavetski et al., 2006), and pareto-based optimization algorithms (Deb et al., 2002; Vrugt, Gupta, et al., 2003).

The issue of process parameterization and system architecture uncertainty has also received attention. Several studies have explored how model structural changes can help to improve the presumed representation of "reality" (e.g., Fenicia, McDonnell, & Savenije, 2008; Freer, Beven, & Ambroise, 1996; Freer, McMillan, et al., 2004; Seibert et al., 2003; Son & Sivapalan, 2007; among many others). Recently, in the context of rainfall-runoff and land-surface models, there have been efforts to modularize the modeling decision process so that both the system architecture and process parameterization hypotheses can be altered to enable investigation of their effects, thereby facilitating the pursuit of multiple working hypotheses (e.g., Clark, Kavetski, & Fenicia, 2011). As examples, MMS (Leavesley et al., 1996), RRMT (Wagener et al., 2001), FUSE (Clark, Slater, et al., 2008), Noah-MP (Niu et al., 2011), FLEX (Fenicia, Kavetski, & Savenije, 2011), SUMMA (Clark, Nijssen, et al., 2015), MARRMoT (Knoben, Freer, Fowler, et al., 2019), and Raven (Craig et al., 2020) enable the model developer to select from prespecified sets of system architectural and PPE choices.

In this context, a significant body of literature has sought to employ diagnostic signatures extracted from the system response data (Gupta, Wagener, & Liu, 2008; Martinez & Gupta, 2010; Yilmaz et al., 2008; among many others) as a means to improve model identifiability and guide proper specification of parameter values and to lesser extent forms of PPEs. Finally, a small amount of work has explored the use of methods to stochastically represent model structural uncertainty, so that model equation forms can be updated/corrected via Bayesian data assimilation (Bulygina & Gupta, 2011; Nearing & Gupta, 2015).

1.2. Formulating the Research Questions

The overarching research questions of this study are driven by the quest for a "*fidelius*" model that represents, as accurately as possible, our understanding of how the natural system works while generating input-state-output simulations that agree with the observed data to within an [often subjective] acceptable level of uncertainty. In particular, we are motivated by the following specific research questions:

Q1) How does each stage of model building encode information into the model and it's simulated outputs?
Q2) Given a particular system architecture, what mathematical forms for the PPEs are most consistent with the information provided by the observed data?
Q3) How uncertain are the inferred PPEs, and how does that uncertainty affect the model generated simulations of system behavior?
Q4) How does information provided in the form of additional constraints on system behavior (beyond that encoded via the aforementioned four levels of modeling hypotheses), affect the model generated simulations of system behavior?

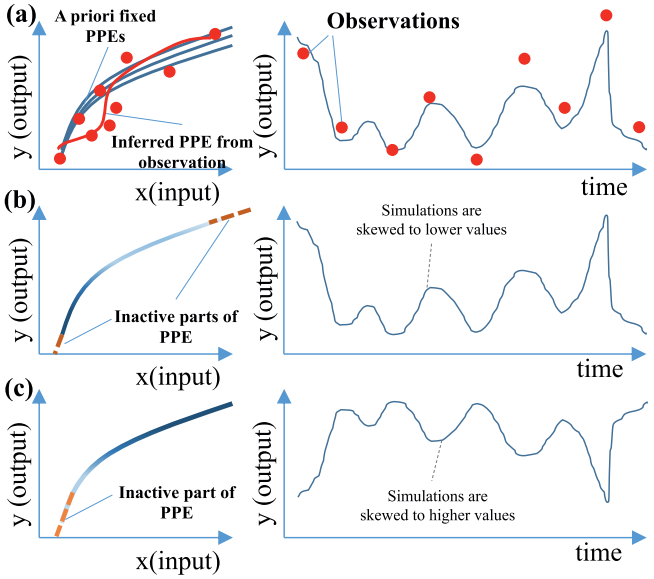

Figure 3. (a) An illustration of how the inferred form (red line) of a process parameterization equation (PPE), that accounts for both system architecture and information provided by the observed data (red dots), may be different from a PPE form that is assumed a priori (blue lines; different lines represent different parameter values). (b) An illustration of PPE activation that is skewed toward the lower portion of the active range, and where there are inactive regions. (c) An example of PPE activation that is skewed toward the upper portion of the active range, and that has inactive regions.

1.2.1. Question 1—How is Information Encoded Into the Model?

As mentioned earlier, the sequence of structural decisions encoded into the model acts to progressively alter (hopefully reduce) our uncertainty regarding its input-state-output behavior. We are interested, therefore, in understanding how the different kinds of structural information act, and interact, to impact the model behaviors and their uncertainty. We are also interested in understanding how model structural errors/inadequacies can result through incorporating incorrect (i.e., bad) or overly strong assumptions. In this regard, we introduce the concept of "*minimally restrictive PPEs*" that facilitate a more structured approach to model development. Our approach generally follows the Information Theoretic perspective suggested by Gupta and Nearing (2014).

1.2.2. Question 2—What PPE Forms are Appropriate Given a Particular System Architecture?

While the so-called "*flexible*" modeling approaches mentioned earlier enable selection from different pre-determined formulations of the PPEs, the available sets of choices are typically granular (discrete), and specified without considering the selected system architecture and its degree of complexity. For example, various models may exploit a fixed process representation such as Jarvis-type stomatal resistance formulation (Jarvis, 1976) for plant transpiration regardless of subsurface representation, assumptions on intercepted water by the plant foliage, and other assumptions made regarding various aspects of the overall system architecture.

Here, we investigate what might be the appropriate forms for the PPE's given a selected system architecture, and how these might differ from the typical forms that are commonly used (Figure 3a). Usually, most of the modeling effort is focused on the parameter specification hypothesis based on the evaluation metric or objective function hypothesis, which in turn means modelers often explore "parameter spaces" via manual or automatic model calibration to develop a behavioral model. We believe there should be a shift in the focus of model calibration from a search over "*parameter spaces*" to a search over "*function spaces*." Exploring the "function spaces" enables modelers to test whether the assumptions on functional forms are reproducible. Although some studies have attempted to explore the "functions spaces" by recreating the PPE's for very simple model architectures from data (Bulygina & Gupta, 2009, 2010, 2011; Kirchner, 2009; Koster & Mahanama, 2012; Lamb & Beven, 1997), our knowledge of how to systematically explore the interactions between data and the PPE forms that are appropriate for a given system architecture is still in its infancy.

This study can be viewed as a preliminary attempt to bridge the physical/process-based and data-driven approaches to Earth System modeling. Whereas the former uses theory to guide the selection of an appropriate model representation, the latter seeks to learn the model representation directly from the data. With increases in computational power, machine learning approaches are now finding increasingly widespread application in hydrology (Goodwell & Kumar, 2017; Jiang et al., 2020; Karpatne et al., 2018; Kratzert et al., 2019; Shen, 2018; Zhao et al., 2019), and attempts are even under way to ensure that such models are able to obey conservation principles such as mass, energy and momentum balance across the system components.

1.2.3. Question 3—How Uncertain are the PPE Forms Inferred From Data?

The uncertainty associated with a model is typically expressed in terms of parameter uncertainties (via marginal distributions) and the consequent output uncertainties. Further, structural uncertainties are typically assessed by running discrete sets of model ensembles, where each member of the ensemble represents a different possible model structural hypothesis. There has been comparatively little investigation of the uncertainties associated with the selected forms of the PPEs.

In this regard, it is important to note that only a small portion of the functional extent of a PPE might actually be activated during any given simulation, and so the available data may only be informative (about the nature of that functional form) across a limited portion of its extent. This affects our ability to properly and unambiguously infer the correct overall form of the PPE, especially when extrapolating beyond the ranges and types of system behaviors represented by the available data. Accordingly, we investigate two aspects of this issue: (a) That only a portion of the functional range of a PPE might actually be activated during a simulation and (b) That the PPEs are not uniformly activated across its full behavioral range and are therefore characterized by varying degrees of structural and functional (and therefore predictive) uncertainty.

For example, in drier climates the activated portion of the PPEs will likely be the water-stressed portions (Figure 3b). Conversely, in humid regions the activated portion of the PPEs will likely be the portions that correspond to no water stress (Figure 3c). The consequence is that the model simulated outputs will have different associated uncertainties (and sensitivities) depending on which portions of the PPEs are being activated. This kind of variation in functional uncertainty is not properly captured by the use of fixed PPE forms where the uncertainties are expressed entirely via parameter uncertainty. The implications are particularly strong when models are used to predict system responses under conditions that have not been part of the data record (such as under future or different climate; Klemeš, 1986).

1.2.4. Question 4—How Do Behavioral Constraints Act to Inform the Model Simulations?

Some kinds of important information about the behavior of a natural system may not be inferable (or be only weakly inferable) from the observed system responses. Such information can take the form of "*expert*" knowledge that expresses conditional constraints on internal system behaviors, relative magnitudes or rates of internal fluxes, relationships between various signature properties of the system responses, and many others. It can also take the form of assumptions that the modeler might wish to make/test.

Such information can be expressed via behavioral constraints on the internal process behavior or the magnitude of state variables. In the context of performance metric design, the information may be expressed using signature indices extracted from the raw data, rather than directly by the raw data. We are interested, therefore, in understanding how much additional information is provided by such expressions of expert (i.e., domain) knowledge, and how this knowledge acts to alter our uncertainty about system behavior (Wagener & Montanari, 2011). To identify simulations that are "*behavioral*," in the spirit of Spear and Hornberger (1980), Schaefli et al. (2011), Gharari, Hrachowitz, et al. (2014), Gharari, Shafiei, et al. (2014), and Bahremand (2016) among many others, we formulate constraints that impose internal restrictions on system component behavior. We exclude, however, any discussion of how the evaluation metrics (or choices of signatures) act as hypotheses regarding the response of the system, as that issue is not directly linked to the four levels of model building addressed in this study.

1.3. Focus, Organization, and Scope of This Paper

The goal of this study is to investigate the four research questions formulated in Section 1.2. Section 2 presents our proposed framework for investigating the information content of modeling hypotheses and outlines the methodology used in Section 3, which presents the results of several numerical experiments. Being exploratory, the scope of these experiments is limited to relatively simple system architectures that can be constructed using a generic bucket, but that encompass the important processes relevant to modeling the catchment scale rainfall-runoff process. Section 4 summarizes and discusses our results and explores their implications for future work. Finally, Section 5 presents our conclusions, and discusses several broader considerations that are suggested by our findings.

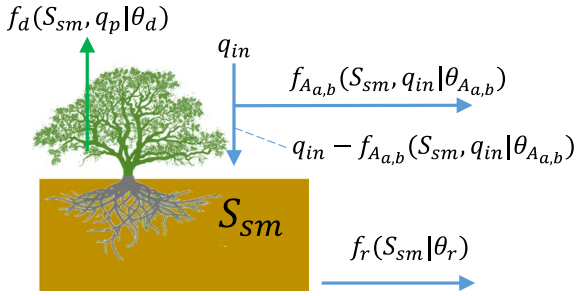

Figure 4. Example of a generic bucket (GB) that can be used for creating bucket style models applied to the topsoil layer, illustrating the processes of absorption (infiltration) and bypass (infiltration excess), release (drainage) and depletion (evapotranspirative fluxes). In this illustration, S_{sm} is the soil moisture of the topsoil layer. q_{in} represents the incoming flux such as precipitation or effective precipitation. $f_{A_{a,b}}$ represent the function that decides which portion of the incoming flux is absorbed by the topsoil layer and which portion is bypassed through the system (similar to the concept of variable source area). q_p represent the potential demand flux that limit the amount of water that can be taken from the soil moisture. This maximum potential flux can be expressed as a gradient of the states of various kind. f_d is the function that related the soil moisture and the potential flux, q_p, to the depletion flux. f_r relates the release from the GB to the amount of state in the system. $\theta_{A_{a,b}}$, θ_d, and θ_r are the parameters specifying the functions $f_{A_{a,b}}, f_d, f_r$, respectively.

2. Methodology to Explore the Nature of the Hierarchy of Model Development Decisions

2.1. Establishing the Building Blocks

2.1.1. Definition of the Generic Bucket Used to Construct Bucket Style Models

We define the simplest building block of bucket style models (e.g., HBV, HYMOD or FLEX, among others) as a generic bucket (GB) to systematically create varieties of different possible system architectural hypotheses. A GB identifies a control volume within the natural system in which mass is to be conserved. In principle, the ideas presented here can be extended to more complex models and processes parameterizations.

We focus on three major processes that can represent most of the processes in a typical bucket style model:

1. ***Absorption/Bypass:*** An incoming flux to the GB can either be absorbed and stored by the GB or can bypass it. For example, part of the incoming water reaching a porous medium, can be absorbed in the soil (both micropore and macropore) and thus stored in the part of the system represent by the GB, while part of it can bypass the soil through preferential flow paths (through macropores) or via surface runoff.
2. ***Depletion:*** This is a process in which the outgoing flux is dependent on both the amount of water stored in the GB and an external [suction] force. The latter may sometimes be parameterized in terms of a maximum (limiting) or "demand" flux rate such as potential evapotranspiration. In hydrological terms, this allows us to simulate mechanisms such as evaporation, transpiration and deep percolation based on a maximum rate, and snowmelt based on temperature forcing (degree day factor formulation of snow melt), whose flux rates depend both on a supply term (available water) and a demand term.
3. ***Release:*** This is a process whereby the flux leaving the GB depends only on the amount of water stored in the GB. A (non-)linear reservoir is the simplest example of this process, in which the outgoing flux is a function of only the amount of water stored in the GB. In hydrological terms, this allows us to simulate drainage mechanisms in their simplest forms.

Figure 4 illustrates the use of a generic bucket to model the hydrological response of few decimeters of the topsoil. The input flux, q_{in}, is partitioned into an infiltration excess portion (represented as $f_{A_{a,b}}\left(S_{sm}, q_{in} \mid \theta_{A_{a,b}}\right)$) that bypasses the GB, and an infiltration portion that is stored (represented as $q_{in} - f_{A_{a,b}}\left(S_{sm}, q_{in} \mid \theta_{A_{a,b}}\right)$). The drainage release flux, $f_r\left(S_{sm} \mid \theta_r\right)$ is dependent on the current state of the soil moisture storage, S_{sm}. The evapotranspirative depletion flux is dependent on both the current soil moisture storage S_{sm} and a driver such as potential evaporation (for example, $f_d\left(S_{sm}, q_p \mid \theta_d\right)$ in which q_p is the demand or potential evaporation in this specific case). The mathematical functions ($f_{A_{a,b}}$, f_d and f_r) represent the PPEs of our generic GB, while the θ's represent corresponding parameters whose values must be specified. We should clarify that the conceptualization expressed here is specific to this study and can be much more complex based on the modeling need.

2.1.2. Strategies to Construct "Minimally Restrictive" Functional Forms for the PPEs

One of the goals of this study is to investigate the role of the *system architectural hypothesis* in determining model behavior. To focus on the influence of system architecture hypothesis on the overall model hypothesis, we need to reduce the dependence of the model simulations on the specific mathematical forms of the PPEs represented by $f_{A_{a,b}}$, f_r and f_d. Referring again to Figure 4, the forms of PPEs can be conceptualized as:

1. Parameterization for the absorption and bypass processes (respectively):

$$q_a = q_{in} \cdot \left[1 - A_{a,b}(S_{sm} \mid \theta_{a,b})\right] \tag{1}$$

$$q_b = q_{in} \cdot A_{a,b}(S_{sm} \mid \theta_{a,b}) \tag{2}$$

2. Process parameterization for the depletion process:

$$q_d = q_p \cdot K_d(S_{sm} \mid \theta_d) \tag{3}$$

3. Process parameterization for the release process:

$$q_r = K_r(S_{sm} \mid \theta_r) \cdot S_{sm} \tag{4}$$

For the rest of this work, we use the units of mm day^{-1} for fluxes and mm for storages. The functions for absorption, bypass, and depletion are unitless; the function for release defines the time constant of the reservoir in unit of day^{-1}.

In current practice, it is typical for the mathematical forms of equations representing, $A_{a,b}$, K_r, and K_d to be rigidly specified, so that any model behavioral flexibility is achieved only by adjustments (over some prespecified feasible ranges) to the *parameters* associated with the functions. Here, we introduce an alternative approach that enables us to work with less restrictive functional forms, thereby enabling evaluating the role of architectural changes in determining model behaviors. Such functional forms should be designed in such a manner that the modeled process behavior:

1. Is constrained by (remains consistent with) the laws and principles of physics, as well as by any well-established hydrological principles that may be pertinent at the system scale and resolution represented by the chosen system architecture.
2. Does not presume any more information about the behavioral nature of the process *than is actually known to be reasonably true*.

In other words, we aim for process parameterization functions that are maximally uncertain (i.e., maximally flexible, or alternatively minimally restrictive) while obeying the aforementioned assumptions of monotonicity and conservation. We generate *monotonically non-decreasing polynomial transmission functions* that provide us with the degrees of freedom required to represent the wide range of behaviors that might be possible in any given situation. Appendix A discusses our proposed strategy, among many other strategies that could be employed, for constructing such equation representations using piecewise linear approximations. This strategy allows for the "*minimally restrictive*" functional representations of the PPEs (herein denoted as MR-PPEs) to be of varying complexity, ranging from constant, to linearly varying over the domain, to having an arbitrarily large number of piecewise linear segments. For practical reasons, we limit ourselves to 1000 MR-PPEs each having a maximum of 20 linear segments, which should be more than sufficient to represent the kind of functional complexity one might encounter in our experiments in this work.

In each of the above representations (Equations 1–4), the terms, $A_{a,b}$, K_r, and K_d are represented by parameters $\theta_x = \left\{\theta_x^{low}, \theta_x^{high}, \theta_x^{scale}\right\}$, where θ_x^{low} and θ_x^{high} specify the minimum and maximum values that the transmission can range over ($0 \le \theta_x^{low} \le \theta_x^{high}$), and θ_x^{scale} specifies the range of storage, S_{sm}, over which that coefficient variation can occur and defines the storage capacity at which the transmission achieves its maximum value (Figure 5). We will refer to these parameters as the MR-PPE parameters.

For the bypass process (Equation 2), if $\theta_{A_{a,b}}^{high}$ is set to 1.0 the GB will represent a finite capacity bucket, meaning that all of the incoming water q_{in} bypasses the bucket when its storage amount reaches the maximum amount specified by $\theta_{A_{a,b}}^{scale}$; accordingly $\theta_{A_{a,b}}^{scale}$ represents the storage capacity of the bucket. Alternatively, if $\theta_{A_{a,b}}^{high}$ is set to 0.0 (accordingly we must also have $\theta_{A_{a,b}}^{low} = 0.0$), this means that *none* of the incoming water q_{in} will bypass the bucket, so that all of it will be absorbed (infiltrated) and stored in the GB; this simulates a formulation where the GB effectively has an infinite capacity (similar to a linear reservoir). If, in addition, we

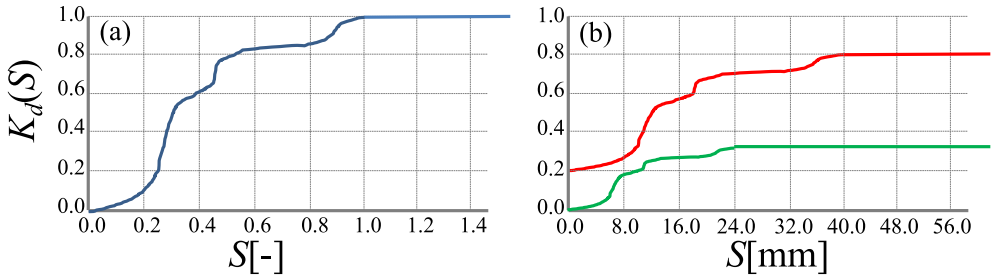

Figure 5. (a) Example of a physically plausible monotonically non-decreasing process parameterization function, and (b) an illustration of how the MR-PPE parameters $\left\{\theta_d^{low}, \theta_d^{high}, \theta_d^{scale}\right\}$ can act to alter the properties of the function. For the red curve $\theta_d^{low} = 0.2, \theta_d^{high} = 0.8$ and $\theta_d^{scale} = 40$ mm, and for the green curve $\theta_d^{low} = 0.0, \theta_d^{high} = 0.3$, and $\theta_d^{scale} = 24$ mm are assumed.

set the drainage/release parameters to be $\theta_r^{low} = \theta_r^{high}$ for the release function f_r, then the GB will represent an infinite capacity "*linear*" reservoir that drains at a constant rate (for more examples see Appendix A). In this study we set the upper limit of θ_x^{high} to 1 for all the three processes; however, the theoretical upper limit can be higher than 1 for release or depletion (for example in case of a linear reservoir, the analytical coefficient can be more than 1, or in case of evapotranspiration, trees can evaporate more than the reference potential evaporation typically calculated for grassland). We also would like to emphasize that higher values of θ_x^{high} can result in lower storage and state variable values, which in turn can transform the behavior of the state variable into that of a flux (reflecting the continuum between state and flux representations in a model)

We would like to remind the readers that the initial parameter values specified before the simulation should be evaluated for consistency at the end of the simulations. As indicated by Figures 3b and 3c, it is possible that only a fraction of the entire functional form of a PPE will be activated during the simulation. This may mean that the effective parameters specified when describing the MR-PPE may need to be different from those specified at the outset. This is important because poorly specified initial parameter values might indicate a scale range of, say, 40 mm, whereas the actual dynamic range of the storage may only cover a range of 20 mm over the course of the simulation. Similarly, while the "*low*" and "*high*" range parameters might initially be set to 0 and 1, it is possible that the function is only activated between 0.3 and 0.8, respectively during the course of the simulation (for more examples see Appendix B). In this study we focus only on the active portions of the MR-PPEs.

2.2. Design of Different System Architectural Hypotheses

For the experiments reported here, we use the GB introduced above (Section 2.1) to construct various system architecture hypotheses. We classify the system architecture into two categories, single- and multi-component.

2.2.1. Single Component Architectural Representations

The simplest system architectural hypothesis can be built by a single GB with the existence of three possible processes (namely absorption/bypass, release and depletion) resulting in the following three different system architectures:

1. ***S1 (Infinite Capacity Draining GB):*** In this case (Figure 6a), we impose no restriction on the ability of the precipitation (or effective precipitation) flux to enter and be stored in the catchment, and so the GB has effectively infinite capacity. The processes represented are the accumulation of soil moisture, and its depletion by evapotranspiration and a release water flux (e.g., lateral and/or vertical drainage).
2. ***S2 (Finite/Infinite Capacity Non-Draining GB):*** In this case (Figure 6b), the ability of precipitation (or effective precipitation) to enter the GB is partially restricted, with the remainder bypassing the GB as a water flux. The moisture stored in the GB is depleted by evapotranspiration only and no drainage release occurs in this model. If $\theta_{A_{a,b}}^{high}$ for the bypass and absorption is set to 1 then the GB is a finite capacity bucket otherwise it can represent an infinite capacity bucket.

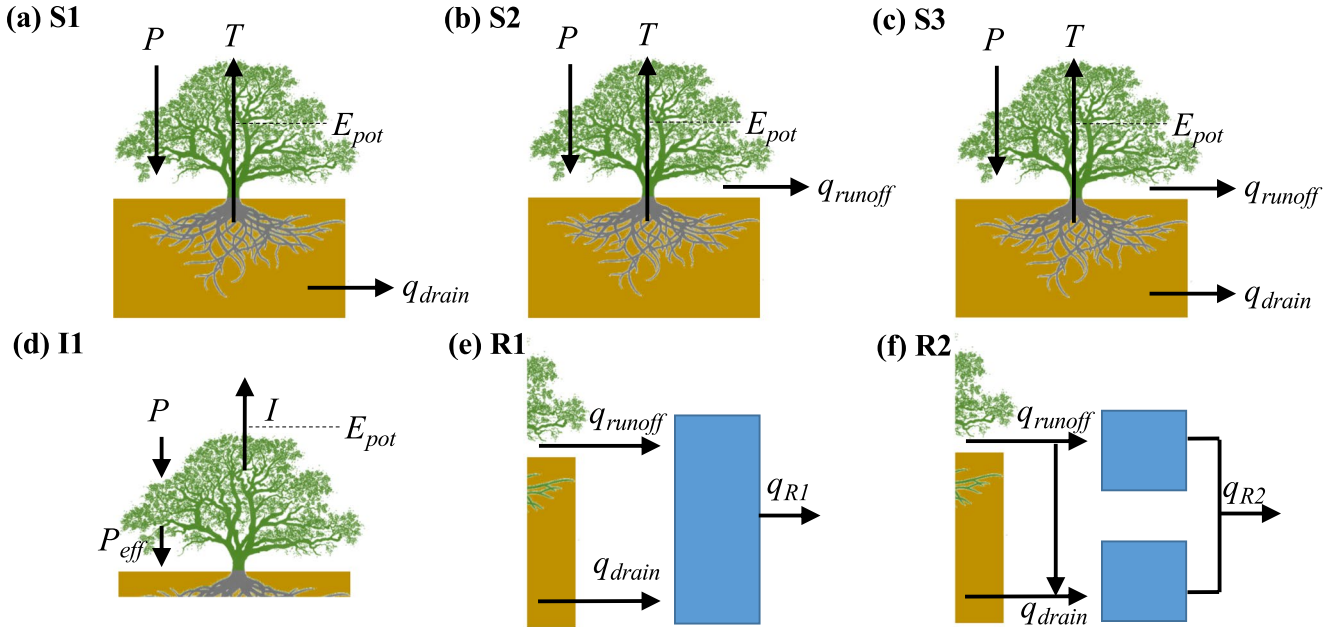

Figure 6. A model consisting of a single *GB* (a) *S1* that represents the partitioning of incoming precipitation flux (*P or* P_{eff}) into soil moisture storage (S_{sm}), evapotranspiration flux (T) and a drainage output flux (q_{drain}), (b) S2 that represents the partitioning of incoming flux into soil moisture storage, evapotranspiration flux, and surface runoff (q_{runoff}) and (c) S*3* that represents the partitioning of incoming flux into soil moisture storage, evapotranspiration flux, surface runoff (q_{runoff}), and lateral or vertical drainage (q_{drain}). (d) A model consisting of an *Interception Storage, I1* component, that partitions incoming precipitation flux into an effective precipitation bypass flux (P_{eff}), interception storage (*I*1), and evaporative flux (I). (e) A model consisting of a *one-pathway Routing, R1,* component representing a single-rate streamflow routing process. (f) A model consisting of a *two-pathway Routing, R2,* component representing parallel fast and slow streamflow routing processes.

3. ***S3 (Finite/Infinite Capacity Draining GB):*** This case (Figure 6c) represents a combination of the processes represented in model structures S1 and S2. The precipitation (or effective precipitation) is partly restricted from entering the GB and bypasses (e.g., as macropore water flow or surface flow), while the portion that accumulates within the GB can be depleted via evapotranspiration and released by lateral and/or vertical drainage. If $\theta^{high}_{A_{a,b}}$ is set to 1 the GB will represent a finite capacity bucket.

2.2.2. Multi-Component Architectural Representations

We further explore several hydrological model hypotheses having different structural architectures, by adding elements that perform the process functions of flow routing and interception to the three versions single-component models (S1, S2, and S3). Flow routing and interception can both be modeled via implementations of the GB as follows:

1. ***I1 (Interception Storage):*** Interception of precipitating water by the water holding capacity of the leaves of plants (and/or by other surfaces) can be modeled by use of a finite capacity non-draining GB (Figure 6d). The intercepted water is depleted by evaporation. Accordingly, an interception storage model component can be achieved similar to model S2.
2. ***R1 (One-Pathway Routing):*** A component that performs simple flow routing (Figure 6e) is achieved by implementing the GB in such a way that all of the incoming water flux enters the bucket (complete absorption with no bypass), no depletion occurs, and water is permitted to leave via drainage release.
3. ***R2 (Two-Pathway Routing):*** Two-pathway (fast and slow) flow routing (Figure 6f) is achieved by implementing two R1-type components in parallel, with the fractional input to each determined by a splitter (splitter has a value between zero and one). As a side note, splitter can be achieved by GB when setting the K_r value to very large numbers while setting the $A_{a,b}$ value to be equal to the splitter value ($\theta^{low}_{A_{a,b}} = \theta^{high}_{A_{a,b}} = \text{constant}$).

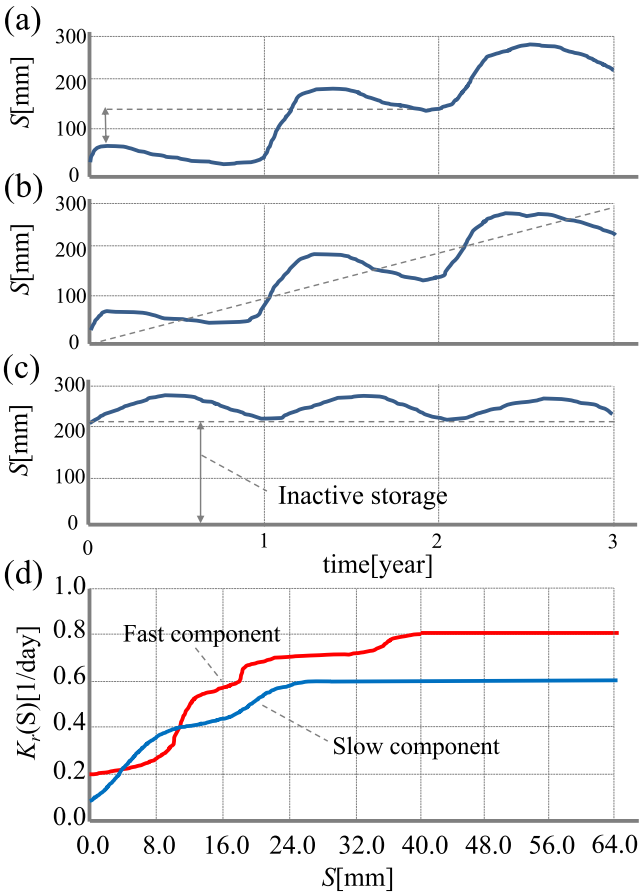

Figure 7. A schematic illustration of various conditions requiring that behavioral constraints be imposed on the model simulations. (a) A situation where the maximum storage value during year one (the first year) is less than the minimum value for another year (the third year). (b) A situation where the storage dynamics indicate a long-term increasing trend. (c) A situation where the ratio of the maximum to minimum storage values experienced during simulation is not sufficiently large to avoid the phenomenon of an inactive storage (200 mm of storage) that does not participate in the model dynamics. (d) A situation where the slow bucket/component in the model react faster than the fast bucket/component for storage values between 4.0 and 11.00 mm.

By combining elements of both kinds, single-component and multi-component, we can construct a variety of more complex structural architectures to represent different hypotheses regarding a catchment (commonly called a conceptual rainfall-runoff or bucket style model). In the following analysis we investigate and compare several different model structural hypotheses—to distinguish them we will use a naming convention as follows: a model that has an interception bucket (component *I1*), a bucket that partitions the incoming precipitation excess flux into soil moisture storage and surface runoff (component *S2*), and that generates streamflow via two-pathway routing (component *R2*) will be referred to as the "*I1-S2-R2*" model hypothesis.

2.3. Imposing Behavioral Constraints on Model Simulations

To ensure that our model simulations are "*behavioral*," we introduce the following constraints that restrict the allowable behaviors of the simulated input-state-output trajectories to be conceptually and physically realistic:

1. ***Non-Accumulating Behavior:*** We will require that the storage values in any of the model buckets should not follow a progressively increasing trend over long periods of time (such as years). This is achieved by imposing the following three different requirements on the storage values outside the period specified for model warm up, as listed below and illustrated using Figures 7a–7c.

(i) To ensure that water does not accumulate in a storage component at any significant rate, the largest yearly minimum water storage value should be less than the corresponding smallest yearly maximum.

(ii) To ensure that the storage values do not slowly increase over long periods of time, we fit a linear regression line to the timeseries of values and confine the slope of this line to be within ranges that satisfy our (subjective) perception of plausible storage dynamics (an increasing or decreasing trend of 100 mm/year).

(iii) To avoid situations wherein a bucket contains a relatively large amount of water but experiences dynamic storage variations over only a small range of high storage values, we constrain the ratio of the simulated maximum to minimum storage to be larger than some specified value (in this study we subjectively required this ratio to be larger than 4.0). This constraint prevents cases wherein significant portions of the accumulated water act as "*inactive storage*." For example, if the storage value only varies between 1,000 and 1,200 mm, then 1,000 mm of stored water does not participate in the dynamics of system behavior while it might affect the model simulation interpretation.

These assumptions can be seen as a stationarity assumptions on cycles of model behavior, over a year for example, and can be relaxed if external information or forcing that drives the system indicates otherwise. However, removing the stationarity assumption should be done with special care, given that most rainfall-runoff models do not possess memory of more than one hydrological cycle (Fowler et al., 2020).

2. ***Constraint on the relative behavior of system components:*** We will require that any storage having a "*long-time-constant*" (e.g., a slow-rate streamflow routing component) must drain more slowly than one having a "*short-time-constant*" (e.g., a fast-rate routing component) for equivalent values of the storage. An example of a violation of this constraint is illustrated in Figure 7d.
3. ***General knowledge about processes in the system of interest:*** We will require that any other known facts (obtained for example via field studies) be imposed as constraints when known. For example, one may have rough estimates of the ratio of transpired to intercepted water over some extended period of time. In the study region, the ratio of yearly interception to total evaporation has been observed to be

about 30%–50% (Gerrits et al., 2010), and we impose this knowledge as a constraint by only retaining model simulations that satisfy this ratio. If more precise information in the form of timeseries data is available, it can instead be used as an additional system response when inferring the forms of the MR-PPEs.

As the architectural structure of the model is made more complex, additional behavioral constraints of various kinds can be imposed. Examples include the relative functioning of riparian zone areas in comparison to hillslopes during dry and wet periods (e.g., see Gharari, Hrachowitz, et al., 2014; Gharari, Shafiei, et al., 2014). Simulations that fulfill these constraints are considered to be *constrained simulations*.

2.4. Measures of Uncertainty and Performance

2.4.1. Measures of Performance

To evaluate the properties of the model simulation ensembles, we use the Kling-Gupta efficiency metric, E_{KG}, proposed by Gupta, Kling, et al. (2009). E_{KG} is calculated as follows:

$$E_{KG} = 1 - \sqrt{O_1 + O_2 + O_3} \tag{5}$$

In which the components are:

$$O_1 = \left(1 - \beta\right)^2, \tag{6}$$

$$O_2 = \left(1 - \alpha\right)^2, \tag{7}$$

$$O_3 = \left(1 - r\right)^2 \tag{8}$$

where β is the ratio of the simulated to observed mean ($\beta = \mu_s/\mu_o$), α is the ratio of simulated to observed standard deviation ($\alpha = \sigma_s/\sigma_o$), and r is the cross-correlation coefficient between the simulations and observations for a deterministic simulation. In this regard, "*optimal models/simulations*" are those that correspond to the Pareto front formed by the three metrics, whereby no other model/simulation exists that can simultaneously provide better performance with regard to all three components. Such models/simulations are technically referred to as non-dominated solutions within the multi-objective framework. We choose to take the Pareto members as optimized simulations to minimize the effect of expert knowledge on the thresholds that are often used to identify behavioral simulations.

2.4.2. Measure of Uncertainty

We use the distance between the 10th and 90th quantile intervals of the simulated streamflow ensemble averaged over all of the simulation period time steps, as an approximate measure of uncertainty of the model response:

$$U = \frac{1}{n}\sum_{t=1}^{n}(Q_{t,90} - Q_{t,10}) \tag{9}$$

in which $Q_{t,10}$ and $Q_{t,90}$ are 10% and 90% quantile values of the ensemble of simulated streamflow values at a specific time, t, and n is the number of time steps. To focus mainly on model structural information, we assume no uncertainty in the input used to force the model (precipitation and potential evapotranspiration), or in the system response (streamflow).

3. Experimental Evaluation of Different Model Structures

3.1. Experiment-1: Single-Component System Architecture

3.1.1. Design of the Experiment

We first investigate the range of behaviors that are achievable using single-component models, S1, S2, and S3 (Section 2.2.1, Figures 6a–6c), that close the water balance (conservation law hypothesis).

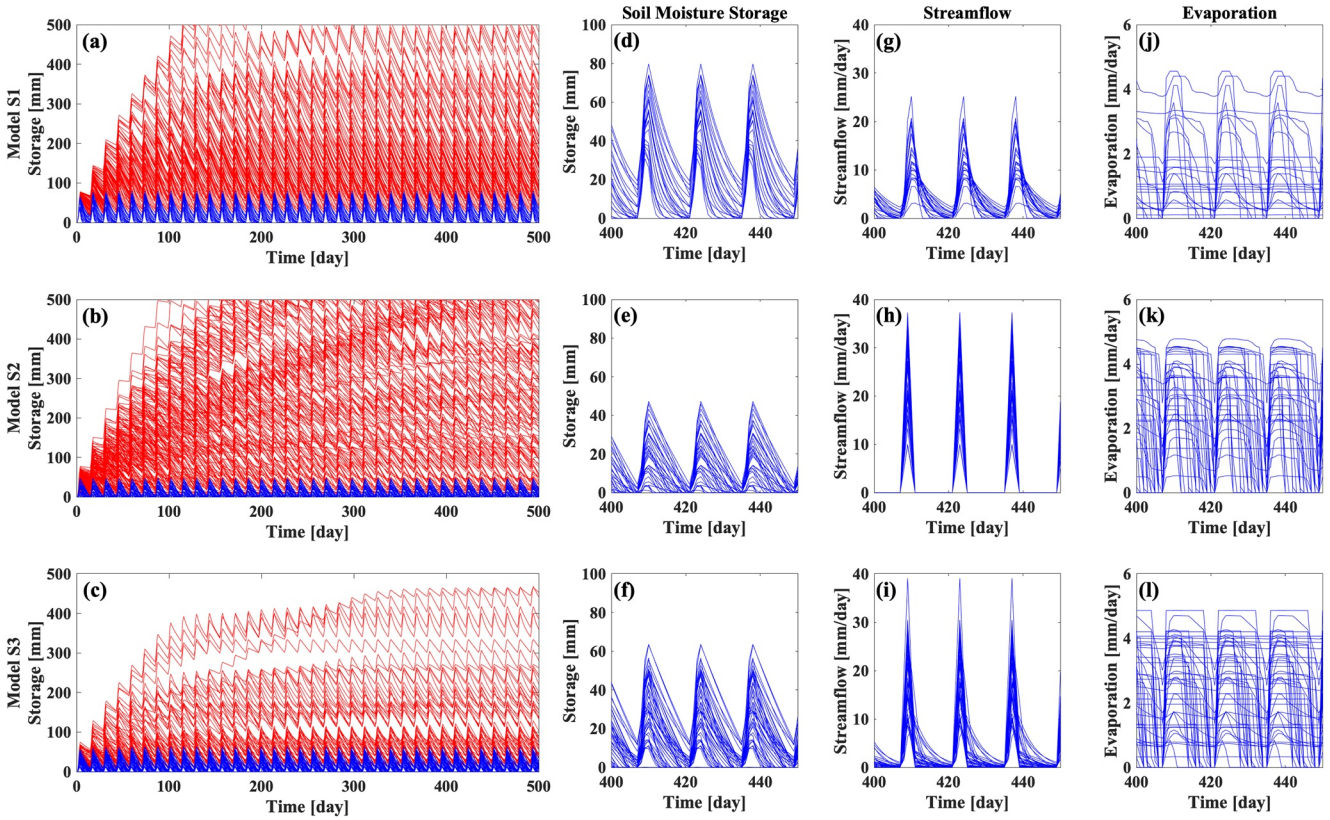

Figure 8. (a–c) Comparison of the storage of the GB for simulations without imposed constraints (red) to simulation with imposed constraints (blue) for S1, S2, and S3. The right three columns zoom into 50 days of the simulations that satisfy the constraints; (d–f) soil moisture storage (or simply storage) (g–i) streamflow ($q_{\text{runoff}} + q_{\text{drain}}$) and (j–l) and evapotranspiration.

To enable understanding of the possible input-state-output behaviors achievable using such model components, we impose a synthetic forcing data set consisting of three precipitation events having successive flux magnitudes of 20, 40, and 20 mm/day, each of 3-day duration, spaced 2 weeks apart (and repeated over and over again). Potential Evaporation is set to a constant value of 5 mm/day. This setup corresponds to approximately 1,600 mm/y of precipitation and 1,800 mm/y of potential evaporation. To assess the informational value of imposing behavioral constraints (Section 2.3), we simulate each model under two different conditions: (a) With no behavioral constraints imposed and (b) With all of the behavioral constraints imposed. We should remind the reader that the synthetic example provided here is conditional on the form of the forcing or input timeseries that we designed to force the model with (all of the inferences here are conditional on the information provided by the synthetic input/forcing).

3.1.2. Results

Figures 8a–8c illustrates the dynamical behavior of soil moisture storage simulated by each of the three different single-component representations (S1, S2, and S3; see Section 2.2.1) when no behavioral constraints are imposed (red) and when all of the behavioral constraints are imposed (blue).

Clearly, the imposition of behavioral constraints limits the range of variation of the soil moisture storage. When the behavioral constraints are not imposed, it becomes possible for storage values to increase to large values, after which the fluctuations are relatively small, depending on the forcing, and a volume of "*inactive storage*" forms that does not participate in the further dynamics of the model. This has significant implications for modeling, because changes in model structure can result in different inferences regarding the possible state of the actual system. The formation of this inactive storage can also result in the PPEs

being poorly activated across their overall range, which can make inference of the corresponding parameter values challenging (Figures 3b and 3c).

Figures 8d–8f compare, for S1, S2, and S3 respectively, the dynamical behavior (over a shorter period of time) of soil moisture storage when the behavioral constraints are imposed. Model structure S1, which is depleted only by evaporation and percolation, achieves the highest dynamic range (0–80 mm) because, unlike the other two structures, no water is permitted to bypass the GB. Model structure S3 shows larger agility in reproducing the range of behaviors, as it includes all the processes represented in S1 and S2. These results illustrate how small increases in process complexity can increase the flexibility of a model to emulate more diverse system behaviors.

Figures 8g–8i compare the corresponding simulated streamflow hydrographs. Model structure S1 is an infinite capacity GB, from which streamflow results only due to drainage from the GB, and so the dynamics of streamflow are damped. In contrast, model structure S2 is a non-draining GB, from which streamflow occurs only due to bypass of rainfall from the GB, and therefore the streamflow response is characterized by periods of rapid (flashy) response interspersed with periods when streamflow is zero. As expected, the composite structure of S3 results in more complex streamflow behaviors (compared to the other model), which clearly shows the informational benefit of structural flexibility in the system.

Finally, Figures 8j–8l compare the evaporation responses of the three models. As an example, model S2 that only depletes by transpiration, shows decreasing actual evaporation with time after the precipitation event. Model S3 however, can generate conditions in which evaporation is almost zero, because the system can be depleted by lateral flow. Overall, as with streamflow, this illustrates that, when a particular mathematical form for the evaporation function is employed with different model structures, one can arrive at different inferences regarding how water-stressed, wet or dry, the system can be.

Figure 9 illustrates the constrained minimally restrictive PPE forms obtained for the three models with single-component architecture. From Figures 9a–9c it is clear that the constrained forms (blue) of the bypass functions for models S2 and S3 are rather different, with S3 allowing a wider range of bypass function behaviors than model S2. Note that model S3 has *two* processes for emptying the storage (evaporation/depletion and lateral flow/release), and more degrees of freedom are therefore expected in the process parameterizations. For the evaporation process, S2 has the most identifiable (constrained) process parameterization, because the storage in model S2 can only be reduced via evapotranspiration process (Figures 9d–9f). These two examples illustrate how adding processes, when there is little knowledge regarding how to constrain them, may result in increased uncertainty associated with the process parameterizations and their related parameter values. For the drainage process parameterization, which is only present in models S1 and S3, Figures 9g and 9i indicate that the maximum storage amount and K_r values defining the release are inversely related. While this may not be surprising, it relates to our discussion in Section 2.1.2 where we mentioned how the state variable of a GB can effectively behave like a flux if the K_r is set to very high values.

We emphasize here that the progressive inclusion of various processes in the three models does not necessarily result in progressively reduced uncertainty bounds (ranges) on the model fluxes and states. Instead, it results in a "change" of uncertainty regarding the model fluxes and states.

3.2. Experiment-2: Exploring the Information Added by More Complex Structural Architectures

3.2.1. Design of the Experiment

In the second experiment, we investigate how the dynamical behavior of a model is shaped by changes to the model system architecture hypothesis (model topology, model structure). While the single-component models do simulate the three major processes of preferential flow, soil moisture accumulation and release, and evapotranspirative fluxes, more behavioral complexity can be achieved by incorporating additional structural components. In this experiment, we make use of a "*non-draining GB*" (S2; Figure 6b) that does not allow drainage release and construct a family of models that incorporate additional processes through a series of four progressively more complex system architectures.

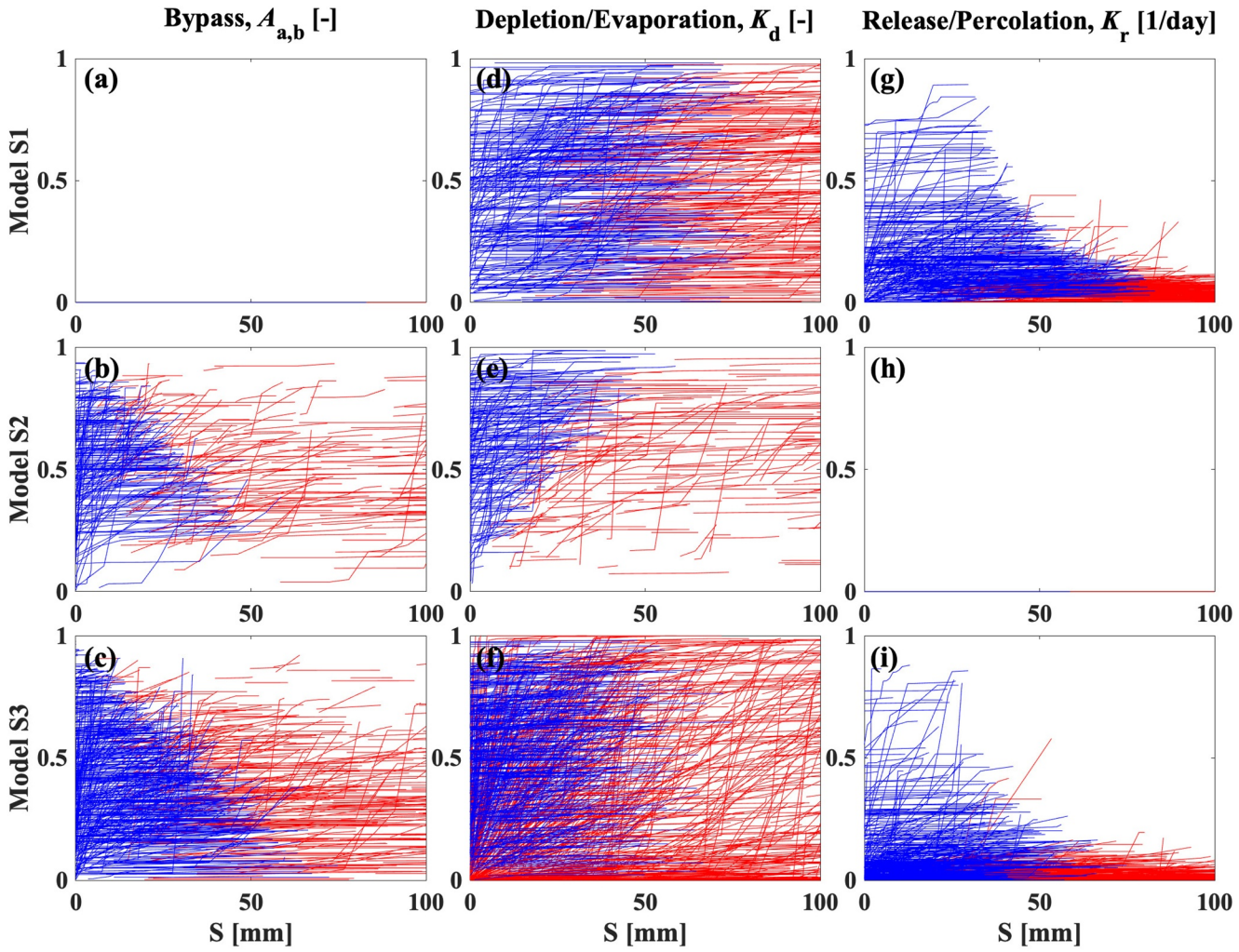

Figure 9. (a–c) The active part of the MR-PPEs for bypassing rainfall, (d–f) for relating evaporation to potential evaporation and (g–i) for percolation from the GB to its storage, for the three single-component models, S1, S2, and S3. Red indicates the PPEs when behavioral constraints are not imposed, and blue represent the PPEs when constraints are imposed.

Specifically, the first model architecture is S2 (as described above) and this is used as our baseline. The second model is S2-R1 in which a simple component that performs flow routing has been added, the third model is S2-R2 in which components that facilitate more complex two-pathway flow routing have been added, and the fourth model is I1-S2-R2 in which a precipitation interception process has been incorporated.

All four of these models are forced with daily precipitation and potential evapotranspiration, calculated based on the Hamon Equation (Hamon, 1960), from the Wark Catchment in the Grand Duchy of Luxembourg for three years 2005–2007 (2005 is used as a warmup period, and no behavioral constraint is applied to this time window). The simulations generated by these models are evaluated against observed streamflow data for the same period, using the performance and uncertainty metrics described in Section 2.4.

To investigate the interplay between the model system architectures and process parameterizations in more detail, we compare the transpiration functions associated with two model structures, S2-R2 and I1-S2-R2, obtained via behavioral constraining and via calibration to observed data.

To achieve these results, we need to tackle three issues as described below:

1. **Evaluating the information provided by 4 level of modeling hypotheses against the information provided by the forcing data:** As described earlier, forcing plays a significant role in providing

information regarding the input-state-output mapping. A model may perform well (or poorly) mainly due to the behavioral nature of the inputs that it is forced with. To characterize the information provided by the forcing (i.e., the information provided by precipitation and potential evaporation based on temperature), we first generated (samples of) all the possible streamflow sequences that can be realized by all possible system architectures constructed using the structural elements introduced in Section 2.2. These system architectures are S1, S2, and S3 and all of the corresponding variations, SX, SX-R1, SX-R2, I1-SX-R2. In principle, there can be an infinite number of possible system architectures, and a variety of minimally restrictive PPE forms that are more complex than those used in this study. The ensemble of all the possible model structures can provide a baseline for assessing what effects the forcing can have.

2. **Contrasting the information provided by conventional and minimally restrictive PPEs:** To compare the difference in information provided by the conventional rigidly pre-specified PPEs (CRP-PPEs) with that provided by the MR-PPEs, we use the MR-PPE framework to recreate the simpler forms of the CRP-PPEs to ensure a consistent numerical implementation. For the threshold-like CRP-PPE of the interception bucket, we design the MR-PPE to fully store water, so that when the maximum capacity is reached the incoming extra fluxes are bypassed (i.e., the bypass is always set to 0 except when storage exceeds the scale parameter, $\theta^{scale}_{A_{a,b}}$). For soil moisture, the bypass function is based on the power function commonly used in bucket style models, in which the ratio of soil moisture to maximum soil moisture capacity is used ($\left[\frac{S}{\theta^{scale}_{A_{a,b}}}\right]^{\beta}$, $0 < \beta < 4$ for $S < \theta^{scale}_{A_{a,b}}$). The fast and slow routing components, R1 and R2, both are constrained to have linear behavior ($\theta^{low} = \theta^{high}$) and are sampled from a logarithmic space (Figure A1f).
3. **Unifying the comparison of process parameterization values across different system architectures:** In model I1-S2-R2, transpiration from S2 is limited by the potential evaporation value depleted by the interception flux (E_p-I). To ensure a one-to-one comparison for models having and not having an interception component, we recompute the ratio of transpiration from S2 based on (E_p-I)/E_p for every time step for models S2-R2 and I1-S2-R2. As an example, for a given time step, if E_p is set to 3 mm day^{-1}, the interception flux is 2 mm day^{-1} and the transpiration to potential evaporation ratio is 1.0 so that, assuming no water limitation, transpiration flux should be 1 mm day^{-1}. The recomputed transpiration ratio is adjusted by a factor of (E_p-I)/E_p which is 1/3 for this example, and hence the "*effective*" value of K_d for transpiration is 1/3 instead of 1. This rescaled transmission values, K_d, are then comparable to that for a model that does not include the interception process. Similarly, to permit a one-to-one comparison of how much water bypasses S2 in the model I1-S2-R2 with that for model S2-R2, we recompute the ratio of bypass to effective precipitation (P_{eff}/P).

3.2.2. Results

Figure 10 illustrates the progressive change in uncertainty of the simulated streamflows, where the darker blue shading indicates the 10%–90% quantile intervals; the benchmark uncertainty associated with the forcing and the ensemble of all possible system architectures, SX, SX-R1, SX-R2, I1-SX-R2, considered here is shown in lighter blue. The red line indicates the observed streamflow values.

Figure 10a shows the results for model S2, which lacks a proper routing component and therefore generates a very flashy streamflow behavior during rainfall events, and no flow on days without precipitation. This architecture reduces the average uncertainty, U (see Equation 9), from 2.8 mm (light blue envelope) to 1.0 mm (dark blue envelope), but clearly does not represent the observed sequence of flows very well. As soon as a routing component is included (model S2-R1), the envelope of model simulations (Figure 10b) more closely reflects the behavior of the observed hydrograph (mainly during higher flow events), while the average uncertainty increases from 1.0 mm (model S2) to 1.9 mm (of course it is still smaller than the 2.8 mm).

This example nicely illustrates the trade-offs between uncertainty and performance that can occur as we vary the system architectural hypothesis. In this case, adding further structural elements (models S2-R2 and I1-S2-R2) progressively reduces the uncertainty from 1.9 to 1.5 mm and 1.3 mm, respectively (Figures 10c and 10d).

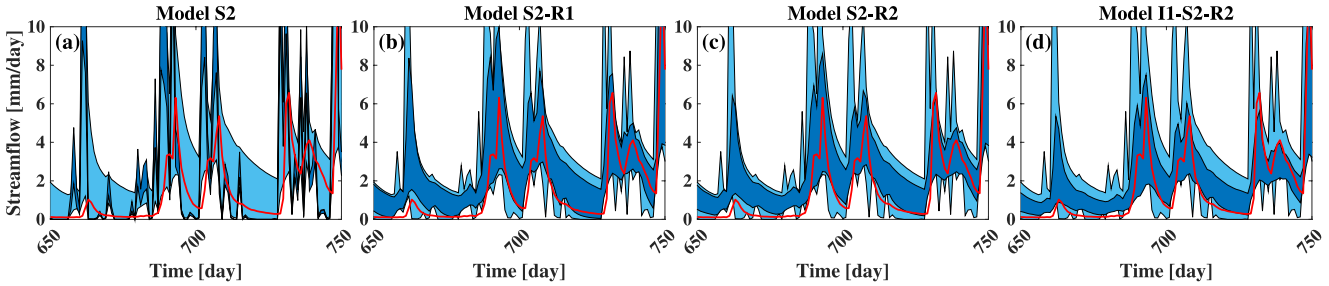

Figure 10. The model simulation envelope (10–90 percentile) for streamflows simulated by models (a) S2, (b) S2-R1, (c) S2-R2, and (d) I1-S2-R2 (darker blue) in comparison with benchmark uncertainty associated with the forcings and all possible system architectures considered here (lighter blue).

Next, we examine the E_{KG} component metrics O_1, O_2, and O_3 discussed in Section 2.4.1. Figures 11a–11c show metric boxplots for the MR-PPE ensembles generated using each model structure. Clearly, the general move is toward better overall performance (0.0 is best for all three metrics) as more structural elements are added. Adding the interception module (I1-S2-R2), clearly results in improved water balance performance (Figure 11a), as the model now has more ways of losing water. While this seems to be accompanied by a slight decrease in the ability to reproduce streamflow variability, the observed decrease is not significant.

Figures 11d–11f show corresponding plots obtained when the conventional approach of using deterministic fixed mathematical forms for the process parameterization equations is employed instead and the parameters are sampled over the full extent of their feasible ranges. Note that use of fixed mathematical forms amounts to providing very strong constraints on possible model behaviors (strong prior information). In this case, the main impact is seen in terms of the abilities of the

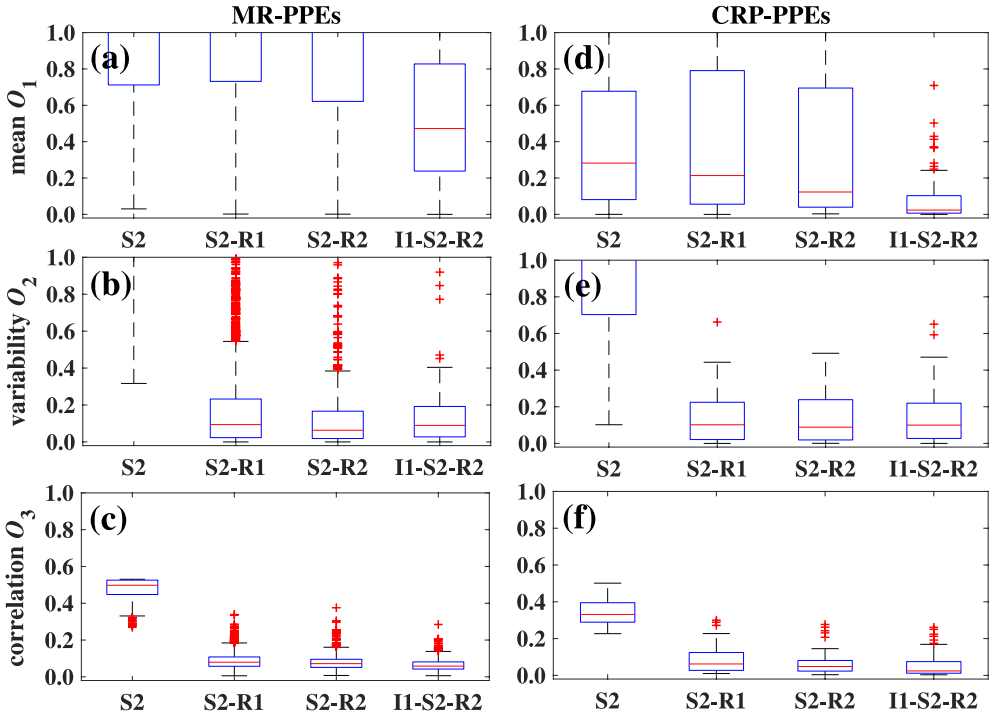

Figure 11. Plots comparing model performance (in terms of the E_{KG} component metrics O_1, O_2, and O_3) obtained using the MR-PPE approach (left column) and a conventional PPE or CRP-PPEs approach (right column). In each case 0.0 indicates best possible performance. The top row presents O_1, which is associated with overall long-term water balance for (a) MR-PPEs and (d) CRP-PPEs. The middle row illustrates O_2 which is associated with streamflow variability for (b) MR-PPEs and (e) CRP-PPEs. The bottom row shows O_3 which is associated with cross-correlation between simulated and observed streamflows for (c) MR-PPE and (f) CRP-PPEs.

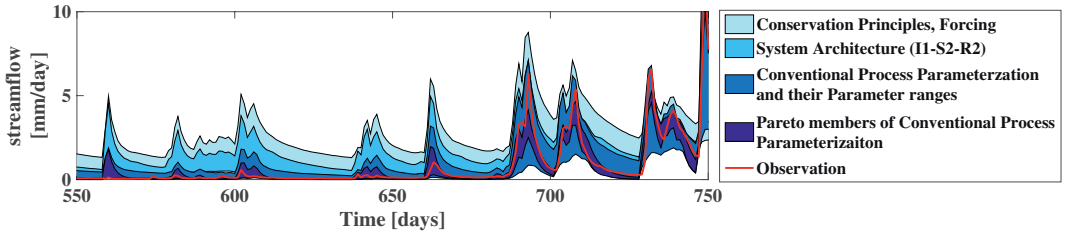

Figure 12. Progressive change in the 10–90 percentile streamflow intervals during the model building process. The lightest blue represents the benchmark ensemble of model simulations provided by the forcing only (conforming with conservation principles using all possible system architectures and MR-PPE forms considered here for models SX, SX-R1, SX-R2, I1-SX-R2). The next darker blue represents model I1-S2-R2 using minimally restrictive PPEs. The next darker blue represents the use of conventional process parameterization equations (CRP-PPEs). The darkest blue represents the behavioral simulations associated with the Pareto members obtained via calibration.

models to simulate the overall long-term catchment scale water balance (compare Figures 11a–11d), with no major impacts on variability and correlation.

It should be borne in mind that the "*improvement*" shown here is mainly a result of the fact that our implementation of the MR-PPE approach included only a limited set of PPE samples (1000) drawn from a very high-dimensional space of possible PPE forms. As such, the sample of CRP-PPE forms being compared with here can be considered as being a subset of the forms that can be constructed using the MR-PPE approach. So, what has actually been gained by using the CRP-PPE forms is that a very strong informational prior has been imposed, restricting the possible PPE forms that to a sub-region of the overall MR-PPE space, thereby improving the chance of selecting PPE forms that provide "good" model performance.

This result illustrates a limitation of our methodology—more exhaustive sampling should result in MR-PPE ensembles that include the CRP- PPE ensembles as a subset, whereas use of CRP-PPE forms reduces the space to be searched and therefore considerably reduces the computational demands associated with selection of PPE form. On the other hand, the MR-PPE approach provides more flexibility in selection of the PPE forms, with the potential to obtain a better representation of process behaviors and consequent overall system response.

It can be argued that our implementation of the MR-PPE approach reported here has been insufficiently well constrained using prior expert knowledge regarding process behavior (we did this for the purposes of illustration and investigation) and that a better compromise could perhaps be achieved by either (a) imposing more constraints on the MR-PPE form, or (b) starting with CRP-PPEs and progressively relaxing their structural forms to enable more flexibility in functional representation. For now, these remain as areas for future investigation.

Figure 12 shows how the overall ensemble envelope of streamflow simulations progressively changes as information is added at each level of model building. The lightest shade of blue represents the benchmark streamflow simulation uncertainty (determined by the forcings and the ensemble of all possible system architectures considered). The I1-S2-R2 structural hypothesis with MR-PPEs (second-lightest shade of blue) results in considerable reduction of the widths of the uncertainty ranges. When strong prior information about the forms of the PPEs is imposed—the MR-PPEs are replaced with CRP-PPEs (second-darkest shade of blue)—the widths of the uncertainty bounds are further significantly reduced during the relatively dry period (days 550–675), but not as much during the relatively wet period (days 675–750).

In general, for all three of these cases the uncertainty bounds bracket the streamflow observations specified in red except a few days around day 730. When the multi-objective calibration method is used to further constrain the parameters of the conventional PPEs (darkest shade of blue), we see that the uncertainty bounds no longer bracket the observations at all times, indicating that the model has become over-constrained (or over-confident).

As a practical matter, while it might be expected that the ensemble of streamflow simulations at each progressive modeling step should be contained within the benchmark uncertainty ensemble (i.e., every

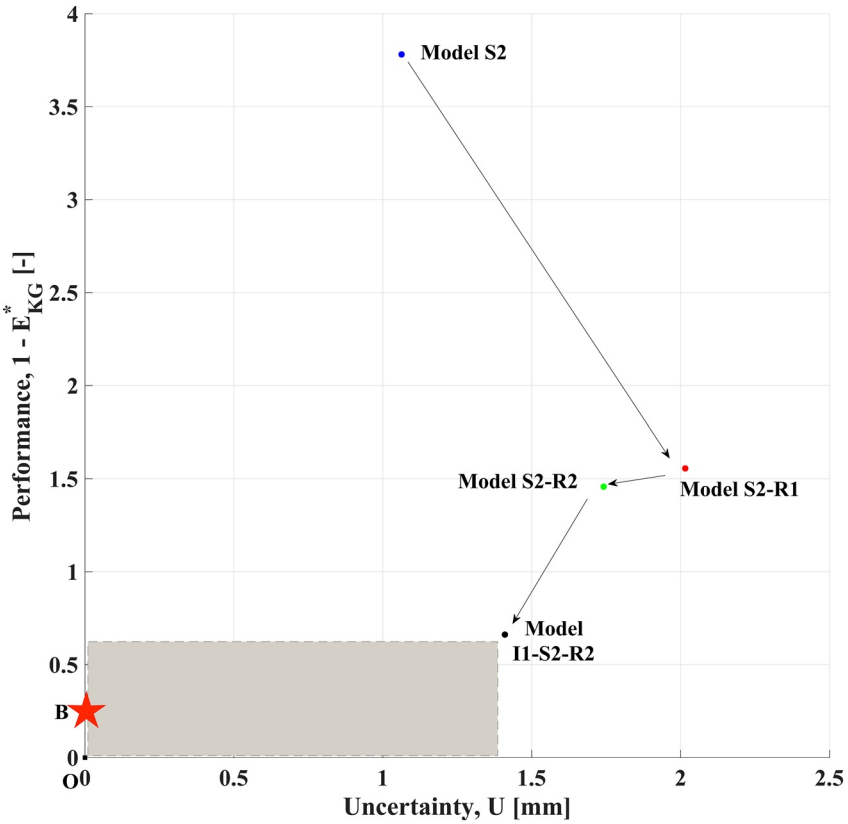

Figure 13. The effect of progressively adding information, in the form of structural elements, on both performance and uncertainty as the model system architecture is changed. If there is no further information to improve the model system architecture or process parameterization beyond that represented by the most complex model, I1-S2-R2, any possible solution in the area identified by the gray rectangle is unsupported by available knowledge and can only be inferred based on calibration to the observations. Point B represent a specific case where a single model simulation is chosen as behavioral via single objective calibration; in this case the apparent uncertainty is *artificially* (and unjustifiably) suppressed to be zero.

system architecture is contained within the ensemble, as illustrated in Figure 2e), the need to sample in high-dimensional spaces makes this difficult to demonstrate perfectly. Of course, this issue also exists when using conventional PPE forms. This should be borne in mind when examining the figures presented.

Figure 13 illustrates the interplay between uncertainty and overall model performance as information is added during the model building process and the model structural hypothesis becomes progressively more complex. Here uncertainty is quantified as the average (over all time steps) width of the 10%–90% quantile intervals, while performance is quantified using $1 - E^*_{KG}$, where E^*_{KG} is the average E_{KG} value taken over all ensemble members (best possible value is 0.0).

Adding a single-rate routing component to the single-component model ***S2***, thereby obtaining model ***S2-R1***, results in larger simulation uncertainty, but considerable improvement in average overall performance. Changing the routing scheme to a more complex one, having fast and slow pathways for flow (***S2-R2***), only slightly improves overall performance but does reduce the simulation uncertainty a bit. Further adding an interception component (***I1-S2-R2***) results in a significant improvement in performance, and some additional reduction in uncertainty.

This model development process could be continued by adding additional information in the form of hypotheses regarding system architecture and/or PPEs, with the goal of moving closer to the origin in Figure 13 (into the lower left-hand region indicated by the light gray box). At any point, the modeler can choose instead to use inverse procedures (calibration) to further constrain the model ensemble.

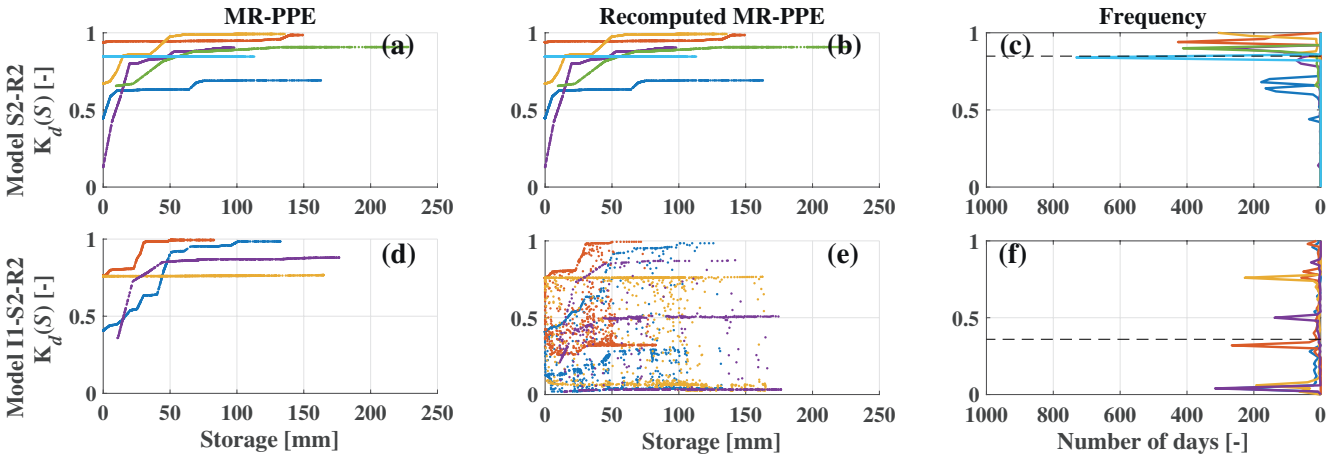

Figure 14. Model S2-R2 Evapotranspiration: (a) The minimally restrictive PPEs that are Pareto members and that satisfy all the constraints for evapotranspiration without recomputing to total potential evaporation; (b) Recomputed MR-PPE evaporation to total evaporation; (c) The frequency of days that a process parameterization for transpiration yields a value for the period of modeling. Model I1-S2-R2 Evapotranspiration: (d) The minimally restrictive PPEs that are Pareto members and that satisfy all the constraints for evapotranspiration without recomputing to total potential evaporation; (e) Recomputed MR-PPE evaporation to total evaporation; (f) The frequency of days that a process parameterization for transpiration yields a value for the period of modeling.

For illustration, we used calibration to select a single "best" model (a single parameter set) that gives the closest match to the observed streamflow data in terms of E_{KG}. Since this results in a deterministic representation of the system, we arrive at the point marked "**B**" where $E_{KG} \approx 0.75$ ($1 - E_{KG} \approx 0.25$**)** and uncertainty is zero. This further illustrates a point that should be kept in mind when interpreting Figure 13, which is that performance and uncertainty clearly do not represent "*independent*" aspects of model behavior; i.e., the fact that model performance is not perfect ($E_{KG} \neq 1$), while uncertainty is reported as being zero, indicates that we have over-constrained the model by adding "*too much information*" (first through the choice of conventional PPE forms and then through calibration on observed timeseries such as streamflow).

Accordingly, while real uncertainty continues to exist (as evidenced by the fact that the model does not perfectly reproduce the observations), the incorporation of "*bad*" information—by imposing overly strong restrictions on the PPE forms and by selecting only a single corresponding parameter set—has reduced the "*apparent*" simulation uncertainty to zero while not continuing to bracket (i.e., exactly match) the observed data. In other words, we have replaced simulation uncertainty with model structural overconfidence (an insidious form of model structural inadequacy).

It should also be kept in mind that the points representing each model in Figure 13 are not associated with "*deterministic*" models. By applying bootstrapping when computing the performance metric, and by considering input and response data uncertainties, neither of which was done here, each such point would actually be replaced by a fuzzy region (probability density) in the performance/uncertainty space.

Finally, we compare the process parameterization equation values obtained for two different system architectures, model structure ***S2-R2*** which has no interception component, and ***I1-S2-R2*** which does. Specifically, we look at the ratio that the process parameterization equation yields over the course of simulation for (a) transpiration from soil moisture and (b) water bypassing the soil moisture storage. In all cases, we examine only the functions obtained for the simulations that both satisfy the behavioral constraints and that lie on the Pareto Optimal frontier obtained by subjecting the simulation to streamflow observation (calibration). To enable one-to-one comparison across system architectures, the ***I1-S2-R2*** actual transpiration to potential evaporation ratio is rescaled based on precipitation and total potential evaporation.

The first row of Figure 14 is for the evapotranspiration process parameterization from the soil moisture GB of model ***S2-R2***. Subplots 14a and 14b are identical as recomputing the ratio of actual transpiration to potential evaporation has no effect for this model structure that lacks the canopy interception process (the ratio of actual evapotranspiration to potential evaporation follows the MR-PPEs). Corresponding to these, subplot 14c shows how often a particular ratio of actual transpiration to potential evaporation has

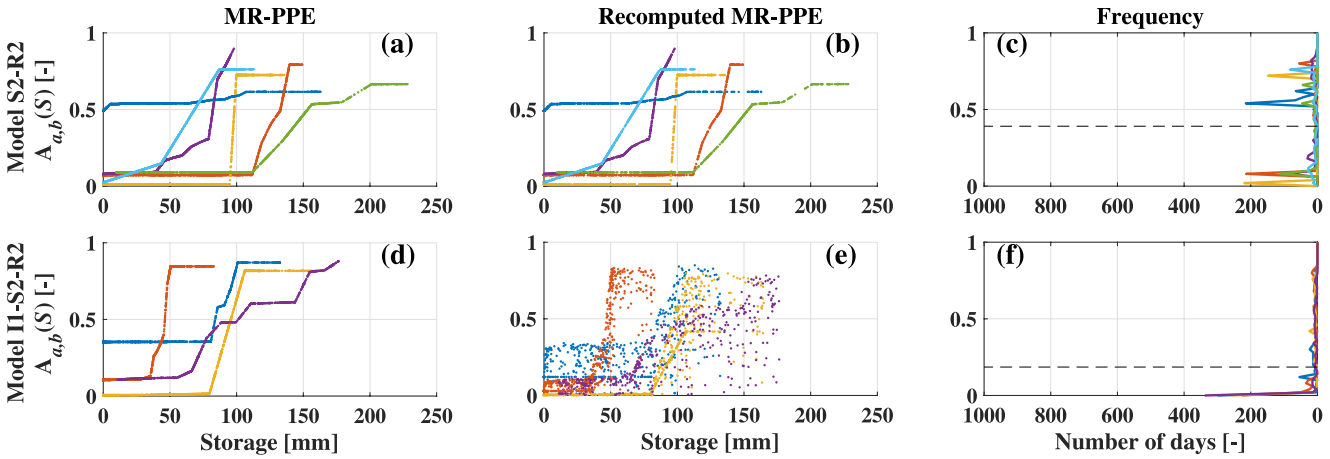

Figure 15. Model S2-R2 Bypass: (a) The minimally restrictive PPEs, that are Pareto members and satisfy all the constraints for the bypass without recomputing to total precipitation; (b) Recomputed MR-PPE bypass to total to total precipitation; (c) The frequency of days that a process parameterization for transpiration yields a value for the period of modeling. Model I1-S2-R2 Bypass: (d) The minimally restrictive PPEs that are Pareto members and that satisfy all the constraints for the bypass without recomputing to total precipitation; (e) Recomputed MR-PPE bypass to total to total precipitation; (f) The frequency of the days that a process parameterization for bypass yields a value for the period of modeling.

been activated; for example, the transpiration ratio to potential evaporation was ~0.8 on about 700 days in the simulation period. The dashed line indicates the average ratio to potential evaporation value over the entire period.

The second row is for the evapotranspiration process parameterization of model ***I1-S2-R2*** in which interception is also present and therefore the ratio of actual evaporation to potential evaporation different from values from MR-PPEs. Subplot 14d shows the MR-PPEs for transpiration, while subplot 14e shows the recomputed profile based on potential evaporation accounting for the effect of interception flux on the evapotranspiration. Notice, from the recomputed profile (Figure 14e), that the effect of interdependence between the transpiration and interception model components is now apparent, whereby a given value of potential evaporation and soil moisture storage does not map to a unique value of computed transpiration—there is an additional dependence on the amount of water evaporated from the interception storage (which reduces the demand on transpiration from the soil moisture storage). The consequence is that the distribution of days with similar ratio of transpiration to potential evaporation values has been dramatically altered, and the average ratio over the period is now only ~0.35. Clearly, this kind of interdependence must be accounted for when comparing different kinds of model structural hypotheses.

The first row of Figure 15 is for the bypass process parameterization of model ***S2-R2***. Again, subplots 15a and 15b are identical, as the recomputed ratio of bypassed water to total precipitation has no effect, while subplot 15c shows how often a particular bypass ratio (ratio of bypassed water to precipitation) has been activated on average (around 0.4). The second row show corresponding results for model ***I1-S2-R2***. Similar to the transpiration process the effect of process interdependence is again apparent, illustrating the fact that the dependence of estimates of bypass flow (macropore flow for example) from different model structural hypotheses should be interpreted in the broader context of processes that have been included/excluded. For example, this dependence of computed flux values on the complexity of the model structural hypothesis can have significant implications when coupling surface hydrology models with subsurface/groundwater models.

Another noteworthy aspect of subplots 15a and 15d (also 15b and 15e) is the apparent threshold-like nature (S-curve shape) of the inferred forms of the bypass process parameterization. This form is compatible with expert knowledge of the Wark Catchment, where a fill and spill mechanism has been observed to occur at the hillslope scale (Matgen et al., 2012; Tromp-van Meerveld & McDonnell 2006; Westhoff et al., 2011). Here we see that a similar mechanism may conceptually be inferred to be operating at the larger catchment scale. The inference becomes increasingly difficult when the system architecture becomes progressively more

complex (for example moving from P2-R2 to I1-P2-R2), and indicates the kind of challenge that a modeler may face when incorporating information into parameters and process parameterization of a complex model via calibration.

4. Summary, Discussion, and Implications for Future Work

4.1. Summary

In this work, we have investigated how the progressive steps in model development (including imposing behavioral constraints on model behavior, and ultimately model calibration to data), can impact the uncertainty and performance of model simulations as compared to the observed system responses. In particular, we have explored how the progressive addition of information (in the form of model structural hypotheses) can increase, decrease, or alter our uncertainty regarding the input-state-output behavior of a system.

This paper may be interpreted as call to shift the focus of model calibration from a search over "*parameter spaces*" to a search over "*function spaces*." It is known that the best performing parameter set based on conventionally used evaluation metrices may not result in the most hydrologically appealing simulation (Andréassian, Le Moine, et al., 2012; Beven, 2006), and hence we should strive to find those parameter sets that are hydrologically "*behavioral*". The same can be said about the process parameterizations, where *a priori* assumptions regarding the forms of the process equations may not represent the most appropriate forms in the context of a model implemented at a given scale. By introducing the concept of "*minimally restrictive process parameterization equations,"* or MR-PPEs, we have shown how the flexibility of process representations in models can be enhanced. This enables an investigation of the role that the system architecture hypothesis plays in determining model behavior, helping to reveal the complex process-level interactions that can occur (even in relatively simple models) as the system architecture is altered. The presented concept of MR-PPEs can be expanded to energy and momentum formulations in more complex models which can have profound implications for model uncertainty and sensitivity analysis. The MR-PPEs can play a significant role in assessment and quantification of, often inseparable, forcing uncertainty, model structural uncertainty and parameter uncertainty (Montanari & Koutsoyiannis, 2012). Whilst this study is not focused on machine learning, the presented MR-PPE can enable process-based modelers to benefit from the recent wave of ML techniques to learn from data.

4.2. Discussion

4.2.1. Implications to Model Implementation at Scale

Our study results have several implications for hydrological modeling practice. First, we have seen that the choice of system architecture hypotheses can substantially influence the forms that should be adopted for the PPEs in order to obtain behavioral input-state-output simulations. In current hydro-meteorological modeling practice, the true nature of our lack of knowledge regarding the appropriate forms of the PPEs is seldom represented. In general, we use *a priori* fixed functional forms for the PPEs, where the only flexibility available to the model is via adjustment of the parameter values within some predefined feasible space.

For example, transpiration is commonly represented in land models using a fixed formulation (such as the Penman-Monteith equations using Jarvis-type stomatal resistance terms), whose parameters values can often be difficult to specify from the available physical information on catchment attributes. Further, this lack of flexibility in functional choice does not take into consideration the manner by which the processes of canopy interception or subsurface flow are represented in the model. Specifying a fixed functional form imposes very strong restrictions on process behavior that may be difficult to defend given current knowledge of the system to which the model is being applied (Mendoza et al, 2015). By implementing a deterministic formulation, we are failing to represent the uncertainty of our knowledge of catchment scale transpiration behavior for the specific catchment of interest. This can result in systematic time-varying biases in the simulated input-state-output trajectories. To compensate for this, we may be forced to alter (increase) our

posterior representation of parameter uncertainty, and even this may not help to resolve the problems with systematic bias.

In this regard, we note that selecting parameter values should be based on the following considerations:

(1) Different models may be based on different system architectural hypotheses, which control the scales at which the system processes are being represented and whether or not various important system processes are being represented.
(2) Appropriate selection of the forms for the PPEs is necessarily conditional on the system architecture (and its resulting effects on operational scale and process description).
(3) The meaning/interpretation of parameters in the PPEs is conditional on the system architecture, and necessarily varies across models with different system architectures.

So, while process modelers should, and do, attempt to parameterize system processes in such a manner that the associated parameters are physically meaningful entities that are related (in principle) to actual physical properties of a given location, the connection between those parameters and meaningful properties of the physical system becomes somewhat less definitive when the equations are implemented into a hydrological model at scale. This issue clearly applies to the practice whereby parameter values for models implemented at some larger-scale are inferred from lookup tables that were in turn inferred from small-scale experiments.

4.2.2. Implications to Specification of Process Parameterization Form

Second, we have seen that imposing reasonable behavioral constraints based on expert knowledge can dramatically constrain the space of feasible PPE forms—and thereby the space of feasible input-state-output solutions. When used in conjunction with minimally restrictive PPEs, we can obtain insights into what functional forms are plausible for the process parameterization (at scale and consistent with the selected system architecture). Such insights can be used to guide the design/selection of conventional (fixed) formulations for the process parameterization hypothesis. Coupled with supportable hypotheses regarding system architecture, this may support better understanding of how the associated processes function at scale, which in turn may help to reduce the problem of systematic time-varying biases in the simulated input-state-output trajectories, and the associated need for model-correction/state-updating via data assimilation. Hopefully this will also pave the way for having flexible system architectures, and also process parameterizations that allow for a model to be used at any place (models of everywhere; Blair et al., 2019).

4.2.3. Implications to the Practice of Model Calibration via Parameter Optimization

Third, we noted during Experiment-2 of this study that a simple model architecture based on only two GB's (representing catchment storage and routing) is able to result in model performance (assessed in terms of the E_{KG} metric components) that is as high as that previously achieved using a more complex model applied to the catchment of interest. This illustrates the fact that use of model calibration/optimization to push the envelope of model performance may result in models with a variety of system architectures being able to excel in terms of performance metrics, while in no way guaranteeing that their internal representation of process dynamics is reliable and can be used to inform decision-making that requires model-based estimates of unobservable quantities. As a relatively simple example, the inclusion/exclusion of an interception component dramatically altered the model-based inferences of evapotranspiration rates from soil moisture storage.

So, while calibration can be a useful step in model development, it is worth questioning whether the resulting models (and associated model parameter values) can be the basis of very strong inference regarding the underlying physical properties of the system. Further, the potential value of using "*uncalibrated*" but behaviorally constrained models in combination with calibrated models should not be overlooked.

4.3. Implications for Future Work

Here, we briefly discuss several potential avenues for exploration that were not pursued in this paper, and which may motivate future studies. These avenues are discussed in the context of the research questions posed in Section 1.2:

4.3.1. What Information Does Each Level of Model Building Add in a Modeling Exercise?

In this study we investigated the nature of added information for a specific type of environmental model—namely the bucket style rainfall-runoff model. Similar strategies can be applied to a range of models constructed based on various modeling philosophies (e.g., process-based models that conserve energy and momentum). Despite existing efforts (see Loritz, Hassler, et al., 2017; Or et al, 2015; Vogel & Ippisch, 2008, as examples), it is of interest to know whether commonly used formulations such as the Richards Equation and its associated uncertainties can be inferred at the scale of interest (and to what degree). Similar concepts can be used to study the effects of system architecture in more complex models (Medici et al., 2012). Preliminary analysis of a semi distributed model using the minimally restrictive PPE concept can be found in Chapter 6b of Gharari (2016); for the sake of brevity, we have limited this current study to spatially lumped models. Given the recent focus on hyper resolution modeling (Maxwell et al., 2015; Wood et al., 2011), it is of interest to investigate, using the concept of MR-PPEs, how much information is actually gained from the use of assumptions imposed as distributions instead of by resolving the processes at finer model resolutions (for a debate on this, please refer to Beven et al. [2015] and Melsen et al. [2016]).

4.3.2. Given a Particular Model Architecture, What Mathematical Forms for the Process Parameterization are Consistent With the Information Contained in the Observed Data?

As shown in the experiments of this study, the choices made regarding system architecture will significantly affect the appropriate forms of process parameterization equations to be used in a model. The MR-PPE concept can enable modelers to investigate whether or not an assumed functional form, such as the widely used van Genuchten formulation (van Genuchten 1980), can in fact successfully represent the intended process, given the system architecture specified for a system. In this regard, MR-PPE formulations can be constructed with various degrees of freedom enabling them to span the spectrum of process parameterizations. Further, while this study only investigated simple representations wherein each process was conditioned on only a single state variable value, in general we might expect that such functional relationships can be multi-variate. Similarly, inclusion of more degrees of freedom when creating the process parameterizations (e.g., incorporating hysteretic components) may enable representation of the effects of small-scale processes (such as wetting and drying of soil) at the scale of the model (Appelbe et al., 2009; Gharari & Razavi, 2018; O'Kane & Flynn, 2007). Given that the PPEs must necessarily reflect uncertainties in knowledge of sub-element scale heterogeneities, one might expect such representations to take probabilistic functional forms (Riihimäki & Vehtari, 2010).

Accordingly, the use of large-sample databases and studies (Addor, Newman, et al., 2017; Duan et al., 2006; Knoben, Freer, Peel, et al., 2020; Mathevet et al., 2020) is necessary to enable such understanding of the process parameterization for larger spatial domains. The growing existence of other sources of information, such as remotely sensed products, may also prove to be helpful to the task of reconstructing PPE forms at the desired scale. While it has been suggested that inclusion of remotely sensed data can help to reduce modeling uncertainties (Crow et al., 2003; Livneh & Lettenmaier, 2012; Nijzink et al., 2018), such reports have typically been based on the development of process parameterizations for a single specific system response (such as runoff). It will be useful to investigate whether or not the information provided by remotely sensed data can, in fact, help to constrain a formulation based on use of minimally restrictive PPE forms and what those forms will become.

This study did not delve into spatially distributed models. In general, parameter specification can be based on a "*Property to Parameters Hypothesis (P2P)*" that relates the parameter values to material and geometrical properties of the system proposed (Götzinger & Bárdossy, 2007; Parajka et al., 2005). This set of hypotheses also imposes a form of model regularization. It is currently poorly understood which catchment characteristics add the most value to a regionalization effort. For example, while Samaniego et al. (2010) proposed a transfer function approach for upscaling model parameters based on soil characteristics, others report not finding such a strong correspondence (Addor, Nearing, et al., 2018; Merz et al., 2020; Oudin et al., 2010; Tafasca et al., 2020). The fact that the P2P hypothesis can be subject to significant uncertainty has been largely ignored in the literature. The concept of MR-PPE's may facilitate investigation into whether or not the often-assumed formulations of property to parameter and subsequent regionalization can be reproduced at the scale of interest (while considering associated uncertainties).

Of course, we expect that machine learning can play an important role in developing strategies to identify the MR-PPEs conforming to a particular model architecture and/or given application. In general, the efforts to incorporate machine learning into the Earth Sciences tend to be focused on: (a) learning directly from the data instead, without presence of a hydrological model (Addor, Nearing, et al., 2018; Stein et al., 2021) and/or (b) replacing an entire physics-based model, or its internal sub-structures or processes, with machine learning components (Bennett & Nijssen, 2020). The latter approach can benefit from the idea of minimally restrictive process parameterization equations (MR-PPEs) presented in this work, and is consistent with recent calls for process modelers to play a more active role in the use of machine learning for scientific discovery (Nearing et al., 2021).

Additionally, using machine learning, modelers can explore the degrees of freedom to which a MR-PPE should be exposed. This can enable more efficient exploration of the range of model structural architectures and process parameterizations and may help avoid unnecessary complexity in both the system architecture and the process parameterization equations. This is aligned with the concept of parsimony (or Occam's razor; Jakeman & Hornberger, 1993; Jakeman et al., 2006; Weijs & Ruddell, 2020). The MR-PPEs and their various form can also help understand the inherent uncertainties in Earth System Modeling.

4.3.3. How Uncertain are the Inferred PPEs, and How Does that Uncertainty Affect the Model Generated Simulations and Model Internal Behavior?

It is also important to develop strategies to account for the fact that the minimally restrictive PPEs will typically not be "properly" sampled over their entire behavioral ranges. Of course, this issue also arises for conventional PPEs where the marginal distribution of parameter uncertainty may result in only portions of the model space being extensively sampled. This will require the development of new families of optimization/search algorithms, that are capable of efficiently searching over functional spaces. In turn, this requires methods for developing, and sampling from, families of minimally restrictive PPEs that are consistent with known principles of physics (conservation, thermodynamics, etc.). Such PPE families will likely need to be constrained via the imposition of smoothness (capacity) and monotonicity constraints, while also permitting thresholding and saturation behaviors to be efficiently represented.

In regard to the information content of data sets, it is important to note that the overall identifiability of a model is strongly conditioned by the nature of the input forcing data and observed system response (Kavetski, et al., 2006; Vrugt, ter Braak, et al., 2008). As an exaggerated example, one can expect that the plausible range of input-state-output trajectories will increase when the data uncertainties are larger. Without loss of generality, the case studies presented here can be revisited by assuming an input uncertainty distribution, and by the use of likelihood measures to assess model performance; for a preliminary investigation see Chapter 6 of Gharari (2016). The MR-PPE concept can provide a platform for investigating the old debates regarding the dominance of model structural uncertainties or input uncertainties. Note that inferring the true nature of input and output data uncertainty cannot be achieved via the modeling exercise itself and must therefore be performed by examining the inherent uncertainties associated with the measurement techniques used (fully separate from the modeling hypotheses).

4.3.4. What is the Effect of Additional Behavioral Constraints (Additional Information Beyond That Provided in the Form of the Aforementioned System Hypotheses), Imposed as Modelers' Decisions, on the Nature of the Model Simulations?

Obtaining data sets that are informative about the forms of PPEs but that do not strongly influence streamflow, and instead control other aspects of the system response, can be a challenge. This is akin to the parameter identifiability problem associated with model calibration to estimate parameter values (Guillaume et al., 2019). For example, data is often not available to unambiguously constrain the partitioning of the precipitation flux into evaporation, transpiration, soil moisture, drainage to groundwater, and streamflow. As pointed out by Andréassian, Perrin, and Michel (2004), evaporative fluxes are generally inferred as the by-product of attempts to properly simulate streamflow (and thereby infer soil moisture). Further, remotely sensed transpiration product that might be used to constrain model responses are themselves the outputs of other models—often land [surface] models with their own hard-coded process parameterizations, parameters and assumptions, which increases the challenge of properly identifying the internal states and fluxes (Khatami et al., 2019).

In this regard, the use of behavioral constraints based on expert knowledge can serve as a form of "soft" regularizing information to enhance identifiability of the model parameterizations resulting in forms that are at least plausible while remaining consistent with the input-output data. Although suggested almost two decades ago by Seibert and McDonnell (2002), the development and use of such behavioral constraints is still largely missing from current hydrological modeling practice. Similarly, the interaction of "*expert knowledge*" in conjunction with various system architecture hypotheses is not very well explored (including, despite recent efforts, a "comprehensive" assessment of uncertainty and feedback processes). It is, for example, desirable to know how the observed threshold behaving fill and spill mechanism is translated into the form of a process parameterization (second experiment from this study). In the process of this evolving perspective on how to develop improved scientific understanding of hydrologic systems, we hope that this paper will inspire modelers and experimentalists to collaborate more closely, thereby learning from each other and improving the transfer of hydrological knowledge among them. Such a dialog will eventually improve understanding of the extent to which various processes can be successfully incorporated into the models we work with, which in turn will provide us with strategies to balance the resources spent on data collection effort alongside the modeling efforts.

5. Conclusions

The problem of designing an appropriate model-based representation of the processes governing the hydrological responses of a watershed is challenging. It can therefore be helpful to characterize the model development process as a hierarchical sequence of conditional hypotheses, beginning with the conservation law and system diagram hypothesis, and proceeding through the steps of system architecture specification, process parameterization specification, and parameter specification. Moreover, viewing the development process as progressively adding information to constrain model structural and behavioral uncertainty can help in diagnosing where and how "*bad*" information is being incorporated, leading to model structural inadequacy (Gupta, Clark, et al., 2012). Given modern computing power, it is (in principle) now possible to acknowledge the uncertainty in our hydrological knowledge and to "*err on the side of caution*" by adopting a maximum entropy approach to model development (Jaynes, 1963; Zehe, Ehret, et al., 2014), wherein we do not impose excessively strong prior information in the form of pre-specified deterministic forms for the process parameterization equations. At the same time, we can acknowledge and exploit the power of soft information in the form of reasonable and plausible behavioral constraints.

Our results suggest the following broader considerations:

1. The system architecture aspect of model structure design, including its scale related and process inclusion/exclusion considerations, can be as important, if not more so, than the issues of process parameterization equation selection and parameter specification. It seems important to reemphasize the value of spending more attention on the problem of system architecture and process parameterization identification and less on the problem of parameter identification. Subsequently, more of the efforts from uncertainty and sensitivity of parameters should be focused on process parameterization equations and system architectures.
2. Process parameterization equations should be treated as conditionally dependent on the system architecture selected for a given application. If there is reason to believe that a particular process parameterization equation form is definitively appropriate at a given spatiotemporal scale, then the system architecture should be selected in a way that is consistent with both the scale and also the process parameterization hypothesis.
3. The relative identifiability of the process parameterization equation forms can be strongly dependent on the degree of activation of the corresponding process, and such activation will rarely (if ever) be uniform across its behavioral range. Accordingly, we may have more confidence regarding the form of the process parameterization equation along some portions of its conceptual range and considerable uncertainty along other parts. Such uncertainty should, in principle be projected into the state and output spaces.
4. The use of "*soft information*" in the form of simple behavioral constraints on internal model behaviors, based in hydrologically informed knowledge, can help to dramatically increase the identifiability of the model state and output spaces and hence facilitate developing a "*fidelius*" model.

5. Model calibration, in the traditional sense of optimization to minimize a cost function, should only be used as an adjunct to the other stages of physical-conceptual model development. However, done correctly, with appropriate attention to proper imposition of prior knowledge in the development of the model structural representation, calibration of parameters can be replaced by inference of the form of the process parameterization equation (e.g., using machine-learning).

It is our hope that this study can inspire future work on how best to combine prior knowledge about the system of interest with the capabilities of machine learning to extract patterns from data. Despite the increasing attention toward application of machine learning in hydrological modeling, we still see a substantial philosophical chasm between those engaged in process-based modeling and those engaged in data science. On the one hand, the process-based modeling approach often results in time-stepping simulation models that are based in strong (and often rigid) hydrological hypotheses, assumptions and empirical formulations (e.g., see Clark, Schaefli, et al., 2016). Consequently, the hypotheses embedded in process-based models may be weak (e.g., not generally applicable), and the process parameterizations and parameters in models may be overly specific (see Mendoza et al., 2015). On the other hand, the machine learning approach focuses on extracting patterns from large datasets (e.g., Kratzert et al., 2019), thereby advancing traditional research in comparative hydrology (Gupta, Perrin, et al., 2014). Arguably, the promise of machine learning—that of learning from data—has yet to be realized, and current data science implementations have yet to demonstrate the ability to improve explanations (theories) of hydrologic processes. Recent developments in interpretable machine learning (e.g., Molnar et al., 2018; 2019) and explainable artificial intelligence (e.g., Arrieta et al., 2020) have the potential to enables data-driven discoveries within the context of process-based hydrologic models, bridging the gaps between data science and process-based hydrologic modeling. We hope that this paper will motivate additional efforts to integrate data science into process-based modeling studies, thereby strengthening the theoretical underpinnings of our models and improving our confidence in model predictions.

Appendix:

Appendix A. The Construction of Minimally Restrictive Process Parameterization Equations (MR-PPE)

A monotonically non-decreasing minimally restrictive process parameterization, MR-PPE, is a randomly generated function that projects $X \in [0\ 1]$ into $Y \in [0\ 1]$. The algorithm used in this study is described below:

1. Divide the X space into a finite number of points from 0:1, in steps of 0.05, and create the set of M which has 21 members from 0 to 1 ($X_0 = 0, X_1 = 0.05, ..., X_{20} = 1$).
2. Assume that $X_0 = 0$ then $Y_0 = 0$ and $X_{20} = 1$ then $Y_{20} = 1$; create the sets $M' = \{X_0, X_{20}\}$ and $M'' = \{Y_0, Y_{20}\}$.
3. Randomly pick a member X_i from set $M = \{X_1,..,X_{19}\}$, remove it from M and add it to M′.
4. Find the closest greater value X_{ig} and smaller value X_{is} to X_i in M′.
5. Find the corresponding Y_{ig} and Y_{is} values in set M″.
6. Generate a random value between Y_{ig} and Y_{is} and add this to the set M″.
7. If $M = \varphi$ then stop else go back to line 3.

We assume that the values of Y for X greater than 1 remain constant at 1 ($Y = 1$). A schematic view of how the MR-PPE is generated is illustrated in Figures A1a–A1d. To allow flexibility in generation of the function, three scaling parameters are introduced, with the role of adjusting the lower and upper limits of the random function and to scale the amount of storage $\theta_x = \left\{\theta_x^{low}, \theta_x^{high}, \theta_x^{scale}\right\}$. Figure A1e illustrates the effect of these three scaling parameters on a random function. For some storage-flux relations that need differentiation at very small values, the random function can be instead be generated using a semi-log space (Figure A1f).

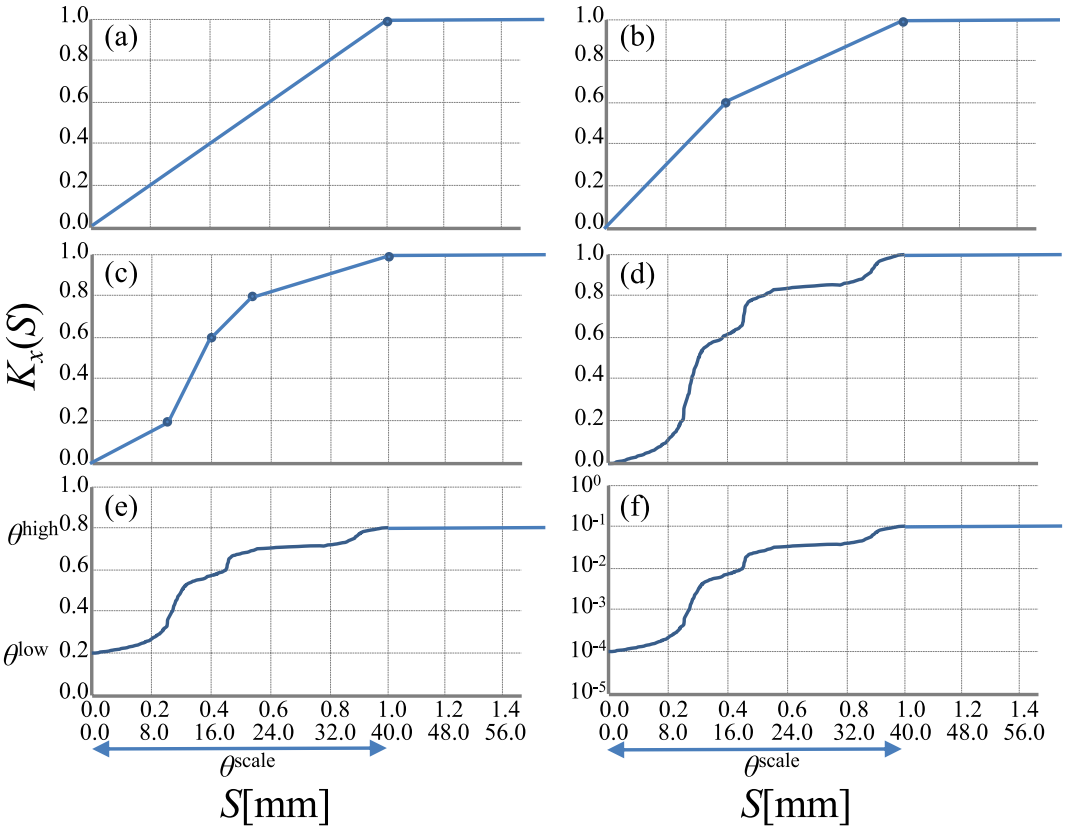

Figure A1. (a–d) An example of developing a monotonically non-decreasing random function and (e) the effect of scale parameters on the shape of the generated random function. In this example the $\theta^{low} = 0.2$ and $\theta^{high} = 0.8$ with $\theta^{scale} = 40$ mm. (f) Projection of the function into a semi-log space with $\theta^{low} = 10^{-4}$, $\theta^{high} = 10^{-1}$, and $\theta^{scale} = 40$ mm.

It is also possible to create (emulate) commonly used PPE forms using this methodology. An MR-PPE that behaves as a threshold can be designed by constraining the random function to always return zero below storage equal to θ^{scale} and return one otherwise; in this case the reservoir will act as a non-leaky bucket. Similarly, by setting θ^{low} and θ^{high} to the same (equivalent) values one can represent the behavior of a linear bucket.

Appendix B

To determine the active portion of a PPE at the end of the model simulation, the MR-PPEs are restricted to being active only in the region between minimum and maximum storage. This results in only the active part of the parameterization being represented. The inactive portions of the parameterization that lie outside of these bounds contain no information that is relevant to the output simulation, as the model never uses them. For the problem of inversely inferring the PPE forms given data (and other hypotheses and assumptions), it is important to use only the active portion of the parameterization to avoid misjudgment about the nature of the PPE. The values for θ^{low}, θ^{high}, and θ^{scale} can be further refined during inference (Figure B1).

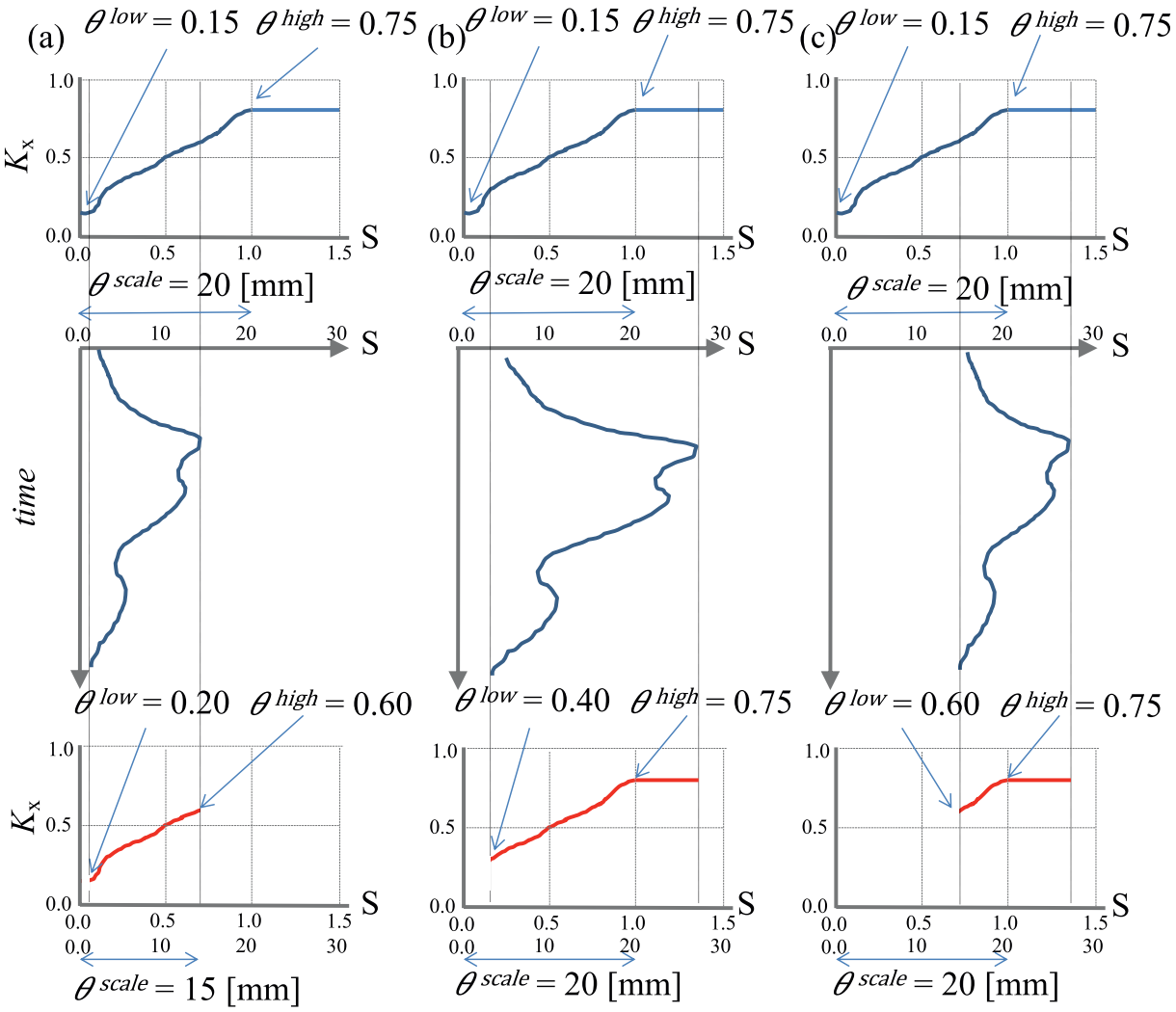

Figure B1. Example illustrating the active portion of a minimally restrictive PPE and its updated parameters.

Data Availability Statement

The data used in this study are made publicly available by administrations of Grand Duchy of Luxembourg. The data belongs to Administration de la Gestion de l'Eau and the Administration de la navigation aérienne (MeteoLux).

Acknowledgments
The authors thank the editor Charles Luce for facilitating the very constructive review process for this work in Water Resources Research. The authors also thank Larry Band, Lieke Melsen, Ralf Loritz, and one anonymous reviewer for their constructive comments over three rounds of revisions.

References

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E., & Rasmussen, J. (1986). An introduction to the European Hydrological System—Systeme Hydrologique Europeen, "SHE", 1: History and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, *87*(1–2), 45–59. https://doi.org/10.1016/0022-1694(86)90114-9

Addor, N., & Melsen, L. A. (2019). Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resources Research*, *55*(1), 378–390. https://doi.org/10.1029/2018wr022958

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, *54*(11), 8792–8812. https://doi.org/10.1029/2018wr022606

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M. H., Oudin, L., Mathevet, T., et al. (2012). All that glitters is not gold: The case of calibrating hydrological models. *Hydrological Processes*, *26*, 2206. https://doi.org/10.1002/hyp.9264

Andréassian, V., Perrin, C., & Michel, C. (2004). Impact of imperfect potential evapotranspiration knowledge on the efficiency and parameters of watershed models. *Journal of Hydrology*, *286*(1–4), 19–35. https://doi.org/10.1016/j.jhydrol.2003.09.030

Antonetti, M., Scherrer, S., Kienzler, P. M., Margreth, M., & Zappa, M. (2017). Process-based hydrological modelling: The potential of a bottom-up approach for runoff predictions in ungauged catchments. *Hydrological Processes*, *31*(16), 2902–2920. https://doi.org/10.1002/hyp.11232

Appelbe, B., Flynn, D., McNamara, H., O'Kane, P., Pimenov, A., Pokrovskii, A., et al. (2009). Rate-independent hysteresis in terrestrial hydrology. *IEEE Control Systems*, *29*(1), 44–69. https://doi.org/10.1109/MCS.2008.930923

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bahremand, A. (2016). HESS Opinions: Advocating process modeling and de-emphasizing parameter estimation. *Hydrology and Earth System Sciences*, *20*(4), 1433–1445. https://doi.org/10.5194/hess-20-1433-2016

Bancheri, M., Serafin, F., & Rigon, R. (2019). The representation of hydrological dynamical systems using Extended Petri Nets (EPN). *Water Resources Research*, *55*(11), 8895–8921. https://doi.org/10.1029/2019WR025099

Bárdossy, A., & Singh, S. K. (2008). Robust estimation of hydrological model parameters. *Hydrology and Earth System Sciences*, *12*(6), 1273–1283. https://doi.org/10.5194/hess-12-1273-2008

Beck, M. B. (1983). Uncertainty, system identification, and the prediction of water quality. In M. B. Beck, G. van Straten, IIASA International Institute for Applied Systems Analysis (Eds.), *Uncertainty and forecasting of water quality*. Springer. https://doi.org/10.1007/978-3-642-82054-0_1

Bennett, A., & Nijssen, B. (2020). *Earth and Space Science Open Archive ESSOAr*. https://doi.org/10.1002/essoar.10505081.1

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K., Cloke, H., Pappenberger, F., Lamb, R., & Hunter, N. (2015). Hyperresolution information and hyperresolution ignorance in modelling the hydrology of the land surface. *Science China Earth Sciences*, *58*(1), 25–35. https://doi.org/10.1007/s11430-014-5003-4

Beven, K., & Westerberg, I. (2011). On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrological Processes*, *25*(10), 1676–1680. https://doi.org/10.1002/hyp.7963

Beven, K. J. (2011). *Rainfall-runoff modelling: The primer*. John Wiley and Sons.

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Bulletin*, *24*(1), 43–69. https://doi.org/10.1080/02626667909491834

Blair, G. S., Beven, K., Lamb, R., Bassett, R., Cauwenberghs, K., Hankin, B., et al. (2019). Models of everywhere revisited: A technological perspective. *Environmental Modelling & Software*, *122*, 104521. https://doi.org/10.1016/j.envsoft.2019.104521

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*(3–4), 251–290. https://doi.org/10.1002/hyp.3360090305

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

Bulygina, N., & Gupta, H. (2010). How Bayesian data assimilation can be used to estimate the mathematical structure of a model. *Stochastic Environmental Research and Risk Assessment*, *24*(6), 925–937. https://doi.org/10.1007/s00477-010-0387-y

Bulygina, N., & Gupta, H. (2011). Correcting the mathematical structure of a hydrological model via Bayesian data assimilation. *Water Resources Research*, *47*(5), W05514. https://doi.org/10.1029/2010WR009614

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9), W09301. https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, *51*(4), 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, *52*(3), 2350–2365. https://doi.org/10.1002/2015wr017910

Clark, M. P., & Slater, A. G. (2006). Probabilistic quantitative precipitation estimation in complex terrain. *Journal of Hydrometeorology*, *7*(1), 3–22. https://doi.org/10.1175/jhm474.1

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2007wr006735

Condon, L. E., Maxwell, R. M., & Gangopadhyay, S. (2013). The impact of subsurface conceptualization on land energy fluxes. *Advances in Water Resources*, *60*, 188–203. https://doi.org/10.1016/j.advwatres.2013.08.001

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J. M., & Jones, P. D. (2018). An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, *123*(17), 9391–9409. https://doi.org/10.1029/2017jd028200

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, W., Jost, G., Lee, K., et al. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling & Software*, *129*, 104728.

Crow, W. T., Wood, E. F., & Pan, M. (2003). Multiobjective calibration of land surface model evapotranspiration predictions using streamflow observations and spaceborne surface radiometric temperature retrievals. *Journal of Geophysical Research*, *108*(D23), 4725. https://doi.org/10.1029/2002JD003292

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Goteti, G., Gupta, H. V., et al. (2006). Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology*, *320*(1–2), 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, *17*(5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, *47*(11), W11510. https://doi.org/10.1029/2010WR010174

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

Fitts, C. (2012). *Groundwater science* (2nd ed.). Academic Press.

Flügel, W.-A. (1995). Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Bröl, Germany. *Hydrological Processes*, *9*(3–4), 423–436. https://doi.org/10.1002/hyp.3360090313

Fowler, K., Knoben, W., Peel, M., Peterson, T., Ryu, D., Saft, M., et al. (2020). Many commonly used rainfall-runoff models lack long, slow dynamics: Implications for runoff projections. *Water Resources Research*, *56*(5), e2019WR025286. https://doi.org/10.1029/2019wr025286

Freer, J., Beven, K., & Ambroise, B. (1996). Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research*, *32*(7), 2161–2173. https://doi.org/10.1029/95WR03723

Freer, J. E., McMillan, H., McDonnell, J. J., & Beven, K. J. (2004). Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology*, *291*(3–4), 254–277. https://doi.org/10.1016/j.jhydrol.2003.12.037

Gao, H., Birkel, C., Hrachowitz, M., Tetzlaff, D., Soulsby, C., & Savenije, H. H. G. (2019). A simple topography-driven and calibration-free runoff generation module. *Hydrology and Earth System Sciences*, *23*(2), 787–809. https://doi.org/10.5194/hess-23-787-2019

Gerrits, A. M. J., Pfister, L., & Savenije, H. H. G. (2010). Spatial and temporal variability of canopy and forest floor interception in a beech forest. *Hydrological Processes*, *24*(21), 3011–3025. https://doi.org/10.1002/hyp.7712

Gharari, S. (2016). *On the role of model structure in hydrological modeling: Understanding models* (PhD thesis).

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., & Savenije, H. H. G. (2014a). Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. *Hydrology and Earth System Sciences*, *18*(12), 4839–4859. https://doi.org/10.5194/hess-18-4839-2014

Gharari, S., & Razavi, S. (2018). A review and synthesis of hysteresis in hydrology and hydrological modeling: Memory, path-dependency, or missing physics? *Journal of Hydrology*, *566*, 500–519. https://doi.org/10.1016/j.jhydrol.2018.06.037

Gharari, S., Shafiei, M., Hrachowitz, M., Kumar, R., Fenicia, F., Gupta, H. V., & Savenije, H. H. G. (2014b). A constraint-based search algorithm for parameter identification of environmental models. *Hydrology and Earth System Sciences*, *18*(12), 4861–4870. https://doi.org/10.5194/hess-18-4861-2014

Goodwell, A. E., & Kumar, P. (2017). Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. *Water Resources Research*, *53*(7), 5920–5942. https://doi.org/10.1002/2016wr020216

Götzinger, J., & Bárdossy, A. (2007). Comparison of four regionalisation methods for a distributed hydrological model. *Journal of Hydrology*, *333*(2–4), 374–384. https://doi.org/10.1016/j.jhydrol.2006.09.008

Guillaume, J. H., Jakeman, J. D., Marsili-Libelli, S., Asher, M., Brunner, P., Croke, B., et al. (2019). Introductory overview of identifiability analysis: A guide to evaluating whether you have the right type of data for your modeling purpose. *Environmental Modelling & Software*, *119*, 418–432. https://doi.org/10.1016/j.envsoft.2019.07.007

Gupta, H., Thiemann, M., Trosset, M., & Sorooshian, S. (2003). Reply to comment by K. Beven and P. Young on "Bayesian recursive parameter estimation for hydrologic models". *Water Resources Research*, *39*(5). https://doi.org/10.1029/2002WR001405

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*(8). https://doi.org/10.1029/2011wr011044

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H. V., & Nearing, G. S. (2014). Debates-the future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science. *Water Resources Research*, *50*(6), 5351–5359. https://doi.org/10.1002/2013WR015096

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*(18), 3802–3813. https://doi.org/10.1002/hyp.6989

Hamon, W. R. (1960). *Estimating potential evapotranspiration* (B.S. dissertation, p. 75). Department of Civil and Sanitary Engineering, Massachusetts Institute of Technology. Retrieved from https://dspace.mit.edu/handle/1721.1/79479

Hrachowitz, M., & Clark, M. P. (2017). HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. *Hydrology and Earth System Sciences*, *21*(8), 3953–3973. https://doi.org/10.5194/hess-21-3953-2017

Jakeman, A. J., & Hornberger, G. M. (1993). How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, *29*(8), 2637–2649. https://doi.org/10.1029/93wr00877

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, *21*(5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

Jarvis, P. G. (1976). The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *273*(927), 593–610.

Jaynes, E. T. (1963). Information theory and statistical mechanics. In K. Ford (Ed.), *Statistical physics* (p. 181). Benjamin.

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, *47*(13), e2020GL088229. https://doi.org/10.1029/2020GL088229

Johnston, P. R., & Pilgrim, D. H. (1976). Parameter optimization for watershed models. *Water Resources Research*, *12*(3), 477–486. https://doi.org/10.1029/WR012i003p00477

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2019). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544–1554. https://doi.org/10.1109/TKDE.2018.2861006

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005wr004376

Khatami, S., Peel, M. C., Peterson, T. J., & Western, A. W. (2019). Equifinality and flux mapping: A new approach to model evaluation and process representation under uncertainty. *Water Resources Research*, *55*(11), 8922–8941. https://doi.org/10.1029/2018wr023750

Kiang, J. E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I. K., et al. (2018). A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, *54*(10), 7149–7176. https://doi.org/10.1029/2018wr022708

Kirchner, J. W. (2009). Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research*, *45*(2). https://doi.org/10.1029/2008wr006912

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

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24.

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., & Woods, R. A. (2019). Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) v1.2: An open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Geoscientific Model Development*, *12*(6), 2463–2480. https://doi.org/10.5194/gmd-12-2463-2019

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resources Research*, *56*(9), e2019WR025975. https://doi.org/10.1029/2019wr025975

Knudsen, J., Thomsen, A., & Refsgaard, J. C. (1986). WATBAL: A semi-distributed, physically based hydrological modelling system. *Hydrology Research*, *17*(4–5), 347–362. https://doi.org/10.2166/nh.1986.0026

Koster, R. D., & Mahanama, S. P. P. (2012). Land surface controls on hydroclimatic means and variability. *Journal of Hydrometeorology*, *13*(5), 1604–1620. https://doi.org/10.1175/jhm-d-12-050.1

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology & Earth System Sciences*, *23*(12). https://doi.org/10.5194/hess-23-5089-2019

Lamb, R., & Beven, K. (1997). Using interactive recession curve analysis to specify a general catchment storage model. *Hydrology and Earth System Sciences*, *1*(1), 101–113. https://doi.org/10.5194/hess-1-101-1997

Leavesley, G. H., Markstrom, S. L., Brewer, M. S., & Viger, R. J. (1996). The modular modeling system (MMS) — The physical process modeling component of a database-centered decision support system for water and power management. *Water, Air, & Soil Pollution*, *90*(1–2), 303–311. https://doi.org/10.1007/bf00619290

Livneh, B., & Lettenmaier, D. P. (2012). Multi-criteria parameter estimation for the Unified Land Model. *Hydrology and Earth System Sciences*, *16*(8), 3029–3048. https://doi.org/10.5194/hess-16-3029-2012

Loritz, R., Gupta, H., Jackisch, C., Westhoff, M., Kleidon, A., Ehret, U., & Zehe, E. (2018). On the dynamic nature of hydrological similarity. *Hydrology and Earth System Sciences*, *22*(7), 3663–3684. https://doi.org/10.5194/hess-22-3663-2018

Loritz, R., Hassler, S. K., Jackisch, C., Allroggen, N., van Schaik, L., Wienhöfer, J., & Zehe, E. (2017). Picturing and modeling catchments by representative hillslopes. *Hydrology and Earth System Sciences*, *21*(2), 1225–1249. https://doi.org/10.5194/hess-21-1225-2017

Martinez, G. F., & Gupta, H. V. (2010). Toward improved identification of hydrological models: A diagnostic evaluation of the "*abcd*" monthly water balance model for the conterminous United States. *Water Resources Research*, *46*(8), W08507. https://doi.org/10.1029/2009WR008294

Matgen, P., Fenicia, F., Heitz, S., Plaza, D., de Keyser, R., Pauwels, V. R. N., et al. (2012). Can ASCAT-derived soil wetness indices reduce predictive uncertainty in well-gauged areas? A comparison with in situ observed soil moisture in an assimilation application. *Advances in Water Resources*, *44*, 49–65. https://doi.org/10.1016/j.advwatres.2012.03.022

Mathevet, T., Gupta, H., Perrin, C., Andréassian, V., & Le Moine, N. (2020). Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds. *Journal of Hydrology*, *585*, 124698. https://doi.org/10.1016/j.jhydrol.2020.124698

Maxwell, R. M., Condon, L. E., & Kollet, S. J. (2015). A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3. *Geoscientific Model Development*, *8*(3), 923–937. https://doi.org/10.5194/gmd-8-923-2015

McMillan, H. (2020). Linking hydrologic signatures to hydrologic processes: A review. *Hydrological Processes*, *34*(6), 1393–1409. https://doi.org/10.1002/hyp.13632

Medici, C., Wade, A. J., & Francés, F. (2012). Does increased hydrochemical model complexity decrease robustness? *Journal of Hydrology*, *440–441*, 1–13. https://doi.org/10.1016/j.jhydrol.2012.02.047

Melsen, L. A., Teuling, A. J., Torfs, P. J. J. F., Uijlenhoet, R., Mizukami, N., & Clark, M. P. (2016). HESS Opinions: The need for process-based evaluation of large-domain hyper-resolution models. *Hydrology and Earth System Sciences*, *20*(3), 1069–1079. https://doi.org/10.5194/hess-20-1069-2016

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., & Gupta, H. (2015). Are we unnecessarily constraining the agility of complex process-based models? *Water Resources Research*, *51*(1), 716–728. https://doi.org/10.1002/2014WR015820

Merz, R., Tarasova, L., & Basso, S. (2020). Parameter's controls of distributed catchment models—How much information is in conventional catchment descriptors? *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019wr026008

Misirli, F., Gupta, H. V., Thiemann, M., & Sorooshian, S. (2003). Bayesian recursive estimation of parameter and output uncertainty for watershed models. In Q. Duan, H. V. Gupta, S. Sorooshian, A. N. Rousseau, R. Turcotte (Eds.), *Advances in calibration of watershed models, AGU Monograph Series on Water Resources*. Water Science and Application. (Vol. 6, pp. 113–124). American Geophysical Union.

Molnar, C. (2019). *Interpretable machine learning – A guide for making black box models explainable*. Retrieved from https://christophm.github.io/interpretable-ml-book/

Molnar, C., Bischl, B., & Casalicchio, G. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software*, *3*(26), 786. https://doi.org/10.21105/joss.00786

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*(9). https://doi.org/10.1029/2011wr011412

Naef, F., Scherrer, S., & Weiler, M. (2002). A process based assessment of the potential to reduce flood runoff by land use change. *Journal of Hydrology*, *267*(1–2), 74–79. https://doi.org/10.1016/s0022-1694(02)00141-5

Nearing, G. S., & Gupta, H. V. (2015). The quantity and quality of information in hydrologic models. *Water Resources Research*, *51*(1), 524–538. https://doi.org/10.1002/2014WR015895

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*(3), e2020WR028091. https://doi.org/10.1029/2020wr028091

Newman, A. J., Clark, M. P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., et al. (2015). Gridded ensemble precipitation and temperature estimates for the contiguous United States. *Journal of Hydrometeorology*, *16*(6), 2481–2500. https://doi.org/10.1175/jhm-d-15-0026.1

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., et al. (2018). Constraining conceptual hydrological models with multiple information sources. *Water Resources Research*, *54*(10), 8332–8362. https://doi.org/10.1029/2017wr021895

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research*, *116*(D12), D12109. https://doi.org/10.1029/2010JD015139

O'Kane, J. P., & Flynn, D. (2007). Thresholds, switches and hysteresis in hydrology from the pedon to the catchment scale: A non-linear systems theory. *Hydrology and Earth System Sciences*, *11*(1), 443–459. https://doi.org/10.5194/hess-11-443-2007

Or, D., Lehmann, P., & Assouline, S. (2015). Natural length scales define the range of applicability of the Richards equation for capillary flows. *Water Resources Research*, *51*(9), 7130–7144. https://doi.org/10.1002/2015WR017034

Oudin, L., Kay, A., Andréassian, V., & Perrin, C. (2010). Are seemingly physically similar catchments truly hydrologically similar? *Water Resources Research*, *46*(11). https://doi.org/10.1029/2009wr008887
Parajka, J., Merz, R., & Blöschl, G. (2005). A comparison of regionalisation methods for catchment model parameters. *Hydrology and Earth System Sciences*, *9*(3), 157–171. https://doi.org/10.5194/hess-9-157-2005
Pitman, A. J. (2003). The evolution of, and revolution in, land surface schemes designed for climate models. *International Journal of Climatology*, *23*(5), 479–510. https://doi.org/10.1002/joc.893
Refsgaard, J. C., & Henriksen, H. J. (2004). Modelling guidelines—terminology and guiding principles. *Advances in Water Resources*, *27*(1), 71–82. https://doi.org/10.1016/j.advwatres.2003.08.006
Reggiani, P., Sivapalan, M., & Majid Hassanizadeh, S. (1998). A unifying framework for watershed thermodynamics: Balance equations for mass, momentum, energy and entropy, and the second law of thermodynamics. *Advances in Water Resources*, *22*(4), 367–398. https://doi.org/10.1016/s0309-1708(98)00012-8
Riihimäki, J., & Vehtari, A. (2010). March. Gaussian processes with monotonicity information. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 645–652).
Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2008wr007327
Santos, L., Thirel, G., & Perrin, C. (2018). Technical note: Pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*, *22*(8), 4583–4591. https://doi.org/10.5194/hess-22-4583-2018
Savenije, H. H. G. (2009). HESS Opinions "The art of hydrology". *Hydrology and Earth System Sciences*, *13*(2), 157–161. https://doi.org/10.5194/hess-13-157-2009
Schaefli, B., Harman, C. J., Sivapalan, M., & Schymanski, S. J. (2011). HESS Opinions: Hydrologic predictions in a changing environment: Behavioral modeling. *Hydrology and Earth System Sciences*, *15*(2), 635–646. https://doi.org/10.5194/hess-15-635-2011
Schulz, K., & Beven, K. (2003). Data-supported robust parameterisations in land surface-atmosphere flux predictions: Towards a top-down approach. *Hydrological Processes*, *17*(11), 2259–2277. https://doi.org/10.1002/hyp.1331
Seibert, J., Bishop, K., Rodhe, A., & McDonnell, J. J. (2003). Groundwater dynamics along a hillslope: A test of the steady state hypothesis. *Water Resources Research*, *39*(1), 1014. https://doi.org/10.1029/2002WR001404,1
Seibert, J., & McDonnell, J. J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, *38*(11), 23-1–23-14. https://doi.org/10.1029/2001WR000978
Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018wr022643
Sivapalan, M., Blöschl, G., Zhang, L., & Vertessy, R. (2003). Downward approach to hydrological prediction. *Hydrological Processes*, *17*, 2101–2111. https://doi.org/10.1002/hyp.1425
Son, K., & Sivapalan, M. (2007). Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data. *Water Resources Research*, *43*(1), W01415. https://doi.org/10.1029/2006WR005032
Spear, R., & Hornberger, G. M. (1980). Eutrophication in peel inlet—II. Identification of critical uncertainties via generalized sensitivity analysis. *Water Research*, *14*(1), 43–49. https://doi.org/10.1016/0043-1354(80)90040-8
Stein, L., Clark, M. P., Knoben, W. J. M., Pianosi, F., & Woods, R. A. (2021). How do climate and catchment attributes influence flood generating processes? A large-sample study for 671 catchments across the contiguous USA. *Water Resources Research*, *57*(4), e2020WR028300. https://doi.org/10.1029/2020wr028300
Tafasca, S., Ducharne, A., & Valentin, C. (2020). Weak sensitivity of the terrestrial water budget to global soil texture maps in the ORCHIDEE land surface model. *Hydrology and Earth System Sciences*, *24*(7), 3753–3774. https://doi.org/10.5194/hess-24-3753-2020
Tang, G., Clark, M. P., Newman, A. J., Wood, A. W., Papalexiou, S. M., Vionnet, V., & Whitfield, P. H. (2020). SCDNA: A serially complete precipitation and temperature dataset for North America from 1979 to 2018. *Earth System Science Data*, *12*(4), 2381–2409. https://doi.org/10.5194/essd-12-2381-2020
Thiemann, M., Trosset, M., Gupta, H., & Sorooshian, S. (2001). Bayesian recursive parameter estimation for hydrologic models. *Water Resources Research*, *37*(10), 2521–2535. https://doi.org/10.1029/2000wr900405
Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006). Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003800
Uhlenbrook, S., Roser, S., & Tilch, N. (2004). Hydrological process representation at the meso-scale: The potential of a distributed, conceptual catchment model. *Journal of Hydrology*, *291*(3–4), 278–296. https://doi.org/10.1016/j.jhydrol.2003.12.038
van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, *44*(5), 892–898. https://doi.org/10.2136/sssaj1980.03615995004400050002x
Vogel, H.-J., & Ippisch, O. (2008). Estimation of a critical spatial discretization limit for solving Richards' equation at large scales. *Vadose Zone Journal*, *7*(1), 112–114. https://doi.org/10.2136/vzj2006.0182
Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., & Sorooshian, S. (2003). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research*, *39*(8), 1214. https://doi.org/10.1029/2002WR001746
Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*(12), W00B09. https://doi.org/10.1029/2007WR006720
Wagener, T., Lees, M. J., & Wheater, H. S. (2001). A toolkit for the development and application of parsimonious hydrological models. *Mathematical Models of Small Watershed Hydrology*, *2*, 1–34.
Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research*, *47*(6). https://doi.org/10.1029/2010wr009469
Weijs, S. V., & Ruddell, B. L. (2020). Debates: Does information theory provide a new paradigm for earth science? Sharper predictions using Occam's digital razor. *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019wr026471
Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., et al. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, *15*(7), 2205–2227. https://doi.org/10.5194/hess-15-2205-2011
Westerberg, I. K., & McMillan, H. K. (2015). Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences*, *19*(9), 3951–3968. https://doi.org/10.5194/hess-19-3951-2015
Westhoff, M. C., Bogaard, T. A., & Savenije, H. H. G. (2011). Quantifying spatial and temporal discharge dynamics of an event in a first order stream, using distributed temperature sensing. *Hydrology and Earth System Sciences*, *15*(6), 1945–1957. https://doi.org/10.5194/hess-15-1945-2011

Winter, T. C. (2001). The concept of hydrologic landscapes. *Journal of the American Water Resources Association*, *37*(2), 335–349. https://doi.org/10.1111/j.1752-1688.2001.tb00973.x

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L. P. H., Bierkens, M. F., Blyth, E., et al. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *47*(5). https://doi.org/10.1029/2010wr010090

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9), W09417. https://doi.org/10.1029/2007WR006716

Young, P. (2003). Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale. *Hydrological Processes*, *17*(11), 2195–2217. https://doi.org/10.1002/hyp.1328

Zehe, E., Ehret, U., Pfister, L., Blume, T., Schröder, B., Westhoff, M., et al. (2014). HESS Opinions: From response units to functional units: A thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments. *Hydrology and Earth System Sciences*, *18*(11), 4635–4655. https://doi.org/10.5194/hess-18-4635-2014

Zehe, E., Loritz, R., Jackisch, C., Westhoff, M., Kleidon, A., Blume, T., et al. (2019). Energy states of soil water—A thermodynamic perspective on soil water dynamics and storage-controlled streamflow generation in different landscapes. *Hydrology and Earth System Sciences*, *23*(2), 971–987. https://doi.org/10.5194/hess-23-971-2019

Zhao, W. L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., et al. (2019). Physics-constrained machine learning of evapotranspiration. *Geophysical Research Letters*, *46*(24), 14496–14507. https://doi.org/10.1029/2019GL085291