AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR028338

Identification of Dominant Hydrological Mechanisms Using Bayesian Inference, Multiple Statistical Hypothesis Testing, and Flexible Models

Cristina Prieto[1,2,3], **Dmitri Kavetski**[4], **Nataliya Le Vine**[3,5], **César Álvarez**[1], **and Raúl Medina**[1]

[1]IHCantabria – Instituto de Hidráulica Ambiental de la Universidad de Cantabria, Santander, Spain, [2]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland, [3]Department of Civil and Environmental Engineering, Imperial College London, London, UK, [4]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, SA, Australia, [5]Swiss Re, Armonk, NY, USA

Key Points:

- Dominant mechanisms for process representation in hydrological models are identified using a Bayesian statistical hypothesis-testing method
- Synthetic and real data case studies are undertaken to investigate the empirical performance of mechanism identification
- Method is reliable: if it identifies a mechanism it is (usually) correct, though its statistical power declines for high streamflow errors

Correspondence to:
C. Prieto,
prietoc@unican.es

Citation:
Prieto, C., Kavetski, D., Le Vine, N., Álvarez, C., & Medina, R. (2021). Identification of dominant hydrological mechanisms using Bayesian inference, multiple statistical hypothesis testing, and flexible models. *Water Resources Research*, *57*, e2020WR028338. https://doi.org/10.1029/2020WR028338

Received 8 JUL 2020
Accepted 17 APR 2021

Abstract In hydrological modeling, the identification of model mechanisms best suited for representing individual hydrological (physical) processes is of major scientific and operational interest. We present a statistical hypothesis-testing perspective on this model identification challenge and contribute a mechanism identification framework that combines: (i) Bayesian estimation of posterior probabilities of individual mechanisms from a given ensemble of model structures; (ii) a test statistic that defines a "dominant" mechanism as a mechanism more probable than all its alternatives given observed data; and (iii) a flexible modeling framework to generate model structures using combinations of available mechanisms. The uncertainty in the test statistic is approximated using bootstrap sampling from the model ensemble. Synthetic experiments (with varying error magnitude and multiple replicates) and real data experiments are conducted using the hydrological modeling system FUSE (7 processes and 2–4 mechanisms per process yielding 624 feasible model structures) and data from the Leizarán catchment in northern Spain. The mechanism identification method is reliable: it identifies the correct mechanism as dominant in all synthetic trials where an identification is made. As data/model errors increase, statistical power (identifiability) decreases, manifesting as trials where no mechanism is identified as dominant. The real data case study results are broadly consistent with the synthetic analysis, with dominant mechanisms identified for 4 of 7 processes. Insights on which processes are most/least identifiable are also reported. The mechanism identification method is expected to contribute to broader community efforts on improving model identification and process representation in hydrology.

1. Introduction

Predictions of streamflow and available water resources are important scientifically and operationally. Such predictions are typically obtained using hydrological models of varying degrees of complexity. Scientifically, models help to understand and communicate catchments functioning and internal process dynamics (e.g., Wheather et al., 1993; Beven, 2010; Gupta et al., 2012). Operationally, hydrological models are used to manage water resources (including designing mitigation measures in anticipation of projected changes in the environment, e.g., due to changes in climate and/or land use), to implement flood early warning systems, and to help design and manage hydraulic structures, among others (e.g., Wagener et al., 2010; Montanari et al., 2013; Srinivasan et al., 2017; Prieto et al., 2020).

In order for a hydrological model to adequately represent catchment function, it must adequately represent the underlying hydrological processes occurring in the catchment. Here, we define a "hydrological process" as the physical phenomenon occurring in a catchment, for example, surface runoff generation. Within a model, hydrological processes are approximated using "hydrological mechanisms," i.e., sets of equations intended to describe that process. A single hydrological model represents a combination of mechanisms (one per hydrological process). In this sense, models represent working hypotheses of the catchment they are applied to, and mechanisms represent working hypotheses of the hydrological processes they are intended to represent.

This paper focuses on hydrological model development from the perspective of using observed data to identify individual process mechanisms rather than complete models, using inference and hypothesis-testing techniques capable of reflecting the inherent data and model uncertainty. The presentation is illustrated using models that represent the temporal dynamics of rainfall-runoff processes.

The selection of processes and mechanisms to be included in models has been a perennial challenge in hydrological and broader environmental sciences. Important questions in hydrological process understanding and representation relate to hydrological laws emergent at the catchment scale, causes of spatial heterogeneity in streamflow, evaporation, groundwater and other environmental fluxes, processes that control surface water–groundwater interactions and catchment connectivity, mechanisms by which climate and land use change impact on the hydrology of arid and semiarid regions, and many others (e.g., see Beven, 1989; McDonnell et al., 2007; Wagener et al., 2007; Clark et al., 2011b; Ehret et al., 2014; Gupta et al., 2014; Blöschl et al., 2019, and many others).

Reliable identification of hydrological mechanisms is important for multiple purposes, including improving predictive performance and gaining insight into catchments' internal functioning, with the latter important in order for the model to "get the right answers for the right reasons" (Kirchner, 2006). The selection of hydrological mechanisms is also generally expected to respect the principle of parsimony, which favors simpler over complex model representations (Jakeman & Hornberger, 1993).

There is a growing interest in approaching hydrological modeling from a hypothesis-testing perspective, using the method of "multiple working hypotheses" (Chamberelin, 1965; Krueger et al., 2010; Clark et al., 2011a; Fenicia et al., 2011; Beven et al., 2012; Coxon et al., 2014; Fenicia et al., 2016; Pfister & Kirchner, 2017, and others). Within this perspective, mechanism identification itself represents hypothesis testing, targeted at specific model components rather than complete models (e.g., Clark et al., 2011a; Fenicia et al., 2011; Clark et al., 2015; Hrachowitz & Clark, 2017; Pfister & Kirchner, 2017, and others). The hypothesis-testing perspective applies to a wide range of "physically motivated" models, including lumped and spatially distributed models, as well as "conceptual" and "physically based" models.

Hypothesis testing in hydrology and broader environmental sciences has been hampered by several challenges, which complicate hypothesis testing both for complete models and individual mechanisms (see, e.g., Clark et al., 2011a; Gupta et al., 2012; Clark et al., 2015 the series of debates papers in Blöschl, 2017; Beven, 2018, 2019; Höge et al., 2019). Notable challenges include lack of access to the true process (e.g., flow through an aquifer), the "uniqueness of place" paradigm (Beven, 2000), broader heterogeneity and variability of environments (Carrera et al., 2005; McDonnell et al., 2007), and resulting uncertainties/scarcity of observation data (Gupta et al., 2008; Wagener & Montanari, 2011). In addition, as noted by Nearing et al. (2016, 2020), model hypothesis testing is necessarily constrained by the Duhem–Quine thesis (Duhem, 1991), which refers to the difficulty or even impossibility of separating individual model hypotheses from their "surrounding" model environment.

The adage "All models are wrong but some are useful" (Box, 1979) is particularly salient in environmental modeling. From this perspective, it may be more constructive to seek "better" and/or "quasi-true" rather than "true" models, where "quasi-true" models are sets of equations that reproduce the entire system or an individual system component with suitable "accuracy" at the scale of interest (Beven, 2018; Gupta et al., 2012; Höge et al., 2019).

Then, how to identify mechanisms representative of specific hydrological processes given that we only have incomplete and inexact information? Hypothesis testing and model selection/development in hydrology have progressed along several directions.

The "fixed modeling approach" seeks a complete model structure that reproduces the observed data (generally streamflow series or hydrological indices) across a wide range of catchments (e.g., PDM, Moore & Clarke, 1981; Moore, 2007; HVB, Lindström et al., 1997, Seibert & Vis, 2012; GR4J, Perrin et al., 2003; Van Esse et al., 2013). Well-performing fixed models can be developed through systematic model comparison (e.g., see the GR4J development history). However, the intermediate analyses conducted as part of these developments are not necessarily structured to provide systematic hypothesis testing of individual components.

The "flexible modeling" approach seeks to formalize the process of constructing (possibly multiple) complete models from basic building blocks, for example, components of existing models and generic elements (e.g., RRMT, Wagener et al., 2001; Framework for Understanding Structural Errors (FUSE), Clark et al., 2008; 2011a; SUPERFLEX, Fenicia et al., 2011; CFM, Kraft et al., 2011; SUMMA, Clark et al., 2015; MARRMoT, Knoben et al., 2019; RAVEN, Craig et al., 2020, and others). Flexible models enable a more targeted approach for comparing different hypotheses about catchments' inner working, because the model hypotheses can be altered "one component at a time" (e.g., Clark et al., 2011a; Wrede et al., 2015; Fenicia et al., 2016).

Bayesian approaches offer a number of important techniques for model selection and hypothesis testing. The well-known Bayes Factor (Raftery, 1993; Kass & Raftery, 1995) quantifies the strength of evidence in favor of one model over another model (Kass & Raftery, 1995; Jeffreys, 1998), which in turn allows assigning "degrees of belief" to the proposed models. Approximations such as the Bayesian Information Criterion (BIC) and Kashyap Information Criterion (KIC) can be used to avoid computationally costly Monte Carlo integration of the likelihood function when computing Bayesian Model Evidence (BME) and Bayes Factors (Ye et al., 2008; Schöniger et al., 2014). Bayesian model selection has seen multiple applications across hydrology, including the selection of rainfall-runoff models based on observed streamflow data (e.g., Marshall et al., 2005), improvement of model predictions (e.g., Vrugt & Robinson, 2007), evaluation of the worth of different types of observations for the selection of crop models (Wöhling et al., 2015), evaluation of uncertainty in posterior model weights due to measurement error in soil plants models (Schöniger et al., 2015); quantification of information provided by a regionalization or hydrological model (Prieto et al., 2019), and many others (see Höge et al., 2019 for a recent review). To our knowledge, previous studies using Bayesian methods have focused primarily on model selection at the level of complete models, and not yet at the level of individual mechanisms.

Statistical hypothesis testing is a method of inference that allows comparing two or more models (hypotheses) describing systems with random (uncertain) behavior (Lumley, 2000; Burnham & Anderson, 2002). Typically, the "lack" of a discovery is taken as the "null" hypothesis (benchmark), for example, "the subsurface flow process is *not* represented by a Fickian diffusion mechanism," and rejection of the null hypothesis indicates a "discovery" (here, that "subsurface flow is Fickian"). Due to the presumed inherent uncertainty in the system being modeled, statistical hypothesis testing is formulated to include a confidence level (e.g., 95%), such that the corresponding significance level (e.g., 5%) represents the chance of reaching erroneous conclusions, notably Type I errors (false positives, i.e., incorrectly rejecting a true null hypothesis). It is also known that stringent control of Type I errors will generally translate into increased frequency of Type II errors (false negatives, i.e., failing to accept a true null hypothesis) (Smith & Bryant, 1975). The initial selection of hypotheses can also bias test results, for example, if some model components/mechanisms are overrepresented in the ensemble under consideration (see, e.g., Elkan, 2001; Saerens et al., 2002).

Many other new approaches for model selection and improvement are emerging in the field of data analytics, including the information-theoretic approach envisioned by Nearing and Gupta (2015, 2018) and Nearing et al. (2020). The approach uses information theory and machine learning to test and refine model hypotheses (e.g., if a data-driven model can extract more information from the data than a conceptual model, then the conceptual model can be improved).

This study develops new Bayesian methods for mechanism identification in hydrology taking advantage of flexible modeling frameworks. A key emphasis is on the identification of process mechanisms rather than complete models, and on the use of statistical techniques to reflect the inherent uncertainty in testing hypotheses related to hydrological mechanisms.

The study aims are as follows:

1. Develop a systematic hypothesis-testing approach for the identification of dominant hydrological mechanisms, using a combination of flexible models, Bayesian inference, and methods for multiple hypothesis testing.
2. Investigate the performance of the proposed mechanism identification method in the presence of error (low, medium, and high magnitude) using synthetic data and using detailed performance metrics computed under replication.

3. Investigate the performance of the proposed method in a real catchment.

The case studies are implemented using the FUSE hydrological modeling system (Clark et al., 2008; 2011a) and daily hydrological data from the Leizarán catchment in northern Spain.

The paper is organized as follows. Section 2 presents theoretical developments, including the derivation of the posterior probability of mechanisms and their use within a hypothesis-testing framework. Section 3 describes the case study setup. Section 4 presents the case study results, which are then discussed in Section 5. Section 6 summarizes the key conclusions.

2. Theoretical Development

2.1. Key Concepts and Terminology

This section defines the key concepts and terminology underlying the proposed method for identifying dominant hydrological mechanisms. Here, it is important to distinguish concepts related to the physical catchment itself from concepts related to its mathematical model.

The term *hydrological process* refers to a physical phenomenon occurring in a catchment, for example, surface runoff generation. The term *hydrological model process*, $\wp$, then refers to a single hydrological process intended to be represented by a hydrological model (irrespective of the accuracy of this representation).

A key concept in this study is *a hydrological model mechanism*, $m^{\wp}$, which we define as a set of model equations intended to represent a hydrological model process $\wp$. Note that a mechanism may or may not contain calibrated model parameters; for this reason, we use it in preference to the term "process parameterization" often used in calibration contexts to denote process equations and associated parameter values.

Given these definitions, a *hydrological model structure* is a specific combination of hydrological model mechanisms intended to represent a number of (preselected) hydrological model processes. For example, the SACRAMENTO and PRMS models are examples of hydrological model structures.

An *ensemble of hydrological model structures* comprises multiple hydrological models that differ in their selection of hydrological model processes and/or in the selection of hydrological model mechanisms used to represent these processes.

A *multihypothesis framework* (MHF) is defined as a framework for generating ensembles of model structures, i.e., a framework for generating model structures from a pool of available hydrological process mechanisms. In this work, the MHF is given by the flexible modeling framework FUSE (Clark et al., 2008; 2011a).

A hydrological model mechanism is defined as *dominant* if, conditionally on the study method and assumptions, it is "substantially" (according to a quantitative criterion) more likely to represent a particular hydrological process than all alternative mechanisms under consideration in that analysis. A *dominant mechanism* should be not confused with a *dominant process*, which is a process that contributes substantially to the overall catchment water balance.

In terms of mathematical behavior, model structures may be *deterministic* or *probabilistic*. In this study, we propose methods for general probabilistic models and illustrate them for the common case in current hydrological practice where probabilistic models comprise a deterministic model (intended for process representation) augmented by a relatively simple error model (intended for predictive uncertainty representation).

2.2. Hydrological Model

Consider a probabilistic model of streamflow at time t, with model structure G representing a function of parameters $\boldsymbol{\theta}$, forcing data $\mathbf{x}_{1:t}$ (up to time t), and initial conditions $\mathbf{s}_0$,

$$Q_t = G_t\left(\boldsymbol{\theta}; \mathbf{x}_{1:t}, \mathbf{s}_0\right) \tag{1}$$

In Equation 1, Q_t is a random variable with probability distribution $p(q_t \mid \boldsymbol{\theta}, G; \mathbf{x}_{1:t}, \mathbf{s}_0)$.

To illustrate a typical construction of a probabilistic model in hydrology, consider a deterministic hydrological model h that computes a (deterministic) estimate of streamflow,

$$q_t^{\boldsymbol{\theta}_h} = h_t(\boldsymbol{\theta}_h; \mathbf{x}_{1:t}, \mathbf{s}_0) \tag{2}$$

If an additive Gaussian error model is used in transformed space, the probabilistic model is given by

$$z(Q_t) = z(q_t^{\boldsymbol{\theta}_h}) + \eta_t \tag{3}$$

$$\eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \tag{4}$$

with additional assumptions that the normalized residuals η_t are independent and identically distributed (i.i.d.) Gaussian with zero mean and constant variance σ_η^2 (e.g., McInerney et al., 2018). The transformation z is typically used to account for the heteroscedasticity and skew of the residual errors (e.g., see Section 3.3).

For the probabilistic model in Equations 2–4, the parameters can be partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_h, \boldsymbol{\theta}_\varepsilon)$, where $\boldsymbol{\theta}_h$ are the "hydrological model" parameters and $\boldsymbol{\theta}_\varepsilon$ are the "error model" parameters. Here, $\boldsymbol{\theta}_\varepsilon = \sigma_\eta^2$, though in general it can include other error parameters such as autocorrelation of residuals, skew, and transformation parameters. For completeness, we denote the parameter domain as Ω.

2.3. Parameter Inference for a Hydrological Model Structure

Here, we consider the inference of model parameters in a given model structure from observed data. Let $\tilde{\mathbf{q}} = (\tilde{q}_t; t = 1,...,N_t)$ denote a time series of observed flows of length N_t, and let $\tilde{\mathbf{x}} = (\tilde{x}_t; t = 1,...,N_t)$ denote the corresponding observed model inputs.

The posterior parameter distribution $p(\boldsymbol{\theta} \mid \tilde{\mathbf{q}}, G)$ is given by Bayes equation,

$$p(\boldsymbol{\theta} \mid \tilde{\mathbf{q}}, G) = \frac{p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}, G) p(\boldsymbol{\theta} \mid G)}{p(\tilde{\mathbf{q}} \mid G)} = \frac{p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}, G) p(\boldsymbol{\theta} \mid G)}{\int_\Omega p(\tilde{\mathbf{q}} \mid \boldsymbol{\varphi}, G) p(\boldsymbol{\varphi} \mid G) \mathrm{d}\boldsymbol{\varphi}} \tag{5}$$

where $p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}, G)$ is the likelihood function associated with the probability model, $p(\boldsymbol{\theta} \mid G)$ is the prior distribution of its parameters over the feasible domain Ω, and $p(\tilde{\mathbf{q}} \mid G)$ is referred to as Bayesian Model Evidence (BME) or Marginal Likelihood. To reduce clutter, the conditioning on observed forcing data $\tilde{\mathbf{x}}$ and initial conditions $\mathbf{s}_0$ is not indicated explicitly in the terms $p(\boldsymbol{\theta} \mid \tilde{\mathbf{q}}, G)$, $p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}, G)$, etc., because in this study these quantities are treated as fixed. The BME term is generally not required for model parameter inference but will be required for model structure inference.

For the illustrative probabilistic model in Equations 3 and 4, the likelihood function is

$$p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}, G) = p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}_h, \sigma_\eta, G) = \prod_{t=1}^{N_t} z'(\tilde{q}_t) \times f_{\mathcal{N}}(z(\tilde{q}_t); z(q_t^{\boldsymbol{\theta}_h}), \sigma_\eta^2) \tag{6}$$

where $z'(\tilde{q}) = \partial z / \partial q \mid_{q=\tilde{q}}$ is the Jacobian of the transformation $z(q)$ evaluated at $q = \tilde{q}$. The notation $f_{\mathcal{N}}(x; \mu, \sigma^2)$ denotes the Gaussian pdf with mean μ and variance σ^2. The complete specification used in the case study is given in Section 3.3.

2.4. Posterior Probability of a Model Structure

We now turn our attention to the estimation of the posterior probability of a model structure within an ensemble of model structures given by an MHF.

Consider the posterior probability $p(G^{(k)} \mid \tilde{\mathbf{q}}, \mathbf{G})$ of a model structure $G^{(k)}$, conditional on observed flows $\tilde{\mathbf{q}}$ and a model structure ensemble $\mathbf{G} = \{G^{(i)}; i = 1,...,N_G\}$ comprising N_G model structures. Let $p(G^{(i)} \mid \mathbf{G})$ denote the prior of model structure $G^{(i)}$ within this ensemble.

Bayes equation yields (see Raftery, 1995; Hoeting et al., 1999; Hsu et al., 2009)

$$p(G^{(k)} \mid \tilde{\mathbf{q}}, \mathbf{G}) = \frac{p\left(\tilde{\mathbf{q}} \mid G^{(k)}\right) p\left(G^{(k)} \mid \mathbf{G}\right)}{\sum_{i=1}^{N_G} p\left(\tilde{\mathbf{q}} \mid G^{(i)}\right) p\left(G^{(i)} \mid \mathbf{G}\right)} = \frac{\int_{\Omega^{(k)}} p\left(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}^{(k)}, G^{(k)}\right) p\left(\boldsymbol{\theta}^{(k)} \mid G^{(k)}\right) \mathrm{d}\boldsymbol{\theta}^{(k)} p\left(G^{(k)} \mid \mathbf{G}\right)}{\sum_{i=1}^{N_G} \int_{\Omega^{(i)}} p\left(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}^{(i)}, G^{(i)}\right) p\left(\boldsymbol{\theta}^{(i)} \mid G^{(i)}\right) \mathrm{d}\boldsymbol{\theta}^{(i)} p\left(G^{(i)} \mid \mathbf{G}\right)} \tag{7}$$

where the notation from Section 2.3 is augmented to explicitly link a model structure $G^{(i)}$ to its model parameters $\boldsymbol{\theta}^{(i)}$. The model structure represents a discrete random variable; hence, the denominator is a sum of discrete probabilities.

The probability $p(G^{(k)} \mid \tilde{\mathbf{q}}, \mathbf{G})$ should be interpreted as $p(G^{\text{true}} = G^{(k)} \mid \tilde{\mathbf{q}}, G^{\text{true}} \in \mathbf{G})$, which highlights the fundamental assumption that the "true" (or at least "quasi-true") model is in the ensemble. In Equation 7, this assumption is reflected in the scaling of the Bayesian Evidence of model $G^{(k)}$ by the sum of BMEs of all models in the ensemble.

The computation of BME and related terms has long been a challenge in Bayesian model selection (e.g., see Ye et al., 2008; Schöniger et al., 2014, for analyses in hydrological contexts). Several approaches are available, including "semianalytical" approximations (often referred to as "information criteria") and numerical approximations (e.g., using Monte Carlo samples from the parameter posterior).

The Bayesian Information Criterion (BIC) offers an attractive avenue to approximate the BME terms in Equation 7. The BIC is computed directly from the likelihood function evaluated at the maximum a posteriori parameter set (which given uniform parameter priors coincides with the maximum likelihood parameter set) and includes a so-called "Occam Razor" term to penalize model complexity (quantified by the number of parameters). The use of BIC to approximate Equation 7 is described in Appendix A, which also includes a brief discussion of potential alternatives.

It is emphasized that the treatment of the BME term in Equation 7 is independent from subsequent derivations, which work solely with $p(G^{(k)} \mid \tilde{\mathbf{q}}, \mathbf{G})$. As such, the modeler is free to compute $p(G^{(k)} \mid \tilde{\mathbf{q}}, \mathbf{G})$ using methods/approximations suitable for their specific application.

2.5. Posterior Probability of a Hydrological Mechanism

We now consider the estimation of probabilities of hydrological process mechanisms, rather than the probabilities of complete models (which represent combinations of mechanisms).

Let the model ensemble $\mathbf{G}$ comprises models that attempt to represent a total of $N^{\wp}$ hydrological model processes, using hydrological model mechanisms $\{m_i^{\wp}; i = 1,...,N_m^{\wp}; \wp = 1,...,N^{\wp}\}$. The notation $N_m^{\wp}$ indicates the number of mechanisms available for process $\wp$. As per earlier definition in Section 2.1, we assume that this model structure ensemble is generated using an MHF.

Let $p_e(m_k^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G})$ denote the "ensemble-specific" posterior probability of mechanism $m_k^{\wp}$ given observed streamflow $\tilde{\mathbf{q}}$ and the ensemble of probabilistic hydrological models $\mathbf{G}$. This posterior probability can be expressed using total probability as follows:

$$
\begin{aligned}
p_e(m_k^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G}) &= \sum_{i=1}^{N_G} p(m_k^{\wp} \mid G^{(i)}, \tilde{\mathbf{q}}, \mathbf{G}) p(G^{(i)} \mid \tilde{\mathbf{q}}, \mathbf{G}) \\
&= \sum_{i=1}^{N_G} \mathcal{I}\left(m_k^{\wp}, G^{(i)}\right) p\left(G^{(i)} \mid \tilde{\mathbf{q}}, \mathbf{G}\right) \\
&= \sum_{i \in S\left(k; \mathbf{G}, \wp\right)} p\left(G^{(i)} \mid \tilde{\mathbf{q}}, \mathbf{G}\right)
\end{aligned} \tag{8}
$$

where $S(k; \mathbf{G}, \wp)$ contains the indices of the subset of model structures within ensemble $\mathbf{G}$ that represent process $\wp$ using mechanism $m_k^{\wp}$; let $N_G^{\wp,k}$ denote the number of models within this subset. The intermediate step in Equation 8 makes use of the indicator function $\mathcal{I}(m_k^{\wp}, G^{(i)})$, which takes the value 1 if model $G^{(i)}$ contains mechanism $m_k^{\wp}$ and takes the value 0 otherwise.

The use of total probability to obtain Equation 8 requires two assumptions:

(i) The MHF provides a "sufficiently" complete coverage of the space of possible model structures. Analogous assumptions are common in model selection methods (e.g., see Hirabayashi et al., 2013; Arnell & Gosling, 2014, in the context of climate modeling).

(ii) The mechanisms are mutually exclusive, i.e., a single model structure cannot use two mechanisms for the same process. This assumption yields the identity $p(m_k^{\wp} \mid G^{(i)}, \tilde{\mathbf{q}}, \mathbf{G}) = p(m_k^{\wp} \mid G^{(i)}) = \mathrm{I}(m_k^{\wp}, G^{(i)})$, i.e., conditionally on $G^{(i)}$ being the "true" model structure, mechanism $m_k^{\wp}$ has posterior probability 1 if it is contained in $G^{(i)}$ and has posterior probability 0 otherwise. Note that the conditioning on the observed streamflow $\tilde{\mathbf{q}}$ and ensemble $\mathbf{G}$ does not contribute any additional information to this probability.

In practice, the model structures within the MHF ensemble $\mathbf{G}$ will be selected in an "unbalanced and opportunistic" way (e.g., Saerens et al., 2002). In other words, the MHF will generally not provide uniform and unbiased—let alone complete—coverage of the "universe" of possible hydrological models and mechanisms. For example, different processes may have a different number of available mechanisms, and some mechanisms for separate processes may be mutually incompatible. As a result, some mechanisms will appear more frequently than others across the model structures provided by the MHF ensemble. In addition, the selection of mechanisms themselves may be biased toward the modeler's expertise, general community preferences, etc.

The inherently subjective (implicit) nature of mechanism selection within any realistic MHF impacts on the estimation of both prior and posterior probabilities of process mechanisms.

For example, the imbalance in the frequency of mechanisms included in an MHF makes the underlying prior distribution of mechanisms nonuniform, with an "effective" ("ensemble-specific") prior distribution of a mechanism $m_i^{\wp}$ being $p_e(m_k^{\wp} \mid \mathbf{G}) = N_G^{\wp,k} \,/\, N_G$. This prior differs from the genuine uniform prior $p_{\text{unif}}(m_k^{\wp} \mid \mathbf{G}) = 1 \,/\, N_m^{\wp}$.

The unbalanced mechanism frequency also affects the posterior probabilities in Equation 8, for example, as shown by Elkan (2001) and Saernes et al. (2002). In particular, due to the probabilities being summed, posterior estimation will be biased toward mechanisms that appear more frequently in the model ensemble.

In order to account for these imbalances, we employ the correction proposed by Saernes et al. (2002), which yields the posterior probabilities $p(m_k^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G})$ corresponding to the prior probabilities $p(m_k^{\wp} \mid \mathbf{G})$ actually desired by the modeler (e.g., $p_{\text{unif}}(m_k^{\wp} \mid \mathbf{G})$ or any other distribution).

The "corrected" posterior probabilities $p(m_k^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G})$ are obtained by weighting the "ensemble-specific" posterior probabilities $p_e(m_k^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G})$ by the ratio of "desired" prior probabilities $p(m_k^{\wp} \mid \mathbf{G})$ over the "ensemble-specific" priors $p_e(m_k^{\wp} \mid \mathbf{G})$,

$$p(m_k^{\wp} \mid \tilde{\mathbf{q}},\mathbf{G}) = \frac{\frac{p(m_k^{\wp} \mid \mathbf{G})}{p_e(m_k^{\wp} \mid \mathbf{G})} p_e(m_k^{\wp} \mid \tilde{\mathbf{q}},\mathbf{G})}{\sum_{i=1}^{N_m^{\wp}} \frac{p(m_i^{\wp} \mid \mathbf{G})}{p_e(m_i^{\wp} \mid \mathbf{G})} p_e(m_i^{\wp} \mid \tilde{\mathbf{q}},\mathbf{G})} \propto \frac{p(m_k^{\wp} \mid \mathbf{G})}{p_e(m_k^{\wp} \mid \mathbf{G})} p_e(m_k^{\wp} \mid \tilde{\mathbf{q}},\mathbf{G}) \tag{9}$$

If we set uniform priors in the mechanism space, $p(m_k^{\wp} \mid \mathbf{G}) = p_{\text{unif}}(m_k^{\wp} \mid \mathbf{G}) = 1 / N_m^{\wp}$ and substitute $p_e(m_k^{\wp} \mid \tilde{\mathbf{q}},\mathbf{G})$ from Equation 9 into Equation 8, we obtain

$$p(m_k^{\wp} \mid \tilde{\mathbf{q}},\mathbf{G}) \propto \frac{1}{N_G^{\wp,k}} p_e(m_k^{\wp} \mid \tilde{\mathbf{q}},\mathbf{G}) \propto \frac{1}{N_G^{\wp,k}} \sum_{i \in S\left(k;\mathbf{G},\wp\right)} p(G^{(i)} \mid \tilde{\mathbf{q}},\mathbf{G}) \tag{10}$$

Equation 10 is used to compute the "unnormalized" posterior probabilities of all mechanisms for process $\wp$; these quantities are then normalized to sum up to 1, yielding $\{p(m_k^{\wp} \mid \tilde{\mathbf{q}},\mathbf{G}); k = 1,...,N_m^{\wp}\}$. This normalization follows from the earlier assumptions that a hydrological process is represented by a single model mechanism and that the true model (and hence the true mechanism) is present in the ensemble.

Equation 10 is intuitive in that the sum of posterior probabilities of model structures with a given mechanism is now scaled by the corresponding number of model structures, i.e., the *sum* of posteriors is replaced by the *average* of posteriors. The implied assumption is that, *on average*, model structures that include highly probable mechanisms have higher posterior probability than model structures that include less-probable mechanisms. This assumption is generally reasonable but can be compromised by interactions between model mechanisms for multiple processes.

Given the considerations presented in this section, Equation 10 will be used in lieu of Equation 8 to compute posterior probabilities of mechanisms in this work. In light of its assumptions, the use of Equation 10 with practical MHF ensembles represents an *approximation* to the probability of a given mechanism rather than a "general" statement of probability theory. For this reason, it is important to embed Equation 10 in a robust hypothesis-testing framework (Sections 2.6–2.9) and to undertake a rigorous verification of its empirical performance (e.g., Section 3.4). We provide additional discussion of these aspects in Sections 5.1.1, 5.2.1, and 5.3.2.

2.6. Multiple Hypothesis-Testing Setup

The posterior probabilities of mechanisms can be used for hypothesis testing. For example, for a given process, we could just search for the mechanism with the highest posterior probability. However, we develop a more comprehensive and reliable hypothesis-testing process, which recognizes the model ensemble uncertainty associated with Equation 10, includes the specification of a prescribed significance level α and reflects the potentially large number of hypothesis tests being carried out (which raises the probability of Type I errors, i.e., false identification of a mechanism as dominant).

The building blocks of the proposed hypothesis-testing method are described next. The method is designed to identify the dominant mechanism for a given process $\wp$. Multiple processes are accommodated by applying the hypothesis-testing method separately to each individual process.

An *individual comparison (or hypothesis test)* is defined as an individual test of whether a mechanism $m_k^{\wp}$ is "dominant," i.e., substantially more likely than all other alternative mechanisms available in $\mathbf{G}$ to represent process $\wp$. The quantitative definition of "substantially" will be introduced shortly.

The *null hypothesis for an individual test*, $\text{H0}_k^{\wp}$, is "mechanism $m_k^{\wp}$ is *not* dominant for process $\wp$."

The *family of comparisons* is then the set of individual comparisons to identify the dominant hydrological model mechanism for model process $\wp$. In this work, given a set of mechanisms, a "family of comparisons" comprises the set of individual tests of each mechanism against all other mechanisms available to describe that process. Note that each mechanism is only tested for dominance once.

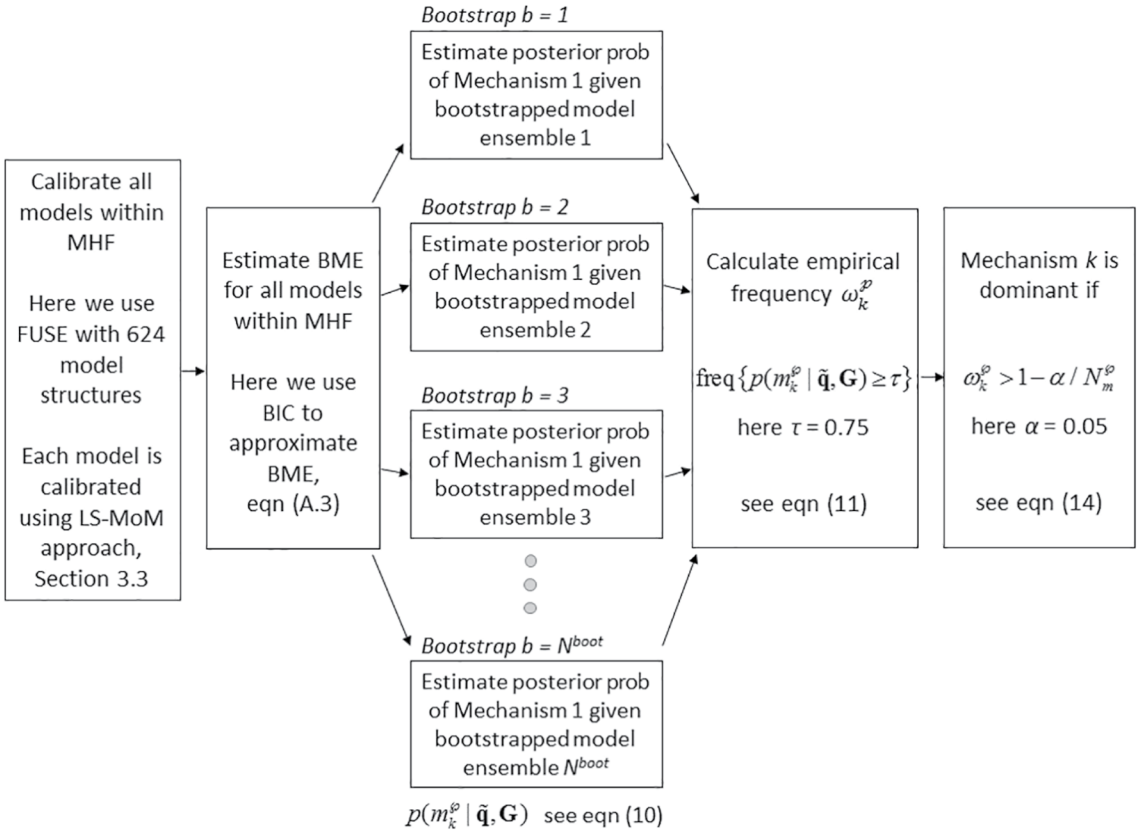

Figure 1. Flowchart of the mechanism identification method. General steps (e.g., "estimate BME of all models") and their specific implementations in this work (e.g., "use BIC approximation") are indicated. The function freq $v = \frac{1}{N} count\ v$ returns the fraction of true elements in Boolean set v of length N. BME, Bayesian Model Evidence; BIC, Bayesian Information Criterion.

The *null hypothesis for a family of comparisons*, $H0^{\wp}$, is defined as "None of the hydrological mechanisms $\{m_k^{\wp}; k = 1,...,N_m^{\wp}\}$ in $\mathbf{G}$ is dominant, i.e., substantially more likely than the set of alternative mechanisms $\{m_{i\neq k}^{\wp}; i = 1,...,N_m^{\wp}\}$ to describe a particular hydrological process $\wp$." If an individual null hypothesis $H0_k^{\wp}$ in the family of comparisons $\{H0_k^{\wp}; k = 1,...,N_m^{\wp}\}$ is rejected, then mechanism $m_k^{\wp}$ is identified as "dominant" for process $\wp$. In other words, the dominant hydrological model mechanism is the mechanism for which the individual null hypothesis is rejected. Note that it is also possible for no mechanism to be identified as dominant.

The *family wise error rate (FWER)* is the probability of making one or more Type I errors in a family of multiple tests, i.e., the probability of *incorrectly* identifying a mechanism as dominant for the given process (after testing all proposed mechanisms). Note that to keep FWER below a prescribed significance level α, stricter significance levels α^* must be imposed in individual tests. In this study, the (conservative) Bonferroni correction is employed for this purpose (Hochberg, 1988).

Hypothesis testing is implemented by defining a test statistic t for the individual null hypothesis (Section 2.7). The empirical probability of the computed test statistic exceeding a prescribed threshold τ is estimated using a bootstrap approach (Section 2.9) and compared to the prescribed significance level α^* (Section 2.8). This approach follows the principles of classical hypothesis testing (Lehmann & Romano, 2005; Triola, 2001).

A flowchart of the hypothesis-testing approach is provided in Figure 1, and detailed descriptions are provided in the following sections.

2.7. Test Statistic Providing the Definition of a "Dominant" Process

The test statistic $t_k^{\wp}$ for (rejecting) the individual null hypothesis $\mathrm{H0}_k^{\wp}$ is taken as the posterior probability of mechanism $m_k^{\wp}$. A mechanism is considered dominant if the test statistic exceeds a threshold value τ,

$$t_k^{\wp} = p(m_k^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G}) > \tau \tag{11}$$

The selection of τ represents a modeling choice that, jointly with the significance level α introduced in Section 2.8, controls the stringency of the hypothesis test. For example, $\tau = 0.5$ represents the weakest requirement, where a mechanism is considered dominant even if it is barely more probable than its alternatives. In this work, we select $\tau = 0.75$, which requires a mechanism to be at least 3 times more probable than all its alternatives. This choice is subjective and may be application specific.

Note that the statistic $t_k^{\wp}$ could be formulated equivalently as $t_k^{\wp *} = p(m_k^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G}) - \sum_{i \neq k} p(m_i^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G})$ or as $t_k^{\wp **} = p(m_k^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G}) \,/\, \sum_{i \neq k} p(m_i^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G})$, with corresponding thresholds $\tau^* = 2\tau - 1$ and $\tau^{**} = \tau \,/\, \left(1 - \tau\right)$, respectively. However, Equation 11 is simpler and hence preferred in this work.

2.8. Hypothesis Testing Using the Test Statistic

The test statistic defined in Equation 11 depends on the model ensemble $\mathbf{G}$. As noted earlier, this model ensemble on its own cannot be expected to provide complete coverage (sampling) of the total space of models and mechanisms. For this reason, we treat $t_k^{\wp}$ as a realization of a random variable $T_k^{\wp}$, with cumulative distribution function $F_{T(\wp,k)}(t)$. The variability (randomness) in $T_k^{\wp}$ is assumed to arise due to limited sampling of model structures within $\mathbf{G}$.

$\mathrm{H0}_k^{\wp}$ is rejected if the test statistic $t_k^{\wp}$ has a probability of exceedance $\omega_k^{\wp} = 1 - F_{T(\wp,k)}(t_k^{\wp})$ larger than the prespecified significance for that individual test,

$$\omega_k^{\wp} \geq 1 - \alpha_{\mathrm{Bonf}}^* \tag{12}$$

where $\alpha_{\mathrm{Bonf}}^* = \alpha \,/\, N_m^{\wp}$ is the Bonferroni correction to the overall significance level α (Hochberg, 1988).

Note that the overall confidence level of the family of comparisons (overall mechanisms) is $1 - \alpha$. In this paper, we set $\alpha = 0.05$ and test the sensitivity of conclusions to this choice.

The next section describes the estimation of the distribution of the test statistic, i.e., the empirical probability of exceedance $\omega_k^{\wp}$.

2.9. Bootstrap Estimation of the (Empirical) Distribution of the Test Statistic

In order to estimate the uncertainty in the test statistic computed from models generated using a MHF, we distinguish the MHF model space from the hypothetical total model space. The MHF model space is given by the set of all distinct model structures available within the framework, i.e., all combinations of mechanisms *available within that framework*. As defined earlier, this sample space is $\mathbf{G}$. In contrast, the total model space $\mathbf{G}^{\mathrm{total}}$ is conceptualized as the set of *all possible* distinct model structures, i.e., all combinations of all mechanisms that might exist "in principle."

Our implementation estimates the uncertainty in the test statistic (more precisely, its probability of exceedance) numerically by applying bootstrapping (Efron & Tibshirani, 1993) to the model ensemble $\mathbf{G}$. The bootstrap approximation is based on the assumption that the uncertainty introduced when the total model space $\mathbf{G}^{\mathrm{total}}$ is reduced to $\mathbf{G}$ is "similar" to the uncertainty introduced when $\mathbf{G}$ is replaced by a resampled subset. This assumption is typical of bootstrap approximations (e.g., Press et al., 1992; see also; Efron & Tibshirani, 1993; Varian, 2005), except here it is applied in the space of model structures.

More specifically, the null hypothesis $H0_k^{\wp}$ that mechanism $m_k^{\wp}$ is *not* dominant for process $\wp$, with significance level α (e.g., 0.05), is tested using the following procedure:

1. Generate a "bootstrapped" ensemble of model structures $\mathbf{G}^{(b)}$ by sampling *with replacement* $N_G^{\wp,k}$ model structures with mechanism $m_k^{\wp}$, where $N_G^{\wp,k}$ is the number of models with mechanism $m_k^{\wp}$ in the sample space $\mathbf{G}$.
2. Calculate $p(m_k^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G}^{(b)})$ using Equation 10.
3. Calculate $p(m_i^{\wp} \mid \tilde{\mathbf{q}}, \mathbf{G}^{(b)})$ for all other mechanisms for process $\wp$, $i = 1,\ldots,N_m^{\wp}$ (and $i \neq k$), also using Equation 10.
4. Repeat steps 1–3 for $b = 1,\ldots,N^{\text{boot}}$, i.e., construct N^{boot} bootstrapped (random) model ensembles. For example, in this study, we set $N^{\text{boot}} = 10{,}000$.
5. Calculate $t_k^{\wp(b)}$ for each bootstrap model ensemble $b = 1,\ldots,N^{\text{boot}}$ using Equation 11.
6. Compute the empirical frequency of $t_k^{\wp} > \tau$ across all bootstrapped model ensembles

$$\omega_k^{\wp} = \frac{1}{N^{\text{boot}}} \text{count}\{t_k^{\wp(b)} > \tau; b = 1,\ldots,N^{\text{boot}}\} \tag{13}$$

   where the function count v is defined as the number of true elements in a Boolean set v.
7. Reject $H0_k^{\wp}$, i.e., identify $m_k^{\wp}$ as dominant, if

$$\omega_k^{\wp} \geq 1 - \alpha_{\text{Bonf}} \tag{14}$$

   where $\alpha_{\text{Bonf}} = \alpha \,/\, N_m^{\wp}$ is the Bonferroni correction to the prescribed significance level α. Otherwise $H0_k^{\wp}$ is not rejected.

The original ensemble $\mathbf{G}$ may be selected randomly one or more times like any other bootstrap ensemble and is not given any special treatment. Due to sampling with replacement, any given bootstrapped ensemble $\mathbf{G}^{(b)}$ may (and almost always will) contain multiple instances of one or more model structures at the expense of excluding one or more other model structures. This setup is analogous to classic bootstrapped data sets, which exchange some of the observed data points with multiple instances of other observed data points.

Steps 6–7 are repeated for all mechanisms $\{m_k^{\wp}; k = 1,\ldots,N_m^{\wp}\}$ proposed for process $\wp$. If none of the individual null hypotheses $\{H0_k^{\wp}; k = 1,\ldots,N_m^{\wp}\}$ are rejected, then the null hypothesis $H0^{\wp}$ for the entire family of comparisons is not rejected. In this case, no mechanism is identified as dominant, i.e., the dominant mechanism is "not identified" or "undefined".

The same hypothesis-testing procedure is then applied to estimate the dominant mechanisms for all other model processes $\wp = 1,\ldots,N^{\wp}$.

3. Case Study Description

This section details the synthetic and real data case studies used to investigate the fundamental properties of the proposed mechanism identification method. A major focus is on the ability of the method to identify dominant mechanisms in the presence of data/model error. In the synthetic case study, the "true" dominant mechanisms are assumed to be known, and the error is added in a controlled way from a distribution that matches the assumed error model, in order to provide a controlled experiment. Multiple (synthetic) data replicates are used to quantify the performance of the mechanism identification method using metrics of statistical reliability and power. In contrast, the real data study investigates the method under conditions when the error assumptions are not met exactly. The following sections provide technical details of the case study procedures. Flowcharts of the synthetic and real data analyses are given in Figure 2.

3.1. Catchment and Data

The case studies use data from Leizarán catchment (c8z1) located in Basque Country in northern Spain. The catchment has an area of 114 km^2 and drains into the Cantabrian Sea. It is a humid catchment (Arora, 2002),

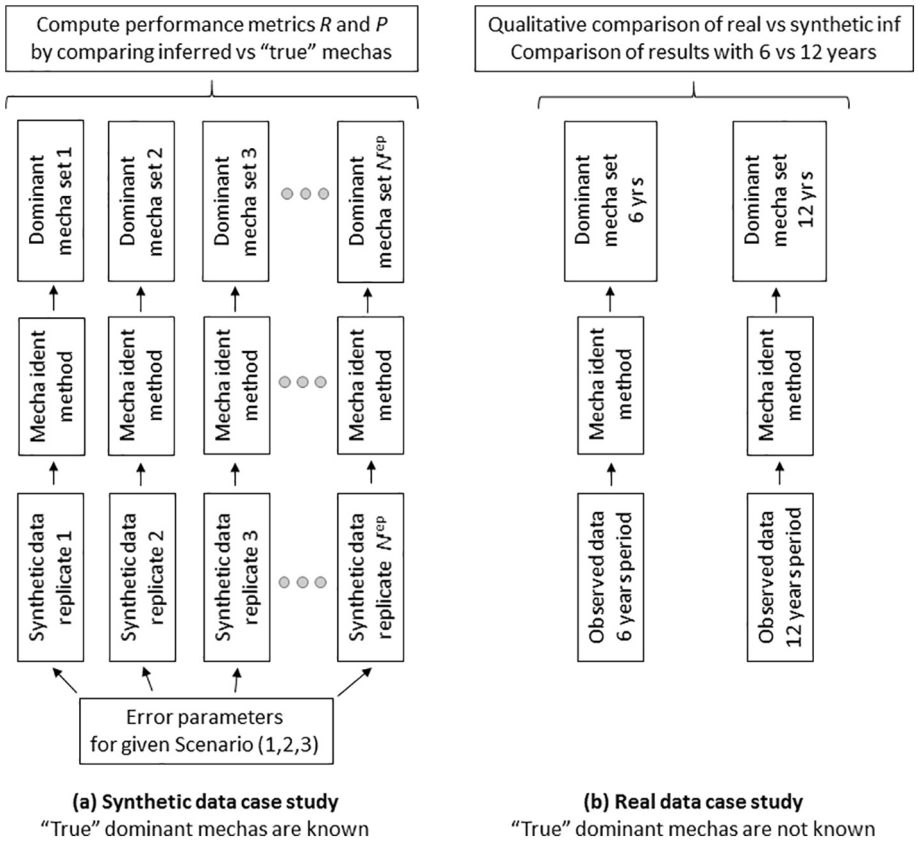

Figure 2. Flowchart of the procedure employed in the empirical case studies: (a) synthetic data analysis (Scenarios 1–3) and (b) real data analysis (Leizarán catchment).

with an annual average rainfall of 1,794 mm, annual potential evapotranspiration of 684 mm, and annual runoff of 1,076 mm (see Prieto et al., 2019 for more details).

Daily time series of precipitation, PET, and streamflow were provided by the Diputación Foral de Guipúzcoa. The annual averages reported above are estimated by arithmetic averaging of daily data.

Data from October 1, 1995 to September 30, 2002 are used for model calibration and mechanism identification, and data from October 1, 1995 to September 30, 1996 are used for model warm-up (to reduce the impact of unknown model initial conditions).

3.2. Modeling Framework for Hypothesizing Hydrological Mechanisms and Deterministic Hydrological Model Structures

The hydrological models and mechanisms for hypothesis testing are generated using the FUSE, a multihypothesis hydrological modeling system designed to facilitate work on model representation and improvement (e.g., Clark et al., 2008; 2011a; Konapala et al., 2020).

FUSE provides a choice of multiple model mechanisms to represent each model process. A total of seven processes are considered in this paper, namely: (1) architecture of the upper soil layer for the storage of water occurring in the unsaturated zone; (2) architecture of the lower soil layer for the storage of water occurring in the saturated zone; (3) evapotranspiration; (4) interflow for the lateral movement of the water into the soil; (5) percolation for the vertical movement of water from the unsaturated zone (upper soil layer) to the saturated zone (lower soil layer); (6) surface runoff generation; and (7) routing for the evolution (shape and time) of the surface runoff hydrograph as the water moves through the river. The total number of mechanisms is 19; see Table 1 for details.

Table 1
Hydrological Processes and Mechanisms in the FUSE Framework. The processes selected as "true" in the synthetic experiments are indicated in **bold** *in the second column*

Processes	Mechanisms
Architecture of the upper soil layer: processes occurring in the unsaturated zone	m_1^1: **single state variable**
	m_2^1: 1 tension storage + 1 free storage
	m_3^1: cascading buckets: tension storage subdivided into recharge and excess
Architecture of the lower soil layer: processes occurring in the saturated zone	m_1^2: 1 baseflow storage of fixed size
	m_2^2: **1 tension storage + 2 parallel tanks**
	m_3^2: 1 baseflow storage of unlimited size, frac rate
	m_4^2: 1 baseflow storage of unlimited size, power recession
Evapotranspiration	m_1^3: root weighting. Evapotranspiration in each soil layer depends on the relative root fraction in the upper and lower soil layers
	m_2^3: **sequential evaporation model**. Evapotranspiration in the upper and lower layers, where evapotranspiration in the lower layer is restricted by the potential evapotranspiration satisfied in the upper layer
Interflow: lateral movement of water in the upper soil layer.	m_1^4: **interflow absent**
	m_2^4: interflow present. Linear function of free storage in the upper layer
Percolation: vertical movement of the water from upper soil layer to lower soil layer	m_1^5: **water from field capacity to saturation is available for percolation**
	m_2^5: percolation defined by moisture content in lower layer (SAC)
	m_3^5: saturated zone control: water from wilting point to saturation is available for percolation
Surface runoff generation	m_1^6: saturated area is related to storage in the unsaturated zone via a Pareto distribution (ARNO/Xzang/VIC)
	m_2^6: **saturated area is a linear function of tension storage in the unsaturated zone (PRMS variant)**
	m_3^6: saturated area is related to storage in the saturated zone via the topographic index (TOPMODEL- only valid for TOPMODEL)
Routing: evolution (shape and time) of surface runoff hydrograph as water moves through the river	m_7^1: routing absent
	m_2^7: **routing present, using Gamma distribution with shape parameter = 2.5**

The FUSE mechanisms are represented by components of existing models, namely, PRMS, SACRAMENTO, TOPMODEL, and ARNO/VIC. Each FUSE model structure is a combination of 7 hydrological mechanisms, with a single mechanism specified for each hydrological process. A total of 624 deterministic model structures are thus considered.

The inputs into FUSE are the (daily) time series of catchment-average observed rainfall and potential evapotranspiration, and the outputs are the (daily) time series of simulated streamflow.

3.3. Probabilistic Model and Parameter Estimation (Single Model Structure)

The probabilistic hydrological model in this work is given by FUSE Equations 3 and 4, in combination with the Box–Cox transformation and the Gaussian error model. The Box–Cox power parameter, λ, is fixed to 0.2;

the offset parameter, A, is fixed to 0.035. The parameter prior in Equation 5 is specified as uniform over the feasible ranges. The corresponding likelihood function and prior are presented in Appendix B.

For pragmatic considerations, the estimates of hydrological and residual model parameters in Equation 5, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_h, \hat{\sigma}_\eta)$, are obtained using a computationally efficient "hybrid" least squares/method-of-moments (LS-MoM) approach, similar to the approach presented by McInerney et al. (2018).

The hybrid approach is summarized below:

Stage 1. The estimated hydrological model parameters, $\hat{\boldsymbol{\theta}}_h$, are obtained using the least squares method, by minimizing the sum of squared errors in Box–Cox space. The robust Gauss–Newton optimization algorithm with 10 multistarts is used (Qin et al., 2018). The estimates are then refined using a quasi-Newton optimization method with higher resolution finite difference gradient estimation and tight convergence tolerance (Kavetski & Clark, 2010).

Stage 2. The estimated standard deviation of normalized residual errors, $\hat{\sigma}_\eta$, is obtained using the method of moments from the time series of normalized residuals $\hat{\boldsymbol{\eta}}$ (computed using the optimum hydrological model parameters $\hat{\boldsymbol{\theta}}_h$).

This approach is equivalent to joint optimization of $\hat{\boldsymbol{\theta}}_h$ and $\hat{\sigma}_\eta$ but is computationally faster (McInerney et al., 2018). The assumption of negligible posterior parameter uncertainty is consistent with the use of the BIC to approximate posterior model probabilities (Section 2.4 and Appendix A). The reduced computational cost is important given the number of model calibrations (parameter estimations) required for the case studies in this paper (Section 3.6).

3.4. Empirical Analysis Using Synthetic Data

The synthetic study enables a thorough investigation (verification) of the fundamental properties of the proposed mechanism identification method. A major focus is on the ability of the method to identify dominant mechanisms in the presence of error. This error is intended to represent the effects of data and model structural errors. Multiple replicates of synthetic data are used to quantify the empirical performance of the inference using metrics of statistical reliability and power.

A flowchart of the synthetic data study is given in Figure 2a; technical details are given in the following sections.

3.4.1. Synthetic Data Generation and Error Scenarios

The synthetic "observed" data are obtained by assuming a set of true mechanisms (yielding a "true" model) and "true" parameter values, shown in Table 1, using the procedure detailed in Appendix C. To achieve a degree of "realism" (representativeness of real conditions), the set of true mechanisms and parameters is taken from a model structure that was well performing in the real case study. The synthetic "observed" data are obtained from the synthetic "exact" data by adding Gaussian noise with variance σ_η^2 in Box–Cox transformed space with $\lambda = 0.2$ and $A = 0.035$. This synthetic error model is consistent with the assumed error model (Section 3.3) and can be interpreted as representing combined "data/model" error.

The following scenarios of data/model errors are considered:

Scenario 1. Low errors: $\sigma_\eta = 0.025$

Scenario 2. Medium errors: $\sigma_\eta = 0.1$

Scenario 3. High errors: $\sigma_\eta = 0.25$

In each scenario, 50 synthetic replicates are generated, shown in Figure 3. The replicates are treated as "statistical trials" when computing the performance metrics detailed in the following section.

3.4.2. Performance Attributes and Metrics for the Hypothesis-Testing Method

The following performance attributes and quantitative metrics are considered:

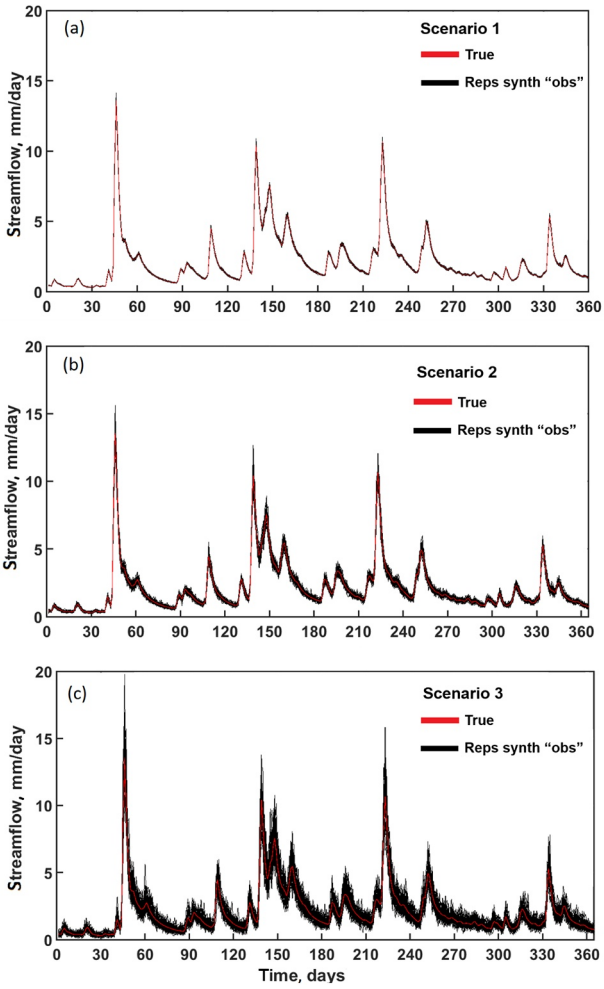

Figure 3. Synthetic streamflow data used in Scenarios 1–3. Red: synthetic "exact" streamflow. The black lines refer to the 50 replicates of synthetic "observed" streamflow.

1. How reliable ("trustworthy") is the mechanism identification method?

In other words, if the method identifies a mechanism as dominant, what is the probability that this mechanism is the true (dominant) mechanism?

We define (empirical) reliability R as

$$R = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{15}$$

where N_{TP} is the number of trials where the true mechanism is identified as dominant and N_{FP} is the number of trials where the wrong mechanism is identified as dominant. This quantity has also been termed "Positive Prediction Value" (Tharwat, 2020). The definition of trials based on the replicates is given in Section 3.4.3.

R ranges from 0 to 1, with higher values indicating better reliability. Given the definition of FWER in Section 2.6, with prescribed significance level of $\alpha = 5\%$, a method can be considered "reliable" if $0.95 \leq R < 1$, as the frequency of false positives is below the imposed significance level.

2. What is the "statistical power" of the method?

Statistical power is defined as the probability of (correctly) rejecting the null hypothesis when it is indeed false (Dekking, 2005; Neyman & Pearson, 1928). For our experimental setup, an empirical estimate of power is given by

$$P = 1 - \frac{N_{\text{FN}}}{N_{\text{trials}}} \tag{16}$$

where N_{FN} is the number of trials where a dominant mechanism was not identified and N_{trials} is the total number of trials.

P ranges from 0 to 1, with higher values indicating better power. Low values of P correspond to the method being "indecisive" and hence of little value to a modeler. In the hypothesis-testing literature, a method is generally considered "powerful/decisive" if $0.8 \leq P < 1$ (e.g., Ellis, 2010).

Note that in general there are trade-offs between the reliability and power of a test: as the probability of making Type I errors (false positives/discoveries) decreases, the probability of making Type II errors (false negatives/rejections) increases (Smith & Bryant, 1975). In the context of hydrological process analysis, it is preferable to be indecisive (i.e., do not identify a dominant mechanism) than wrong (i.e., identify a wrong mechanism as dominant), because identifying a wrong mechanism as dominant is misleading and confuses our catchment understanding.

3.4.3. Use of Performance Metrics Across the Synthetic Error Scenarios

The performance metrics are used to analyze the method for three different stratifications/pooling setups.

1. Scenario-specific metrics: metrics listed in Section 3.4.2 computed separately for each data/model error scenario defined in Section 3.4.1. We pool together the results of identifying all model processes, so that the number of synthetic trials per scenario is $N_{\text{trials}}^{\text{scen}} = 350$ (7 processes × 50 replicates). This analysis tells us how the performance of the hypothesis-testing method depends on the error magnitude.
2. Process/scenario-specific metrics: metrics computed separately for the identification of each process in each scenario. In this case, the number of (synthetic) trials is $N_{\text{trials}}^{\text{process}} = 50$ (50 replicates). This analysis provides an indication of how the hypothesis-testing method performs for the identification of different

processes, which processes are the most/least identifiable (i.e., for which processes are the true mechanisms identified as dominant with highest/lowest reliability), and what is the statistical power of these identifications.

3. Overall metrics: metrics computed by pooling together the results of identifying all processes in all scenarios, so that the number of total trials is $N_{\text{trials}}^{\text{overall}} = 1{,}050$ (3 scenarios × 7 processes × 50 replicates). These metrics can be used to compare the overall performance of the hypothesis-testing method.

The analysis above is carried out for a significance level $\alpha = 0.05$. In addition, we report results obtained when α is relaxed to 0.1 and when α is tightened to 0.01. This analysis provides an indication into the sensitivity of the inference to the significance level.

3.5. Empirical Analysis Using Real Data

The performance of the proposed method is illustrated in a real data case study using observed data from the Leizarán catchment (see Section 4.2). Since the "true" mechanisms in this catchment are not known, our analysis here is limited to the following:

- Evaluation of broad similarities in the behavior of the mechanism identification method under synthetic versus real conditions. In particular, we compare how the number and type of dominant mechanisms identified using real observations compared to the number and type of dominant mechanisms identified using synthetic data.
- Comparison of findings regarding dominant mechanisms to the existing knowledge of the hydrology of the Leizarán catchment.
- Appraisal of consistency of mechanism identification based on 6 and 12 years of data, and sensitivity to the choice of significance level α.

A flowchart of the real data study is given in Figure 2b.

3.6. Computational Platform and Costs

The numerical experiments carried out in this study are computationally expensive.

In total, across all scenarios, replicates, error magnitude levels, and FUSE model structures, (approximately) 95,000 individual calibrations were required, each implemented using 10 optimizations (corresponding to 10 initial seeds). The complete analysis consumed approximately 13 months of CPU time on the IHCantabria supercomputer cluster Neptuno.

4. Results

4.1. Synthetic Data Experiments

This section reports the performance of the mechanism identification method in Scenarios 1–3 as the magnitude of (synthetic) error is increased. Given the synthetic nature of the analysis, the true mechanisms are known; hence, we can compute the reliability and power metrics.

Figure 2 shows the synthetic "exact" streamflow and the synthetic "observations" of streamflow generated in Scenarios 1–3 (see Section 3.4.1). The true mechanisms are highlighted in bold in Table 1. Figure 4 shows the estimated empirical reliability and power of the method for Scenarios 1–3 (low, medium, and high error, respectively). The performance metrics are shown both for each process individually and averaged across all processes (upper row). In addition, we report mechanism identifiability averaged across all processes (right-most column).

We begin by considering reliability and power achieved for the "default" significance level of 0.05. The sensitivity to this specification is examined in Section 4.1.5.

4.1.1. Scenario 1: Low Errors

Figure 4 shows that, in the presence of low errors, the mechanism identification method achieves a reliability $R_{\text{scen1}} = 1$ and $P_{\text{scen1}} = 0.92$. In other words, a determination is made in 92% of the $N_{\text{trials}}^{\text{scen}} = 350$ replicates

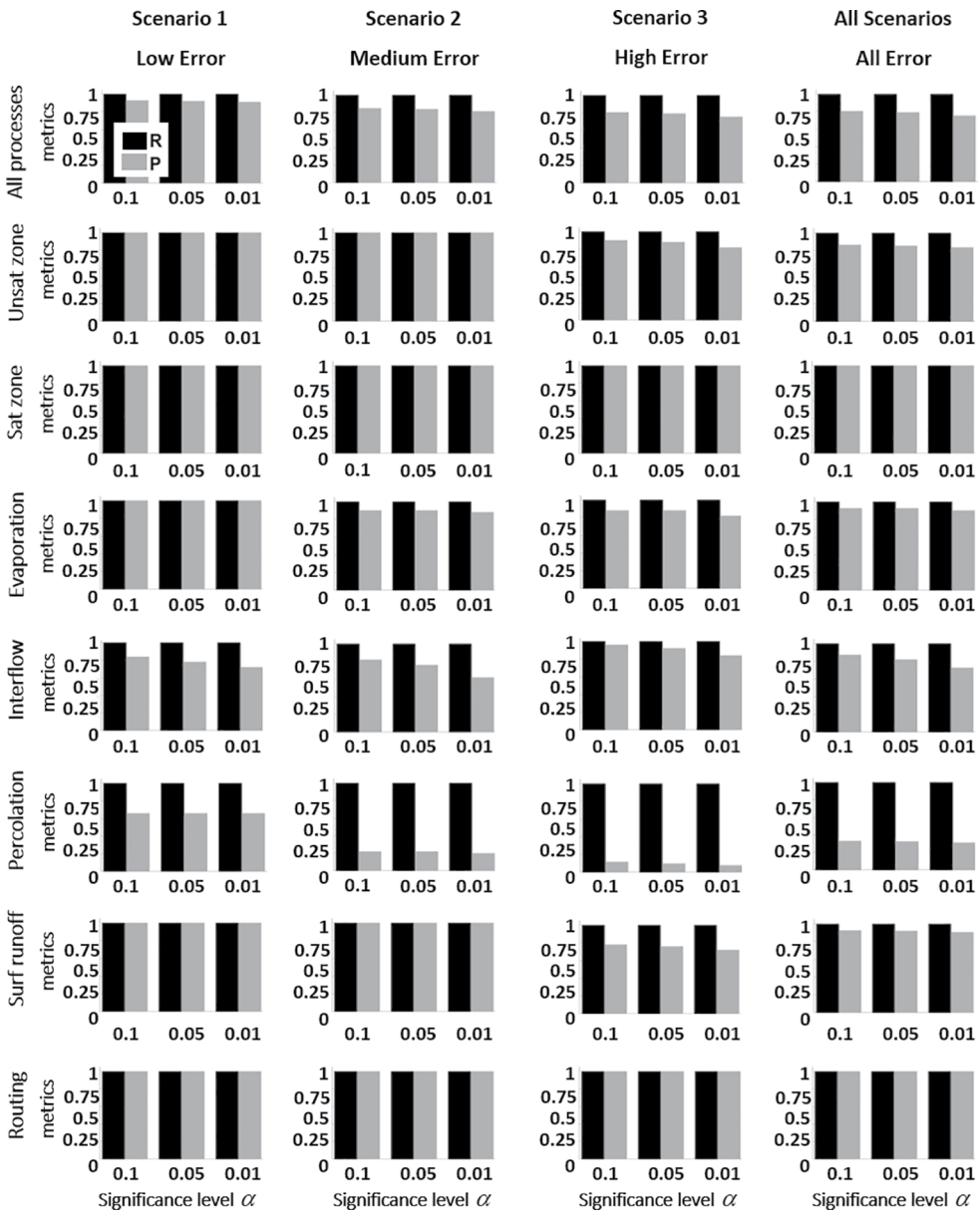

Figure 4. Reliability and power of mechanism identification in synthetic Scenarios 1–3. Process-averaged and scenario-averaged metrics are shown in Row 1 and Column 4, respectively. Reliability shown with black bars and power shown with gray bars. Results reported for significance levels $\alpha = 0.1$, 0.05, and 0.01.

(yielding $P = 0.92$), and the true mechanism is (correctly) identified as dominant in 100% of these trials (yielding $R = 1$).

Looking at each process independently, Figure 4 also shows "perfect" mechanism identification for 5 processes: processes related to the storage in the upper and lower soil layers, evapotranspiration, surface runoff generation, and routing. For these 5 processes, the dominant mechanisms are identified correctly in all 50 replicates ($N_{\text{trials}}^{\text{process}} = 50$), i.e., $R = 1$ and $P = 1$.

For the remaining 2 processes, interflow and percolation, the dominant mechanisms are identified with perfect reliability ($R_{\text{evapotranspiration1}} = R_{\text{interflow1}} = R_{\text{percolation1}} = 1$). However, these mechanisms are identified with lower power, $P_{\text{interflow1}} = 0.78$ and $P_{\text{percolation1}} = 0.66$.

4.1.2. Scenario 2: Medium Errors

Figure 4 (row 1) shows that, in the presence of medium errors, the mechanism identification achieves a reliability $R_{scen2} = 1$ and $P_{scen2} = 0.84$.

Looking at each process independently, Figure 4 shows that the reliability and power in Scenario 2 remain the same as in Scenario 1 for processes related to the storage in the upper and lower soil layers, surface runoff generation and routing. For these 4 processes, the dominant mechanisms are identified correctly in all 50 replicates ($N_{trials}^{process} = 50$), i.e., $R = 1$ and $P = 1$.

For the remaining 3 processes, perfect reliability is maintained, $R_{evapotranspiration2} = R_{interflow2} = R_{percolation2} = 1$, but once again with a notable reduction in power, for example, $P_{evapotranspiration2} = 0.9$ and $P_{interflow2} = 0.76$. The identification of percolation suffers the largest loss of power, with $P_{percolation2} = 0.22$ (down from $P_{percolation1} = 0.66$ in Scenario 1).

4.1.3. Scenario 3: High Errors

Figure 4 (row 1) shows that, in the presence of high errors, the mechanism identification maintains same reliability as Scenario 2, $R_{scen3} = 1$, while its power decreases from $P_{scen2} = 0.84$ to $P_{scen3} = 0.8$.

Looking at each process independently, Figure 4 shows that the method maintains perfect identifiability for processes related to the storage in the lower soil layer and for the routing process ($R_{lower\ soil\ layer3} = R_{routing3} = 1$ and $P_{lower\ soil\ layer3} = P_{routing3} = 1$). These results are the same as in Scenario 2.

For the remaining processes, reliability remains perfect ($R_{upper\ soil\ layer3} = R_{evapotranspiration3} = R_{surface\ runoff3} = R_{percolation3} = 1$), but there is a notable loss of power. For example, for the upper soil layer and evapotranspiration, $P_{upper\ soil\ layer3} = P_{evapotranspiration3} = 0.88$, and for surface runoff generation, $P_{surface\ runoff3} = 0.76$. For percolation, the loss of mechanism identification power is almost complete, $P_{percolation3} = 0.1$.

4.1.4. Comparison Across the Processes for All Error Levels

Pooling the results across all scenarios and processes (Figure 4, upper right hand corner), mechanism identification achieves $R_{overall} = 1$ and $P_{overall} = 0.86$.

The most identifiable processes are those related to the storage in the lower soil layer and routing. For these processes, mechanism identification achieves perfect reliability and power in all three scenarios, i.e., regardless of streamflow errors.

In contrast, the least identifiable process is the percolation process, especially as streamflow errors increase. For this process, the power of mechanism identification deteriorates from $P = 0.66$ in Scenario 1 to $P = 0.22$ in Scenario 2 and then to $P = 0.1$ in Scenario 3.

Interestingly, the identifiability of (mechanisms for) the interflow process decreases from $P = 0.78$ in Scenario 1 to 0.76 in Scenario 2 but then increases to 0.92 in Scenario 3. This pattern of change is accompanied by several other processes becoming poorly identifiable. For example, the identification of mechanisms for the surface runoff generation process deteriorates to $P_{surface\ runoff\ generation3} = 0.76$ (down from $P_{surface\ runoff\ generation2} = 1$ in Scenario 2). A similar deterioration is seen in the identification of mechanisms for the storage in the upper soil layer. The concurrent increase of power in the identification of interflow mechanisms and the drop of power in the identification of mechanisms for other processes suggests a "compensatory/interaction behavior," discussed in Section 5.4.

4.1.5. Sensitivity to the Prescribed Significance Level

We now consider the sensitivity of mechanism identification to the prescribed significance level α. Figure 4 shows that the results are in general stable. For example, tightening α from 0.05 to 0.01 does not impact R in any of the scenarios and does not impact P in Scenarios 1 and 2 for processes in the unsaturated and saturated zones, surface runoff generation, and routing.

The value of α makes the most impact when errors are high. For example, in Scenario 3, tightening α from 0.05 to 0.01 results in a degradation in the identification of interflow, with $P_{interflow3}$ decreasing from 0.92 to 0.84 (i.e., a loss in P of 0.08). For percolation, the loss in P is only 0.02; however, power is already very low

Table 2
Dominant Mechanism Identification in the Leizarán Catchment (Real Catchment Case Study) Using 6 and 12 Years of Data, and Prescribed Significance Levels of 0.01, 0.05, and 0.1

	6 years			12 years		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Architecture of the upper soil layer	m_1^1	m_1^1	m_1^1	–	–	–
Architecture of the lower soil layer	m_2^2	m_2^2	m_2^2	m_2^2	m_2^2	m_2^2
Evapotranspiration	m_1^3	m_1^3	m_1^3	m_1^3	m_1^3	m_1^3
Interflow	–	–	–	–	–	m_1^4
Percolation	–	–	m_1^5	–	–	–
Surface runoff generation	–	–	–	–	m_1^6	m_1^6
Routing	m_2^7	m_2^7	m_2^7	m_2^7	m_2^7	m_2^7

in the identification of this process, for example, $P_{\text{percolation3}} = 0.1$ when $\alpha = 0.05$ which reduces further to $P_{\text{percolation3}} = 0.08$ when $\alpha = 0.01$.

Conversely, the value of α has the least impact when errors are low. For example, in Scenario 1, tightening α from 0.05 to 0.01 does not impact the identification of the dominant mechanisms in the upper and lower soil layers, evapotranspiration, percolation, surface runoff generation and routing, where $P = 1$ both when $\alpha = 0.05$ and when $\alpha = 0.01$. Indeed, the only process affected substantially by the change in the significance level is the interflow, where the tightening of α from 0.05 to 0.01 results in a reduction of P from 0.78 to 0.72.

4.2. Real Data

Table 2 reports the mechanisms identified from real data in the Leizarán catchment. In this analysis, the true mechanisms are unknown and hence we cannot reliably establish whether the mechanisms identified as dominant are the "true" (or "quasi-true") mechanisms. Estimates of reliability and power are unavailable, and we focus instead on qualitative results.

We begin by considering mechanism inference from 6 years of data, with significance level $\alpha = 0.05$.

Dominant mechanisms are identified for 4 of the 7 processes represented in FUSE: storage in the unsaturated zone (single state variable), storage in the saturated zone (tension storage plus two parallel tanks), evapotranspiration (proportional to the depth of the roots in each layer), and routing (routing present). No dominant mechanisms are identified for the remaining 3 processes, namely interflow, percolation, and surface runoff.

These process identification patterns are broadly similar to those found in the synthetic scenarios. In terms of identifiable processes/mechanisms: processes related to the storage in the lower soil layer and routing have well identifiable dominant processes—similar to synthetic Scenarios 2 and 3. In terms of nonidentifiable mechanisms: notably percolation is not identifiable in the real catchment, just as it was in the synthetic study. This similarity is unsurprising given that the synthetic studies were set up using real data findings but does indicate that no unexpected artifacts are introduced by the inference.

Table 2 reports the sensitivity of mechanism identification to changes in the significance level α. For the 6-year inference period, tightening α from 0.05 to 0.01 has no impact on mechanism identifiability for any of the processes. However, loosening α from 0.05 to 0.1 leads to a dominant mechanism being identified for the percolation process.

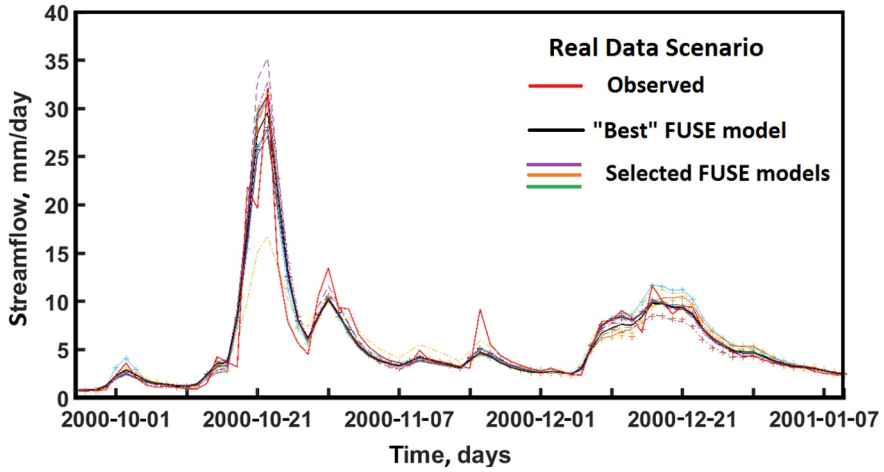

Figure 5. Observed (red) hydrograph in the Leizarán catchment and simulated (multiple colors) hydrographs generated by selected FUSE model structures. The following subset of FUSE model structures is shown: (1) models with mechanisms identified as dominant for processes in the upper and lower soil layer, evapotranspiration, and routing; and (2) models with all possible mechanisms for those processes where a mechanism is not identified as dominant (interflow, percolation, and surface runoff generation).

Table 2 also shows the impact of data length on mechanism identification. There are no *direct* contradictions between the mechanisms identified using 6 and 12 years of data: in all processes for which a mechanism is identified in both periods, the same mechanism is identified in both periods. More specifically, mechanisms for processes in the saturated zone, evapotranspiration, and routing of surface runoff remain identifiable in both periods, and the same mechanisms are identified for these processes.

However, the change in data length does impact identifiability, with some switches in the processes for which dominant mechanisms are identified. In particular, there are switches in the identifiability of processes in the unsaturated zone and processes for surface runoff generation. For the 6 years period, the dominant mechanism is identified for the process of storage in the saturated zone (singe state variable mechanism), whereas for the 12 years period, a dominant mechanism is not identified for this process. The opposite is found for the surface runoff generation process: a dominant mechanism is not identified from the 6 years period but is identified from the 12 years period (the mechanism where saturated area is related to the storage in the unsaturated zone via a Pareto distribution).

Finally, Figure 5 compares the observed and simulated hydrographs from a selected subset of FUSE model structures. The following 18 model structures are selected: (1) models with mechanisms identified as dominant for processes in the upper and lower soil layer, evapotranspiration, and routing; and (2) models with all possible mechanisms for those processes where a mechanism is not identified as dominant (interflow, percolation, and surface runoff generation). Figure 5 shows that the FUSE models provide a generally accurate approximation to the observed streamflow, with NSE values (not labelled) of around 0.82–0.88 (and as high as 0.91–0.93 when computed with BC0.2-transformed streamflow).

5. Discussion

The discussion section is organized as follows. Key insights from the synthetic and real data case studies are discussed in Sections 5.1 and 5.2, respectively. Broader concepts of mechanism identification in the presence of data and model error are discussed in Section 5.3. Future research directions to overcome present study limitations are outlined in Section 5.4.

5.1. Insights From the Synthetic Study

5.1.1. Higher Reliability Gives the Modeler Confidence in the Mechanisms Identified as Dominant

The mechanism identification method is "reliable": if it identifies a mechanism as dominant, this mechanism is the true (or, more generally, "quasi-true") mechanism. Perfect reliability is achieved for all processes and scenarios, $R_{\text{overall}} = 1$ across a total number of trials $N_{\text{trials}}^{\text{overall}} = 1{,}050$, which is obviously well within the imposed confidence level of 95%.

However, high reliability comes at the cost of lower power, i.e., ability to make an identification. The power when pooled over all scenarios is $P_{\text{overall}} = 0.86$, indicating that the method does not identify a dominant mechanism in 14% of the trials. However, for some processes and error levels, the power dropped to as low as 0.22 and 0.10 (see Section 5.1.3 for further discussion of process identifiability).

The trade-off between reliability (i.e., low frequency of false positives) and power (i.e., low frequency of false negatives) is known from theoretical considerations. In particular, tightening the significance level α reduces the probability of Type I errors but increases the probability of Type II errors (e.g., Smith & Bryant, 1975). It is important to note that such lack of identification power is at least not misleading to the modeler—in contrast to the case of lack of reliability where the wrong mechanism is identified as dominant. A method with lower reliability, *especially* in conjunction with high power, has less practical value: the relatively high frequency of identifying a wrong mechanism will confound the modeler's interpretation of catchment functioning. It may also result in worse predictive performance when the identified mechanisms are used in predictive modeling contexts.

Overall, the high reliability of the mechanism identification method gives the modeler confidence in the identification of a mechanism as dominant, and cases where no mechanism is identified as dominant may point the modeler to seek more and/or higher quality data and/or to hypothesize new process mechanisms (see Section 5.1.3).

5.1.2. Performance in the Presence of Increased Errors in the Model/Data

Mechanism identification tends to deteriorate as the magnitude of data/model error is increased. In this synthetic study, this error represents a combined effect of data/model error (as the residual error model used to obtain the synthetic "observed" data does not distinguish multiple sources of error). While the deterioration in performance is unsurprising, the deterioration pattern is relatively "benign." In particular, the deterioration manifests as a loss of power, i.e., an increased probability of Type II errors (here, not identifying any mechanism as dominant): no mechanism is identified as dominant in 8%, 16%, and 20% of the trials for Scenarios 1, 2 and 3, respectively. In line with earlier arguments, we consider a loss of power to be a lesser limitation than a loss of reliability.

5.1.3. Which Processes Are the Most and Least Identifiable?

Processes related to storage of water in the saturated zone, and routing, appear well identifiable. Dominant mechanisms for these processes can be identified with perfect reliability and power even from streamflow corrupted with high errors. In contrast, interflow and percolation processes are the least identifiable. Specifically, no percolation mechanism is identified as dominant in as many as 78% and 90% of trials in Scenarios 2 and 3, respectively.

Process identifiability necessarily depends on aspects such as the contribution of the process to the model response used for mechanism identification. For example, a mechanism for a process that contributes a negligible amount of streamflow is unlikely to be identified from streamflow alone. The degree of difference in the competing mechanisms is also of clear relevance. For example, distinguishing between three mechanisms will be much harder if they employ similar equations (e.g., see earlier study by Gupta and Sorooshian 1983). This effect can be seen for the percolation process, for which FUSE provides three options all of which are power functions of (different) model storages. Mechanism identification clearly suffers in these circumstances, with power dropping to as low as 22% and 10% for medium and high errors, respectively.

The dependence of identifiability on mechanism similarity is not surprising. As we consider more and more subtle differences between mechanisms, our ability to establish a mechanism as dominant will necessarily

become more limited, especially in the presence of error. Increasingly accurate data would be needed to continue refining process representation.

Other considerations of data uncertainty are also relevant here, for example, estimated values of low flows can be sensitive to data errors (Westerberg et al., 2016), complicating the identification of mechanisms operating during low flow and ephemeral conditions.

5.2. Insights From Real Data Study (Leizarán Catchment)

5.2.1. Are Findings in the Real Data Study Similar to the Synthetic Scenarios?

In this section, we compare the findings in the synthetic and real case studies to look for broad similarities and differences. Section 4.2 indicates that process identifiability is generally consistent across the real data study and Scenarios 2 and 3. For example, the saturated zone process is always identifiable, with the dominant mechanism being tension storage plus two parallel tanks, and the surface runoff routing process is also always identifiable, with the dominant mechanism being "routing present." A minor exception is the evapotranspiration process, where a dominant mechanism is always identified in the real data study but is occasionally not identified in Scenarios 2 and 3 (P is 0.9 and 0.88).

5.2.2. Connection Between the Mechanisms Identified as Dominant and Existing Process Understanding in the Leizarán

The mechanisms identified as dominant can be interpreted from a process-oriented perspective that is available, albeit in a limited way, in the Leizarán (Basque Water Agency, personal communication, March 12, 2020):

Storage in the upper soil layer. Approximated by a single state variable (i.e., without a tension storage). This mechanism, which presumes low tension storage, is plausible because Leizarán catchment has a low clay content (3%).

Storage of water in the saturated zone. Two parallel tanks and one tension storage. This mechanism might be plausibly linked to the combination of the geological and topographic conditions of the catchment, which favors a subsurface flow component. Geology is composed of 28% calcareous rocks, 28% sands, and 37% siliceous rocks; the catchment slope is high (elevation change over horizontal length is 0.42).

Evapotranspiration. Proportional to the depth of roots in each soil layer. This mechanism is plausible given the catchment has a riparian forest whose vegetation is oak and alder.

Routing. Present. The inference method finds that it is more probable that the surface runoff hydrograph is propagated through the river to the catchment outlet rather than being directly delivered to the outlet, i.e., the "routing mechanism is dominant." This mechanism, which specifies a lag between precipitation and streamflow generation, is plausible because the catchment has meanders across its area of 114 km^2.

For interflow, percolation, and surface runoff generation (the remaining processes), no mechanism is identified as dominant.

We note that these interpretations are necessarily tentative, in view of the currently limited understanding of the hydrology of the Leizarán catchment. These interpretations can be pursued in more depth in future work, along the research lines on the correspondence between models and catchments (Fenicia et al., 2014; Wrede et al., 2015; Carrer et al., 2019).

The general finding that the interflow and percolation processes are the hardest to identify aligns with previous work on hydrological process identification, where processes related to soil, geology, and vegetation were found harder to characterize than those related to climatic attributes (Beck et al., 2015; Addor et al., 2018). This raises the question of how to make the best use of soil and geological data, including signatures, for hydrological modeling (Gupta et al., 2008; Fenicia et al., 2018), and how to represent the continuum of response dynamics in the unsaturated and saturated zones (Silva et al., 2009).

The sensitivity of the identified dominant mechanism to the choice of significance level α appears low: only a single additional mechanism becomes identifiable if α is relaxed to 0.1 (percolation for the 6 years period and interflow for the 12 years period). However, this finding is likely to be case specific.

Available data length is expected to impact mechanism identification, both by providing information for such identification and by potentially introducing variability into the identification if the catchment undergoes hydrological change. In the real catchment case study using 6 and 12 years of data, 4 out of 7 mechanisms are identified in both periods. An encouraging finding is that the mechanism inference is consistent: for those processes where a mechanism is identified as dominant with both lengths of data, the same mechanism is identified. In other words, there was no change in the estimation from one mechanism to another as more data were added—the only changes were from a mechanism being identified to a mechanism not being identified, and vice versa. Naturally the temporal consistency of mechanism identification also depends on the catchment not undergoing any genuine major changes. These findings also align with the previous literature on hydrological model component identification, including on the number of model components identifiable from a streamflow time series (e.g., Jakeman & Hornberger, 1993) and on the length of data needed to calibrate a model in a humid catchment (e.g., Gupta & Sorooshian, 1983; Sorooshian et al., 1983; Yapo et al., 1996; Li et al., 2010).

5.3. Connection to Current Hypothesis Testing and Model Selection in Hydrology

5.3.1. Methods Proposed in This Work Focus on Individual Model Processes/Mechanisms, in Contrast to Existing Methods Which Focus on Complete Models

Flexible modular models have facilitated important advances for hypothesis testing in hydrology, enabling the decomposition of models into multiple testable hypotheses about mechanisms for individual processes (e.g., Clark et al., 2008; 2011a; 2015; Fenicia et al., 2011; Kraft et al., 2011; Wrede et al., 2015; Fenicia et al., 2016; Addor and Melsen, 2019; Knoben et al., 2019; Craig et al., 2020, and others).

Previous applications of Bayesian model selection in hydrology have focused on comparing models (or model parameters) but to our knowledge have not considered the question of individual *mechanism* identification. Our study builds on previous work on Bayesian model selection (e.g., Marshall et al., 2005; Vrugt & Robinson, 2007; Almeida et al., 2014; Schöniger et al., 2014; Wöhling et al., 2015; Prieto et al., 2019) and develops methods for the identification of dominant hydrological mechanisms by making hypotheses about the model mechanisms and their uncertainty. These advances are presented in Section 2 and represent the major contribution of this study.

The proposed mechanism identification method is general: it is derived for an ensemble of general probabilistic models without assuming a particular probabilistic model composition. For example, the case studies use a probabilistic model constructed from a combination of a deterministic model of hydrological processes and a residual error model for uncertainty characterization. This composition is typical in contemporary hydrological models. More general probabilistic (stochastic) models as well as probabilistic models constructed by forcing deterministic models with ensemble inputs could also be used, provided the probability density function of their outputs is known or can be approximated. As discussed next, the probabilistic model construction is less important than the coverage by the MHF of the space of mechanism hypotheses.

5.3.2. Toward Accounting for Uncertainty in the Identification of Dominant Hydrological Mechanisms

A key challenge in the hydrological community is to develop identification methods that perform reliably in the presence of incomplete and inexact information (i.e., uncertain streamflow observations, approximate model components and structures, limited coverage of hypothesis space, etc.).

Bayesian model selection, via posterior model probabilities (Hoeting et al., 1999), is well posed for process identification if a (quasi) true model is in the ensemble (Höge et al., 2019). Bayesian methods have been used for model selection, ranking, and elimination (Wöhling et al., 2015), as well as for model structure estimation (Bulygina & Gupta, 2011).

In this paper, Bayesian model inference is applied in the distinct (though related) context of identification of dominant hydrological mechanisms rather than complete models. The cornerstone of this method is the ability to estimate the posterior probabilities of individual mechanisms, which is based on the posterior probabilities of models containing these mechanisms. Hence, the mechanism identification method proposed in this work builds on previous applications of Bayesian inference to model selection in hydrology (Höge et al., 2019), including Bayes Factors (Marshall et al., 2005), and the use of information criteria including Occam Razor terms to approximate the BME term (Ye et al., 2008; Schöniger et al., 2014).

The inference of the dominant mechanisms can be blurred by interactions between mechanisms used for (multiple) model processes. Hypothesis testing is necessarily constrained by the Duhem–Quine thesis (Duhem, 1991), which highlights the difficulty or even impossibility of separating individual model hypotheses from their "surrounding" model environment (Nearing et al., 2016, 2020). Conceptually, we see mechanism interactions as similar to parameter interactions in traditional inference: for example, two "poor" mechanisms (parameter values) can compensate for each other's weaknesses and produce a model with similar or even higher posterior probability than a model with two "good" mechanisms (parameter values). Such interactions can be problematic, as seen from the derivation of the mechanism identification equations (Section 2.5). Scenario 3 presents some empirical evidence of mechanism interactions (see end of Section 4.1.4), where they manifest in a loss of power—though strong interactions may eventually manifest in loss of reliability. As individual model hypotheses are (gradually) improved, we expect to see corresponding improvement in the identification of dominant mechanisms. Interactions between model mechanisms can also be reduced by using observations of model outputs other than streamflow, for example, actual ET and groundwater levels. The benefits of using multivariate data for model identification and refinement are vividly seen from previous studies (e.g., Fenicia et al., 2008; Gupta et al., 2008; Wagener & Montanari, 2011; Euser et al., 2013).

In terms of requirements for the MHF, a key requirement is sufficiently wide coverage of the mechanism hypothesis space, including a diverse range of simple and complex candidate mechanisms, so that the resulting ensemble is more likely to include the true (or at least quasi-true) mechanisms. The presence of nested models and/or nested model mechanisms could lead to multiple models having the same likelihood function value, though the inclusion of Occam Razor terms in the computation of posterior model probabilities can help steer the inference toward the simpler representation (see Appendix A). Another important consideration is to avoid the potential trap of all models in the ensemble being wrong for the same reason (Clark et al., 2011a).

Finally, important simplifications undertaken in the case studies are the exclusion of parametric uncertainty within each model and the (related) use of the BIC to approximate the BME. These simplifications are motivated by computational considerations, with limitations and future work outlined next in Section 5.4.

5.4. Limitations and Future Work

Important directions for future research emerge in the application of the proposed mechanism identification method to more complex modeling scenarios, both in the context of synthetic tests and real data applications. This section lists the most salient opportunities.

The case studies in this work did not consider parametric uncertainty in the hydrological model. This assumption may be reasonable in many circumstances, for example, when the uncertainty associated with a (relatively) parsimonious hydrological model is represented using a simple residual error model, and a suitable long observational record is available for parameter inference (Kavetski, 2018; McInerney et al., 2018). However, there are many modeling situations where parametric uncertainty can be substantial, notably, in poorly identifiable models (e.g., Renard et al., 2010). An important and logical next step is to apply the mechanism identification method in a more complete Bayesian context that explicitly considers model parameter uncertainty, for example, using importance or Markov Chain Monte Carlo sampling.

The synthetic studies in this paper have used relatively simple error models to corrupt the synthetic "observed" data, and, importantly, the true mechanisms were included in the model space. But what happens if a true mechanism is not included in the ensemble? It is well known that many statistical model identification methods assume that the true model is present, which may not be realistic in practical hydrological

contexts. An important practical question is then how "good" should the "best available" mechanism/model representation be for it to be identified as "dominant" by the method?

Another important uncertainty-related consideration that remains largely unexplored in this study is the interaction between identified mechanisms (see Duhem, 1991; Nearing et al., 2016, 2020). In most catchments, specialized measurements to test each hypothesis/mechanism (e.g., "observations" of storage of water in the unsaturated and saturated zone, percolation, actual evapotranspiration, etc.) are not available, and hypothesis testing is necessarily limited to tests against observed streamflow. Given the dependence of streamflow on multiple and often interacting hydrological processes, model predictions can be expected to depend on multiple interdependent hypotheses (Pfister & Kirchner, 2017). This aspect of mechanism identification also requires substantial further investigation.

Furthermore, as flexible frameworks continue to improve and expand their coverage, so will their ability to meet the mechanism identification requirements. However, as more candidate mechanisms are hypothesized, it is likely that differences between them will become smaller. We anticipate a limit on the degree of mechanism detail that can be inferred from highly uncertain environmental data. But more research will be needed before meaningful quantitative statements can be made.

Future work will apply the method proposed in this paper to a more diverse range of catchments, as well as to a range of different time periods in the same catchment, analogous to split-sampling and cross-validation. These analyses can help elucidate the pattern, consistency, and variability of the identified dominant mechanisms and can help establish model performance when faced with data unseen during mechanism identification. Of particular interest are applications in experimental catchments, where more extensive fieldwork and multivariate data other than streamflow time series (e.g., storages of water) is available. Application of Bayesian mechanism identification in experimental catchments could extend previous studies where model comparison was done on the basis of existing perceptual knowledge (Carrer et al., 2019; Fenicia et al., 2014; Wrede et al., 2015). Such work could yield insights into correspondence of models and reality, as well as into the variability of hydrological mechanisms in space (across catchments, e.g., due to differences in topography, geology, and land use) and time (within catchments, e.g., across seasons).

Finally, the mechanism selection method will be extended to work with hydrological signatures, in order to enable the identification of dominant mechanisms in ungauged catchments (e.g., Sivapalan et al., 2003; Bulygina et al., 2012; Westerberg et al., 2016; McMillan et al., 2017; Prieto et al., 2019).

6. Conclusions

The development of hydrological models that provide an accurate representation of catchment dynamics and produce accurate streamflow predictions represents a formidable model identification challenge. In this work, we approach this model identification challenge with a focus on identifying individual model components (hydrological mechanisms) for each hydrological process that is included in the model.

The proposed hydrological mechanism identification method takes advantage of flexible hydrological models, Bayesian inference, and statistical hypothesis testing. A "dominant" mechanism is defined as a mechanism with (substantially) higher posterior probability than the sum of posterior probabilities of all other proposed mechanisms; here, we set the probability threshold at 0.75 (i.e., 3 times more a posteriori probable than any alternative mechanism). A test statistic is constructed for the null hypothesis that "none of the proposed mechanisms is dominant", and its estimated probability is compared against an a priori confidence level (here, 95%, corresponding to the classic significance level of 5%).

The method is evaluated empirically using a synthetic and real data case study based on daily data from the Leizarán catchment (Basque Country, Spain). The hydrological modeling system FUSE is employed to represent seven hydrological processes using 2–4 mechanisms per process, yielding a total of 624 feasible model hypotheses. Synthetic scenarios with 3 levels of error magnitude (low to high) and 50 data replicates each are used to establish the performance of the proposed method in the presence of data/model error. Metrics of statistical reliability and power are used to quantify and then rate the method performance. Real data are used to investigate the generality of key qualitative findings of the synthetic experiments.

The following conclusions are obtained:

1. Hydrological mechanism identification can be based on posterior probabilities of mechanisms estimated from observed rainfall-runoff data using a Bayesian approach, with a correction to account for the unbalanced frequency of mechanism occurrence in the multihypothesis-testing method. A key underlying assumption of the mechanism identification method is that, on average, models with highly probable mechanisms have a higher posterior probability than models with less-probable mechanisms. The uncertainty in the test statistic arising from incomplete coverage of the mechanism hypothesis space is approximated by bootstrapping the ensemble of model structures.
2. Empirical verification indicates that the mechanism identification method is statistically reliable: If the method identifies a mechanism as dominant, this mechanism is usually the true mechanism. Pooling the results across all processes and error levels, the method identifies the true mechanism in all synthetic trials. As expected, the statistical power of the test (ability to make a determination) decreases when data/model errors are high, with no mechanism being identified as dominant in 14% of all trials.
3. In the synthetic study, the following insights are obtained into process identification:
   a) The most identifiable processes, i.e., the processes for which dominant mechanisms are most identifiable, are those related to storage of water in the lower soil layer (saturated zone) and the movement of surface water (i.e., routing). Dominant mechanisms for these processes can be identified with perfect reliability and power even from streamflow corrupted with high (synthetic) errors.
   b) The least identifiable processes, i.e., the processes for which dominant mechanisms are least identifiable, are those related to the movement of water into the soil (interflow and percolation). In particular, the dominant mechanism for percolation is identified with power as low as 22% and 10% when streamflow errors are medium and high, respectively. In addition, the mechanisms representing percolation in the FUSE model are comparatively similar, making them harder to distinguish.
4. The real data study demonstrates how the proposed mechanism identification method is implemented in practice. Overall, the behavior in the real case study is comparable to the behavior in the synthetic case study. For example, a similar number and type of mechanisms were identified in both cases (e.g., the storage in the saturated zone is better approximated by a tension storage combined with two parallel tanks). This consistency provides a degree of confidence in the robustness of the findings as a proof-of-concept demonstration. Generally consistent mechanism identification is obtained using inference based on 6 and 12 years of data, as well as when the significance level is tightened from 10% to 1%.

More generally, this study contributes to broader community efforts on improving model identification and catchment understanding, by combining ideas from flexible models, Bayesian inference, and statistical hypothesis testing. Future research directions include a more complete treatment of uncertainty (in particular parameter uncertainty) in the context of hydrological mechanism identification, better understanding of the impact of the best models being "quasi-true" rather than "true," detection and where possible mitigation of mechanism interactions, as well as applications to wider sets of catchments and modeling contexts (including ungauged catchments).

Appendix A: Simplification of Posterior Probability of a Model Structure Using Maximum A Posteriori Estimation

This appendix briefly discusses the approximation of the Bayesian Model Evidence (BME) in the context of the mechanism identification method developed in this study.

For a given model structure, BME is defined by the integral in the denominator of Bayes equation, namely

$$p(\tilde{\mathbf{q}} \mid G^{(k)}) = \int_{\Omega^{(k)}} p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}^{(k)}, G^{(k)}) p(\boldsymbol{\theta}^{(k)} \mid G^{(k)}) \mathrm{d}\boldsymbol{\theta}^{(k)} \tag{A1}$$

Direct evaluation of this high-dimensional integral is a formidable computational task (Ye et al., 2008; Schöniger et al., 2014).

An attractive pragmatic alternative is to approximate Equation A1 using information from the most probable parameter set. Such approximations are appropriate when posterior parameter uncertainty is (relatively) low, which is often the case in hydrological modeling applications when the effects of model and data

uncertainty are represented using a residual error model and long observed data time series are used (e.g., Kuczera et al., 2006; Yang et al., 2007; Sun et al., 2017; Kavetski, 2018).

The most probable parameter set is defined as the parameter set that maximizes the posterior distribution of model parameters in Equation 5,

$$\hat{\boldsymbol{\theta}}^{(k)} = \arg\max_{\boldsymbol{\theta}^{(k)}} p(\boldsymbol{\theta}^{(k)} \mid \tilde{\mathbf{q}}, G^{(k)}) \tag{A2}$$

Equation A1 can then be approximated using several analytical approaches.

The BIC, also known as the Schwarz Information Criterion (SIC), is defined as

$$\mathrm{BIC}^{(k)} = \mathrm{BIC}(G^{(k)}, \tilde{\mathbf{q}}) = -2 \log p(\tilde{\mathbf{q}} \mid \hat{\boldsymbol{\theta}}^{(k)}, G^{(k)}) + N_{\boldsymbol{\theta}}^{(k)} \log N_t \tag{A3}$$

where $N_{\boldsymbol{\theta}}^{(k)}$ is the total number of model parameters in the probabilistic model $G^{(k)}$ (Ye et al., 2008; Schöniger et al., 2014). The last term in Equation A3 is a so-called "Occam Razor" term, which penalizes model complexity (here measured by the number of parameters). Occam Razor terms are essential to distinguish between models (and eventually mechanisms) that yield comparable predictive performance but differ in their degree of complexity.

The BIC approximation of the BME in Equation 7 is then

$$p(G^{(k)} \mid \tilde{\mathbf{q}}, \mathbf{G}) \approx \frac{\exp\left(-\frac{1}{2}\mathrm{BIC}^{(k)}\right) p(G^{(k)} \mid \mathbf{G})}{\sum_{i=1}^{N_G} exp\left(-\frac{1}{2}\mathrm{BIC}^{(i)}\right) p(G^{(i)} \mid \mathbf{G})} = \frac{\exp\left(-\frac{1}{2}\left(\mathrm{BIC}^{(k)} - \mathrm{BIC}_{\max}\right)\right) p(G^{(k)} \mid \mathbf{G})}{\sum_{i=1}^{N_G} \exp\left(-\frac{1}{2}\left(\mathrm{BIC}^{(i)} - \mathrm{BIC}_{\max}\right)\right) p(G^{(i)} \mid \mathbf{G})} \tag{A4}$$

The last term in Equation A4 uses the shift by $BIC_{\max} = \max_{i=1,\ldots,N_G} BIC^{(i)}$ to achieve a robust computational implementation that avoids numerical underflows or overflows.

Unless specific prior information is available, the prior over the hydrological models is set as uniform,

$$p(G^{(k)} \mid \mathbf{G}) = 1 / N_G \; \forall \, k \tag{A5}$$

where N_G is the total number of models under consideration.

The BIC is derived by applying the Laplace approximation to the integral in Equation A1 and only retaining the terms dependent on the data length N_t, i.e., assuming the number of observations is large (Schwarz, 1978; Konishi & Kitagawa, 2008). There are several alternative information criteria. For example, the Akaike Information Criterion (AIC) is derived from information theory and penalizes the number of parameters according to $2N_{\boldsymbol{\theta}^{(k)}}$ instead of $N_{\boldsymbol{\theta}^{(k)}} \log N_t$ (Akaike, 1974). The Kashyap Information Criterion (KIC) is derived by assuming the parameter posterior is Gaussian and has additional Occam Razor terms, including a term containing the determinant of the Fisher information matrix (Fisher & Russell, 1922; Kashyap, 1982).

All three information criteria, AIC, BIC, and KIC, converge to the integral in Equation A1 as the number of observations used in the analysis increases, by the virtue of the likelihood term eventually dominating all other terms. The KIC approach is theoretically more accurate; however, the approximation of the Fisher matrix (e.g., using finite differences) is noisy, and it is difficult to ensure comparable numerical accuracy across many (hundreds to thousands) model structures as required in the mechanism identification framework. For these reasons, the BIC approach offers a better balance of numerical accuracy and computational robustness in the specific context of this study.

It is also emphasized that the simplification in Equation A4 is separate from the derivation of the mechanism identification equations, which requires solely $p(G^{(k)} \mid \tilde{\mathbf{q}}, \mathbf{G})$. As such, the modeler is free to compute $p(G^{(k)} \mid \tilde{\mathbf{q}}, \mathbf{G})$ using methods/approximations suitable for their specific application.

Appendix B: Probabilistic Model and Its Inference (Likelihood and Prior) for a Single Hydrological Model Structure

A single probabilistic hydrological model in this work is given by Equations 2–4. It is constructed using a given FUSE configuration (model structure) as the deterministic model in Equation 2 in combination with the Box–Cox transformation (Box & Cox, 1964) in Equation 3,

$$z(q) = z(q;\lambda,A) = \begin{cases} \dfrac{(q+A)^{\lambda}-1}{\lambda} & \text{when } \lambda \neq 0 \\ \log(q+A) & \text{otherwise} \end{cases} \tag{B1}$$

The transformation parameters are fixed a priori: the power parameter $\lambda = 0.2$ and the offset parameter $A = 0.035$. To reduce clutter, these fixed parameters are omitted from the equations.

The likelihood function for this probabilistic model is given in Equation 6 and makes use of the Jacobian of the Box–Cox transformation, $z'(q) = z'\left(q;\lambda,A\right) = \left(q+A\right)^{\lambda-1}$.

In order to improve numerical robustness in the computation of posterior model probabilities via Equations A3 and A4, the likelihood function is computed directly in log-space,

$$p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}) = p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}_h,\sigma_\eta) = \prod_{t=1}^{N_t}(z'(\tilde{q}_t) \times f_{\mathcal{N}} z(\tilde{q}_t); z(q_t^{\boldsymbol{\theta}_h}),\sigma_\eta^2) \tag{B2}$$

$$\begin{aligned} \log p(\tilde{\mathbf{q}} \mid \boldsymbol{\theta}_h,\sigma_\eta,G) &= \log\left[\prod_{t=1}^{N_t}(\tilde{q}_t + A)^{\lambda-1} \times f_{\mathcal{N}}(z(\tilde{q}_t); z(q_t^{\hat{\boldsymbol{\theta}}_h}),\hat{\sigma}_\eta^2)\right] = \\ &= \sum_{t=1}^{N_t}\left[(\lambda-1)\log(\tilde{q}_t + A) + \log f_{\mathcal{N}}(z(\tilde{q}_t); z(q_t^{\hat{\boldsymbol{\theta}}_h}),\hat{\sigma}_\eta^2)\right] \end{aligned} \tag{B3}$$

and then used to compute the BIC as described in Appendix A.

The prior $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_h,\sigma_\eta)$ in Equation 5 is specified as uniform over the feasible parameter ranges (defined by min and max parameter bounds),

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_h,\sigma_\eta) = \prod_{i=1}^{N_{\boldsymbol{\theta}h}}(\boldsymbol{\theta}_{h,i}^{\max} - \boldsymbol{\theta}_{h,i}^{\min})^{-1} \times (\sigma_\eta^{\max} - \sigma_\eta^{\min})^{-1} \tag{B4}$$

where $N_{\boldsymbol{\theta}h}$ is the number of parameters in the deterministic hydrological model h.

Appendix C: Generation of Data for Synthetic Experiments

C.0. Generation of Synthetic "Exact" Model and Data

The synthetic "exact" hydro mechanisms, synthetic "exact" hydrological parameters $\breve{\boldsymbol{\theta}}_h$, and synthetic "exact" streamflow time series $\breve{\mathbf{q}}$ are generated as follows:

1. Calibrate all models within the FUSE ensemble to observed streamflow from the Leizarán catchment. A hybrid parameter estimation method similar to the LS-MoM approach of McInerney et al. (2018) is used (Section 3.3).
2. Set the synthetic "exact" model and mechanisms.

(a) The "exact" deterministic hydrological model, $\breve{h}$, is set to the best-performing FUSE structure (highest posterior density at the optimal parameter set).
(b) The "exact" hydro mechanisms, $\breve{\mathbf{m}}$, are set to the mechanisms that comprise the exact model structure $\breve{h}$.
(c) The "exact" parameters, $\breve{\boldsymbol{\theta}}_h$, are set to the estimated parameters of the exact model $\breve{h}$.

3. Generate "exact" streamflow $\breve{\mathbf{q}}$.
   A time series of 6 years (2,190 daily time steps) of synthetic "exact" daily streamflow $\breve{\mathbf{q}}$ is generated by running the exact deterministic model (with exact parameters) forced with the observed daily precipitation and potential evapotranspiration from the Leizarán catchment (see Section 3.1),

$$\breve{q}_t = \breve{h}_t(\breve{\boldsymbol{\theta}}_h; \tilde{\mathbf{x}}_{1:t}, \mathbf{s}_0) \tag{C1}$$

This procedure produces synthetic "exact" streamflow data that broadly resemble the real streamflow data from the Leizarán catchment.

C.1. Generation of Synthetic "Observed" Data in Scenarios 1–3

The synthetic replicates of "observed" streamflow for Scenarios 1–3 are generated using the following procedure.

In a given scenario, the magnitude of synthetic noise in the replicates is specified by σ_η.

1. Compute transformed "exact" streamflow $\breve{\boldsymbol{\gamma}} = z(\breve{\mathbf{q}})$.
2. Generate the ith replicate of "observed" streamflow, $\breve{\mathbf{q}}^{(r)}$:

(a) Sample the replicate in transformed space

$$\boldsymbol{\gamma}^{(r)} \leftarrow \mathcal{N}(\breve{\boldsymbol{\gamma}}, \sigma_\eta^2) \tag{C2}$$

(b) Back-transform to streamflow space

$$\breve{\mathbf{q}}^{(r)} = z^{-1}(\boldsymbol{\gamma}^{(r)}) \tag{C3}$$

3. Generate multiple replicates, $\{\breve{\mathbf{q}}^{(r)}; r = 1,...,N^{\text{rep}}\}$, where N^{rep} is the total number of replicates.

The error levels are specified as follows: Scenario 1 uses $\sigma_\eta = 0.025$, Scenario 2 uses $\sigma_\eta = 0.1$, and Scenario 3 uses $\sigma_\eta = 0.25$.

The Box–Cox transformation parameters in the generation of synthetic data are set the same values as in the assumed probability model, namely $\lambda = 0.2$ and $A = 0.035$.

A total of $N^{\text{rep}} = 50$ synthetic replicates are generated.

Acknowledgments
The authors from IHCantabria acknowledge the financial support from the Government of Cantabria through the FÉNIX Program (ID 2020.03.03.322B.742.09). Computations were performed on the Neptuno cluster at IHCantabria. We thank the Diputación Foral de Guipúzcoa for providing the observed hydrological data. We thank David del Prado and Gloria Zamora from IHCantabria's IT team for their assistance with the supercomputing facilities and Claudia Vitolo from ECWMF for her initial help with the study. The Fortran implementation of FUSE provided by Martyn Clark is employed in the case study. We are grateful to Fabrizio Fenicia, Andreas Scheidegger, Ben Renard, and Grey Nearing for insightful discussions on the topic of model inference and selection. We thank the Associate Editor Thorsten Wagener, Anneli Guthke, and the two anonymous reviewers for their insightful comments and constructive feedback.

Data Availability Statement

The data presented in this paper are deposited in https://doi.org/10.5281/zenodo.4744400.

References

Addor, N., & Melsen, L. A. (2019). Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resources Research*, *55*, 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*, 8792–8812. https://doi.org/10.1029/2018WR022606

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Almeida, S. (2014). *The value of regionalised information for hydrological modelling* (PhD thesis). London, UK: Imperial College London.

Arnell, N. W., & Gosling, S. N. (2014). The impacts of climate change on river flood risk at the global scale. *Climatic Change*, *134*(3), 1–15.

Arora, V. K. (2002). The use of the aridity index to assess climate change effect on annual runoff. *Journal of Hydrology*, *265*(1–4), 164–177.

Beck, H. E., De Roo, A., & van Dijk, A. I. J. M. (2015). Global maps of streamflow characteristics based on observations from several thousand catchments. *Journal of Hydrometeorology*, *16*(4), 1478–1501. https://doi.org/10.1175/JHM-D-14-0155.1

Beven, K. (1989). Changing ideas in hydrology—The case of physically-based models. *Journal of Hydrology*, *105*(1–2), 157–172.

Beven, K. (2010). *Environmental modelling: An uncertain future?*. Boca Raton, FL: CRC Press.

Beven, K. (2012). Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience*, *344*(2), 77–88. https://doi.org/10.1016/j.crte.2012.01.005

Beven, K. (2019). Towards a methodology for testing models as hypotheses in the inexact sciences, Proceedings of the Royal Society A: Mathematical. *Physical and Engineering Sciences*, *475*(2224), 20180862. https://doi.org/10.1098/rspa.2018.0862

Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, *4*(2), 203–213. https://doi.org/10.5194/hess-4-203-2000

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. *Wiley Interdisciplinary Reviews-Water*, *5*(3), e1278. https://doi.org/10.1002/wat2.1278

Blöschl, G. (2017). Debates—Hypothesis testing in hydrology: Introduction. *Water Resources Research*, *53*3, 1767–1769. https://doi.org/10.1002/2017WR020584

Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., & Zhang, Y. (2019). Twenty-three unsolved problems in hydrology (UPH)—A community perspective. *Hydrological Sciences Journal*, *64*(10), 1141–1158. https://doi.org/10.1080/02626667.2019.1620507

Box, G. E. P. (1979). All models are wrong, but some are useful. *Robustness in Statistics*, *202*, 549.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Bulygina, N., Ballard, C., McIntyre, N., O'Donnell, G., & Wheater, H. (2012). Integrating different types of information into hydrological model parameter estimation: Application to ungauged catchments and land use scenario analysis. *Water Resources Research*, *48*, W06519. https://doi.org/10.1029/2011WR011207

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., & Slooten, L. J. (2005). Inverse problem in hydrogeology. *Hydrogeology Journal*, *13*(1), 206–222.

Carrer, G. E., Klaus, J., & Pfister, L. (2019). Assessing the catchment storage function through a dual-storage concept. *Water Resources Research*, *55*, 476–494. https://doi.org/10.1029/2018WR022856

Chamberlin, T. C. (1965). The method of multiple working hypotheses: With this method the dangers of parental affection for a favorite theory can be circumvented. *Science*, *148*(3671), 754–759. https://doi.org/10.1126/science.148.3671.754

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

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

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

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

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrological Processes*, *28*(25), 6135–6150. https://doi.org/10.1002/hyp.10096

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., et al. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling & Software*, *129*, 104728. https://doi.org/10.1016/j.envsoft.2020.104728

Dekking, F. M. (2005). *A modern introduction to probability and statistics: Understanding why and how*. New York: Springer.

Duhem, P. M. M. (1991). *The aim and structure of physical theory* (Vol. 13). Princeton, NJ: Princeton University Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY/London: Chapman & Hall.

Ehret, U., Gupta, H. V., Sivapalan, M., Weijs, S. V., Schymanski, S. J., Blöschl, G., & Bogaard, T. (2014). Advancing catchment hydrology to deal with predictions under change. *Hydrology and Earth System Sciences*, *18*(2), 649–671.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence* (Vol. 17, No. 1, pp. 973–978). Mahwah, NJ: Lawrence Erlbaum Associates Ltd.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

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., Reichert, P., & Albert, C. (2018). Signature-domain calibration of hydrological models using approximate Bayesian computation: Empirical analysis of fundamental properties. *Water Resources Research*, *54*, 3958–3987. https://doi.org/10.1002/2017WR021616

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*, W11510. https://doi.org/10.1029/2010WR010174

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., & Freer, J. (2014). Catchment properties, function, and conceptual model representation: Is there a correspondence? *Hydrological Processes*, *28*(4), 2451–2467. https://doi.org/10.1002/hyp.9726

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

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*, W06419. https://doi.org/10.1029/2007WR006386

Fisher, R. A., & Russell, E. J. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *222*(594–604), 309–368. https://doi.org/10.1098/rsta.1922.0009

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*, W08301. https://doi.org/10.1029/2011WR011044

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

Gupta, V. K., & Sorooshian, S. (1983). Uniqueness and observability of conceptual rainfall-runoff model parameters: The percolation process examined. *Water Resources Research*, *19*(1), 269–276. https://doi.org/10.1029/WR019i001p00269

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al. (2013). Global flood risk under climate change. *Nature Climate Change*, *3*(9), 816–821.
Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802. https://doi.org/10.1093/biomet/75.4.800
Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401. https://doi.org/10.1214/ss/1009212519
Höge, M., Guthke, A., & Nowak, W. (2019). The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, *572*, 96–107. https://doi.org/10.1016/j.jhydrol.2019.01.072
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
Hsu, K. L., Moradkhani, H., & Sorooshian, S. (2009). A sequential Bayesian approach for hydrologic model selection and prediction. *Water Resources Research*, *45*, W00B12. https://doi.org/10.1029/2008WR006824
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
Jeffreys, H. (1998). *The theory of probability*. Oxford: Oxford University Press.
Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *4*(2), 99–104. https://doi.org/10.1109/tpami.1982.4767213
Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572
Kavetski, D. (2018). Parameter estimation and predictive uncertainty quantification in hydrological modelling. In Q. Duan, et al. (Eds.), *Handbook of hydrometeorological ensemble forecasting* (chap. 25-1). Berlin, Germany: Springer-Verlag. https://doi.org/10.1007/978-3-642-40457-3_25-1
Kavetski, D., & Clark, M. P. (2010). Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. *Water Resources Research*, *46*, W10511. https://doi.org/10.1029/2009WR008896
Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*, W03S04. https://doi.org/10.1029/2005WR004362
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
Konapala, G., Kao, S.-C., & Addor, N. (2020). Exploring hydrologic model process connectivity at the continental scale through an information theory approach. *Water Resources Research*, *56*, e2020WR027340. https://doi.org/10.1029/2020WR027340
Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. New York: Springer Science & Business Media.
Kraft, P., Vaché, K. B., Frede, H.-G., & Breuer, L. (2011). CMF: A hydrological programming language extension for integrated catchment models. *Environmental Modelling & Software*, *26*(6), 828–830. https://doi.org/10.1016/j.envsoft.2010.12.009
Krueger, T., Freer, J., Quinton, J. N., Macleod, C. J. A., Bilotta, G. S., Brazier, R. E., & Haygarth, P. M. (2010). Ensemble evaluation of hydrological model hypotheses. *Water Resources Research*, *46*, W07516. https://doi.org/10.1029/2009WR007845
Kuczera, G., Kavetski, D., Franks, S., & Thyer, M. (2006). Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, *331*(1), 161–177. https://doi.org/10.1016/j.jhydrol.2006.05.010
Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York: Springer.
Li, C.-z., Wang, H., Liu, J., Yan, D.-h., Yu, F.-l., & Zhang, L. (2010). Effect of calibration data series length on performance and optimal parameters of hydrological model. *Water Science and Engineering*, *3*(4), 378–393. https://doi.org/10.3882/j.issn.1674-2370.2010.04.002
Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1), 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3
Lumley, T. (2000). In A. Stuart K. Ord & S. Arnold(Eds.), *Kendall's advanced theory of statistics. Volume 2A: Classical inference and the linear model*. London: Arnold.
Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. *Water Resources Research*, *41*, W10422. https://doi.org/10.1029/2004WR003719
McDonnell, J. J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., & Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*, W07301. https://doi.org/10.1029/2006WR005467
McInerney, D., Thyer, M., Kavetski, D., Bennett, B., Lerat, J., Gibbs, M., & Kuczera, G. (2018). A simplified approach to produce probabilistic hydrological model predictions. *Environmental Modelling & Software*, *109*, 306–314. https://doi.org/10.1016/j.envsoft.2018.07.001
McMillan, H., Westerberg, I., & Branger, F. (2017). Five guidelines for selecting hydrological signatures. *Hydrological Processes*, *31*(26), 4757–4761. https://doi.org/10.1002/hyp.11300
Montanari, A., Young, G., Savenije, H., Hughes, D., Wagener, T., Ren, L., & Grimaldi, S. (2013). "Panta Rhei—Everything flows": Change in hydrology and society—The IAHS scientific decade 2013–2022. *Hydrological Sciences Journal*, *58*(6), 1256–1275.
Moore, R. J. (2007). The PDM rainfall-runoff model. *Hydrology and Earth System Sciences*, *11*(1), 483–499. https://doi.org/10.5194/hess-11-483-2007
Moore, R. J., & Clarke, R. T. (1981). A distribution function approach to rainfall runoff modeling. *Water Resources Research*, *17*(5), 1367–1382. https://doi.org/10.1029/WR017i005p01367
Nearing, G. S., & Gupta, H. V. (2015). The quantity and quality of information in hydrologic models. *Water Resources Research*, *51*, 524–538. https://doi.org/10.1002/2014WR015895
Nearing, G. S., & Gupta, H. V. (2018). Ensembles vs. information theory: Supporting science under uncertainty. *Frontiers of Earth Science*, *12*(4), 653–660. https://doi.org/10.1007/s11707-018-0709-9
Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & Gupta, H. V. (2020). Does information theory provide a new paradigm for Earth Science? Hypothesis testing. *Water Resources Research*, *56*, e2019WR024918. https://doi.org/10.1029/2019WR024918
Nearing, G. S., Tian, Y. D., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal/Journal Des Sciences Hydrologiques*, *61*(9), 1666–1678. https://doi.org/10.1080/02626667.2016.1183009
Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika*, *20A*, 175–240. https://doi.org/10.1093/biomet/20A.1-2.175
Perrin, C., Michel, C., & Andreassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1–4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Pfister, L., & Kirchner, J. W. (2017). Debates—Hypothesis testing in hydrology: Theory and practice. *Water Resources Research*, *53*, 1792–1798. https://doi.org/10.1002/2016WR020116

Press, W. H., Fannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in Fortran 77: The art of scientific computing* (2nd ed.). New York: Cambridge University Press.

Prieto, C., Le Vine, N., Kavetski, D., Garcia, E., & Medina, R. (2019). Flow prediction in ungauged catchments using probabilistic Random Forests regionalization and new statistical adequacy tests. *Water Resources Research*, *55*, 4364–4392. https://doi.org/10.1029/2018WR023254

Prieto, C., Patel, D., & Han, D. (2020). Preface: Advances in flood risk assessment and management. *Natural Hazards and Earth System Sciences*, *20*(4), 1045–1048. https://doi.org/10.5194/nhess-20-1045-2020

Qin, Y., Kavetski, D., & Kuczera, G. (2018). A robust Gauss–Newton algorithm for the optimization of hydrological models: Benchmarking against industry-standard algorithms. *Water Resources Research*, *54*, 9637–9654. https://doi.org/10.1029/2017WR022489

Raftery, A. E. (1993). *Bayesian model selection in structural equation models* (Vol. 154, p. 163). Sage Focus Editions.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

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

Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, *14*(1), 21–41. https://doi.org/10.1162/089976602753284446

Schöniger, A., Wöhling, T., & Nowak, W. (2015). A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. *Water Resources Research*, *51*, 7524–7546. https://doi.org/10.1002/2015WR016918

Schöniger, A., Wöhling, T., Samaniego, L., & Nowak, W. (2014). Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research*, *50*, 9484–9513. https://doi.org/10.1002/2014WR016062

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Seibert, J., & Vis, M. J. P. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, *16*(9), 3315–3325. https://doi.org/10.5194/hess-16-3315-2012

Silva, O., Carrera, J., Dentz, M., Kumar, S., Alcolea, A., & Willmann, M. (2009). A general real-time formulation for multi-rate mass transfer problems. *Hydrology and Earth System Sciences*, *13*(8), 1399–1411. https://doi.org/10.5194/hess-13-1399-2009

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., et al. (2003). IAHS decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421

Smith, R. J., & Bryant, R. G. (1975). Metal substitutions in carbonic anhydrase: A halide ion probe study. *Biochemical and Biophysical Research Communications*, *66*(4), 1281–1286. https://doi.org/10.1016/0006-291X(75)90498-2

Sorooshian, S., Gupta, V. K., & Fulton, J. L. (1983). Evaluation of Maximum Likelihood Parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility. *Water Resources Research*, *19*(1), 251–259. https://doi.org/10.1029/WR019i001p00251

Srinivasan, V., Sanderson, M., Garcia, M., Konar, M., Blöschl, G., & Sivapalan, M. (2017). Prediction in a socio-hydrological world. *Hydrological Sciences Journal*, *62*(3), 338–345.

Sun, W. C., Wang, Y. Y., Wang, G. Q., Cui, X. Q., Yu, J. S., Zuo, D. P., & Xu, Z. X. (2017). Physically based distributed hydrological model calibration based on a short period of streamflow data: Case studies in four Chinese basins. *Hydrology and Earth System Sciences*, *21*(1), 251–265. https://doi.org/10.5194/hess-21-251-2017

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Triola, M. (2001). *Elementary statistics* (8th ed., p. 388). Boston: Addison-Wesley.

Van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C. M., Fenicia, F., Kavetski, D., & Lobligeois, F. (2013). The influence of conceptual model structure on model performance: A comparative study for 237 French catchments. *Hydrology and Earth System Sciences*, *17*(10), 4227–4239. https://doi.org/10.5194/hess-17-4227-2013

Varian, H. (2005). Bootstrap tutorial. *The Mathematica Journal*, *9*, 768–775.

Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, *43*, W01411. https://doi.org/10.1029/2005WR004838

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., & Sorooshian, S. (2001). A framework for development and application of hydrological models. *Hydrology and Earth System Sciences*, *5*(1), 13–26. https://doi.org/10.5194/hess-5-13-2001

Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research*, *47*, W06301. https://doi.org/10.1029/2010WR009469

Wagener, T., Sivapalan, M., Troch, P. A., McGlynn, B. L., Harman, C. J., Gupta, H. V., & Wilson, J. S. (2010). The future of hydrology: An evolving science for a changing world. *Water Resources Research*, *46*, W05301. https://doi.org/10.1029/2009WR008906

Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment classification and hydrologic similarity. *Geography Compass*, *1*(4), 901–931. https://doi.org/10.1111/j.1749-8198.2007.00039.x

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., & Freer, J. (2016). Uncertainty in hydrological signatures for gauged and ungauged catchments. *Water Resources Research*, *52*, 1847–1865. https://doi.org/10.1002/2015WR017635

Wheater, H., Jakeman, A., & Beven, K. (1993). Progress and directions in rainfall-runoff modeling. In A. J. Jakeman, M. B. Beck, & M. J. McAleer (Eds.), *Modeling change in environmental systems* (pp. 101–132). John Wiley & Sons.

Wöhling, T., Schöniger, A., Gayler, S., & Nowak, W. (2015). Bayesian model averaging to explore the worth of data for soil–plant model selection and prediction. *Water Resources Research*, *51*, 2825–2846. https://doi.org/10.1002/2014WR016292

Wrede, S., Fenicia, F., Martínez-Carreras, N., Juilleret, J., Hissler, C., Krein, A., & Pfister, L. (2015). Towards more systematic perceptual model development: A case study using 3 Luxembourgish catchments. *Hydrological Processes*, *29*(12), 2731–2750. https://doi.org/10.1002/hyp.10393

Yang, J., Reichert, P., & Abbaspour, K. C. (2007). Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur River basin (Switzerland). *Water Resources Research*, *43*, W10401. https://doi.org/10.1029/2006WR005497

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1996). Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *Journal of Hydrology*, *181*(1), 23–48. https://doi.org/10.1016/0022-1694(95)02918-4

Ye, M., Meyer, P. D., & Neuman, S. P. (2008). On model selection criteria in multimodel analysis. *Water Resources Research*, *44*, W03428. https://doi.org/10.1029/2008WR006803