AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR030705

An Exploration of Bayesian Identification of Dominant Hydrological Mechanisms in Ungauged Catchments

Cristina Prieto[1,2,3], Nataliya Le Vine[3,4], Dmitri Kavetski[5], Fabrizio Fenicia[2], Andreas Scheidegger[2], and Claudia Vitolo[3,6]

[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]Swiss Re, Armonk, NY, USA, [5]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, SA, Australia, [6]European Centre for Medium-range Weather Forecasts (ECMWF), Reading, UK

Key Points:

- Dominant mechanisms for process representation in ungauged catchments can be identified from regionalized flow indices via a Bayesian method
- The reliability of identification is impacted by the limited and uncertain information available in ungauged catchments
- With real data, the most identifiable process is routing; the least identifiable processes are percolation and unsaturated zone processes

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. Prieto,
prietoc@unican.es

Citation:

Prieto, C., Le Vine, N., Kavetski, D., Fenicia, F., Scheidegger, A., & Vitolo, C. (2022). An exploration of Bayesian identification of dominant hydrological mechanisms in ungauged catchments. *Water Resources Research*, *58*, e2021WR030705. https://doi.org/10.1029/2021WR030705

Received 28 JUN 2021
Accepted 1 MAR 2022

Abstract Hydrological modeling of ungauged catchments, which lack observed streamflow data, is an important practical goal in hydrological sciences. A major challenge is to identify a model structure that reflects the hydrological processes relevant to the catchment of interest. This study contributes a Bayesian framework for identifying individual model mechanisms (process representations) from flow indices regionalized to the catchment of interest. We extend a method previously introduced for mechanism identification in gauged basins, by formulating the inference equations in the space of (regionalized) flow indices and by accounting for posterior parameter uncertainty. A flexible hydrological model is used to generate candidate mechanisms and model structures, followed by statistical hypothesis testing to identify "dominant" (more a posterior probable) model mechanisms. The proposed method is illustrated using real data and synthetic experiments based on 92 catchments from northern Spain, from which 16 catchments are treated as ungauged. 624 hydrological model structures from the flexible framework FUSE are employed. In real data experiments, the method identifies a dominant mechanism in 27% of 112 trials (processes and catchments). The most identifiable process is routing, whereas the least identifiable processes are percolation and unsaturated zone processes. In synthetic experiments, where "true" mechanisms are known, the reliability of method varies from 60% to 95% depending on the combined regionalization and hydrological error; the probability of making an identification remains stable at around 25%. More broadly, the study contributes perspectives on hydrological mechanism identification under data-scarce conditions; limitations and opportunities for improvement are outlined.

1. Introduction

Hydrological modeling and streamflow prediction in ungauged catchments is a challenging but practically important branch of hydrological science. The defining feature of ungauged catchments is the lack of observed (measured) streamflow data, which poses a stark challenge for modeling endeavors expressly intended to simulate and predict streamflow. The lack of observed streamflow also complicates the development of hydrological models (structures and parameters) that provide a suitable representation of internal catchment processes.

Given that the majority of catchments worldwide are ungauged (Goswami et al., 2007; Sivapalan et al., 2003), model identification and application in these catchments has been a notable challenge of the hydrological community for several decades (e.g., Almeida et al., 2016; Hrachowitz et al., 2013; Kratzert et al., 2019; Prieto et al., 2019; Sivapalan, 2003). The present study focuses on the problem of model identification in ungauged catchments from the perspective of identifying appropriate process representations from a range of competing alternatives, combining and extending several recent modeling advances.

Previous research with respect to model identification in *gauged* catchments has suggested that hydrological model structures tend to be catchment specific (e.g., Beven, 2000; Clark et al., 2008; Coxon et al., 2014; Perrin et al., 2001). This perception motivated the hydrological community to move from seeking a single "one-size-fits-all" (Fenicia et al., 2008; McDonnell, 2003) hydrological model to finding hydrological models that represent "uniqueness-of-the-place" (Addor & Melsen, 2019; Beven & Lane, 2019; Craig et al., 2020; Knoben et al., 2020). Flexible modeling frameworks allow to compare multiple working hypotheses describing catchment processes and their mechanisms in a controlled and systematic way (e.g., FUSE [Clark et al., 2008]; 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]).

Formal approaches for model identification include Bayesian Model Selection (e.g., Höge et al., 2019; Marshall et al., 2005; Prieto et al., 2021; Schöniger et al., 2014; Wöhling et al., 2015; Ye et al., 2008), information-theoretic approaches (Nearing et al., 2020), and optimization approaches (Spieler et al., 2020).

This study approaches model identification from the Bayesian perspective, which is widely used in statistical hypothesis testing, model identification, and uncertainty quantification (Kass & Raftery, 1995; Marshall et al., 2005; Prieto et al., 2021; Raftery, 1995; Schöniger et al., 2015; Vrugt & Robinson, 2007; Wöhling et al., 2015). In Bayesian Model Selection, model probabilities are determined from the Bayesian Model Evidence (BME) term, which is defined as the integral of the model likelihood over the parameter space. Direct evaluation of the BME is usually difficult or impossible, and in practice it is often approximated via information criteria or (Monte Carlo) numerical integration (e.g., see Schöniger et al., 2015). The treatment of posterior parameter uncertainty within the BME is expected to become particularly important in ungauged basins, where comparatively less information is available for parameter estimation and model identification.

In gauged catchments, model identification is typically based on calibration to streamflow time series (Knoben et al., 2020; Lane et al., 2019; Spieler et al., 2020). Alternatively, streamflow time series can be replaced by flow indices, also referred to as "signatures" or "hydrological indices" (Gupta et al., 2008). When carefully chosen, these indices can encapsulate a substantial quantity of information from the streamflow time series and help guide model identification (e.g., Clark et al., 2011; Coxon et al., 2014). Typical flow indices include, for example, average and monthly flows, runoff coefficients, quantiles and slope of flow duration curves, baseflow index, etc. However, calibration to flow indices is challenged by the corresponding information loss (Fenicia et al., 2018; McMillan et al., 2017; Westerberg et al., 2016), which in some cases can be mitigated by using a large number of indices, especially in multiobjective approaches (e.g., Shafii & Tolson, 2015).

In ungauged catchments, streamflow time series are not available. One approach to overcome this limitation is to calibrate the hydrological models to "regionalized" flow indices (Wagener & Montanari, 2011). These regionalization models are typically constructed by establishing approximate relationships between flow indices (e.g., mean annual flow) and catchment descriptors (e.g., climate, topography, geology) from comparable or nearby gauged "donor" catchments, which are then extrapolated to ungauged "target" catchments of interest. Such models are usually implemented using regression models—traditionally, linear regression (e.g., Almeida et al., 2016; Yadav et al., 2007; Zhang et al., 2008)—and more recently machine learning techniques such as random forests (RF; e.g., Addor et al., 2018; Prieto et al., 2019; Snelder et al., 2013). In order to identify the most informative subsets of indices and to minimize redundancy, these relationships have also been formulated in principal component (PC) space (e.g. Olden & Poff, 2003; Peñas et al., 2014; Prieto et al., 2019).

Compared to model calibration on observed time series, calibration on regionalized flow indices poses additional challenges, as these indices, besides being potentially less informative than the full time series, are subject to substantial uncertainty due to their extrapolation from the donor catchments (Westerberg et al., 2016). Thus, multiple studies have incorporated flow index regionalization into a Bayesian framework, which allowed to combine the information from multiple indices and to incorporate rigorous uncertainty quantification (e.g., Almeida et al., 2016; Bulygina et al., 2009, 2011; Prieto et al., 2019; Westerberg et al., 2016).

Previous studies on identifying models best suited to simulate specific ungauged catchments used the "fixed model" approach ("one-size-fits-all"). These studies did not explore the identification of individual model components and did not allow for catchment specific models. In this article, we focus on hydrological model identification in ungauged catchments from the perspective of identifying "process mechanisms" rather than complete models. In this context, a "hydrological process" is the physical phenomenon occurring in a catchment, for example, surface runoff generation, and each hydrological process is approximated via a hydrological mechanism. Therefore, a "hydrological mechanism" is the set of equations intended to describe such process.

Recently, Prieto et al. (2021) proposed a method to identify hydrological mechanisms in gauged catchments using a statistical hypothesis-testing perspective. Their framework combines: (a) Bayesian estimation of posterior probabilities of individual mechanisms from a given ensemble of model structures; (b) a test statistic that defines a "dominant" mechanism as a mechanism more probable than all its alternatives given observed data; and (c) a

flexible modeling framework to generate model structures using combinations of available mechanisms. In that work, the BME was approximated using the Bayesian Information Criterion, which assumes that the observed data (streamflow time series) is of sufficiently long duration that the posterior parameter uncertainty is small.

In this article, we advance model identification in ungauged catchments by extending the hypothesis testing method proposed by Prieto et al. (2021) to use regionalized flow indices as the inference data and to account for posterior parameter uncertainty.

The study aims are:

1. Develop a Bayesian method for identification of dominant mechanisms in ungauged catchments, using a multiple working-hypothesis approach (modular flexible models) and regionalized flow index PCs
2. Assess the method empirically using real data from multiple catchments, as well as using synthetic experiments where the true mechanisms are known
3. Gain insights into which hydrological model processes are most and least identifiable in the case study catchments

The case study is based on 92 catchments in northern Spain, from which 16 are treated as "ungauged" (target). Experiments using both real and synthetic data are employed to gain insights into how the proposed method is impacted by errors in the hydrological model and/or regionalization model.

The article is organized as follows. Section 2 presents theoretical developments, Section 3 describes the case study setup, and Section 4 presents the case study results, which are then discussed in Section 5. Finally, Section 6 summarizes the key conclusions.

2. Theoretical Development

This section presents the theoretical basis of the mechanism identification framework for ungauged catchments. Section 2.1 provides the basic definitions of processes, models, and mechanisms. Sections 2.2 and 2.3 formulate, respectively, the hydrological and regionalization models. Section 2.4 describes the methods for estimating the posterior probabilities of mechanisms. Section 2.5 details the hypotheses testing setup and methods.

The key steps of the mechanism identification procedure are schematized in Figure 1. Detailed descriptions are provided in the following sections.

2.1. Terminology and General Overview

The modeling concepts underpinning the model identification framework used in this study are defined as follows (see Prieto et al., 2021 for additional details):

1. A hydrological process is a physical phenomenon occurring in a catchment, for example, surface runoff generation
2. A hydrological model process $\wp$ is a single hydrological process intended to be represented by a hydrological model
3. A hydrological model mechanism, $m^{\wp}$ is the set of equations (and associated parameters) to represent a process $\wp$
4. A hydrological model structure is the combination of mechanisms to represent $N^{\wp}$ preselected hydrological processes
5. An ensemble of hydrological model structures is the set of hydrological models that differ in their processes and/or mechanisms. In this work, the ensemble is generated using FUSE (Clark et al., 2008). FUSE (Clark et al., 2008) is a modular hydrological model that can be used for hypothesis testing

Given these definitions, a mechanism $m^{\wp}$ is here considered "dominant" if it is "substantially" more likely (a posteriori probable) to represent a particular process $\wp$ than all alternative mechanisms under consideration. Note that a "dominant mechanism" should be not confused with a "dominant process", which is typically intended as a process that contributes substantially to the overall catchment water balance.

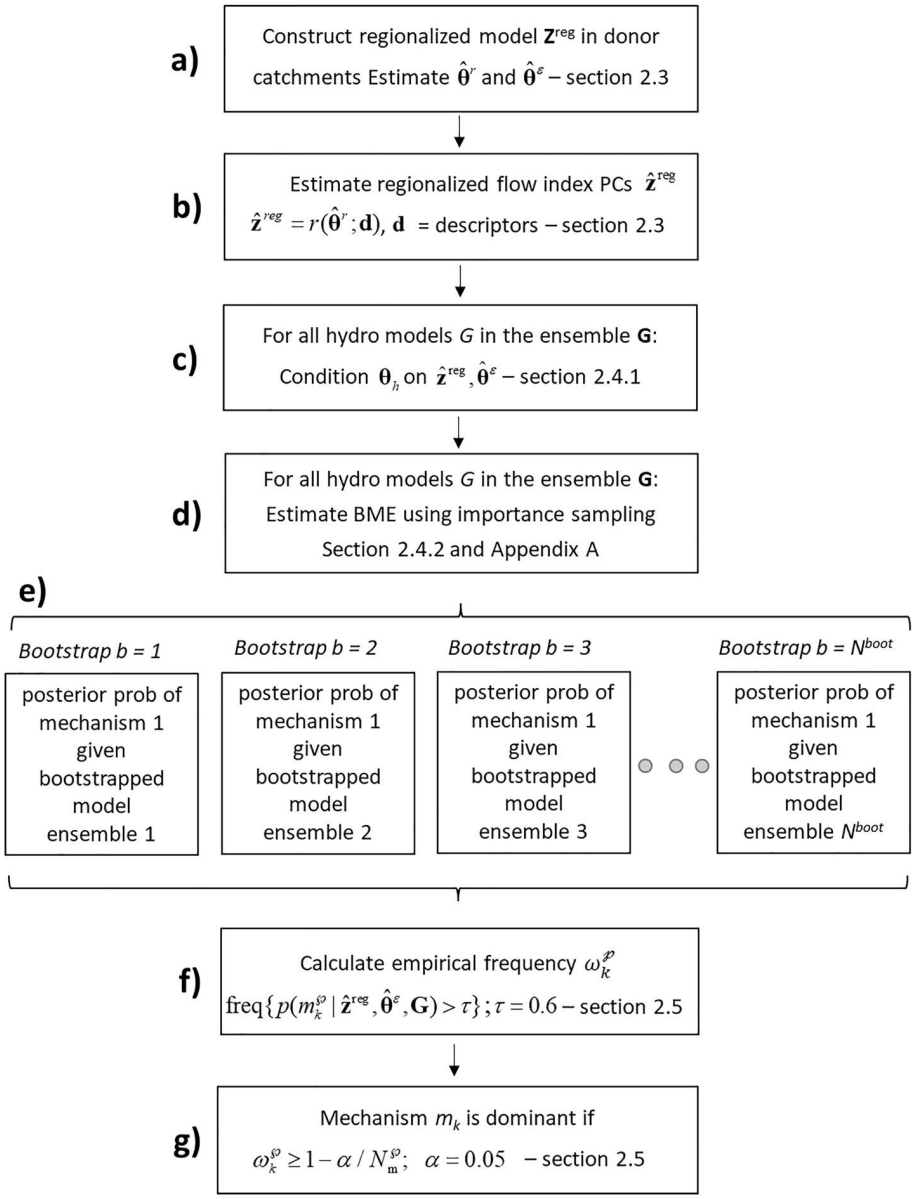

Figure 1. Flowchart of the mechanism identification method for ungauged catchments.

2.2. Hydrological Model

Consider a deterministic hydrological model h that estimates streamflow $q_t^{\theta_h}$ at time t,

$$q_t^{\theta_h} = h(\boldsymbol{\theta}_h; \mathbf{x}_{1:t}, \mathbf{s}_{0(h)}) \tag{1}$$

where $\boldsymbol{\theta}_h$ are the hydrological model parameters, $\mathbf{x}_{1:t}$ are the forcing data, $\mathbf{s}_{0(h)}$ are the initial conditions, and let $\boldsymbol{\Omega}$ denote the domain of model parameters.

In an ungauged catchment, hydrological model identification and parameter inference must proceed from flow indices rather than from streamflow time series. Moreover, the flow indices are themselves estimated from catchment attributes rather than computed from streamflow data, as described next.

2.3. Probabilistic Regionalization Model

Let $\tilde{\mathbf{z}} = \{\tilde{z}_i; i = 1, ..., N_z\}$ denote the PCs of observation-based flow indices in a given catchment. Note that the flow index PCs are dimensionless because all flow indices are normalized (centered and scaled) before the Principal Component Analysis (PCA) is applied. These flow index PCs are derived as follows: (a) compute the observation-based flow indices $\tilde{\boldsymbol{\omega}} = \{\tilde{\omega}_i; i = 1, ..., N_\omega\}$ from observed streamflow time series $\tilde{\mathbf{q}} = \{\tilde{q}_t; t = 1, ..., N_t\}$, $\tilde{\boldsymbol{\omega}} = \mathbf{f}_\omega(\tilde{\mathbf{q}})$; (b) "compress" $\tilde{\boldsymbol{\omega}}$ into $\tilde{\mathbf{z}}$ using a rotation matrix $\boldsymbol{\varsigma}$, $\tilde{\mathbf{z}} = \boldsymbol{\varsigma}\tilde{\omega}$; note that $N_z < N_\omega$. The rotation matrix is constructed a priori from the set of observation-based flow indices $\left\{\tilde{\boldsymbol{\omega}}^{\text{donor}(i)}; i = 1, ..., N_{\text{donor}}\right\}$ in the gauged donor catchments, using PCA and the number of flow index PCs to be retained (N_z) is determined prior to regionalization using the Broken Stick criterion (Prieto et al., 2019). In a practical context, $\tilde{\mathbf{z}}$ is available only in the donor catchments, but in a case study experiment where gauged catchments are treated as ungauged, the procedure above can also be used to calculate $\tilde{\mathbf{z}}$ in a target catchment as part of verification and analysis of results. Note that we use the "tilde" to indicate quantities computed from "observed" data. For example, $\tilde{\boldsymbol{\omega}}$ refers to flow indices computed from observed streamflow, $\tilde{\mathbf{z}}$ refers to flow index PCs computed from observed streamflow, and so on.

The probabilistic regionalization model of the observation-based flow index PCs, $\tilde{\mathbf{z}}$ is constructed by combining a deterministic regionalization model with a random error term (Prieto et al., 2019),

$$\mathbf{Z}^{\text{reg}} = R\left(\theta^R; \mathbf{d}\right) = \mathbf{z}^{\text{reg}} + \varepsilon\left(\theta^\varepsilon\right) = \mathbf{r}\left(\theta^r; \mathbf{d}\right) + \varepsilon(\theta^\varepsilon) \tag{2}$$

where $\mathbf{z}^{\text{reg}}$ refers to the deterministic estimate of the "regionalized" flow index PCs in a given catchment. This estimate is obtained using the deterministic regionalization model $\mathbf{r}\left(\theta^r; \mathbf{d}\right)$, which relates the flow index PCs to the catchment attributes $\mathbf{d}$. The parameters of this model are denoted by θ^r. The residual error model $\boldsymbol{\varepsilon}\left(\theta^\varepsilon\right)$, with parameters θ^ε, is used to characterize the uncertainty in the deterministic model $\mathbf{r}$. The complete set of parameters of the probabilistic regionalization model is denoted by $\theta^R = \left\{\theta^r, \theta^\varepsilon\right\}$. The dimension of $\mathbf{z}$, and hence of $\mathbf{Z}^{\text{reg}}$, $\mathbf{z}^{\text{reg}}$ and $\boldsymbol{\varepsilon}(\theta^\varepsilon)$, are all equal to N_z.

The probabilistic regionalization model R is calibrated using observed flow index PCs $\tilde{\mathbf{z}}^{\text{donor}}$ and catchment descriptors $\mathbf{d}^{\text{donor}}$ from the donor catchments (see Peñas et al., 2014 and Prieto et al., 2019 Section 3.2). Rather than having a fixed set of gauged and ungauged catchments, the flow index PCs in a given ungauged catchment are estimated from all remaining catchments. Since we treat 16 catchments as ungauged (one catchment at a time), we have 16 regionalization models. This is the same approach as employed in Almeida et al. (2016) and Prieto et al. (2019). Following previous work, the deterministic model r can be constructed using RF regression, and the residual error model can be constructed using a jack-knife approach with subsequent fitting of a parametric distribution with parameters θ^ε (Prieto et al., 2019). This stage of the analysis is illustrated schematically in Figure 1 row a, and detailed further in Section 3.

The regionalized flow index PCs in the ungauged target catchments, $\hat{\mathbf{z}}^{\text{reg}}$, are obtained as follows,

$\hat{\mathbf{z}}^{\text{reg}} = r\left(\hat{\theta}^r; \mathbf{d}^{\text{target}}\right)$ where $\hat{\theta}^r$ are the estimated parameters of the deterministic regionalization model and $\mathbf{d}^{\text{target}}$ are the catchment descriptors in the target catchment (see Peñas et al., 2014; Prieto et al., 2019). This stage of the analysis is depicted in Figure 1 row b.

2.4. Bayesian Inference of Mechanisms Given Regionalized Flow Index PCs

2.4.1. Parameter Inference for a Hydrological Model Structure Given Regionalized Flow Index PCs

We now consider the inference of hydrological model parameters in a given hydrological model structure from regionalized flow index PCs $\hat{\mathbf{z}}^{\text{reg}}$ estimated in Section 2.3. It is assumed that observed rainfall and other hydrological model forcings $\tilde{\mathbf{x}}$, are available in the ungauged catchments.

We formulate a probabilistic model for flow indices $\mathbf{Z}$ based on the (deterministic) hydrological model h and a (random) residual error term $\boldsymbol{\eta}$ representing the combined errors of the hydrological and regionalization models,

$$\mathbf{Z}^{\text{sim}} = G(\theta_h, \theta^c) = \mathbf{z}^{\text{sim}(\theta_h)} + \boldsymbol{\eta}(\theta^c) \tag{3}$$

where $\mathbf{z}^{\text{sim}(\theta_h)}$ are the flow index PCs computed from the streamflow time series simulated by the hydrological model with parameters θ_h, that is, $\mathbf{z}^{\text{sim}(\theta_h)} = \mathbf{z}\left[h(\theta_h; \tilde{\mathbf{x}}_{1:t}, \mathbf{s}_0)\right] = \boldsymbol{\varsigma}\,\boldsymbol{\omega}^{\theta_h}$ with $\boldsymbol{\omega}^{\theta_h = \mathbf{f}_\omega(\mathbf{q}^{\theta_h})}$ and $\mathbf{q}^{\theta_h} = \left\{q_t^{\theta_h}; t = 1, ..., N_t\right\}$.

The term $\boldsymbol{\theta}^c$ denotes all parameters of the combined error model. The number of simulation time steps N_t used to calculate $\mathbf{z}^{\mathrm{sim}(\boldsymbol{\theta}_h)}$ need not equal the number of observed time steps $N_{\tilde{t}}$ used to calculate $\hat{\mathbf{z}}^{\mathrm{reg}}$ and/or $\tilde{\mathbf{z}}$ in Section 2.3, and indeed these time steps may span different periods.

The combined error is difficult to characterize for several reasons. For example, even if the hydrological model errors were known in the streamflow space, they would require propagation into the space of PCs, leading to an error model without a closed form probability distribution (Kavetski et al., 2018). Estimating the combined error directly in PC space is also difficult because the length of $\mathbf{z}$ is short (unlike for streamflow time series); it would also be limiting because in practice hydrological models are intended to make predictions in the streamflow space.

In light of these challenges, we make the pragmatic approximation that the combined error is dominated by the regionalization model error, that is, $\boldsymbol{\eta}(\boldsymbol{\theta}^c) \sim \boldsymbol{\varepsilon}(\boldsymbol{\theta}^\varepsilon)$. A similar assumption was effectively made in previous works (e.g., Almeida et al., 2016; Bulygina et al., 2012, 2009; Prieto et al., 2019; Yadav et al., 2007). Under this assumption, the only parameters requiring inference are $\boldsymbol{\theta}_h$, with $\boldsymbol{\theta}^\varepsilon$ kept fixed at values $\hat{\boldsymbol{\theta}}^\varepsilon$ estimated earlier during the regionalization.

The posterior distribution of parameters $\boldsymbol{\theta}_h$ of the probabilistic model G given $\hat{\mathbf{z}}^{\mathrm{reg}}$ and $\hat{\boldsymbol{\theta}}^\varepsilon$ is obtained from Bayes equation as follows,

$$\begin{aligned} p\left(\boldsymbol{\theta}_h|\hat{\mathbf{z}}^{\mathrm{reg}},\hat{\boldsymbol{\theta}}^\varepsilon,G\right) &= \frac{p\left(\hat{\mathbf{z}}^{\mathrm{reg}}|\boldsymbol{\theta}_h,\hat{\boldsymbol{\theta}}^\varepsilon,G\right)p\left(\boldsymbol{\theta}_h|G\right)}{p\left(\hat{\mathbf{z}}^{\mathrm{reg}}|\hat{\boldsymbol{\theta}}^\varepsilon,G\right)} = \frac{p\left(\hat{\mathbf{z}}^{\mathrm{reg}}|\boldsymbol{\theta}_h,\hat{\boldsymbol{\theta}}^\varepsilon,G\right)p\left(\boldsymbol{\theta}_h|h\right)}{\int_\Omega p\left(\hat{\mathbf{z}}^{\mathrm{reg}}|\boldsymbol{\varphi},\hat{\boldsymbol{\theta}}^\varepsilon,G\right)p(\boldsymbol{\varphi}|h)\mathrm{d}\boldsymbol{\varphi}} \\ &= \frac{p\left(\hat{\mathbf{z}}^{\mathrm{reg}}|\mathbf{z}^{\mathrm{sim}(\boldsymbol{\theta}_h)},\hat{\boldsymbol{\theta}}^\varepsilon,G\right)p\left(\boldsymbol{\theta}_h|h\right)}{\int_\Omega p\left(\hat{\mathbf{z}}^{\mathrm{reg}}|\mathbf{z}^{\mathrm{sim}(\boldsymbol{\varphi})},\hat{\boldsymbol{\theta}}^\varepsilon,G\right)p(\boldsymbol{\varphi}|h)\mathrm{d}\boldsymbol{\varphi}} \end{aligned} \tag{4}$$

The likelihood function $p\left(\hat{\mathbf{z}}^{\mathrm{reg}}|\boldsymbol{\theta}_h,\hat{\boldsymbol{\theta}}^\varepsilon,G\right) = p\left(\hat{\mathbf{z}}^{\mathrm{reg}}|\mathbf{z}^{\mathrm{sim}(\boldsymbol{\theta}_h)},\hat{\boldsymbol{\theta}}^\varepsilon,G\right)$ describes the relationship between regionalized and simulated flow index PCs, as given by the probabilistic model G. The prior on the hydrological parameters, $p\left(\boldsymbol{\theta}_h|h\right)$, is set to uniform over the feasible parameter ranges unless specific prior information is available. This stage of the analysis is schematized in Figure 1 row c.

Equation 4 explicitly indicates conditioning on the entire probabilistic model G, which in addition to the deterministic hydrological model includes the regionalization residual error model. Note also that the conditioning on $\hat{\mathbf{z}}^{\mathrm{reg}}$ in equation Equation 4 corresponds to conditioning on the deterministic regionalization model and its parameters estimated during flow index PC regionalization. These quantities are kept fixed during the identification of dominant mechanisms.

2.4.2. Posterior Probability of a Single Hydrological Model Structure

Consider an ensemble of N_G model structures, $\mathbf{G} = \left\{G^{(i)}; i = 1, \ldots, N_G\right\}$, for example, generated using a modular modeling framework.

The posterior probability $p(G^{(k)}|\hat{\mathbf{z}}^{\mathrm{reg}},\hat{\boldsymbol{\theta}}^\varepsilon,\mathbf{G})$ of a model structure $G^{(k)}$, given the regionalized flow index PCs $\hat{\mathbf{z}}^{\mathrm{reg}}$, the error model parameters $\hat{\boldsymbol{\theta}}^\varepsilon$ and the ensemble $\mathbf{G}$ is:

$$\begin{aligned} p(G^{(k)}|\hat{\mathbf{z}}^{\mathrm{reg}},\hat{\boldsymbol{\theta}}^\varepsilon,\mathbf{G}) &= \frac{p(\hat{\mathbf{z}}^{\mathrm{reg}}|\hat{\boldsymbol{\theta}}^\varepsilon,G^{(k)})p(G^{(k)}|\mathbf{G})}{\sum_{i=1}^{N_G} p(\hat{\mathbf{z}}^{\mathrm{reg}}|\hat{\boldsymbol{\theta}}^\varepsilon,G^{(k)})p(G^{(k)}|\mathbf{G})} = \\ &= \frac{\int_{\Omega_{(k)}} p(\hat{\mathbf{z}}^{\mathrm{reg}}|\boldsymbol{\theta}_{h(k)},\hat{\boldsymbol{\theta}}^\varepsilon,G^{(k)})p(\boldsymbol{\theta}_{h(k)}|G^{(k)})\mathrm{d}\boldsymbol{\theta}_{h(k)}p(G^{(k)}|\mathbf{G})}{\sum_{i=1}^{N_G}\int_{\Omega_{(i)}} p(\hat{\mathbf{z}}^{\mathrm{reg}}|\boldsymbol{\theta}_{h(i)},\hat{\boldsymbol{\theta}}^\varepsilon,G^{(i)})p(\boldsymbol{\theta}_{h(i)}|G^{(i)})\mathrm{d}\boldsymbol{\theta}_{h(i)}p(G^{(i)}|\mathbf{G})} \end{aligned} \tag{5}$$

where $p(G^{(i)}|\mathbf{G})$ denotes the prior of model structure $G^{(i)}$.

The integral terms in Equation 5 are often referred to as the BME or Marginal Likelihood. There are several approaches for their computation, including "semi-analytical" and numerical approximations (e.g., see Schöniger

et al., 2014; Ye et al., 2008 for analysis in hydrological contexts). In this work, we use Monte Carlo integration with importance sampling as discussed in Appendix A. This stage of the analysis is schematized in Figure 1 rows d–e.

Note that in this work, the residual error term in the probabilistic model G is kept fixed at the structure identified during the flow index PC regionalization stage. For this reason, for a given ungauged catchment, the model structures G within the ensemble $\mathbf{G}$ differ solely in the structure used for the hydrological model h; see later discussion in Section 5.4.

2.4.3. Posterior Probability of a Hydrological Model Mechanism

Suppose the model ensemble $\mathbf{G}$ comprises model structures that represent $N^{\wp}$ hydrological model processes, using hydrological model mechanisms $\left\{m_k^{\wp}; k = 1, ..., N_{\mathrm{m}}^{\wp}; \wp = 1, ..., N^{\wp}\right\}$. The number of available mechanisms to represent process $\wp$ is $N_{\mathrm{m}}^{\wp}$.

The posterior probability of a mechanism $m_k^{\wp}$ can be approximated as the average of the posterior probabilities of all model structures that contain mechanism $m_k^{\wp}$,

$$p(m_k^{\wp}|\hat{\mathbf{z}}^{\mathrm{reg}}, \hat{\theta}^{\epsilon}, \mathbf{G}) \propto \frac{1}{N_{\mathrm{G}}^{\wp,k}} \sum_{i \in S(k;\mathbf{G},\wp)} p(G^{(i)}|\hat{\mathbf{z}}^{\mathrm{reg}}, \hat{\theta}^{\epsilon}, \mathbf{G}) \tag{6}$$

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 $m_k^{\wp}$ and $N_{\mathrm{G}}^{\wp,k}$ denote the number of models within this subset (see details in Prieto et al., 2021).

The assumptions behind Equation 6 are: (a) on average, highly probable mechanisms are those appearing in highly probable model structures and vice versa, (b) we assume that the multiple hypothesis framework provides a sufficient coverage of the space of possible mechanisms. This assumption is more likely to be reasonable if we use a large ensemble of models since such an ensemble is expected to provide a relative complete representation of the major hydrological fluxes (e.g., Addor & Melsen, 2019; Clark et al., 2011, 2008; Kavetski & Fenicia, 2011), (c) mechanisms are mutually exclusive—a hydrological process is represented by a single model mechanism, (d) the prior on the mechanisms is uniform, and $p_{\mathrm{unif}}(m_k^{\wp}|\mathbf{G}) = 1/N_{\mathrm{m}}^{\wp}$.

This stage of the analysis is depicted in Figure 1 row e. Detailed derivation of Equation 6 and its use of corrections to account for unbalanced (opportunistic) distribution of mechanisms when the model ensemble is biased toward particular mechanisms, can be found in (Prieto et al., 2021); see also Elkan (2001) and Saerens et al. (2002).

2.5. Multiple Hypothesis Testing to Identify Hydrological Mechanisms

2.5.1. Key Definitions

The posterior probabilities of mechanisms are used to identify the dominant mechanism for a process $\wp$. The identification is conditioned on the regionalized flow index PCs $\hat{\mathbf{z}}^{\mathrm{reg}}$, representing an extension of the method introduced by Prieto et al. (2021) for mechanism identification from observed streamflow time series. In addition, the method is extended to account for hydrological parameter uncertainty, which is likely to be more pronounced when working with the first few flow index PCs instead of long streamflow time series. The mechanism identification method includes the specification of a significance level α, and accounts for the potentially large number of hypothesis tests being carried out (which raises the probability of identifying a wrong mechanism as the dominant mechanism, i.e., of Type 1 errors).

The key definitions of the multiple hypothesis testing methods are as follows (Prieto et al., 2021):

1. An individual comparison is an individual test of whether a mechanism $m_k^{\wp}$ is "dominant", that is, substantially more likely than alternative mechanisms available in $\mathbf{G}$ to represent $\wp$
2. The null hypothesis for an individual test $\mathrm{H0}_k^{\wp}$ is "mechanism $m_k^{\wp}$ is not the dominant mechanism for process $\wp$"
3. A family of comparisons for a process $\wp$ is the set of individual tests of each mechanism against all other mechanisms that represent that process
4. The null hypothesis for a family of comparisons $\mathrm{H0}^{\wp}$ is "none of the proposed hydrological mechanism is dominant for process $\wp$"

5. The family wise error rate (FWER) is the probability of making one or more type I errors in the family of multiple tests, that is, the probability of incorrectly identifying a mechanism as dominant for a given process. The (conservative) Bonferroni's correction is employed to keep the FWER below a prescribed significant level α, by imposing a stricter α^* in each individual test (Hochberg, 1988)
6. A test statistic t is required for each individual null hypothesis $H0_k^{\wp}$. As detailed in Prieto et al. (2021), the test statistic has a probability distribution without a closed analytical form, and hence the probability that the statistic exceeds a prescribed threshold τ is estimated using a bootstrap approach. The estimated exceedance probability is then compared to a significance level α (see next section for more details)

2.5.2. Test Statistic and Identification of Dominant Mechanisms

The test statistic for the individual null hypothesis $H0_k^{\wp}$ is taken as the posterior probability of mechanism $m_k^{\wp}$. A mechanism is considered dominant if this test statistic exceeds a threshold value τ that is,

$$t_k^{\wp} = p\left(m_k^{\wp} | \hat{\mathbf{z}}^{\text{reg}}, \hat{\boldsymbol{\theta}}^{\epsilon}, \mathbf{G}\right) > \tau \tag{7}$$

The hypothesis $H0^{\wp}$ is rejected if one of the proposed hydrological mechanisms is found to be dominant for process $\wp$ (if $\tau > 0.5$ then only a single process can be dominant).

The threshold τ and the significance level α control the stringency of the hypothesis test. In the previous study on mechanism identification in gauged catchments (Prieto et al., 2021), we used $\tau = 0.75$ because the inference was conditioned on observed streamflow data with thousands of time steps (data points). In the present article, where the inference is based on just a few flow index PCs, the threshold is relaxed to $\tau = 0.6$, that is, a mechanism is considered dominant if it is at least 1.5 times more probable than all its alternatives. See Figure 1 row f.

The test statistic $t_k^{\wp}$ depends on the model ensemble $\mathbf{G}$ and is treated as a realization of a random variable $T_k^{\wp}$, with cumulative distribution function (cdf) $F_{T(\wp,k)}(t)$. An individual hypothesis $H0_k^{\wp}$ is rejected if the probability $\omega_k^{\wp}$ that $t_k^{\wp}$ exceeds the threshold τ, that is, $\omega_k^{\wp} = 1 - F_{T(\wp,k)}\left(t_k^{\wp}\right)$, is no less than the pre-specified significance for an individual test, that is, $\omega_k^{\wp} \geq 1 - \alpha_{\text{Bonf}}^*$, where $\alpha_{\text{Bonf}}^* = \alpha / N_{\text{m}}^{\wp}$ and $N_{\text{m}}^{\wp}$ is the number of mechanisms available for process $\wp$ (Section 2.4.3). The overall significance level is here chosen as $\alpha = 0.05$. Since the cdf $F_{T(\wp,k)}(t)$ has no closed form, $\omega_k^{\wp}$ is estimated by applying bootstrapping (Efron & Tibshirani, 1986) to the ensemble $\mathbf{G}$. The bootstrap re-sampling represents (approximately) the uncertainty associated with a particular choice of model structures within the ensemble; see Prieto et al. (2021) for a detailed discussion. The null hypothesis $H0_k^{\wp}$ is then tested using the procedure in Prieto et al. (2021) as given in Appendix B. These stages of the analysis are illustrated schematically in Figure 1 row g.

3. Case Study Description

3.1. Catchments

The case study employs 92 gauged catchments in northern Spain, shown with black stars in Figure 2. To evaluate the method, we treat 16 of the 92 catchments as "ungauged" (one at a time); these catchments are indicated with red circles and labels in Figure 2. The 16 catchments are selected because they have a sufficiently long (at least 8 years) concomitant record of observed daily rainfall and streamflow. Daily potential evapotranspiration (PET) is estimated from monthly PET. The data set is the same as in the earlier studies of (Peñas et al., 2014; Prieto et al., 2019).

Land cover of the 16 catchments is dominated by pastures, broadleaf forests, and coniferous forests; urbanized zones comprise less than 8% of the catchment areas. Catchment areas range from 22 to 623 km^2, elevations from 483 to 1,505 m, slopes of main river channels from 21% to 53%, annual average temperature from 7°C to 12°C, surface runoff coefficient from 0.24 to 0.95, annual average rainfall from 681 mm/year to 1,809 mm/year, and annual average PET from 564 mm/year to 962 mm/year. In Figure 2, the catchments are colored according to their aridity index (Arora, 2002) and minimum monthly average temperature as described in Prieto et al. (2019).

The full list of the 103 flow indices includes, for example, the mean and standard deviation of the annual flow, maximum and minimum monthly annual flow, quantiles of the flow duration curve, etc. These indices can be found in (Peñas et al., 2014).

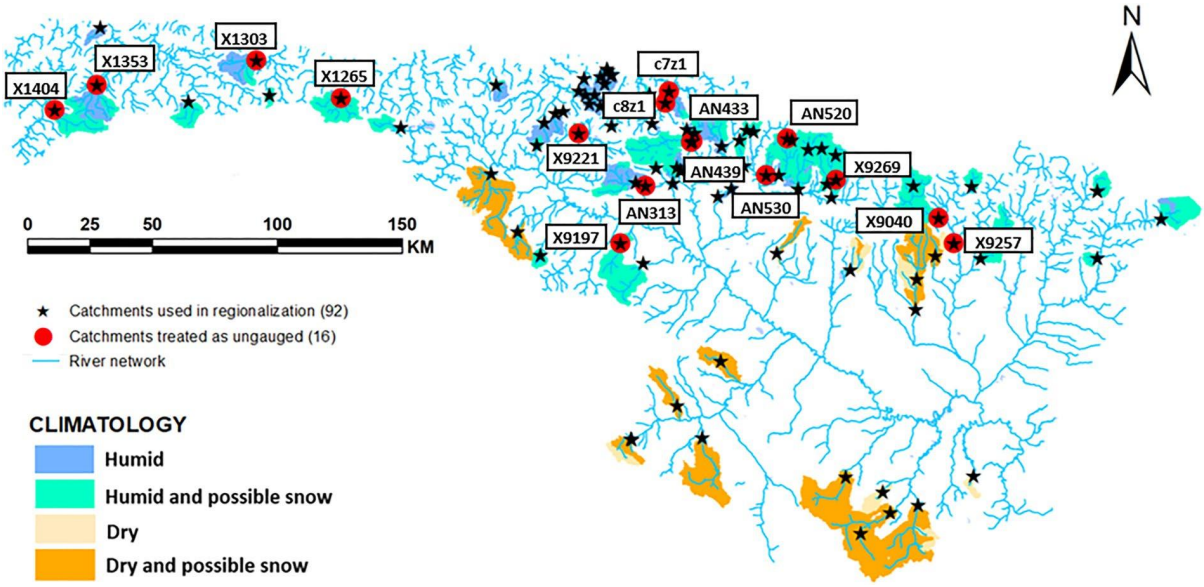

Figure 2. Case study catchments in northern Spain. Adapted from Prieto et al. (2019).

3.2. Regionalized Flow Indices

The first 4 PCs collectively explain 87% of observed variability in the 103 flow indices (see supplementary material in Peñas et al., 2014) across all 92 case study catchments and were used to construct the regionalization model using RF regression. This section reports earlier results from Prieto et al. (2019) to provide context. When working with a given target catchment, the residual errors of the regionalization model were estimated across the remaining 91 catchments using a jack-knife strategy and approximated by a joint Gumbel-Gauss distribution (Prieto et al., 2019). More specifically, the residuals of the first PC follow (empirically) a Gumbel distribution (Generalized Extreme Value Distribution Type I); the residuals for the second and third PCs are correlated and follow a bivariate Gaussian distribution; the residuals for the fourth PC have a Gaussian distribution,

$$\varepsilon_1^{\text{ucat}} \sim \text{G}\left(\mu_1^{\text{obs}}, \beta_1^{\text{obs}}\right) \qquad \sigma_1^{\text{obs}} = \pi\beta_1^{\text{obs}}/\sqrt{6} \tag{8}$$

$$\varepsilon_{2:3}^{\text{ucat}} \sim \text{N}(\boldsymbol{\mu}_{2:3}^{\text{obs}}, \Sigma^{\text{obs}}) \qquad \Sigma^{\text{obs}} = \begin{bmatrix} \left(\sigma_2^{\text{obs}}\right)^2 & \rho\sigma_2^{\text{obs}}\sigma_3^{\text{obs}} \\ \rho\sigma_2^{\text{obs}}\sigma_3^{\text{obs}} & \left(\sigma_3^{\text{obs}}\right)^2 \end{bmatrix} \tag{9}$$

$$\varepsilon_4^{\text{ucat}} \sim \text{N}\left(\mu_4^{\text{obs}}, \sigma_4^{\text{obs}}\right) \tag{10}$$

where μ denotes the mean parameters, β denotes the Gumbel dispersion parameter, and σ (and Σ) denote the standard deviation parameter. The ranges of estimated parameters in Equations 8–10 are reported in Prieto et al. (2019), and reproduced in Appendix C (Table C1).

The "quantity" of information I conveyed by the selected number of flow index PCs is defined by the fraction of variance explained with respect to the full set of available flow indices,

$$I = \frac{\sum_{i=1}^{N_z} (\tilde{\mathbf{z}}_i)^2}{\sum_{i=1}^{N_\omega} (\tilde{\mathbf{z}}_i)^2} \tag{11}$$

Table 1
Information Content of the First Four Flow Index PCs in the 16 "Ungauged" Catchments—Experiments 1 and 2

Catchments	Information content	FID Experiment 1	FID Experiment 2
X9269	41%	0.14	0.43
X9197	53%	0.29	0.14
X1265	54%	0.29	0.00
AN313	58%	0.29	0.57
C7Z1	65%	0.14	0.14
C8Z1	66%	0.29	0.14
X9257	69%	0.29	0.29
AN530	73%	0.29	0.43
X9221	79%	0.14	0.29
AN520	81%	0.29	0.00
X1404	86%	0.29	0.29
AN433	87%	0.43	0.43
X1303	87%	0.14	0.29
AN439	91%	0.14	0.14
X9040	94%	0.57	0.29
X1353	96%	0.29	0.14

and is calculated separately in each target catchment. Given this definition, Table 1 shows that the quantity of flow index information conveyed by the first four flow index PCs in the 16 catchments varies from 41% to 96%.

As such, the "quantity" of information is controlled by the number of flow index PCs included in the analysis, and the "quality" of information is controlled by the bias and dispersion of the errors.

3.3. Hydrological Model Structures

The hydrological models and mechanisms for hypothesis testing are generated using the Framework for Understanding Structural Errors (FUSE; Clark et al., 2011, 2008).

FUSE allows choosing among multiple model mechanisms to represent each model process. In this article, we consider 7 processes: (a) (water) storage in the unsaturated zone, defined by the architecture of the upper soil layer; (b) storage of water occurring in the unsaturated zone, defined by the architecture of the lower soil layer; (c) evaporation; (d) interflow for the lateral movement of the water into the soil; (e) percolation for the vertical movement of water from the unsaturated zone (upper soil layer) to the saturated zone (lower soil layer); (f) surface runoff generation; and (g) routing for the evolution (shape and time) of the surface runoff hydrograph as the water moves through the river. The number of mechanisms for each process ranges from 2 to 4, with a total of 19 mechanisms (Table 1 in Prieto et al., 2021). In this article, we use the implementation of FUSE in the R language (Vitolo et al., 2016).

3.4. Approximation of the BME of a Hydrological Model Structure

The mechanism identification framework requires estimates of the BME of each model structure in the ensemble **G**. BME is approximated via Monte Carlo integration, using an importance sampling algorithm; see Appendix A. The importance sampling makes use of 1,000 parameter sets sampled from a uniform prior using the Latin Hypercube method.

The simulated streamflow time series, needed to estimate the BME of a given model structure in a given catchment, are computed using the FUSE model structure, forced by observed daily rainfall and (estimated) daily PET. The first year of daily data is used as warm-up and the corresponding streamflow is not used in the BME estimation.

This stage of the analysis is by far the most computationally expensive, requiring a total of 9,984,000 simulations (16 basins × 624 models × 1,000 parameter sets), with at least 8 yr of daily data each. Parallel computing was undertaken; the total CPU runtime is estimated at approximately 11 year.

3.5. Performance Metrics

The performance of the mechanism identification framework in the case study experiments is evaluated using two metrics, namely "Fraction of Identifications" F_{id} and "Reliability" R. The fraction of identifications characterizes the probability of making an identification, irrespective of whether the identification is correct or incorrect. The reliability metric characterizes the probability that an identification, once made, is correct. In the real data experiments, which do not use replication, the metrics are calculated across the 16 catchments treated as ungauged. In the synthetic experiments, the metrics are additionally averaged across multiple replicates of (synthetic) inference data.

3.5.1. Fraction of Identifications

The fraction of identifications metric, F_{id}, seeks to quantitatively answer the following question: How likely is the method to identify a mechanism as dominant? The metric is defined as follows,

$$F_{\text{id}} = \frac{N_{\text{id}}}{N_{\text{trials}}} \tag{12}$$

where N_{id} is the number of trials where dominant mechanisms are identified, and N_{trials} is the total number of trials. The trials are represented by the application of the mechanism identification method across multiple synthetic replicates, multiple catchments, and multiple processes (real data experiments), or across multiple synthetic replicates and multiple catchments (synthetic experiments).

In each experiment, we define two types of this metric:

1. the fraction of identifications across all hydrological processes. The number of trials for this metric is 112 (16 catchments × 7 processes) in the real data experiments, and 11,200 (16 catchments × 7 processes × 100 replicates) in the synthetic experiments
2. the fraction of identifications for each individual process. The number of trials for this metric is 16 (16 catchments × 1 process) in the real data experiments and 1,600 (16 catchments ×1 process × 100 replicates) in the synthetic experiments

The F_{id} metric ranges from 0 to 1; a value of 0 indicates that a dominant mechanism is never identified, while 1 indicates that a dominant mechanism is always identified (though the identification may be incorrect). F_{id} is broadly similar to the power metric P used in Prieto et al. (2021), except that F_{id} can be calculated for any experiment while P can be calculated only when the "true" model is known.

3.5.2. Reliability

The reliability metric R seeks to answer the following question: How reliable ("trustworthy") is the mechanism identification method? Is the mechanism identified as dominant the actual true mechanism? The metric is defined as follows,

$$R = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{13}$$

where N_{TP} is the number of trials where a true mechanism is identified as dominant, and N_{FP} is the number of trials where the wrong mechanism is identified as dominant.

Reliability can be calculated only in synthetic experiments where the true model is known.

In each experiment, we define two types of this metric:

1. the reliability of mechanism identification across all hydrological processes. The number of trials here is 11,200
2. the reliability of mechanism identification for each individual process. The number of trials here is 1,600

The reliability metric ranges from 0 to 1; a value of 0 indicates that the identification (when made) is always incorrect, while value of 1 indicates that the identification (when made) is always correct. Reliability has also been termed "Positive Prediction Value" (Tharwat, 2020) and was used in the earlier study by Prieto et al. (2021).

3.6. Empirical Analysis Using Real and Synthetic Data

Mechanism identification is affected by multiple sources of error: (a) hydrological model error (structure, input, and parameter uncertainties), (b) regionalization model error (including structure, catchment descriptors, and flow indices), (c) the quantity of information (including the number of flow index PCs and the fraction of variance they explain, the use of flow indices instead of observed time series; and the use of PCs to represent the flow indices) and (d) algorithmic aspects, in particular, the (relatively) limited number of samples of parameter sets used to estimate the BME.

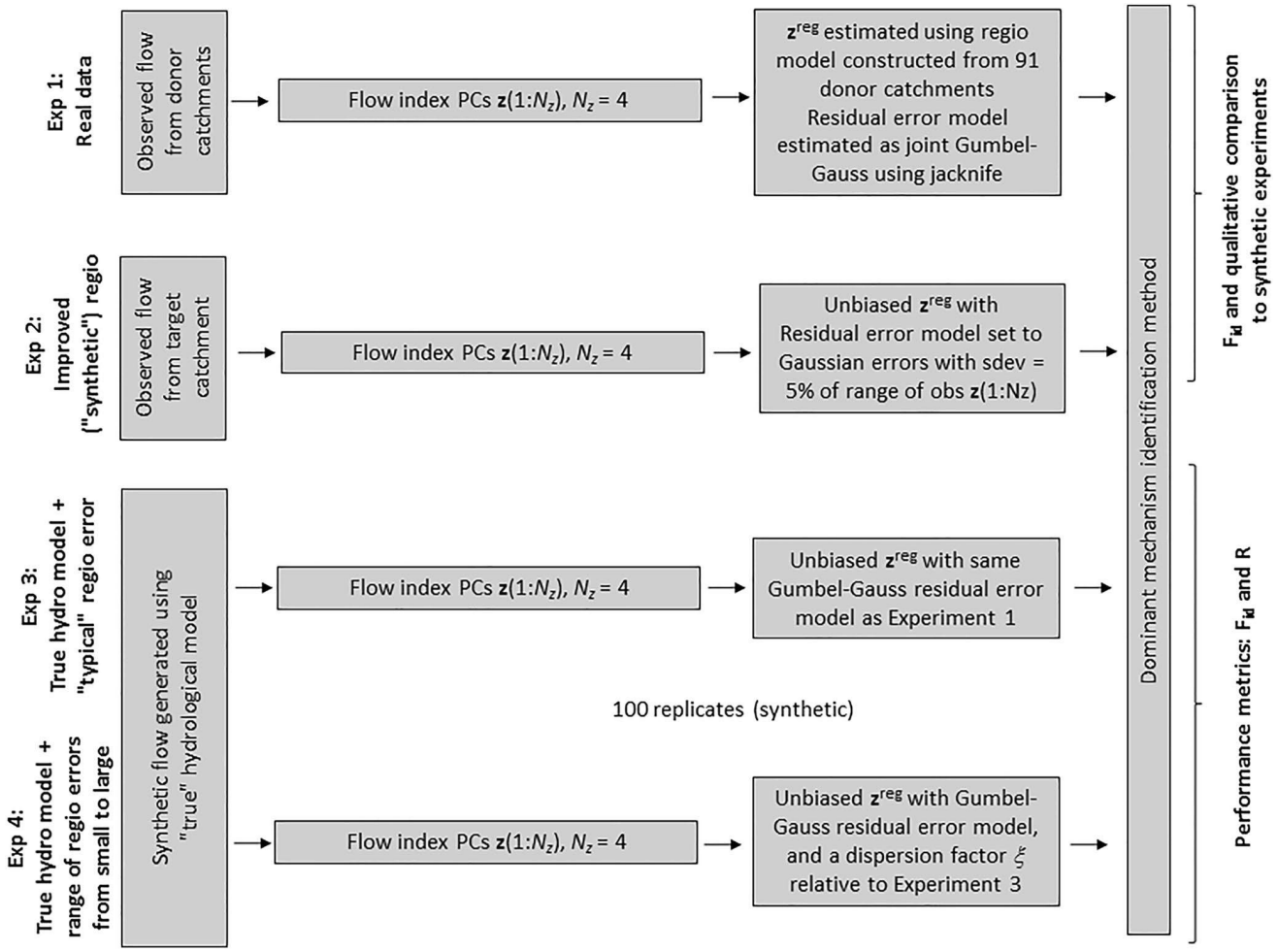

Figure 3. Flow chart of the case study experiments.

Four experiments are carried out to investigate the impact of the first two sources of error on mechanism identification, holding the number of PCs equal to the number used in the real data experiments. Flow charts of the Experiments are given in Figure 3.

3.6.1. Experiment 1: Real Regionalized Data

This experiment illustrates the intended usage of the proposed mechanism identification method in ungauged catchments, where observed-based flow index PCs are not available. We investigate the following questions: (a) What is the method performance in terms of fraction of identifications? (b) Which are the most identifiable processes and most identified mechanisms? The most identifiable processes are those for which a dominant mechanism is most frequently identified. The most identified mechanism is the mechanism that is most frequently identified for a given process.

In a given target catchment, we estimate the first 4 regionalized flow index PCs $\hat{\mathbf{z}}^{reg(1)}$ using the regionalization model constructed from the remaining 91 (donor) catchments, condition FUSE on these flow index PCs (Prieto et al., 2019), and undertake mechanism identification. This procedure is repeated for each of the 16 catchments treated as ungauged (target). The flow chart for Experiment 1 is schematized in Figure 3 row 1.

Since the true mechanisms are not known, we can only calculate the fraction of identifications metric, F_{id}. Limitations in identifying a mechanism (F_{id} below 1) can be attributed to regionalization error and/or hydrological model error and/or insufficient quantity of information. Note also that the first four flow index PCs cover a different quantity of information in each catchment, as low as 41% in some cases.

3.6.2. Experiment 2: "Accurate" Regionalized Data (Observation-Based Flow Index PCs)

This experiment evaluates mechanism identification when the regionalization model error is lower than in Experiment 1. Are dominant mechanisms identified for more processes? Are the same mechanisms identified as in Experiment 1?

The experiment is devised such that the regionalization has low error (i.e., nearly exact). In each target catchment, the synthetic "regionalized" flow index PCs $\hat{\mathbf{z}}^{\mathrm{reg}(2)}$ used to condition the mechanism identification are set equal to the observation-based flow index PCs $\tilde{\mathbf{z}}$. The same number (4) of PCs is used as in Experiment 1.

The deterministic term in the regionalization model is also set equal to $\tilde{\mathbf{z}}$, that is, the regionalization model is unbiased. The residual error term is set to a Gaussian distribution with mean equal to 0 and a standard deviation equal to 5% of the full range of observation-based flow index PCs computed for all 92 case study catchments. This value is intended to reflect an accurate regionalization model based on values reported in the earlier literature (Almeida et al., 2016). The flow chart for Experiment 2 is given in Figure 3 row 2.

This experiment is "synthetic", though without the concept of "true" hydrological mechanisms. Hence we can only calculate F_{id}. Deficiencies in mechanism identification can be attributed to hydrological model error, inaccuracies in the observed streamflow used to compute the flow index PCs, and/or insufficient quantity of "regionalized" information (since only four PCs are employed, which once again convey a different quantity of information across the catchments).

An important benefit of Experiment 2 is that it helps appraise the assumption made in Section 2.4.1 that the combined model error is dominated by regionalization error. Specifically, in the 16 target catchments, we compare three estimates of flow index PCs: (a) the regionalized estimate $\hat{\mathbf{z}}^{\mathrm{reg}}$, (b) the observation-based estimate $\tilde{\mathbf{z}}$, and (c) the "best" estimate $\breve{\mathbf{z}}$ from the full ensemble of hydrological models calibrated to the observation-based values. If regionalization model errors are (considerably) larger than hydrological model errors, we would see a (considerably) larger discrepancy between $\hat{\mathbf{z}}^{\mathrm{reg}}$ and $\tilde{\mathbf{z}}$ than between $\breve{\mathbf{z}}$ and $\tilde{\mathbf{z}}$. Conversely, if hydrological model errors dominate, the opposite relationship would hold. Note that this analysis is also impacted by measurement errors in the observed streamflow data, which we assume to be smaller than the other two sources of error.

3.6.3. Experiment 3: Synthetic Experiment With "Typical" Model Error

This experiment evaluates mechanism identification in a synthetic scenario where the "true" mechanisms are known, under the condition that the combined error $\boldsymbol{\eta}$ in Equation 3 is "typical" (here taken as similar to the error estimated in Experiment 1). Note that, as per the inference setup in Section 2.4.1, the combined error represents the joint effects of errors in the regionalization and hydrological models, and in this synthetic setup we do not attempt to assign or infer them individually.

We investigate the following questions: Are the mechanisms identified as dominant the "true" mechanisms? Does mechanism identification improve with respect to the previous experiments? For which process is a dominant mechanism most and least frequently identified? For which process is the "true" mechanism most and least frequently identified?

The "true" hydrological model, which comprises the "true" mechanisms for a given catchment, is selected as the model structure that provided the best match to the observation-based flow index PCs in Experiment 2; see Appendix C2 for details. The flow index PCs computed using the true model are treated as the "true" flow index PCs, $\breve{\mathbf{z}}$.

In each target catchment, the synthetic "regionalized" flow index PCs $\hat{\mathbf{z}}^{\mathrm{reg}(3)}$ are obtained by corrupting the "true" flow index PCs with known ("synthetic") error with magnitude based on the combined error in Experiment 1, that is, $\hat{\mathbf{z}}^{\mathrm{reg}(3)} = \breve{\mathbf{z}} - \mathbf{e}^{\mathrm{synth}}$ where the error term $\mathbf{e}^{\mathrm{synth}}$ is sampled from the same joint Gumbel-Gauss distribution from Experiment 1. The negative sign is used because the Gumbel distribution used within the residual error model is asymmetric.

The same number of flow index PCs is used as in Experiments 1–2; the quantity of regionalized information varies across catchments.

The deterministic term in the regionalization model is set equal to $\breve{\mathbf{z}}$, which in this setup corresponds to both the regionalization and hydrological models being unbiased. The combined error model is set to the joint Gumbel-Gauss distribution with catchment-specific parameters equal to those used in Experiment 1.

Experiment 3 allows us to investigate mechanism identification performance in the hypothetical case where the error model used in Experiment 1 is statistically consistent with the actual errors. While this naturally represents a considerable idealization, we argue it provides a useful appraisal of the proposed method. Moreover, given the limitations of the jack-knife estimation of the regionalization residual error model (which itself operates in very data-scarce conditions), it is difficult to design synthetic experiments that reflect its characteristics better than Experiment 3 without making other strong assumptions about the nature of the errors (in particular their bias and random components).

The flow chart for Experiment 3 is given in Figure 3 row 3. Note that 100 replicates of the "conditioning data" $\hat{\mathbf{z}}^{\mathrm{reg}(3)}$ are generated, in order to obtain meaningful estimates of the performance metrics.

As the true mechanisms are known in this synthetic experiment, we calculate the reliability metric R in addition to the fraction of identifications metric F_{id}. Deficiencies in mechanism identifiability (F_{id} below 1) and/or discrepancies between identified vs. "true" mechanisms (R below 1) can be attributed to (synthetic) combined error and/or insufficient quantity of information.

3.6.4. Experiment 4: Dependence of Method Performance on Model Error (Synthetic)

This experiment examines the performance of the method for a range of values of the combined (hydrological and regionalization) model error. We expect to see improved (though not perfect) performance when the combined error is low, and conversely worse performance when the combined error is large.

The experiment is constructed in the same way as Experiment 3, but we scale the random errors by a "synthetic" factor ξ. When $\xi = 1$ the errors have the same dispersion as in Experiment 3 (which itself was set according to Experiment 1); $\xi > 1$ yield (on average) larger errors and $\xi < 1$ yield (on average) smaller errors. We report results for $\xi = 2$, 1, 0.5, 0.1 and 0.05.

The flow chart for Experiment 4 is given in Figure 3 row 4. Note that 100 replicates of the "conditioning data", $\hat{\mathbf{z}}^{\mathrm{reg}(4)}$, are generated for each value of the error factor ξ.

We report both the fraction of identifications and reliability. Deficiencies in mechanism identification are attributed to the (synthetic) combined error and incomplete regionalized information (as not all flow index PCs are used, which as mentioned earlier provides a varying quantity of information across the catchments).

4. Results

Figure 4 reports the fraction of identifications for Experiments 1–3, and the reliability for Experiment 3. Figure 5 provides a map of catchments with results from Experiments 1 and 2, distinguishing the locations where mechanisms are identified (circles colored according to mechanism) and not identified (empty circles). Additionally, Figure 5 distinguishes the identification in Experiment 1 (large circles) and Experiment 2 (smaller circles).

4.1. Experiment 1: Real Regionalized Data

Figure 4 shows that the overall fraction of identifications in Experiment 1, that is, F_{id} computed across all processes, is 0.27 (i.e., identification made in 30 out of 112 trials). When computed for individual processes, F_{id} ranges between 0 and 0.94. The processes with most identifiable mechanisms are routing ($F_{\mathrm{id}} = 0.94$, i.e., identifications made in 15 out of 16 trials), surface runoff ($F_{\mathrm{id}} = 0.31$), evaporation ($F_{\mathrm{id}} = 0.25$), and interflow ($F_{\mathrm{id}} = 0.25$). The processes with least identifiable mechanisms are those related to the unsaturated zone ($F_{\mathrm{id}} = 0.0$), percolation ($F_{\mathrm{id}} = 0.0$), and saturated zone ($F_{\mathrm{id}} = 0.12$).

The most frequently identified mechanisms (when a dominant mechanism is identified) are as follows: a routing component (i.e., there is a routing delay) is identified in 15 out of 15 trials; surface runoff generation controlled by the topographic index is identified in 5 out of 5 trials; sequential evaporation is identified in 3 out of 4 trials; no interflow mechanism is identified as dominant in 4 out of 4 trials; and a tension saturated storage with two

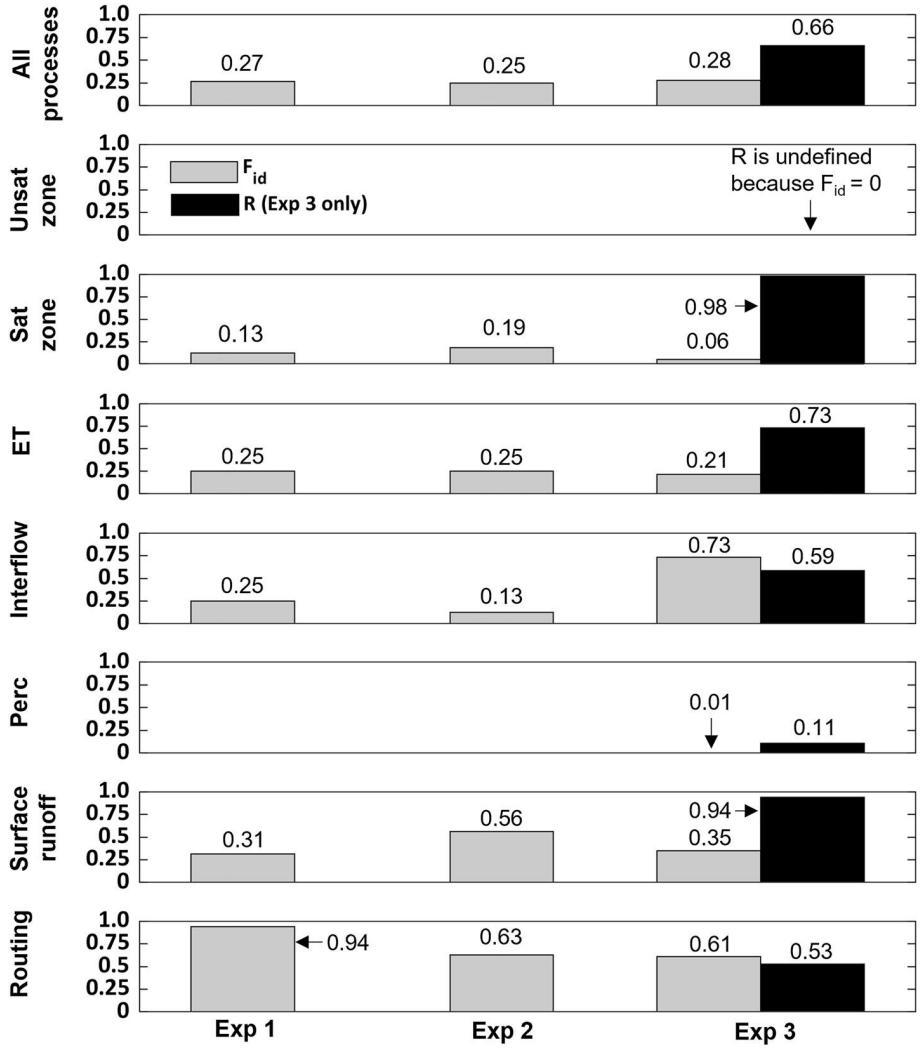

Figure 4. Performance of the mechanism identification method in Experiments 1–3. Fraction of identifications is reported for all three experiments; reliability is reported for Experiment 3 only. Metric values are reported as overall across all processes (row 1) and for each individual process (rows 2–8).

parallel tanks is identified in 2 out of 2 trials. Figure 5 provides a general sense of spatial aspects of mechanism identification. Apart from the consistent identification of mechanism 3 for surface runoff (controlled by topographic index, large pink circles) and mechanism 2 for routing (routing present, large blue circles), no obvious spatial patterns in the mechanism distribution are noted. For evaporation, a few different mechanisms are identified in different catchments.

4.2. Experiment 2: "Accurate" Regionalized Data (Observation-Based Flow Index PCs)

4.2.1. Fraction of Identifiability and Processes

Figure 4 shows that the overall fraction of identifications in Experiment 2, $F_{id} = 0.25$, remains largely unchanged with respect to Experiment 1. For individual processes, F_{id} ranges from 0 to 0.63 (0–10 out of 16 trials). The same processes are among the most/least identifiable as in Experiment 1, namely surface runoff and routing are most identifiable while unsaturated zone and percolation are the least identifiable. However, for individual processes, the fraction of mechanisms identified as dominant may increase (e.g., for surface runoff process) or decrease (interflow process) with respect to Experiment 1.

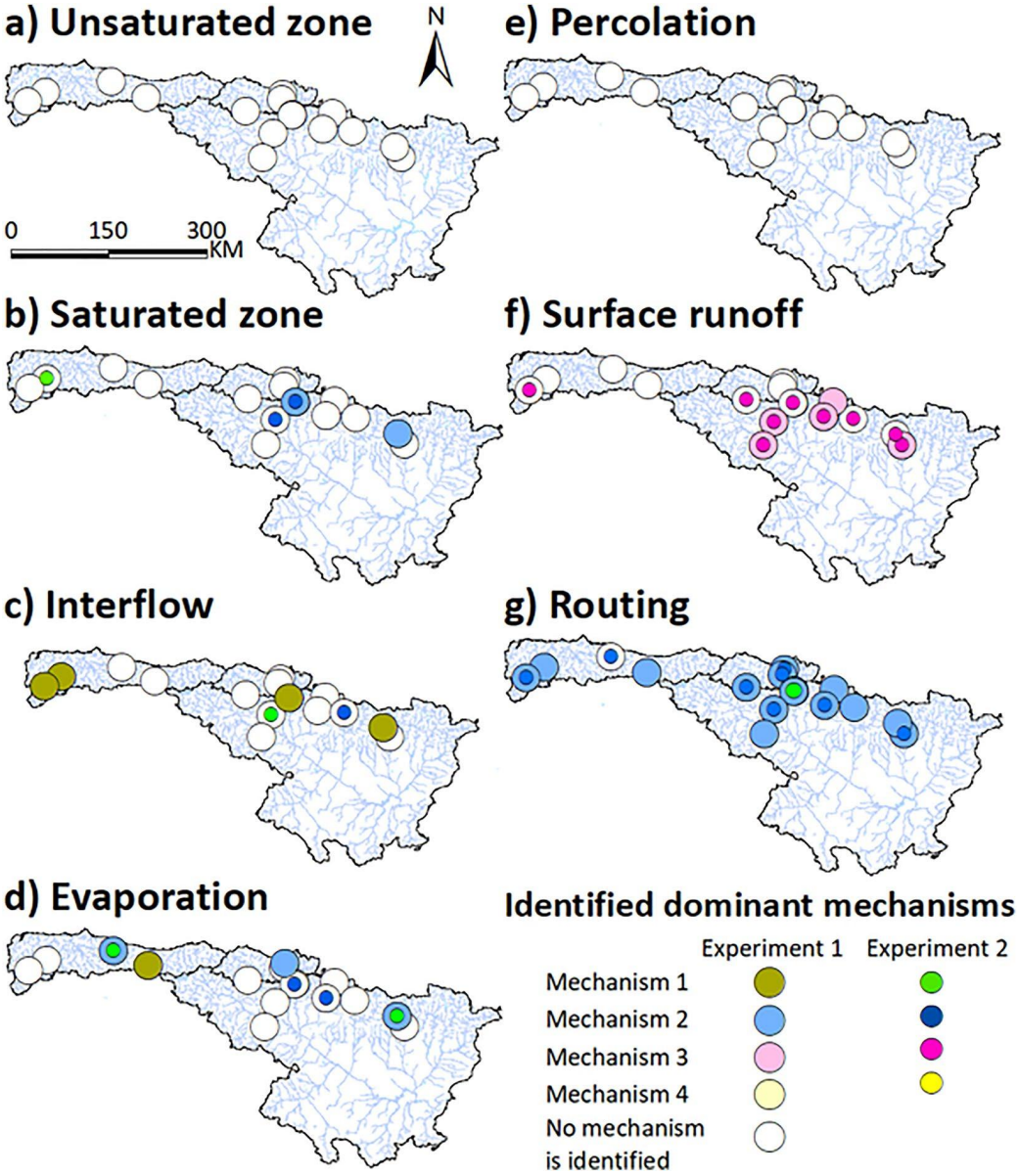

Figure 5. Identified mechanisms in Experiment 1 vs. Experiment 2. Outer colored circles show the identification for Experiment 1, inner colored circles show the identification for Experiment 2, blank circles show where no mechanism is identified. Catchments without the "inner" circle in this figure are those where a dominant mechanism is identified in Experiment 1 but not in Experiment 2.

Figure 5 shows that Experiment 2 is consistent with Experiment 1 in terms of the most frequently identified type of mechanisms for the processes of routing and surface runoff generation (a routing component is preferred and the surface runoff generation is controlled by the topographic index). In addition, there are 2 "switches" (i.e., a different mechanism identified for 2 processes) in Experiment 2 with respect to Experiment 1, and 10 new identifications (i.e., a mechanism was not identified for process in Experiment 1 but it is in Experiment 2).

4.2.2. Exploration of Assumption That Regionalization Errors Dominate Hydrological Errors

Panels a–e in Figure 6 show, for each of the 16 target catchments, the observation-based flow index PCs (black circles, computed), regionalized flow index PCs (blue triangles), and flow index PCs estimated using the best of the 624 hydrological models calibrated in the respective catchment (red crosses). Panel e in Figure 6 shows the

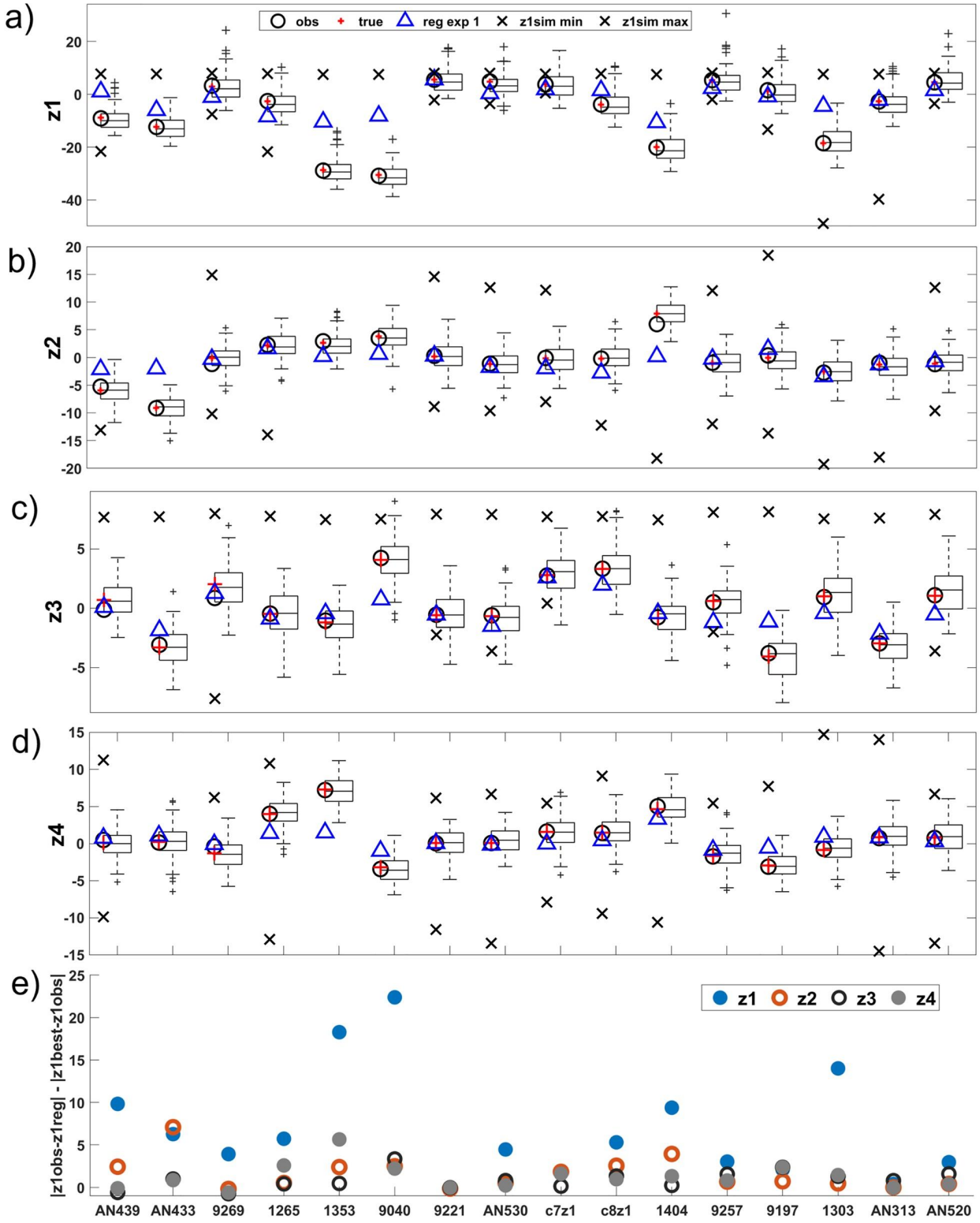

Figure 6. Regionalized vs. hydrological model error. Panels (a–e) compare flow index PCs estimated based on regionalization (Experiment 1), observations (Experiment 2), and the best hydrological model fit (Experiment 2); note that the latter is taken as the true model (in Experiment 3). Box plots of the synthetic regionalized flow index PCs used in Experiment 3 are also shown. Panel (e) shows the difference between $|\hat{\mathbf{z}}^{reg} - \tilde{\mathbf{z}}|$ and $|\bar{\mathbf{z}} - \tilde{\mathbf{z}}|$. Catchments where this difference is positive (i.e., $|\hat{\mathbf{z}}^{reg} - \tilde{\mathbf{z}}| > |\bar{\mathbf{z}} - \tilde{\mathbf{z}}|$) correspond to catchments where regionalization model errors dominate hydrological model errors. The error distributions of $\hat{\mathbf{z}}^{reg}$ for Experiment 1 are shown in Figure 4 in Prieto et al. (2019).

difference between $|\hat{\mathbf{z}}^{reg} - \tilde{\mathbf{z}}|$ and $|\bar{\mathbf{z}} - \tilde{\mathbf{z}}|$. Catchments where this difference is positive correspond to catchments where regionalization model errors exceed hydrological model errors.

This comparison enables an appraisal of the assumption that regionalization errors dominate hydrological errors. In the majority of cases, the regionalized estimates are clearly *further away* from observation-based estimates than the hydrological model estimates (e.g., z_1 in catchments 1,353, 9,040, and 1,303). In the majority of cases (54 of 64), $|\hat{\mathbf{z}}^{reg} - \tilde{\mathbf{z}}|$ is indeed larger than $|\bar{\mathbf{z}} - \tilde{\mathbf{z}}|$.

4.3. Experiment 3: Synthetic Experiment With "Typical" Model Error

As seen from Figure 4, the overall fraction of identification in Experiment 3, $F_{id} = 0.28$, is once again similar to the values in Experiments 1 and 2. For individual processes, F_{id} ranges from 0 to 0.73 (0–1,166 out of 1,600 trials). The processes for which a dominant mechanism is most frequently identified are interflow ($F_{id} = 0.73$) and routing ($F_{id} = 0.61$), which is in partial disagreement with Experiments 1 and 2 (where interflow had a low $F_{id} = 0.25$). The processes for which a dominant mechanism is least frequently identified are those in the unsaturated zone ($F_{id} = 0.0$) and percolation ($F_{id} = 0.006$), which are consistent with Experiment 1.

For this experiment, Figure 4 also reports the reliability R. The overall reliability is 0.66 (correct identifications in 2070 out of 3,149 trials where an identification is made). For individual processes, reliability varies from 0.53 to 0.98, except for percolation where it is very low 0.11. Note that reliability is undefined for the unsaturated zone because no identifications are made. Reliability is close to perfect for mechanisms in the saturated zone and surface runoff generation processes ($R = 0.98$ and 0.94 respectively) but is notably lower for evaporation ($R = 0.73$), interflow ($R = 0.59$), and routing ($R = 0.53$).

Figure 6 provides additional context for this experiment, by displaying the range of synthetic flow index PCs generated across the 100 replicates. The synthetic replicates generally reproduce (on average) the distances between the regionalized estimates and the observation-based values.

4.4. Experiment 4: Dependence of Method Performance on Model Error (Synthetic)

Figure 7 shows a plot of F_{id} and reliability R as a function of the synthetic error dispersion factor ξ. The overall trends are that F_{id} stays relatively constant with values around 0.25, whereas R improves as ξ decreases. For example, $R = 0.6$ when $\xi = 2$ which is improved to $R = 0.95$ when $\xi = 0.05$.

For individual processes, two distinct trends emerge for F_{id}. For the unsaturated zone, percolation, saturated zone, surface runoff, F_{id} increases monotonically as ξ is reduced. In contrast, for interflow and routing, the opposite trend takes place, with F_{id} decreasing as ξ is reduced. These results are interpreted in Section 5.3.3 in terms of process identifiability and interactions. The behavior for R is more consistent: it increases as ξ is reduced—except for saturated zone and surface runoff, for which R is already very high (around 0.95) even when $\xi = 2$.

As a consequence of the divergent trends in F_{id}, the processes that are most identifiable change depending on the value of ξ. When the model error is large (high values of ξ), the most identifiable processes are interflow ($F_{id} = 0.77$) and routing ($F_{id} = 0.53$). As the model error is reduced (lower values of ξ), surface runoff emerges as the most identifiable process, though its identifiability stays relatively low at $F_{id} = 0.46$.

5. Discussion

5.1. Connection to Other Studies on Hypothesis-Testing and on Prediction in Ungauged Catchments

This work focuses on identification of dominant mechanisms in ungauged catchments, which is in contrast to existing methods focusing directly on flow prediction (e.g., Almeida et al., 2016; Bulygina et al., 2011; Prieto et al., 2019), or identification of dominant mechanisms in gauged catchments (Prieto et al., 2021). The study builds on previous work in Bayesian modeling and mechanism identification in gauged catchments (Prieto et al., 2021; Schöniger et al., 2014; Wöhling et al., 2015) and extends the methodology to ungauged catchments. The key ingredient to achieve this is the conditioning of the inference on regionalized flow indices (e.g., Blöschl et al., 2013; Coxon et al., 2014; Hrachowitz et al., 2013; Prieto et al., 2019; Westerberg et al., 2016) instead of

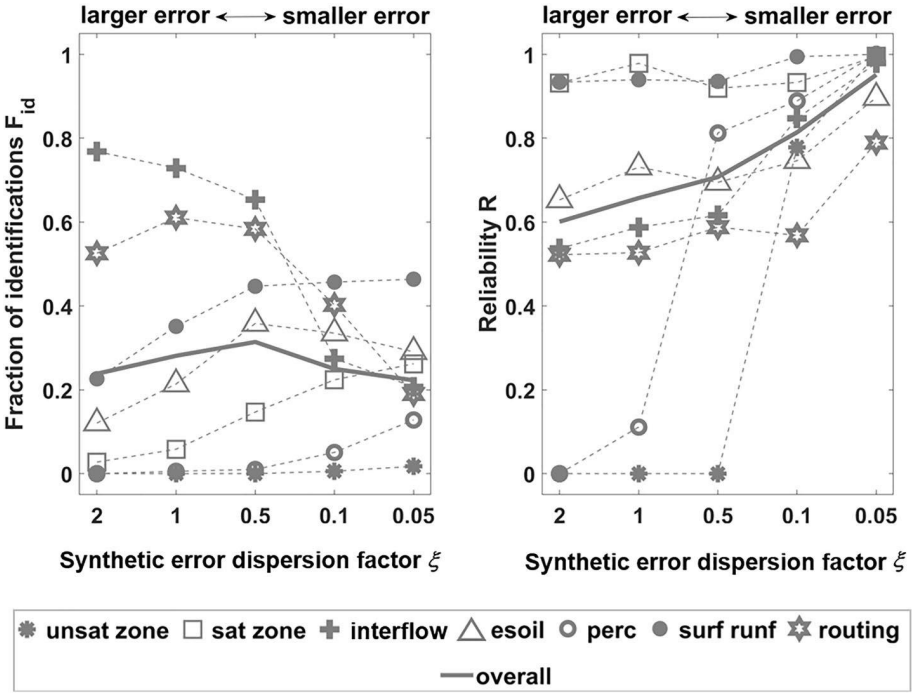

Figure 7. Performance of mechanism identification in synthetic Experiment 3, where the dependence of fraction of identifications and reliability metrics on the synthetic error factor ξ is examined. Results shown for individual processes and for the overall estimate. Analysis based on 100 synthetic replicates.

streamflow time series. In addition, parameter uncertainty is incorporated into the estimation, as it is expected to be considerably larger in ungauged catchments than in gauged catchments. These advances represent the major contribution of this study, and enable mechanism identification in a broader range of catchments.

5.2. Do Regionalization Errors Dominate Hydrological Errors?

An important finding from Experiments 1 and 2 concerns a key simplifying assumption made when deriving the mechanism identification method in the context of regionalized flow indices, namely that regionalization errors dominate hydrological errors (Section 2.2). As shown empirically in Section 3.6.2 and Figure 6, the flow index PCs obtained by the regionalization model are generally further away from the observation-based flow index PCs than flow index PCs obtained by the hydrological model. Therefore, at least for the models and data used in this work, the assumption appears justified. This finding is perhaps unsurprising because extrapolating flow indices from donor to target catchments is (arguably) an inherently harder challenge than estimating flow indices by the best of many (here, 624) hydrological models calibrated directly in the catchment of interest. This assumption was already common (explicitly or implicitly) in previous regionalization work (e.g., Almeida et al., 2016; Bulygina et al., 2012, 2009; Prieto et al., 2019; Yadav et al., 2007), and corroborating it empirically represents another contribution of this work.

It is nevertheless worth noting how close are all three estimates of the flow index PCs in many of the catchments, giving credence to the RF regionalization model. Note also that the ability to match the observation-based flow index PCs does not by itself guarantee that the underlying streamflow time series are also simulated with high accuracy (Prieto et al., 2019).

5.3. Insights on Method Performance/Interpretation of Real Scenarios Using Synthetic Insights

5.3.1. Real Data Experiments

As the true mechanisms are unknown in the experiments based on real observed data (Experiments 1 and 2), we focus on the fraction of identifications.

In Experiment 1, where the inference proceeds from regionalized flow index PCs, the first four flow index PCAs allow identifying dominant mechanisms in 27% out of 112 total trials. As such, a dominant mechanism is not identified for the majority of the processes. As will be elaborated below in Section 5.3.3, relatively poor identifiability is unsurprising in ungauged catchments, where the mechanism identification method has comparatively little data to work with, and indeed the data itself is extrapolated according to the regionalization model and hence are likely to contain appreciable uncertainty.

Mechanisms for routing are found to be the most identifiable. Mechanisms for water storage in unsaturated zone and percolation are the least identifiable, followed by mechanisms for saturated zone processes. These findings are generally consistent with the existing literature. In particular, previous work has suggested that the hardest identifiable processes are processes in the subsurface (e.g., Massmann, 2020; van Esse et al., 2013), baseflow and percolation (e.g., Coxon et al., 2014; Spieler et al., 2020).

Routing strongly affects lag time and flow variability. These characteristics are captured by flow indices such as the duration of high pulses within each year and number of days with increasing flow. In the FUSE models employed in our analysis, routing is either enabled or disabled, as opposed to other mechanisms for which alternative formulations are provided. Enabling vs. disabling routing has arguably a strong effect on the model ability to simulate flow delays and damping effects—hence routing tends to be well identified compared to other processes.

Experiment 2, where observation-based flow index PCs are used, achieves a similar fraction of identifications as Experiment 1 (25% vs. 27%) and finds the same processes as the most/least identifiable. The overall similarity between the results of Experiments 1 and 2, despite employing very different data sources, is re-assuring and increases confidence in the empirical findings in terms of method performance and process identifiability.

5.3.2. Synthetic Experiments

In Experiments 3 and 4, the true mechanisms are known and therefore both the fraction of identification F_{id} and reliability R are evaluated. The key findings are that, as the magnitude of (random) model errors decreases, overall R improves to above 0.9 for most processes, but overall F_{id} stays relatively constant at around 0.25.

The overall value of $F_{id} \approx 0.25$ is seen as relatively low and is tentatively attributed to using only 4 of 103 flow index PCs (which inherently limits the total quantity of information available). There may also be inherent limitations in mechanism identification due to factors such as the similarity of competing mechanisms, algorithm assumptions, limited number of importance samples in the estimation of BME, and so forth. Moreover, the stability in overall F_{id} masks appreciable internal variations in the identifiability of individual processes: as the error decreases the identifiability of interflow (and routing) decreases but the identifiability of unsaturated and saturated processes increases.

We attribute these variations to poor identifiability and process/mechanism interactions. For example, when multiple processes and/or mechanisms can mimic each other in reproducing the flow index PCs being fitted, we expect to see a low F_{id}. This interpretation is very plausible given conditioning on only four data points per catchment. Moreover, the presence of large errors, $\xi = 1 - 2$, makes it likelier that a particular set of flow index PCs is matched by the "wrong" mechanism(s), in which case we expect a higher rate of false positives, that is, an artificially high F_{id} and a low R. As model error is reduced, $\xi < 1$, the rate of false positives decreases, which can manifest in a reduced F_{id} and increased R. A combination of such trends is indeed seen in Figure 7. See further discussion in Section 5.3.3.

The reliability of mechanism identification for several processes is very high ($R \geq 0.95$) even when errors are large ($\xi = 1 - 2$). For the remaining processes, reliability exceeds 0.8 once errors are reduced below a certain level (here $\xi = 0.05$), that is, the method identifies the true dominant mechanism with a relatively high probability, which is seen as a valuable empirical check.

The findings also highlight that a high fraction of identifications is not necessarily indicative of correct identifications. For example, in Experiment 3, for routing, a dominant mechanism is identified in 61% of trials, but that mechanism is correct in only 53% of these latter trials. Conversely, for the saturated zone process, a dominant mechanism is identified in only 6% of trials, but this identification is correct in the vast majority, 98%, of these trials. For the saturated zone, high reliability is maintained at all error levels.

There is a broad, though not full, consistency between the most/least identifiable processes in the real vs. synthetic experiments. Percolation and unsaturated zone are always the least identifiable process. The most identifiable processes are not always the same: routing in Experiment 1 and interflow in Experiment 3 (though here routing is the second most identifiable process). The process identifiability in Experiment 2 (observation-based flow index PCs) is generally consistent with Experiment 4 when the errors are small ($\xi = 0.1 - 0.05$), with surface runoff being the most identifiable process.

5.3.3. Limited Reliability and Identifiability as a Consequence of Limited Information

The fraction of identifications and reliability found in this work is notably lower than in previous study on mechanism identification from streamflow time series in a gauged catchment (Prieto et al., 2021), where power (similar to fraction of identifications) and reliability were as high as 1 for most processes (in synthetic experiments with replication).

The relatively lower reliability and fraction of identifications found in Experiments 3 and 4, even when errors are (relatively) small, $\xi << 1$, can be attributed to two potential causes: (a) the limited quantity and type of information (here the flow index PCs); and/or (b) a limited number of importance samples when estimating the BME (Section 2.4.2 and Appendix A). These potential causes are elaborated below.

The quantity of information is highly relevant because, when only a few flow index PCs are used, it is more likely that hydrological structures with different mechanisms can generate streamflow time series that match these flow index PCs, which in turn makes it difficult or impossible to discriminate between multiple hypotheses (Ley et al., 2016) and leads to poor identifiability. Note that similar values of flow indices could be obtained by two genuinely similar time series as well as by two time series with some characteristics that are similar and others that are different. In general, a diverse set of multiple indices is needed to fully characterize catchment behavior (Euser et al., 2013). In the present case study, we regionalized the first 4 PCs of 103 flow indices, which appears insufficient to achieve an overall fraction of identifiability above about 0.25–0.3.

Also note that, depending on the catchment, the same flow indices can be more or less representative of the underlying streamflow data, as seen in Table 1. Information content of the first four flow index PCs in the 16 "ungauged" catchments—Experiments 1 and 2. In other words, these flow index PCs represent different quantity of information across the 16 study catchments, which adds further variability into the analysis. Therefore, the selection of "informative" flow indices (including type and number) may be catchment and process-dependent. Such considerations relate to the difficulty of quantifying the data information content (here flow indices), and how to extract such information in ungauged catchment lacking streamflow observations (Gupta et al., 2008; Wagener and Montanari, 2011). Note that regionalization as an overall concept effectively relies on the presumption that the available physical characteristics of a catchment are sufficient to estimate their hydrological characteristics without direct observations of streamflow time series; these considerations fall within the broader theme on "uniqueness of the place" (Beven, 2000) and continue to attract research attention.

The limited number of importance samples when approximating the BME can also contribute to poor identifiability, as well as to poor reliability. The BME estimation in our case study uses 1,000 sampled sets of hydrological model parameters per hydrological model structure, in order to manage computational costs. This number of parameter sets might be insufficient to accurately reproduce the hydrographs (and hence the flow indices and their PCs) across all catchments. For example, a set of flow index PCs might not be reproduced because the parameter set that would lead to a similar underlying hydrograph has not been sampled. In this case, the likelihood of all sampled flow index PCs might be similarly low (leading to poor identifiability) or a flow index PC spuriously close to the regionalized value (leading to poor reliability). Preliminary tests suggest that the variability of the BME estimation with 1,000 samples has a small influence on the results. For example, for catchment X1404, a bootstrap analysis over 1,000 model runs, where each run is the result of a different parameter set, suggests that the standard deviation in the fraction of identifications is $\leq$1% and the standard deviation of the reliability is $\leq$1% except for the evapotranspiration case where it is 6%; see Supporting Information S1.

Finally, as discussed in Prieto et al. (2021), the degree of difference in the competing mechanisms is also of clear relevance. If two mechanisms are very similar, it is harder to distinguish them. However, mechanism similarity does not decrease the reliability of mechanism identification, but rather reduces the fraction of identification, that is, the chance to identify a dominant mechanism. For example, distinguishing between three mechanisms will be much harder if they employ similar equations (e.g., see earlier study by Gupta & Sorooshian, 1983).

5.4. General Limitations and Future Work

The mechanism identification method assumes that the true mechanisms are included in the model ensemble **G**. This assumption is typical in statistical model identification methods (Höge et al., 2019). However, in practice, the method will be applied with imperfect models to understand real catchments, infer mechanisms and make predictions. A key question is then, how "good" should be the "best available" mechanisms/model representation to be identified as "dominant"? For example, consider a model with the (quasi) true mechanism for a single process and a poor choice of mechanisms for other processes. This model may have a low posterior probability, which in turn will lower the posterior probability of the quasi-true mechanism (since the posterior probability of a mechanism is approximated as the average of all model structures that contain such mechanism). However, if the posterior probability of the remaining models containing the (quasi) true mechanism is high, this mechanism will be identified as dominant.

A related question is the design of synthetic experiments in a way most consistent with the errors of real regionalization models. A complicating factor is that the regionalization error model is constructed by replicating the errors across multiple catchments but is applied to describe errors at a single catchment. This approximation, corresponding to the jack-knife procedure, is difficult to avoid when working with flow indices, which unlike streamflow time series have a very short length. The approximation results in difficulties separating biases and random error components in ungauged locations, as well as challenges in designing synthetic experiments that reproduce these types of error.

The assumption that the regionalization model error dominates the hydrological model error avoids the considerable challenge of disaggregating hydrological and regionalization uncertainties. Relaxing this limitation may require formulating a hydrological residual error model in the time domain but inferring it from the (estimated) flow index PCs. This kind of "mixed-domain" decompositional inference is computationally challenging and could be implemented using techniques such as Approximate Bayesian Computation (Albert et al., 2015; Kavetski et al., 2018; Nott et al., 2012; Sadegh & Vrugt, 2014). Note also that precipitation error is not considered in this work but may be considered in follow up studies.

Another key assumption in this work and many other regionalization studies (e.g., Almeida et al., 2016; Bulygina et al., 2012, 2009; Prieto et al., 2019) is that flow indices do not vary significantly in time and can be treated as an internal catchment property. This assumption can of course be limiting; however, this is a general challenge of inference and prediction that is not specific to our work. If there is a substantial temporal change in the flow indices (and hence their PCs)—or in any observed data used for model selection—then model selection for the future will necessarily be less reliable and will require extrapolation along some estimated trends. Such analysis is not within the scope of this work. Furthermore, in this article, we assume the mechanisms are stationary over time. A further step would be to update the regionalization model to reflect any climate change differences. For example, this could be implemented by applying the methods presented in this article independently to multiple time periods.

The generalization of the analysis to use different numbers of flow index PCs for each catchment, and to vary these in a similar way to the errors in Experiment 4 would shed additional light into the reasons for the relatively low fraction of identifiability. However, such analysis is challenging to implement in a way that maintains correspondence to the real data analysis and may require constructing dozens or hundreds of residual errors models using the jack-knife approach for different numbers of flow index PCs. Hence this analysis is deferred to future work.

The synthetic studies suggest that dominant mechanisms can be identified (relatively) reliably even from limited and highly uncertain information, in particular from a limited number of regionalized flow index PCs. The following questions arise: Which flow index PCs contain information allowing to constrain process representation in a given catchment? How much and what hydrological information (in the form of flow index PCs) needs to be assimilated into a hydrological model (Markstrom et al., 2016)? What is the impact of using a limited number of PCs vs. using the streamflow time series in the identification of dominant mechanisms and vs. using the flow indices with highest weights in the PCs?

There is an inherent difficulty in identifying processes from flow indices alone since these quantities mask short-scale variability. For example, vegetation dynamics is expected to affect the storage and flux of water in

the unsaturated zone; however, this dynamic is not represented by the flow indices. Augmenting the flow indices with remote sensed data may increase model identification in this case (Wagener & Montanari, 2011). For example, estimates of evaporation based on satellite data could be used in the conditioning procedure (Winsemius et al., 2009).

Finally, PUB is predicated on the idea of reducing uncertainty through hydrological process understanding (Wagener & Montanari, 2011). Before using a fixed pre-selected model, it seems beneficial to estimate the model structure. In this work, we present an approach that tries to help in that direction, by identifying dominant mechanisms most likely to represent specific hydrological processes. An interesting study to be undertaken in future work is a comparison between the predictive capacity of model structures that include the dominant mechanisms with respect to common (fixed) model structures.

6. Conclusions

This study explores the representation of hydrological processes in ungauged catchments, where streamflow observations are not available. The focus is on the identification of dominant hydrological mechanisms, that is, mechanisms that are more a posteriori probable to represent a given process. A new method is proposed by combining a Bayesian mechanism identification method introduced in recent work on gauged catchments with advances in methods for flow prediction in ungauged catchments.

The method is illustrated using real and synthetic experiments using data from 92 catchments in northern Spain, from which 16 catchments are treated as ungauged. We use 624 hydrological model structures from the hydrological modeling system "Framework for Understanding Structural Error" (FUSE), which represent a total of 7 hydrological processes. The synthetic experiments illustrate how the magnitude of hydrological and regionalization model error impact mechanism identification.

The key findings are as follows:

1. Bayesian identification of dominant mechanisms can be based on flow indices regionalized from gauged to ungauged catchments. Here, following previous work on regionalization, the flow indices are represented in PC space, keeping the first 4 PCs out of a total of 103 based on information content analysis. In its current form, the mechanism identification method is implemented under the assumption that regionalization model errors are larger than hydrological model errors, in order to avoid an uncertainty decomposition problem. This assumption is shown to be reasonable at least for the current selection of models and data
2. In the real data experiment, the average fraction of identifications across the 16 catchments and 7 hydrological processes is 0.27. The process for which a dominant mechanism is most identified is routing, with an average fraction of identifications of 0.94; the processes for which a mechanism is least identified are percolation and the unsaturated zone, for which dominant mechanisms are not identified in any of the 16 catchments
3. In synthetic experiments, where the true mechanisms are known and the same error model is used as in the real data experiment, the mechanism identification method achieves a reliability of 0.66. As expected a priori, this value is lower than in previous work in gauged catchments, where streamflow time series are available Prieto et al. (2021). The loss in reliability is attributed to the fundamentally reduced and uncertain quantity of information: four data points per catchment (first 4 flow index PCs), moreover corrupted with (random) errors. The overall fraction of identifiability, estimated at 0.28, is very close to the value achieved in the real data experiment
4. The magnitude of model error impacts primarily on reliability, which increases from 0.6 when model error is "large" (error dispersion multiplier of 2 relative to Experiment 1) to 0.95 when model error is "small" (error dispersion multiplier of 0.05). The overall fraction of identifications remains stable at around 0.22–0.31, though for individual processes it varies depending on the error level, suggesting interactions between the identified mechanisms. The relatively low fraction of identifications and the evidence of interactions are attributed to the reduced quantity of information

Future work envisages a more complete treatment of uncertainty using a probabilistic hydrological model (i.e., explicitly distinguishing between hydrological vs. regionalization model errors), understanding the impact of even the "best" mechanisms being approximations of the actual hydrological processes, and understanding the limits on mechanism identification imposed by the use of (a limited number of) regionalized flow index PCs

instead of streamflow time series. These research questions are of particular importance for model identification in ungauged catchments, where fundamentally less information is available than in gauged catchments.

Appendix A: Estimation of Bayesian Model Evidence (BME) Using Importance Sampling

The mechanism identification framework requires estimates of BME. This section describes the estimation on this quantity using Monte Carlo integration.

For a given hydrological model structure and catchment, BME is:

$$p\left(\hat{\mathbf{z}}^{\text{reg}}|h_k\right) = \int_{\Omega_{h(k)}} p\left(\hat{\mathbf{z}}^{\text{reg}}|\boldsymbol{\theta}_{h(k)}, h_k\right) p\left(\boldsymbol{\theta}_{h(k)}|h_k\right) \mathrm{d}\boldsymbol{\theta}_{h(k)} \tag{A1}$$

In this work, we approximate BME using Monte Carlo integration with importance sampling (Kuczera & Parent, 1998; Prieto et al., 2019; Schöniger et al., 2014).

The following computations are implemented:

1. Draw N_{imp} parameter sets $\left\{\boldsymbol{\theta}^s_{h(k)}; s = 1, \ldots, N_{\text{imp}}\right\}$ from the uniform prior distribution $p\left(\boldsymbol{\theta}_{h(k)}|h_k\right)$
2. Run the hydrological model h_k with each sampled parameter set $\boldsymbol{\theta}^s_{h(k)}$ to generate N_{imp} streamflow time series $\left\{\mathbf{q}^{\text{sim}}_s; s = 1, \ldots, N_{\text{imp}}\right\}$ and project the latter to PC space to obtain the corresponding flow index PCs, $\left\{\mathbf{z}^{\text{sim}}_s; s = 1, \ldots, N_{\text{imp}}\right\}$
3. Compute the un-scaled weight $\pi_s = p\left(\hat{\mathbf{z}}^{\text{reg}}|\boldsymbol{\theta}^s_{h(k)}\right)$ for each parameter set $\boldsymbol{\theta}^s$ using the likelihood function in Equation 5
4. Sum π_s over all sampled parameter sets

$$\Pi_k = \sum_{s=1}^{N_{\text{imp}}} \pi^k_s \tag{A2}$$

5. The BME is then given by scaling (normalizing) the weights so they add up to 1

$$p\left(\hat{\mathbf{z}}^{\text{reg}}|h_k\right) = \frac{\Pi_k}{\sum_{i=1}^{N_h} \Pi_i} \tag{A3}$$

1. The approximation error in this procedure depends on the number of parameter sets, N_{imp}, sampled in step a
2. The procedure is used to estimate the BME of each hydrological model h_k in the ensemble $\mathbf{G}$. The estimated BMEs are then used to estimate mechanism probabilities as described in Section 2.4.3

Appendix B: Identification of a Dominant Mechanism Using a Bootstrap Approach

This section describes the steps used to test whether a mechanism is dominant. The null hypothesis is that there is no dominant mechanism.

The null hypothesis $\text{H0}^{\wp}_k$ is tested using the procedure in (Prieto et al., 2021) as follows:

1. Sample with replacement $N^{\wp}_k$ model structures with mechanism $m^{\wp}_k$, where $N^{\wp}_k$ is the number of models with mechanism $m^{\wp}_k$ in the sample space $\mathbf{G}$
2. Denote this "bootstrapped" ensemble of model structures as $\mathbf{G}^{(b)}$
3. Calculate $p(m^{\wp}_k|\mathbf{z}^{\text{reg}}, \mathbf{G}^{(b)})$ using Equation 6
4. Calculate $p(m^{\wp}_i|\mathbf{z}^{\text{reg}}, \mathbf{G}^{(b)})$ for all other mechanisms available for process $\wp$, $i = 1, \ldots, N^{\wp}_m \cap i \neq k$, also using Equation 6
5. Repeat steps a–c for $b = 1, ..., N^{\text{boot}}$ as illustrated schematically in Figure 1 row f. In this study, we set $N^{\text{boot}} = 10,000$, that is, 10,000 bootstrapped model ensembles are generated
6. Calculate $t^{\wp(b)}_k$ for each bootstrapped ensemble $b = 1, ..., N^{\text{boot}}$ using Equation 7
7. Calculate the empirical frequency of $t^{\wp}_k > \tau$ across all bootstrapped ensembles

$$\omega_k^{\wp} = \frac{1}{N^{\text{boot}}} \text{count} \left\{ t_k^{\wp(b)} > \tau; b = 1, \ldots, N^{\text{boot}} \right\} \tag{B1}$$

where the function count$v$ is defined as the number of true elements in a Boolean set v;

8. Reject $\text{H0}_k^{\wp}$, that is, identify $m_k^{\wp}$ as dominant, if:

$$\omega_k^{\wp} \geq 1 - \alpha_{\text{Bonf}} \tag{B2}$$

where $\alpha_{\text{Bonf}} = \alpha / N_m^{\wp}$ is the Bonferroni correction to the prescribed significance level α. Otherwise $\text{H0}_k^{\wp}$ is not rejected.

Steps 6–8 are repeated for all mechanisms $\left\{ m_k^{\wp}; k = 1, \ldots, N_{\text{m}}^{\wp} \right\}$ proposed for process $\wp$. If none of the individual null hypotheses $\left\{ \text{H0}_k^{\wp}; k = 1, \ldots, N_{\text{m}}^{\wp} \right\}$ are rejected, then the null hypothesis $\text{H0}^{\wp}$ for the entire family of comparisons is not rejected, and 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 $\wp = 1, \ldots, N^{\wp}$. See Prieto et al. (2021) for further details.

Appendix C: Case Study Details

C1. Regionalization Residual Error Model Parameters

Table C1 lists the ranges of mean and standard deviation of the residual error distributions of the regionalization model from Section 3.2. These values are reproduced from Table 4 in Prieto et al. (2019).

C2. Selection of the "True" Hydrological Model

The true model for Experiments 3–5 is selected separately for each ungauged catchment, as follows. Consider all model structures, and all 1,000 parameter sets per model structure generated as described in Section 3.4. For each hydrological model structure and parameter set, we calculate the normalized distance between observed and simulated flow index PCs,

$$\zeta(\boldsymbol{\theta}_h; h) = \sqrt{\sum_{i=1}^{N_z} \frac{\left(\tilde{\mathbf{z}}_i - \mathbf{z}_i^{\text{sim}(\boldsymbol{\theta}_h)} \right)^2}{\text{var}[\mathbf{z}_i^{\text{sim}(\boldsymbol{\theta}_h)}]}} \tag{C1}$$

where $\text{var}[\mathbf{z}_i^{\text{sim}(\boldsymbol{\theta}_h)}]$ denotes the variance of the ith flow index PC computed from the 1,000 streamflow time series simulations obtained using model structure h. The true hydrological model structure $\breve{\mathbf{h}}$ is taken as the model structure that (for one of its parameter sets) achieves the lowest distance ζ across all models and their respective parameter sets. The "best" parameter set $\breve{\boldsymbol{\theta}}_h$ is treated as the "true" parameter set.

Table C1
Ranges of Mean and Standard Deviation of the Residual Error Distributions of the Regionalization Model for the 16 "Ungauged" Catchments

Flow index PC	Range of values of the mean (μ)	Range of values of the standard deviation (σ)
z_1	4.13–4.72	4.51–4.86
z_2	0.14–0.20	1.34–2.37
z_3	0.02–0.05	1.65–2.71
z_4	3.52–3.79	1.84–1.93

Note. Table adapted from Table 4 in Prieto et al. (2019).

Data Availability Statement

The study data are deposited in https://doi.org/10.5281/zenodo.5774699.

Acknowledgments
The lead author acknowledges the financial support from the Government of Cantabria through the FÉNIX Program (ID 2020.03.03.322B.742.09). The authors thank the Spanish Meteorological Agency (AEMET), Confederación Hidrográfica del Cantábrico, Confederación Hidrográfica del Ebro, Agencia Vasca del Agua, Agencia Catalana del Agua, Diputación Foral, and Gobierno de Navarra for providing the observed hydrological data used in this work. The authors thank Francisco Jesús Peñas Silva for sharing the hydrological indices and catchments descriptors used in this study. The authors thank Nans Addor for insightful discussions on the topic of mechanism identification based on catchment descriptors and hydrological indices. The authors thank Raúl Medina for feedback on an earlier version of the manuscript.

References

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

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

Albert, C., Künsch, H. R., & Scheidegger, A. (2015). A simulated annealing approach to Approximate Bayes Computations. *Statistics and Computing*, *25*(6), 1217–1232. https://doi.org/10.1007/s11222-014-9507-8

Almeida, S., Le Vine, N., McIntyre, N., Wagener, T., & Buytaert, W. (2016). Accounting for dependencies in regionalized signatures for predictions in ungauged catchments. *Hydrology and Earth System Sciences*, *20*(2), 887–901. https://doi.org/10.5194/hess-20-887-2016

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. https://doi.org/10.1016/S0022-1694(02)00101-4

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

Beven, K., & Lane, S. (2019). Invalidation of models and fitness-for-purpose: A rejectionist approach. In C.Beisbart, & N. J.Saam (Eds.), *Computer simulation validation: Fundamental concepts, methodological frameworks, and philosophical perspectives* (pp. 145–171). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-70766-2_6

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., & Savenije, H. H. G. (2013). *Runoff prediction in ungauged basins: Synthesis across processes, places, and scales*. Cambridge: Cambridge University Press.

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*(6). https://doi.org/10.1029/2011wr011207

Bulygina, N., McIntyre, N., & Wheater, H. (2009). Conditioning rainfall-runoff model parameters for ungauged catchments and land management impacts analysis. *Hydrology and Earth System Sciences*, *13*(6), 893–904. https://doi.org/10.5194/hess-13-893-2009

Bulygina, N., McIntyre, N., & Wheater, H. (2011). Bayesian conditioning of a rainfall-runoff model for predicting flows in ungauged catchments and under land use changes. *Water Resources Research*, *47*(2), W02503. https://doi.org/10.1029/2010wr009240

Clark, M. P., McMillan, H. K., Collins, D. B. G., Kavetski, D., & Woods, R. A. (2011). Hydrological field data from a modeler'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., et al. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, *51*(4), 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2007WR006735

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrological behavior 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 modeling framework. *Environmental Modeling & Software*, *129*, 104728. https://doi.org/10.1016/j.envsoft.2020.104728

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Elkan, C. (2001). The foundations of cost-sensitive learning. *Paper presented at International Joint Conference on Artificial Intelligence*. Lawrence Erlbaum Associates Ltd.

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: Theory and comparison to existing applications. *Water Resources Research*, *54*(6), 4059–4083. https://doi.org/10.1002/2017wr020528

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, *47*(11). https://doi.org/10.1029/2010wr010174

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

Goswami, M., O'Connor, K. M., & Bhattarai, K. P. (2007). Development of regionalization procedures using a multi-model approach for flow simulation in an ungauged catchment. *Journal of Hydrology*, *333*(2–4), 517–531. https://doi.org/10.1016/j.jhydrol.2006.09.018

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

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

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., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

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., & Fenicia, F. (2011). Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research*, *47*(11), https://doi.org/10.1029/2011WR010748

Kavetski, D., Fenicia, F., Reichert, P., & Albert, C. (2018). Signature-domain calibration of hydrological models using Approximate Bayesian Computation: Empirical analysis of fundamental properties. *Water Resources Research*, *54*(6), 3958–3987. https://doi.org/10.1002/2017wr021616

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

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

Kraft, P., Vaché, K. B., Frede, H.-G., & Breuer, L. (2011). CMF: A hydrological programming language extension for integrated catchment models. *Environmental Modeling & Software*, *26*(6), 828–830. https://doi.org/10.1016/j.envsoft.2010.12.009

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kuczera, G., & Parent, E. (1998). Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm. *Journal of Hydrology*, *211*(1), 69–85. https://doi.org/10.1016/S0022-1694(98)00198-X

Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., et al. (2019). Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1,000 catchments in Great Britain. *Hydrology and Earth System Sciences*, *23*(10), 4011–4032. https://doi.org/10.5194/hess-23-4011-2019

Ley, R., Hellebrand, H., Casper, M. C., & Fenicia, F. (2016). Comparing classical performance measures with signature indices derived from flow duration curves to assess model structures as tools for catchment classification. *Hydrology Research*, *47*(1), 1–14. https://doi.org/10.2166/nh.2015.221

Markstrom, S. L., Hay, L. E., & Clark, M. P. (2016). Towards simplification of hydrologic modeling: Identification of dominant processes. *Hydrology and Earth System Sciences*, *20*(11), 4655–4671. https://doi.org/10.5194/hess-20-4655-2016

Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. *Water Resources Research*, *41*(10). https://doi.org/10.1029/2004wr003719

Massmann, C. (2020). Identification of factors influencing hydrologic model performance using a top-down approach in a large number of U.S. catchments. *Hydrological Processes*, *34*(1), 4–20. https://doi.org/10.1002/hyp.13566

McDonnell, J. J. (2003). Where does water go when it rains? Moving beyond the variable source area concept of rainfall-runoff response. *Hydrological Processes*, *17*(9), 1869–1875. https://doi.org/10.1002/hyp.5132

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

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*(2), e2019WR024918. https://doi.org/10.1029/2019wr024918

Nott, D. J., Marshall, L., & Brown, J. (2012). Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resources Research*, *48*(12). https://doi.org/10.1029/2011WR011128

Olden, J. D., & Poff, N. L. (2003). Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. *River Research and Applications*, *19*(2), 101–121. https://doi.org/10.1002/rra.700

Peñas, F. J., Barquín, J., Snelder, T. H., Booker, D. J., & Álvarez, C. (2014). The influence of methodological procedures on hydrological classification performance. *Hydrology and Earth System Sciences*, *18*(9), 3393–3409. https://doi.org/10.5194/hess-18-3393-2014

Perrin, C., Michel, C., & Andréassian, V. (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, *242*(3), 275–301. https://doi.org/10.1016/S0022-1694(00)00393-0

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

Prieto, C., Le Vine, N., Kavetski, D., García, E., & Medina, R. (2019). Flow prediction in ungauged catchments using probabilistic random forests regionalization and new statistical adequacy tests. *Water Resources Research*, *55*(5), 4364–4392. https://doi.org/10.1029/2018wr023254

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

Sadegh, M., & Vrugt, J. A. (2014). Approximate Bayesian Computation using Markov chain Monte Carlo simulation: DREAM$_{(ABC)}$. *Water Resources Research*, *50*(8), 6767–6787. https://doi.org/10.1002/2014WR015386

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*(9), 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*(12), 9484–9513. https://doi.org/10.1002/2014WR016062

Shafii, M., & Tolson, B. A. (2015). Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research*, *51*(5), 3796–3814. https://doi.org/10.1002/2014WR016520

Sivapalan, M. (2003). Prediction in ungauged basins: A grand challenge for theoretical hydrology. *Hydrological Processes*, *17*(15), 3163–3170. https://doi.org/10.1002/hyp.5155

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

Snelder, T. H., Datry, T., Lamouroux, N., Larned, S. T., Sauquet, E., Pella, H., & Catalogne, C. (2013). Regionalization of patterns of flow intermittence from gauging station records. *Hydrology and Earth System Sciences*, *17*(7), 2685–2699. https://doi.org/10.5194/hess-17-2685-2013

Spieler, D., Mai, J., Craig, J. R., Tolson, B. A., & Schütze, N. (2020). Automatic model structure identification for conceptual hydrologic models. *Water Resources Research*, *56*(9), e2019WR027009. https://doi.org/10.1029/2019wr027009

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

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

Vitolo, C., Wells, P., Dobias, M., & Buytaert, W. (2016). FUSE: An R package for ensemble hydrological modeling. *Journal of Open Source Software*, *1*, 52. https://doi.org/10.21105/joss.00052

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*(1). https://doi.org/10.1029/2005WR004838

Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research*, *47*(6). https://doi.org/10.1029/2010wr009469

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*(3), 1847–1865. https://doi.org/10.1002/2015wr017635

Winsemius, H. C., Schaefli, B., Montanari, A., & Savenije, H. H. G. (2009). On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, *45*(12). https://doi.org/10.1029/2009WR007706

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*(4), 2825–2846. https://doi.org/10.1002/2014wr016292

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, *30*(8), 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

Ye, M., Meyer, P. D., & Neuman, S. P. (2008). On model selection criteria in multimodel analysis. *Water Resources Research*, *44*(3). https://doi.org/10.1029/2008WR006803

Zhang, Z., Wagener, T., Reed, P., & Bhushan, R. (2008). Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2008wr006833