Water Resources Research

RESEARCH ARTICLE
10.1029/2018WR023254

**Key Points:**
- Regionalized principal components (PCs) of flow indices are used in Bayesian framework to condition flow predictions in ungauged catchments
- A priori adequacy of hydrological model with respect to regionalized flow index PCs is indicative of the ability to predict flow time series
- In a 16-catchment study, the hydrological model and its inputs contribute more uncertainty than the flow index PC regionalization model

**Supporting Information:**
- Supporting Information S1

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

**Citation:**
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*, 4364–4392. https://doi.org/10.1029/2018WR023254

Received 7 MAY 2018
Accepted 4 MAR 2019
Accepted article online 13 MAR 2019
Published online 29 MAY 2019

# Flow Prediction in Ungauged Catchments Using Probabilistic Random Forests Regionalization and New Statistical Adequacy Tests

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

[1]Environmental Hydraulics Institute "IHCantabria", Universidad de Cantabria, Santander, Spain, [2]Department of Civil and Environmental Engineering, Imperial College London, London, UK, [3]Department of Civil Engineering, Bristol University, Bristol, UK, [4]Swiss Re, Armonk, NY, USA, [5]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia

**Abstract** Flow prediction in ungauged catchments is a major unresolved challenge in scientific and engineering hydrology. This study attacks the prediction in ungauged catchment problem by exploiting advances in flow index selection and regionalization in Bayesian inference and by developing new statistical tests of model performance in ungauged catchments. First, an extensive set of available flow indices is reduced using principal component (PC) analysis to a compact orthogonal set of "flow index PCs." These flow index PCs are regionalized under minimal assumptions using random forests regression augmented with a residual error model and used to condition hydrological model parameters using a Bayesian scheme. Second, "adequacy" tests are proposed to evaluate a priori the hydrological and regionalization model performance in the space of flow index PCs. The proposed regionalization approach is applied to 92 northern Spain catchments, with 16 catchments treated as ungauged. It is shown that (1) a small number of PCs capture approximately 87% of variability in the flow indices and (2) adequacy tests with respect to regionalized information are indicative of (but do not guarantee) the ability of a hydrological model to predict flow time series and are hence proposed as a prerequisite for flow prediction in ungauged catchments. The adequacy tests identify the regionalization of flow index PCs as adequate in 12 of 16 catchments but the hydrological model as adequate in only 1 of 16 catchments. Hence, a focus on improving hydrological model structure and input data (the effects of which are not disaggregated in this work) is recommended.

## 1. Introduction

Flow prediction in ungauged catchments remains an elusive challenge in hydrological sciences and engineering, even with the advances achieved during the "predictions in ungauged basins decade" (Hrachowitz et al., 2013; Sivapalan et al., 2003; Smith et al., 2014). Meeting this challenge largely depends on the ability to successfully extrapolate hydrological information from gauged to ungauged catchments, a process often referred to as "regionalization" in the hydrological literature (e.g., Blöschl & Sivapalan, 1995; Oudin et al., 2010; Gottschalk, 1985; Riggs, 1973; Wagener & Wheater, 2006; Young, 2006).

Traditionally, regionalization proceeds in terms of model parameters, which are calibrated in gauged catchments and then transferred to ungauged catchments according to assumed relationships between model parameters and catchment characteristics (see Pechlivanidis et al., 2010; Samaniego et al., 2010; Wagener et al., 2004; see He et al., 2011, for a review of the corresponding methods). Parameter regionalization has several drawbacks, including the following: (i) hydrological model parameters often suffer from poor identifiability and strong interdependencies (Bulygina et al., 2012; McIntyre et al., 2005) and (ii) parameter regionalization relationships are difficult or impossible to derive (Blöschl et al., 2013), in no small part due to difficulties in establishing correspondence between model parameters and physical catchment attributes (Duan et al., 2006; Koren et al., 2003; see also Fenicia et al., 2014).

A more recent approach to regionalization seeks to extrapolate the hydrological characteristics of a catchment rather than its fitted hydrological model parameters (see Wagener & Montanari, 2011, and Razavi & Coulibaly, 2013, for a review). In this approach, catchment characteristics (descriptors) such as climate, topography, geology, soils, and vegetation are related via a regionalization model to a set of hydrological (flow) indices or signatures calculated in gauged catchments from observed data (Almeida et al., 2013; Almeida, 2014; Almeida et al., 2016; Blöschl et al., 2013; Bulygina et al., 2009, 2011; Bulygina et al., 2012; Gupta et al., 2008; Hrachowitz et al., 2013; Hrachowitz et al., 2013; Olden & Poff, 2003; Sawicz et al., 2011; Wagener & Montanari, 2011; Yadav et al., 2007; Zhang et al., 2008). Examples of flow indices include average annual and monthly flows (Peñas et al., 2014), runoff coefficient (Almeida et al., 2016), quantiles and slope of the flow duration curves (Westerberg et al., 2014; Yilmaz et al., 2008), and the base flow index (Bulygina et al., 2009). The regionalized relationship is used to estimate flow indices in an ungauged location based on its catchment descriptors, and these estimated indices are then used to infer (condition) both hydrological model parameters and predictions. This regionalization strategy is based on the assumption that catchments that are similar physiographically and climatologically are also similar hydrologically.

Regionalized indices are affected by several sources of uncertainties (Almeida, 2014; Westerberg et al., 2016), due to the limited number of gauged catchments, limited quantity and quality of both dynamic observations and catchment descriptors, and the simplified nature of the regionalization model relationships. The uncertainty in the regionalized flow indices results in uncertainty in the estimated hydrological model parameters and predictions. Hence, reliable and precise regionalization of flow indices is recognized as one of the main challenges of the prediction in ungauged catchment (Hrachowitz et al., 2013; Sivapalan et al., 2003). If such regionalized indices were available, the Bayesian paradigm offers promising ideas and techniques to quantify and reduce the uncertainty in the hydrological parameters and flow predictions (Blöschl et al., 2013; Bulygina et al., 2009; Singh, 2013; Yadav et al., 2007).

Like any inference technique, Bayesian methods rely on multiple modeling choices, including the specification of (1) flow indices, (2) regionalization procedure, and (3) hydrological model. From the multitude of possible indices, the hydrologist must select those that capture best the key characteristics of the flow regime. For example, Olden and Poff (2003) used 171 indices that represent five aspects of the flow regime: average annual and monthly flows, high and low flows, duration and frequency of high flows, rate of change in flows, and time of maximum and minimum flow events. They demonstrated the ability of principal component (PC) analysis (PCA) to efficiently summarize the information (variability) in the observed indices. More recently, Yadav et al. (2007) considered 39 indices, which were divided into seven classes by means of linear and Spearman rank correlation coefficients; Coxon et al. (2014) used 3 signatures to evaluate modeled flow behavior over decadal, annual, and monthly time scales; Westerberg et al. (2016) used 15 flow indices to quantify the uncertainty coming from the discharge data and propagated these uncertainties into the regionalization of the indices; Almeida et al. (2012, 2016) combined information from multiple regionalized signatures to condition rainfall-runoff models in ungauged catchments.

To regionalize flow indices to ungauged catchments, linear regression relationships are usually fitted between flow indices (e.g., runoff ratio, base flow index, flow elasticity, slope of flow duration curve, high pulse count, etc) and catchment characteristics (e.g., average annual flow, average annual precipitation, average annual potential evapotranspiration [PET], aridity index, average elevation, etc) in gauged catchments (Almeida, 2014; Almeida et al., 2016; Yadav et al., 2007; Zhang et al., 2008). As the regionalization relationships are usually nonlinear, regression models based on the random forests (RF) technique have been proposed (Peñas, 2013; Snelder et al., 2009, 2013). RF regression extends the concept of a regression tree (Breiman et al., 1984), a machine learning technique that can be used to relate a set of predictors (here catchment descriptors) to a predictand (here a single flow index), from a single tree (Breiman et al., 1984) to a set of trees (Breiman, 2001). RF regression retains the advantages of a single regression tree (flexibility to accommodate different data patterns and error distributions), offers an improvement in accuracy, and is more robust with respect to the selection of predictors than a single regression tree (Snelder et al., 2013). RF regression has been successfully applied by the ecohydrological community to regionalize flow indices and to explain variations in hydrological patterns (e.g., Booker & Snelder, 2012; Booker, 2013; Peñas, 2013; Snelder et al., 2013). It has also proven effective for predicting combinations of flow indices compared with other machine learning algorithms (Peñas et al., 2014) and physically based approaches (Booker & Woods, 2014). However, the previous studies have not explored the use of RF regression within a probabilistic framework to predict flow dynamics in ungauged catchments.

The use of hydrological models conditioned on regionalized flow indices rests on the following two assumptions: (i) the regionalization model is capable of estimating flow characteristics from other sources of information and (ii) the hydrological model is capable of characterizing hydrological dynamics in the catchment of interest (Yadav et al., 2007; Bulygina et al., 2009, 2011; Almeida et al., 2012; Almeida, 2014). Such modeling assumptions require careful testing, as demonstrated in previous studies that explored model structure adequacy (Bulygina & Gupta, 2010; Clark et al., 2008; Clark et al., 2011; Fenicia et al., 2008; Fenicia et al., 2011; Wagener et al., 2001), the quality of hydrological observations used to drive and evaluate models (Beven & Westerberg, 2011; Kavetski et al., 2002, 2006; McMillan et al., 2012; Renard et al., 2011; Westerberg & Birkel, 2015), and the quality of regionalization procedures (Beven, 2000; Wagener & Montanari, 2011). A key complication in estimating the quality of flow predictions in ungauged catchments is that observed flow data typically used in posterior diagnostics and model verification (e.g., Gneiting et al., 2007; McInerney et al., 2017; Schoups & Vrugt, 2010, and many others) are not available. Hence, practical methods for flow prediction in ungauged catchments must either (i) assume error characteristics estimated using leave-one-out strategies on gauged sites are applicable at the ungauged site of interest (Almeida et al., 2016) or (ii) employ performance tests based solely on the conditioning data available (e.g., regionalized flow indices) and assume that these tests are indicative of the quality of predicted flow time series (e.g., Gupta et al., 2008). In addition, for practical reasons, it is of interest to develop model tests that are applied a priori rather than a posteriori with respect to the conditioning of hydrological model parameters—that is, being able to assess the adequacy of a model before estimating its parameters. If a model is found inadequate a priori and rejected, the modeler is spared the effort in conditioning the model parameters, which can be a substantial saving when the conditioning is implemented using computationally expensive Monte Carlo techniques. The ability to diagnose model adequacy a priori can hence help inform the selection of models for specific sites, help identify dominant sources of uncertainty, and so forth.

This study has the following objectives:

1. to incorporate PC-based methods, as well as RF regression techniques for regionalization, into a Bayesian framework to condition hydrological model parameters and flow predictions in ungauged catchments;
2. to develop statistical tests to evaluate a priori the adequacy of hydrological and regionalization models in representing flow indices, in PC space, available in a catchment; and
3. to empirically assess the behavior of the proposed model adequacy tests under different model quality and data availability scenarios and explore the extent to which model performance in flow index PC space is representative of the model's ability to predict flow time series.

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

## 2. Theoretical Development

The proposed approach for flow prediction in ungauged catchments using a hydrological model comprises the following key steps. First, a set of flow indices available in gauged catchments is summarized in PC space. Second, a regionalization model that relates catchment descriptors to catchment flow indices (in PC space) is developed using the RF regression technique supplemented with a residual error model to describe regionalization model uncertainty. Third, the regionalization model is applied to ungauged locations to predict flow index PCs from available catchment descriptors. Fourth, the combination of the regionalization model and the hydrological model is evaluated in flow index PC space via new proposed statistical adequacy tests, which are assumed to provide an indirect indication of potential model performance in the flow time series space. The term *model* includes the model structure, parameters, and inputs. In the specific case of the RF model, the term *model* refers to its functional form, parameters, and the set of selected predictors (here the catchment descriptors). Fifth, if the adequacy tests are passed, the hydrological model parameters are conditioned on the regionalized flow index PCs and the model used for flow prediction; conversely, if the adequacy tests are failed, the model cannot be expected to provide trustworthy flow predictions and should be replaced or enhanced. This section provides a detailed description of each step above; Figure 1 provides a summary illustration.

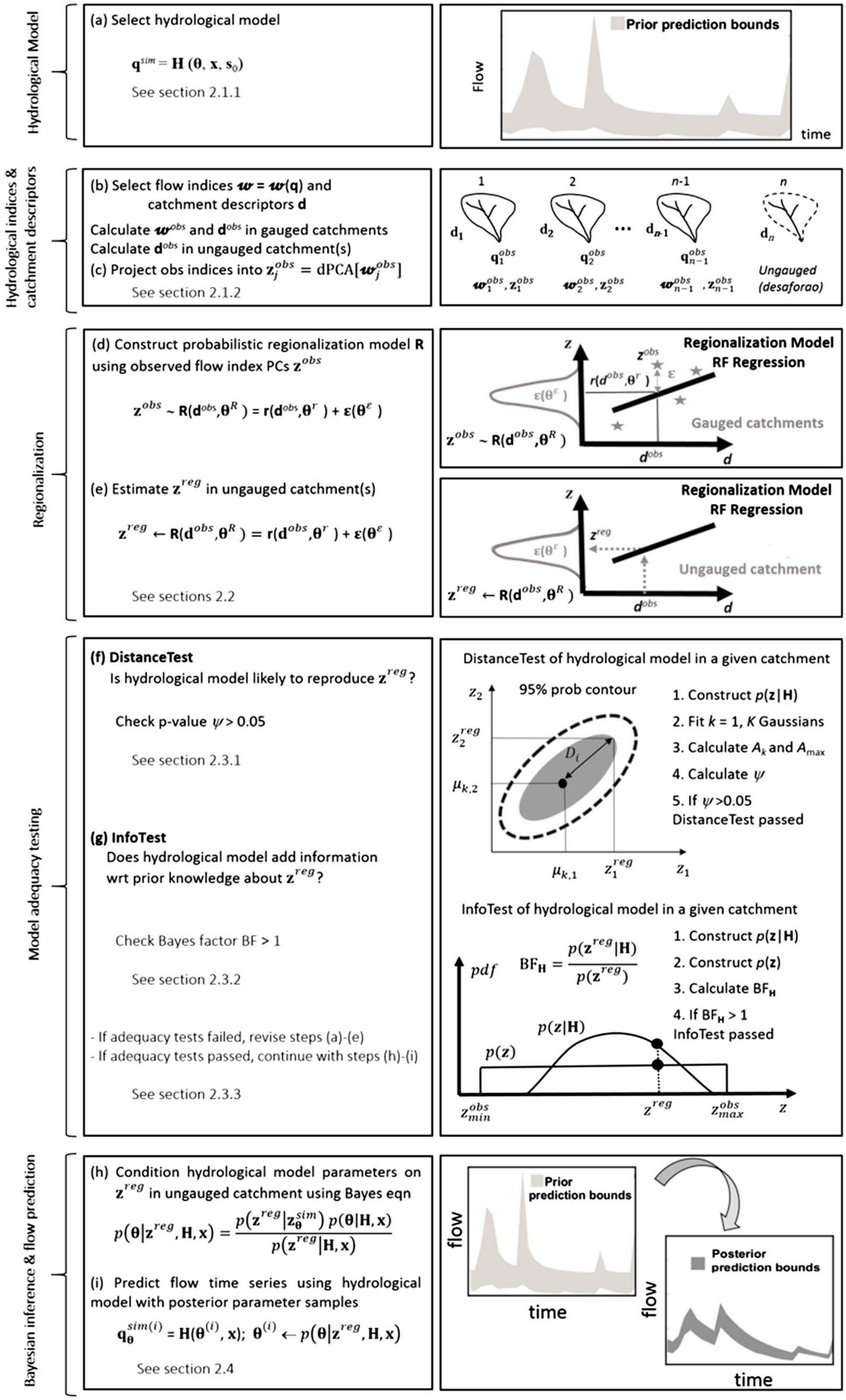

Hydrological Model
(a) Select hydrological model
$\mathbf{q}^{sim} = \mathbf{H}(\boldsymbol{\theta}, \mathbf{x}, \mathbf{s}_0)$
See section 2.1.1
Prior prediction bounds
Flow
time
Hydrological indices & catchment descriptors
(b) Select flow indices $\boldsymbol{w} = \boldsymbol{w}(\mathbf{q})$ and catchment descriptors $\mathbf{d}$
Calculate $\boldsymbol{w}^{obs}$ and $\mathbf{d}^{obs}$ in gauged catchments
Calculate $\mathbf{d}^{obs}$ in ungauged catchment(s)
(c) Project obs indices into $\mathbf{z}_j^{obs} = \mathrm{dPCA}[\boldsymbol{w}_j^{obs}]$
See section 2.1.2
1
2
n-1
n
$\mathbf{d}_1$
$\mathbf{d}_2$
$\mathbf{d}_{n\text{-}1}$
$\mathbf{d}_n$
$\mathbf{q}_1^{obs}$
$\mathbf{q}_2^{obs}$
$\mathbf{q}_{n-1}^{obs}$
$\boldsymbol{w}_1^{obs}, \mathbf{z}_1^{obs}$
$\boldsymbol{w}_2^{obs}, \mathbf{z}_2^{obs}$
$\boldsymbol{w}_{n-1}^{obs}, \mathbf{z}_{n-1}^{obs}$
Ungauged (desaforao)
Regionalization
(d) Construct probabilistic regionalization model $\mathbf{R}$ using observed flow index PCs $\mathbf{z}^{obs}$
$\mathbf{z}^{obs} \sim \mathbf{R}(\mathbf{d}^{obs}, \boldsymbol{\theta}^R) = \mathbf{r}(\mathbf{d}^{obs}, \boldsymbol{\theta}^r) + \boldsymbol{\varepsilon}(\boldsymbol{\theta}^\varepsilon)$
(e) Estimate $\mathbf{z}^{reg}$ in ungauged catchment(s)
$\mathbf{z}^{reg} \leftarrow \mathbf{R}(\mathbf{d}^{obs}, \boldsymbol{\theta}^R) = \mathbf{r}(\mathbf{d}^{obs}, \boldsymbol{\theta}^r) + \boldsymbol{\varepsilon}(\boldsymbol{\theta}^\varepsilon)$
See sections 2.2
Regionalization Model RF Regression
z
$\mathbf{z}^{obs}$
$r(\mathbf{d}^{obs}, \boldsymbol{\theta}^r)$
ε
$\varepsilon(\theta^\varepsilon)$
Gauged catchments
$\mathbf{z}^{obs} \sim \mathbf{R}(\mathbf{d}^{obs}, \boldsymbol{\theta}^R)$
$\mathbf{d}^{obs}$
d
Regionalization Model RF Regression
z
$\mathbf{z}^{reg}$
$\varepsilon(\theta^\varepsilon)$
Ungauged catchment
$\mathbf{z}^{reg} \leftarrow \mathbf{R}(\mathbf{d}^{obs}, \boldsymbol{\theta}^R)$
$\mathbf{d}^{obs}$
d
Model adequacy testing
(f) DistanceTest
Is hydrological model likely to reproduce $\mathbf{z}^{reg}$?
Check p-value $\psi > 0.05$
See section 2.3.1
(g) InfoTest
Does hydrological model add information wrt prior knowledge about $\mathbf{z}^{reg}$?
Check Bayes factor BF > 1
See section 2.3.2
- If adequacy tests failed, revise steps (a)-(e)
- If adequacy tests passed, continue with steps (h)-(i)
See section 2.3.3
DistanceTest of hydrological model in a given catchment
$z_2$
95% prob contour
$z_2^{reg}$
$D_i$
$\mu_{k,2}$
$\mu_{k,1}$
$z_1^{reg}$
$z_1$
1. Construct $p(\mathbf{z}|\mathbf{H})$
2. Fit $k = 1, K$ Gaussians
3. Calculate $A_k$ and $A_{\max}$
4. Calculate $\psi$
5. If $\psi > 0.05$ DistanceTest passed
InfoTest of hydrological model in a given catchment
$pdf$
$\mathrm{BF}_{\mathbf{H}} = \dfrac{p(\mathbf{z}^{reg}|\mathbf{H})}{p(\mathbf{z}^{reg})}$
$p(\mathbf{z}|\mathbf{H})$
$p(\mathbf{z})$
$z_{min}^{obs}$
$z^{reg}$
$z_{max}^{obs}$
$z$
1. Construct $p(\mathbf{z}|\mathbf{H})$
2. Construct $p(\mathbf{z})$
3. Calculate $\mathrm{BF}_{\mathbf{H}}$
4. If $\mathrm{BF}_{\mathbf{H}} > 1$ InfoTest passed
Bayesian inference & flow prediction
(h) Condition hydrological model parameters on $\mathbf{z}^{reg}$ in ungauged catchment using Bayes eqn
$p(\boldsymbol{\theta}|\mathbf{z}^{reg}, \mathbf{H}, \mathbf{x}) = \dfrac{p(\mathbf{z}^{reg}|\mathbf{z}_{\boldsymbol{\theta}}^{sim})\, p(\boldsymbol{\theta}|\mathbf{H}, \mathbf{x})}{p(\mathbf{z}^{reg}|\mathbf{H}, \mathbf{x})}$
(i) Predict flow time series using hydrological model with posterior parameter samples
$\mathbf{q}_{\boldsymbol{\theta}}^{sim(i)} = \mathbf{H}(\boldsymbol{\theta}^{(i)}, \mathbf{x});\ \boldsymbol{\theta}^{(i)} \leftarrow p(\boldsymbol{\theta}|\mathbf{z}^{reg}, \mathbf{H}, \mathbf{x})$
See section 2.4
Prior prediction bounds
flow
time
Posterior prediction bounds
flow
time

**Figure 1.** Key steps of proposed procedure for prediction in ungauged catchments.

### 2.1. General Model Setup

#### 2.1.1. Hydrological Model Formulation

A deterministic hydrological model $\mathbf{H}$ simulates flow time series $\mathbf{q}^{\text{sim}}$ given model parameters $\theta$, inputs $\mathbf{x}$, and initial conditions $\mathbf{s}_0$ (Figure 1a):

$$\mathbf{q}^{\text{sim}} = \mathbf{H}(\theta, \mathbf{x}, \mathbf{s}_0) \tag{1}$$

The influence of $\mathbf{s}_0$ is minimized using a warm-up period; $\mathbf{s}_0$ is hence not inferred and is excluded from subsequent notation.

In ungauged catchments, observed flow data $\mathbf{q}^{\text{obs}}$ typically used to condition parameters $\theta$ are not available. In this work, the hydrological parameter estimation problem is informed by the PCs of flow indices, which are estimated at ungauged locations using a regionalization model (section 2.2).

Note that the term *hydrological model* will be used to refer to the combination of a hydrological model structure and a transformation of observations of hydrological drivers to hydrological model inputs (e.g., for rainfall, using Thiessen polygons, nearest neighbor, areal average, etc.). In the absence of more detailed information about the input data and their associated uncertainty (e.g., Renard et al., 2011), it is not possible to distinguish between these two sources of uncertainty.

#### 2.1.2. Flow Indices and PCs

Let $\boldsymbol{w} = \{w_i; i = 1, N_{\boldsymbol{w}}\}$ denote a set of flow indices computable from a flow time series (Figure 1b):

$$\boldsymbol{w} = \boldsymbol{w}\,(\mathbf{q}) \tag{2}$$

For example, this work makes use of indices such as mean annual and monthly flows, timing of events and so forth (see section 3.2).

The set of flow indices $\boldsymbol{w}$ can be transformed into a set of uncorrelated (orthogonal) PCs via PCA and the subset of dominant components $\mathbf{z} = \{z_i; i = 1, N_{\mathbf{z}}\}$ selected using a technique such as the broken stick method (Jackson, 1993; Peres-Neto et al., 2005; Figure 1c).

$$\mathbf{z} = \text{dPCA}(\boldsymbol{w}) \tag{3}$$

where the prefix “d” denotes the dominant PCs.

The quantity $\mathbf{z}$ will be referred to as “flow index PCs.” Flow index PCs computed from observed flows $\mathbf{q}^{\text{obs}}$ are denoted $\mathbf{z}^{\text{obs}}$, while those computed from simulated flows $\mathbf{q}^{\text{sim}}$ are denoted $\mathbf{z}^{\text{sim}}$. In addition, the notation $\mathbf{z}^{\text{reg}}$ is used to refer to the flow index PCs estimated by regionalization (see section 2.4.1).

Note that earlier work (Peñas et al., 2014) has suggested that applying additional transformations (such as dividing by the mean annual flow) to the flow series prior to calculating the flow indices and applying the PCA procedure can be advantageous. In addition, the selection of flow indices might be affected by the purpose of the application (e.g., high flow characteristics might be selected in preference, and/or given particular weight, if the model is intended for flood analysis). Although such flow index transformations or weights are not applied in this paper's case study, the procedures are general and can be applied with or without additional transformations and for any selection of flow indices.

### 2.2. Regionalization of Hydrological Information

#### 2.2.1. Regionalization Model Structure

The probabilistic regionalization model for the flow index PCs is denoted by $\mathbf{R}(\mathbf{d},\theta^R)$. It is constructed to estimate a set of flow index PCs $\mathbf{z}$ from a set of catchment descriptors $\mathbf{d}$ such as catchment area, climate, and topography (section 3.2), as schematized in Figure 1d.

$$\mathbf{z} \;= \mathbf{R}\left(\mathbf{d}, \theta^R\right) = \mathbf{r}(\mathbf{d}, \theta^r) + \varepsilon(\theta^\varepsilon) \tag{4}$$

where $\mathbf{r}(\mathbf{d},\theta^r)$ is a deterministic regionalization model with parameters $\theta^r$ and $\varepsilon(\theta^\varepsilon)$ is a (random) residual error model with parameters $\theta^\varepsilon$ intended to represent all sources of uncertainty in the deterministic model $\mathbf{r}$. The complete set of parameters of the regionalization model is $\theta^R = \{\theta^r, \theta^\varepsilon\}$. Note that the number (dimension) of models $\mathbf{R}$ and $\mathbf{r}$ and the dimension of $\varepsilon(\theta^\varepsilon)$ are equal to $N_{\mathbf{z}}$ (i.e., $\mathbf{R} = \{R_i; i = 1, N_{\mathbf{z}}\}$, etc.).

The number of parameters in the residual error model within the regionalization model is denoted by $N_{\Theta^{\varepsilon}}$, that is, $\Theta^{\varepsilon} = \{\theta_j^{\varepsilon};\ j = 1\ \ N_{\Theta^{\varepsilon}}\}$.

The regionalization model $\mathbf{R}$ is estimated using $\mathbf{z}^{\mathrm{obs}}$ and observed catchment descriptors $\mathbf{d}^{\mathrm{obs}}$ derived from available data from $n$ gauged catchments, as follows:

1. The deterministic term $\mathbf{r}$ is estimated using the RF regression technique as described in section 2.2.2.
2. The uncertainty in each regionalized flow index PC is estimated using a jackknife strategy followed by parametric distribution fitting, as described in section 2.2.3.

Once the regionalization model is constructed, it is used to estimate the flow index PCs $\mathbf{z}^{\mathrm{reg}}$ in an ungauged catchment of interest (Figure 1e); in turn, $\mathbf{z}^{\mathrm{reg}}$ is used in the model adequacy tests (section 2.3) and to condition hydrological model parameters (section 2.4).

Similarly to section 2.1.1, the term *regionalization model* will be used to refer to the combination of the regionalization model structure and its fixed inputs (catchment descriptors), as it is not possible to distinguish between these two sources of uncertainty given the available data. There are several publications that demonstrate the difficulties of separating model structure uncertainty and input uncertainty in the rainfall-runoff model community, for example, Renard et al. (2010, 2011), Beven and Westerberg (2011), Beven and Smith (2015), and others. Similar reasoning suggests difficulties in the separation of uncertainties due to regionalization model structure and its inputs.

### 2.2.2. RF Regression Model for Regionalization

The deterministic term $\mathbf{r}$ in the regionalization model in equation (4) is constructed using the RF regression technique from the machine learning community (Liaw & Wiener, 2002; Snelder et al., 2012). A separate model $r_i$ is built for each flow index PC $z_i$. The RF algorithm resamples (with replacement) an ensemble of regression "trees" to create a "forest." Each tree relates predictors (catchment descriptors) to the predictand (a flow index PC). The trees "grow" so that combinations of multiple catchment descriptors are randomly sampled at each node, and the combinations providing the lowest mean square error in the predictands are retained (Liaw & Wiener, 2002; Snelder et al., 2012). The resampling of predictors introduces randomness into the regression model built using RF (Breiman, 2001), in contrast to the single regression tree approach (Breiman et al., 1984). The model prediction is computed as the expected value of all individual predictions from each tree in the forest. Note that once the deterministic term of the RF model is estimated and the selected set of parameters are specified, the resulting RF model is deterministic, in the sense that given the same input, the model will always produce the same numerical results. In addition, although the RF model is flexible, it still incurs model structural error even after it is estimated.

The RF technique offers many theoretical and practical benefits compared with traditional linear regression techniques (Breiman, 2001), including the following:

1. flexible model structure that does not require preselecting a model equation form or preselecting predictors from the available set of candidate predictors; and
2. relatively low computational cost (see Peñas, 2013, for a comparison of machine learning algorithms for regionalization).

This work uses the RF algorithm implementation in the R package "randomForest v4.6.7" (Liaw & Wiener, 2002).

The deterministic term in the regionalization model cannot be expected to reproduce the observed flow index PCs exactly: (i) the relationship between catchment descriptors and flow index PCs is unlikely to be deterministic, especially given a limited set of catchment descriptors; (ii) RF regression can provide only an approximate representation of data relationships, especially when estimated from a finite set of samples (catchments); and (iii) observed flow index PCs and catchment attributes differ from their "true values" due to observation errors in the underlying data. To characterize regionalization uncertainty, a residual error model is constructed as described next.

### 2.2.3. Uncertainty Characterization in the Regionalization Model

The uncertainty in each regionalized flow index PC is estimated using a jackknife strategy, followed by parametric distribution fitting (Almeida et al., 2016), as follows: (i) leave out a single catchment from the $n$ gauged catchments, (ii) use the remaining $(n - 1)$ catchments to estimate the regionalization model $\mathbf{r}$ as described in section 2.2.2; (iii) use the model **r** to estimate ("regionalize") the flow index PCs in the left-out catchment; (iv) compute the residual error vector of the regionalized flow index PCs using the observed flow index PCs (available in the left-out catchment); (v) repeat steps (i)–(iv) for each catchment, resulting in a set of $n$ residual error vectors; and (vi) fit a parametric joint probability distribution $\varepsilon(\theta^{\varepsilon})$ to the set of residuals from step (v).

The parametric joint distribution is fitted to the residual errors of the regionalization model as follows: (i) estimate the cross-correlation structure using the Pearson correlation between all pairings of the residual errors of individual flow index PCs; (ii) hypothesize a particular parametric distribution for the residual errors—for example, the distributions considered in this work include the Gaussian, extreme value type 1 (also known as the Gumbel distribution), and generalized extreme value distributions; (iii) check the hypothesized distribution against the actual residuals using the $\chi^2$, Lillietest (Lilliefors, 1967, 1969), Jbtest (Jarque & Bera, 1987), and Mardia statistical tests (Mardia, 1970; Trujillo-Ortiz & Hernandez-Walls, 2003).

The parameters of the residual error model reflect the quality of the regionalization model. For example, location parameters such as the mean of the residual errors provide a characterization of bias in the regionalization model, whereas scale parameters such as the standard deviation of residual errors are indicative of the typical magnitude of random errors. To facilitate their interpretation, the estimated residual error parameters of the regionalization model are averaged and normalized by the range of observed or simulated PCs of the flow indices, as described next. Note that because the residual error model parameters are estimated separately for each flow index PC, the auxiliary notation $\theta_{i,j}^{\varepsilon(u)}$ is introduced to refer to the $j$th parameter of the residual error model for the regionalization of the $i$th flow index PC in ungauged catchment $u$.

When the normalization is done using the range of PCs based on observed data, equation (5) is applied:

$$\theta_{i,j}^{\varepsilon(\text{scaled-obs})} = \left(N_u\left[z_i^{\max} - z_i^{\min}\right]\right)^{-1} \sum_{u=1}^{N_u} \theta_{i,j}^{\varepsilon(u)} \tag{5}$$

where $\theta_{i,j}^{\varepsilon(\text{scaled-obs})}$ is the average normalized value of the $j$th parameter of the residual error model for the regionalization of the $i$th flow index PC; $z_i^{\max} = \max\{z_i^{\text{obs}(k)}; k = 1, N_{\text{cat}}\}$ and $z_i^{\min} = \min\{z_i^{\text{obs}(k)}; k = 1, N_{\text{cat}}\}$ are, respectively, the largest and smallest values of the $i$th flow index PC across $N_{\text{cat}}$ case study catchments; and $N_u$ is the number of ungauged catchments.

When the normalization is done using the range of PCs simulated by the hydrological model **H**, equation (6) is applied:

$$\theta_{i,j}^{\varepsilon(\text{scaled-sim})} = (N_u)^{-1} \sum_{u=1}^{N_u} \theta_{i,j}^{\varepsilon(u)} \left[z_i^{\max,\text{sim}(u)} - z_i^{\min,\text{sim}(u)}\right]^{-1} \tag{6}$$

where $z_i^{\max,\text{sim}(u)} = \max\{z_i^{\text{sim}(u)(k)}; k = 1, N_{\text{sim}}\}$ and $z_i^{\min,\text{sim}(u)} = \min\{z_i^{\text{sim}(u)(k)}; k = 1, N_{\text{sim}}\}$ are, respectively, the largest and smallest values of the $i$th flow index PC simulated by the hydrological model **H** in ungauged catchment $u$; and $N_{\text{sim}}$ is the number of model simulations (see sections 2.1.2 and 2.4.2**)**.

### 2.3. Model Adequacy Tests in Flow Index PC Space

This section introduces two model tests to scrutinize the quality of the hydrological and regionalization models in the flow index PC space:

1. "DistanceTest", which quantifies the ability of a model (hydrological or regionalization) to reproduce the set of flow index PCs (denoted by **z**) in a catchment; and
2. "InfoTest", which quantifies the information added by a model (hydrological or regionalization) over prior knowledge about flow index PCs **z** in a catchment.

A model (hydrological or regionalization) is considered "adequate" if it passes both DistanceTest and InfoTest, as described in the following sections.

#### 2.3.1. Ability to Reproduce Flow Index PCs: DistanceTest

DistanceTest evaluates the hypothesis that a model (hydrological or regionalization) is statistically likely to reproduce a given set of flow index PCs $\mathbf{z}$ in a catchment. A DistanceTest $p$ value $\psi(\mathbf{z})$ is calculated as follows (see Figure 1f):

1. Construct the distribution of $\mathbf{z}$, $p(\mathbf{z}|\bullet)$, from the model being tested. When testing the hydrological model, $p(\mathbf{z}|\mathbf{H})$ is constructed by propagating the *prior* parameter distribution through the hydrological model in equation (1); when testing the regionalization model, $p(\mathbf{z}|\mathbf{R})$ is constructed by sampling residual errors within the regionalization model in equation (4).
2. Approximate $p(\mathbf{z}|\bullet)$ by a mixture of $K$ Gaussian distributions (components) with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ for $k = 1, K$ (Bulygina & Gupta, 2011; Muller et al., 1996).
3. Calculate the probability mass $A_k$ of the confidence region associated with the value $\mathbf{z}$ being tested, within each Gaussian component $k$:

$$A_k = F^{-1}_{\chi2}\left(D_k^2; N_{\mathbf{z}}\right) \tag{7}$$

$$D_k = \sqrt{\left(\mathbf{z}-\boldsymbol{\mu}_k\right)^{\text{transp}}\ \boldsymbol{\Sigma}_k^{-1}\ \left(\mathbf{z}-\boldsymbol{\mu}_k\right)} \tag{8}$$

where $D_k$ is the Mahalanobis distance, $F^{-1}_{\chi2}(x; m)$ is the inverse cumulative distribution function of the $\chi^2$ distribution with $m$ degrees of freedom, and transp denotes the vector transpose.

Equations (7) and (8) can be derived by considering that elliptical contours in $\mathbf{z}$-space equidistant in Mahalanobis distance from $\boldsymbol{\mu}_k$ represent Gaussian equidensity contours and define the most compact confidence region with probability mass $A_k$ (Gallego et al., 2013; Ribeiro, 2004). Assuming $\mathbf{z}$ represents a sample from the $k$th Gaussian component, $\mathbf{z}$~$N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, the quantity $D_k^2$ follows a $\chi^2$ distribution with $N_{\mathbf{z}}$ degrees of freedom (Gallego et al., 2013).

4. The DistanceTest $p$ value $\psi$ over all $K$ Gaussian components is defined as

$$\psi = 1\text{–}A_{\max} \tag{9}$$

where $A_{\max} = \max\{A_k; k = 1, K\}$. DistanceTest is passed if $\psi > \psi^*$, where $\psi^*$ is a prescribed significance level; in this work, $\psi^* = 0.05$ is used.

DistanceTest can be used in the following two ways:

i The main purpose of DistanceTest is to test the *combined* regionalization and hydrological model under "real operating" conditions in an ungauged catchment, where observed flow is not available. In this case, DistanceTest is applied with $\mathbf{z} = \mathbf{z}^{\text{reg}}$. This setup tests whether the hydrological model is likely to reproduce the flow index PCs estimated using the regionalization model, so that the test outcome is affected by deficiencies in the regionalization *and/or* hydrological models (which as noted in sections 2.1.1 and 2.2.1 include the respective model structures and forcing inputs). To allow for regionalization uncertainty, DistanceTest is conducted using replication as follows: draw a sample $\mathbf{z}^{\text{reg}(i)}$ from the distribution of regionalized flow index PCs, compute its $p$ value $\psi^{(i)}$, repeat $N_{\text{sam}}$ times, and compute the average $p$ value $\psi = \frac{1}{N_{\text{sam}}}\sum_{i=1}^{N_{\text{sam}}}\psi^{(i)}$. In this study, $N_{\text{sam}} = 10{,}000$ draws are used.

ii DistanceTest can also be used to test an individual model—hydrological or regionalization—under "verification" conditions when $\mathbf{z}^{\text{obs}}$ is available. In this case, $\mathbf{z}^{\text{obs}}$ can be used in DistanceTest either directly as $\mathbf{z} = \mathbf{z}^{\text{obs}}$ (no replication) or indirectly by constructing "synthetic" scenarios (using replication when testing the hydrological model). In this work, such verification is conducted in scenario 0 and scenarios 2–4 in section 3.5.

#### 2.3.2. Model Informativeness About Flow Index PCs: InfoTest

InfoTest quantifies how much information about catchment flow index PCs is added by a model (hydrological or regionalization) over prior knowledge. In other words, it quantifies the added value of using a model to predict the dominant flow characteristics as represented by the flow index PCs. The test is defined via the Bayes Factor (BF; Gelman et al., 2013; see Figure 1g).

**Table 1**
*Interpretation of Bayes Factor (BF) Values*

| BF (M1/M0) | Strength of the evidence |
|---|---|
| <1 | Negative: Reject hypothesis M1 |
| 1–$10^{1/2}$ | Not worth more than a bare mention |
| $10^{1/2}$–10 | Substantial evidence favoring hypothesis M1 |
| 10–$10^{3/2}$ | Strong evidence favoring hypothesis M1 |
| $10^{3/2}$–$10^{2}$ | Very strong evidence favoring hypothesis M1 |
| >$10^{2}$ | Decisive evidence favoring hypothesis M1 |

*Note*. Adapted from Jeffreys (1961).

For a hydrological model, the Bayes Factor $\mathrm{BF}_{\mathbf{H}}$ is defined as

$$\mathrm{BF}_{\mathbf{H}}(\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{H})}{p(\mathbf{z})} \tag{10}$$

where $p(\mathbf{z}|\,\mathbf{H})$ is the probability density function (pdf) of the distribution of flow index PCs given the model $\mathbf{H}$ with parameters sampled from the prior distribution $p(\theta)$. The term $p(\mathbf{z})$ denotes the pdf of the prior distribution of observed flow index PCs (see below).

For the regionalization model, the Bayes factor $\mathrm{BF}_{\mathbf{R}}$ is defined as

$$\mathrm{BF}_{\mathbf{R}}(\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{R})}{p(\mathbf{z})} \tag{11}$$

where $p(\mathbf{z}|\,\mathbf{R})$ is the pdf of the distribution of flow index PCs associated with the probabilistic regionalization model $\mathbf{R}$ defined in section 2.2.1.

In this work, the prior density $p(\mathbf{z})$ is set as uniform over the range of flow index PCs observed across the available gauged catchments:

$$p(\mathbf{z}) = \begin{cases} \prod_{i=1}^{N_z}\left(z_i^{\max} - z_i^{\min}\right), \text{when } z_i^{\min} \le z_i \le z_i^{\max} \text{ for } \quad i = 1, N_{\mathbf{z}} \\ 0, \text{otherwise} \end{cases} \tag{12}$$

where $N_z$ is the number of PCs retained for the analysis and $z_i^{\max} = \max\{z_i^{\mathrm{obs}(k)}; k = 1, N_{\mathrm{cat}}\}$ and $z_i^{\min} = \min\{z_i^{\mathrm{obs}(k)}; k = 1, N_{\mathrm{cat}}\}$ are, respectively, the largest and smallest values of the $i$th PC across $N_{\mathrm{cat}}$ case study catchments.

The pdf $p(\mathbf{z}|\,\mathbf{H})$ is approximated by fitting a mixture of Gaussians to the set of flow index PCs generated using the hydrological model with parameters sampled from the prior (same as in DistanceTest). The pdf $p(\mathbf{z}|\,\mathbf{R})$ is obtained as described in section 2.2.3.

Table 1 provides a qualitative interpretation of BF values. BF = 1 indicates that the model provides the same information as the prior, BF < 1 indicates that the prior provides more information than does the model (i.e., the model is not informative), and BF > 1 indicates that the model adds information beyond the prior. BF can be interpreted quantitatively, for example, BF = 10 indicates that the model provides "10 times more information" than the prior, and so forth. While other choices can be made, the study considers InfoTest "passed" when BF > 1.

Similar to DistanceTest, InfoTest can be used in two ways:

i InfoTest can be used to test the *combined* regionalization and hydrological models under "real operating" conditions, that is, when observed flow is not available, and $\mathbf{z}$ is estimated via regionalization. In effect, this tests the added information value of the hydrological model in reproducing the regionalized flow index PCs, as compared with prior estimates of observed flow index PCs. If the test is failed, it could be due to deficiencies in the regionalization *and/or* hydrological models (including their structures and inputs). To allow for regionalization uncertainty, InfoTest uses a replication similar to that in DistanceTest, drawing from the distribution of regionalized flow index PCs, computing the BF using equation (10) and averaging over multiple draws to get the overall BF.

ii InfoTest can also be used to test an individual model—hydrological or regionalization—under "verification" conditions, when $\mathbf{z}^{\mathrm{obs}}$ is available (scenarios 0 and 2–4 in section 3.5).

### 2.3.3. Practical Usage of Adequacy Tests

The use of DistanceTest and InfoTest relies on the assumption that model adequacy in the space of flow index PCs is at least broadly indicative of model performance in predicting flow time series. This assumption is necessary given that ungauged catchments do not have observed flow time series available for model verification; its validity will be appraised in the empirical case study (sections 3 and 4).

The adequacy tests are carried out prior to any conditioning of hydrological model parameters. If the combined regionalization/hydrological model passes the adequacy tests, the modeler can proceed to condition the hydrological model parameters $\theta$ on the regionalized flow index PCs and produce flow predictions (next section). If the adequacy tests are failed, the model cannot be expected to provide trustworthy flow predictions and should be replaced or enhanced. Note that the proposed adequacy tests represent necessary but insufficient conditions for a model to be considered capable of predicting streamflow. The tests are a priori necessary because the model must be able to represent the set of streamflow indices (e.g., annual and monthly flows), before being used to predict detailed flow dynamics (hydrograph) in the ungauged catchment. However, the tests are not sufficient on their own, because even if the model can reproduce the streamflow indices, it might not be able to reproduce the full dynamics.

### 2.4. Hydrological Model Inference and Prediction

#### 2.4.1. Bayesian Inference Using Regionalized PCs

Given a set of regionalized flow index PCs $\mathbf{z}^{\mathrm{reg}}$, a hydrological model $\mathbf{H}$, and inputs $\mathbf{x}$, the posterior distribution of hydrological model parameters $p\left(\theta|\mathbf{z}_{\theta}^{\mathrm{reg}},\mathbf{H},\mathbf{x}\right)$ is given by Bayes equation (Bulygina et al., 2009; see Box & Tiao, 1973, for basic theory; Figure 1h).

$$p(\theta|\mathbf{z}^{\mathrm{reg}},\mathbf{H},\mathbf{x}) = \frac{p(\mathbf{z}^{\mathrm{reg}}|\theta,\mathbf{H},\mathbf{x})\,p(\theta|\mathbf{H},\mathbf{x})}{p(\mathbf{z}^{\mathrm{reg}}|\mathbf{H},\mathbf{x})} = \frac{p\left(\mathbf{z}^{\mathrm{reg}}|\mathbf{z}_{\theta}^{\mathrm{sim}}\right)p(\theta|\mathbf{H},\mathbf{x})}{p(\mathbf{z}^{\mathrm{reg}}|\mathbf{H},\mathbf{x})} \tag{13}$$

The likelihood function $p(\mathbf{z}^{\mathrm{reg}}|\theta,\mathbf{H},\mathbf{x})=p\left(\mathbf{z}^{\mathrm{reg}}|\mathbf{z}_{\theta}^{\mathrm{sim}}\right)$ describes the statistical relationship between regionalized and simulated flow index PCs (see section 2.2); $\mathbf{z}_{\theta}^{\mathrm{sim}} = \mathrm{dPCA}(\mathbf{q}^{\mathrm{sim}}) = \mathrm{dPCA}(\mathbf{H}(\theta,\mathbf{x}))$ denotes flow index PCs simulated by model $\mathbf{H}$ with input $\mathbf{x}$ and parameters $\theta$.

Under the assumption that the errors of the regionalization model dominate the errors of the hydrological model, the likelihood function can be constructed using the same probability distribution as estimated for the residual errors of the regionalization model in section 2.2.3. This assumption follows published work on conditioning of hydrological parameters to streamflow statistics (e.g., Almeida et al., 2016; Bulygina et al., 2009, 2012; Yadav et al., 2007), though it can be questioned (see section 6).

Unless specific prior information is available, $p(\theta|\,\mathbf{H},\mathbf{x})$ can be specified as uniform over the feasible parameter ranges. The denominator in equation (13) is a normalizing constant and is not required explicitly by many sampling algorithms. The posterior distribution in equation (13) is sampled as described in section 2.4.2. The predictive distribution of flows is then constructed by running the hydrological model $\mathbf{H}(\theta,\,\mathbf{x})$ with the posterior parameter samples $\theta$ (Figure 1i).

#### 2.4.2. Posterior Distribution Sampling

The posterior parameter distribution in equation (13) is approximated using importance sampling (Doucet et al., 2000; see also Kuczera & Parent, 1998), as follows:

1. Draw $S$ parameter sets $\{\theta_i;\, i = 1,\, S\}$ from the uniform prior distribution $p(\theta|\,\mathbf{H},\mathbf{x})$ using the Latin hypercube method ($S = 1{,}000$ in the study).
2. Run the hydrological model (with fixed inputs and initial conditions) with each parameter set $\theta_i$ to generate $S$ flow time series $\{\mathbf{q}_i^{\mathrm{sim}};\, i = 1,\, S\}$.
3. Calculate the flow index PCs $\mathbf{z}_i^{\mathrm{sim}}$ for each simulated flow time series.
4. Compute the (unscaled) weight $w_i$ for each parameter set using the likelihood function in equation (13); each parameter set $\theta_i$ generated in step 1 is assigned a weight based on the likelihood function value of its corresponding flow index $z_i^{\mathrm{sim}}$ computed in step 3.
5. Scale the weights $w_i$ so they add up to 1.

The parameter sets and their weights, $\{\theta_i,\, w_i;\, i = 1,\, S\}$, provide an approximation to the posterior parameter distribution in equation (13) (Doucet et al., 2000).

## 3. Case Study Description

### 3.1. Case Study Catchments

The case study is based on a set of 92 catchments in northern Spain (Figure 2) selected from the larger set of 156 catchments used by Peñas et al. (2014). The selected catchments are characterized by a "natural" hydrological regime as defined by the European Water Framework Directive (European Commission, 2000). From the selected catchments, 62 catchments drain into the Cantabrian Sea, and the remaining 30 catchments

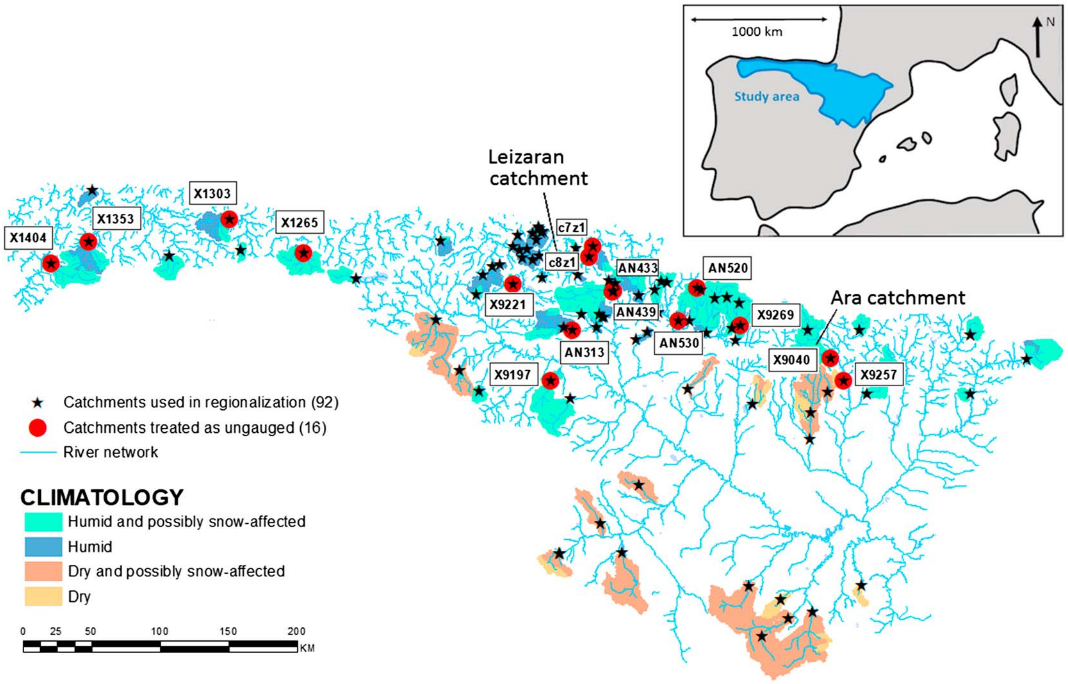


**Figure 2.** Case study catchments.

drain into the Mediterranean Sea. The climatic characteristics described below are derived from monthly climate series calculated from a 1 km × 1 km grid map developed by the Centre for Hydrographic Studies using data from more than 5,000 weather stations across Spain (CEDEX, Ministry of Public Works and Ministry of Agriculture, Food and Environment, 2013, Spain). Topography and catchment geometry are derived using a 25-m-resolution digital elevation model. Land use is derived from the Soil Occupancy Information System (in Spanish "SIOSE") at a 1:25,000 scale developed by the National Geographic Institute of the Spanish Government. The geological variables are derived from the lithostratigraphic and permeability maps at scale 1:200,000, developed by the Spanish Geologic and Mining Institute (in Spanish "IGM") of the Spanish Government.

The selected catchments exhibit a wide variety of geologies, soils, topographies, land uses, and climatic conditions. Dominant lithological groups in the catchments draining into the Mediterranean Sea are clay, sand, and gravel; from these catchments, those in the Pyrenees also contain some siliceous and calcareous rock. The western catchments draining into the Cantabrian Sea are composed primarily of slates, while the eastern catchments are dominated by calcareous rock (Geological and Mining Institute of Spain, 2013). In each catchment, urbanized zones comprise less than 8% of the total area, with land cover dominated by pastures, broadleaf forests, and coniferous forests. Table 2 reports the ranges of average catchment elevation, slopes of main river channels, catchment areas, annual average rainfall, annual average PET, aridity index, annual average flows, annual runoff coefficients, and annual average temperatures. Figure 2 classifies the catchments based on their aridity index (Arora, 2002) and minimum monthly average temperatures (snowfall is likely below 0 °C). In this work the aridity index is defined as average annual PET divided by average annual precipitation; dry and humid climatic conditions are described, respectively, by aridity indices above and below 1 (Arora, 2002).

A subset of 16 catchments (out of the 92 catchments) is selected and treated as "ungauged" for the purposes of evaluating the proposed prediction methods and adequacy tests. These catchments have synchronized daily data of rainfall, daily flow, and daily PET (estimated from monthly PET) of sufficient length (at least 8 years). Such data are not available in other catchments, making it impossible to implement complete leave-one-out cross-validation (Figure 2 and Table 3). See Peñas et al. (2014) for detailed information. Daily precipitation data for the catchments are provided by the Spanish Meteorological Agency (AEMET), while daily flow data are provided by different Spanish water agencies and regional governments.

**Table 2**
*Average Characteristics of the 92 Case Study Catchments*

| Catchment set | | Average | | | | Rainfall-runoff coefficient | Annual rainfall (mm/year) | Annual PET (mm/year) | Temperature maximum-monthly minimum (°C) | Permeability | Average rock hardness (scale 1–5) | Geology classes[a] (% area occupied) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Area ($km^2$) | Elevation (m) | Slope (1%) | Temperature (°C) | | | | | | | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| 92 catchments | Min | 2 | 241 | 9 | 5 | 0.03 | 450 | 492 | −6 | Very low | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Median | 106 | 898 | 32 | 10 | 0.5 | 1,256 | 700 | 0 | Mean | 3 | 18 | 0 | 1 | 9 | 2 | 21 | 0 | 0 | 0 | 0 |
| | Max | 1,038 | 2,218 | 63 | 15 | 0.97 | 1,809 | 987 | 5 | Very high | 4 | 100 | 33 | 79 | 90 | 22 | 92 | 74 | 97 | 22 | 1 |
| 16 catchments | Min | 22 | 483 | 21 | 7 | 0.2 | 681 | 564 | −3 | Very low | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Median | 18 | 822 | 37 | 10 | 0.6 | 1,364 | 681 | 0 | Low | 3 | 17 | 0 | 0 | 13 | 1 | 21 | 0 | 0 | 0 | 0 |
| | Max | 623 | 1,505 | 53 | 12 | 1 | 1,809 | 962 | 4 | High | 3 | 90 | 10 | 19 | 80 | 11 | 73 | 35 | 72 | 4 | 0 |

*Note.* PET = potential evapotranspiration.
[a]Geology classes: (1) calcareous rocks, (2) clay, (3) conglomerates rocks, (4) sand, (5) sedimentary rocks, (6) shale (sedimentary) rocks, (7) siliceous rocks, (8) slates, (9) volcanic rocks, and (10) wetlands.

## 3.2. Catchment Flow Indices and Catchment Descriptors

The case study uses the same 103 flow indices $\boldsymbol{w}$ and 16 catchment descriptors used by Peñas et al. (2014). Here a brief summary is provided; for more details, see Peñas et al. (2014).

The 103 flow indices were computed from flow records available at the 92 gauged locations, using at least 8 years of daily data in the period 1976–2009 (see Appendix A in Peñas et al., 2014, for a complete description). The flow indices comprise the mean and standard deviation of the following quantities: (1) annual and monthly flows; (2) high and low flows; (3) the duration and frequency of high flows; (4) the rate of change in the flows; and (5) the timing of maximum and minimum flow events.

The following 16 catchment descriptors $\mathbf{d}$ are used (Peñas et al., 2014): area, climate (mean annual precipitation and PET, and ratio of minimum quarterly precipitation to maximum quarterly precipitation), topography (average catchment elevation and gradient), catchment geometry (drainage density and number of river confluences), land use (area covered by agricultural land, broadleaf forest, coniferous forest, bare land, pasture, and urban areas), and geology (average rock density and permeability). These catchment descriptors are the least correlated from a larger set of catchment descriptors, with Pearson correlation coefficients below 0.7 (Peñas et al., 2014).

## 3.3. Hydrological Model

The conceptual rainfall-runoff model used in this paper is the Probability Distributed Model (PDM) (Moore, 2007). This model has a simple structure and is used widely across the world, including in the United Kingdom (Lee et al., 2005; Pechlivanidis et al., 2010), Europe (Arnell, 1999; Cabus, 2008; Willems et al., 2014), the United States (Kollat et al., 2012), Southeast Asia (Thompson et al., 2013), Southern Africa (MacKellar et al., 2013), and Australia (Srikanthan et al., 2007). Here PDM is used as a lumped model over each catchment and run on a daily time step.

A schematic of PDM is shown in Figure 3. Spatial variability of soil moisture storage capacity is represented using a Pareto distribution, and runoff routing is represented using two linear reservoirs in parallel. The model has a total of five parameters, namely, the Pareto distribution parameters $C_{\mathrm{max}}$ and $b$ (which control storage capacity and its variability, respectively), parameter $\alpha_{\mathrm{PDM}}$ (which controls the split of effective rainfall into quick flow and slow flow), and two routing parameters $T_{\mathrm{q}}$ and $T_{\mathrm{s}}$ (which control the residence time of the quick flow and slow flow reservoirs, respectively). A more detailed model description can be found in Moore (2007). Prior parameter ranges are adapted from Kollat et al. (2012) and shown in Figure 3. The initial soil moisture storage is set to 0, and the first year of the simulations is used for a model warm-up.

## 3.4. Analyses and Evaluations

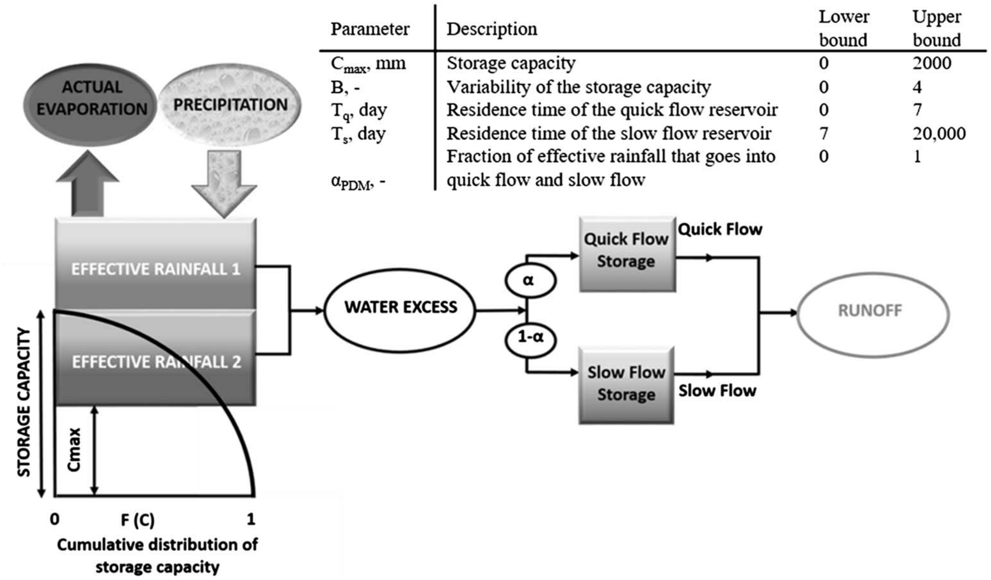
### 3.4.1. Analysis 1. Selection of Dominant Flow Index PCs

This analysis applies PCA to the complete set of flow indices, selects the dominant PCs using the broken stick method, and reports the fraction of variance in the flow indices explained by these selected PCs.

**Table 3**
*Characteristics of 16 Catchments Treated as "Ungauged" in the Evaluation Case Study*

| | | Average | | | | | | | | | Geology classes[a] (% area occupied) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Flow gauge name | Area ($km^2$) | Elevation (m) | Slope (%) | Temperature (°C) | Rainfall-runoff coefficient | Annual rainfall (mm/year) | Annual PET (mm/year) | Temperature maximum-monthly minimum (°C) | Permeability | Average rock hardness (scale 1–5) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| X1353 | 529 | 1,047 | 52 | 10 | 0.61 | 1,459 | 660 | 1 | Low | | 2 | 0 | 10 | 11 | 4 | 0 | 1 | 72 | 0 | 0 |
| X1404 | 293 | 1,068 | 50 | 10 | 0.81 | 1,449 | 638 | 0 | Very low | | 0 | 0 | 5 | 16 | 11 | 0 | 0 | 68 | 0 | 0 |
| X1303 | 377 | 500 | 39 | 11 | 0.64 | 1,342 | 734 | 4 | Low | | 19 | 10 | 0 | 2 | 3 | 0 | 35 | 31 | 0 | 0 |
| X1265 | 294 | 1,188 | 53 | 7 | 0.57 | 1,041 | 716 | −1 | Very low | | 18 | 0 | 19 | 4 | 8 | 0 | 0 | 50 | 0 | 0 |
| X9257 | 80 | 696 | 24 | 12 | 0.33 | 814 | 962 | 0 | Low | | 1 | 0 | 2 | 49 | 0 | 48 | 0 | 0 | 0 | 0 |
| X9040 | 623 | 1,505 | 52 | 7 | 0.69 | 1,356 | 705 | −3 | Low | | 17 | 0 | 5 | 47 | 1 | 22 | 6 | 0 | 2 | 0 |
| X9269 | 75 | 1,385 | 48 | 8 | 0.39 | 1,333 | 619 | −1 | Low | | 7 | 0 | 0 | 73 | 0 | 20 | 0 | 0 | 0 | 0 |
| X9197 | 283 | 1,147 | 33 | 8 | 0.24 | 681 | 655 | −1 | Low | | 17 | 5 | 3 | 0 | 1 | 73 | 0 | 0 | 0 | 0 |
| X9221 | 22 | 839 | 24 | 8 | 0.79 | 1,134 | 564 | −0 | Low | | 65 | 0 | 0 | 0 | 0 | 35 | 0 | 0 | 0 | 0 |
| AN439 | 152 | 702 | 27 | 10 | 0.95 | 1,469 | 701 | 0 | Low | | 90 | 2 | 0 | 1 | 1 | 5 | 0 | 0 | 1 | 0 |
| AN433 | 554 | 789 | 21 | 9 | 0.41 | 1,376 | 679 | 0 | High | | 53 | 0 | 0 | 6 | 11 | 30 | 0 | 0 | 0 | 0 |
| AN520 | 73 | 1,071 | 24 | 8 | 0.35 | 1,683 | 609 | −2 | Low | | 11 | 0 | 0 | 55 | 0 | 33 | 0 | 0 | 0 | 0 |
| AN530 | 95 | 805 | 36 | 10 | 0.39 | 1,371 | 693 | −0 | Low | | 0 | 0 | 0 | 80 | 0 | 20 | 0 | 0 | 0 | 0 |
| AN313 | 477 | 790 | 21 | 11 | 0.35 | 1,038 | 650 | 1 | High | | 57 | 0 | 1 | 0 | 10 | 21 | 10 | 0 | 0 | 0 |
| c8z1 | 114 | 598 | 42 | 11 | 0.66 | 1,794 | 684 | 2 | High | | 28 | 3 | 0 | 28 | 0 | 37 | 0 | 1 | 2 | 0 |
| c7z1 | 28 | 483 | 40 | 12 | 0.65 | 1,809 | 731 | 3 | Low | | 5 | 1 | 0 | 21 | 0 | 18 | 0 | 51 | 4 | 0 |

*Note*. See section 3.1 for details. PET = potential evapotranspiration.
[a] Geology classes: (1) calcareous rocks, (2) clay, (3) conglomerates rocks, (4) sand, (5) sedimentary rocks, (6) shale (sedimentary) rocks, (7) siliceous rocks, (8) slates, (9) volcanic rocks, and (10) wetlands.



| Parameter | Description | Lower bound | Upper bound |
|---|---|---|---|
| $C_{max}$, mm | Storage capacity | 0 | 2000 |
| B, - | Variability of the storage capacity | 0 | 4 |
| $T_q$, day | Residence time of the quick flow reservoir | 0 | 7 |
| $T_s$, day | Residence time of the slow flow reservoir | 7 | 20,000 |
| $\alpha_{PDM}$, - | Fraction of effective rainfall that goes into quick flow and slow flow | 0 | 1 |

**Figure 3.** Schematic of the hydrological model PDM (probability distributed model; adapted from Pechlivanidis et al., 2010). The inset shows the parameter ranges used in this study.

### 3.4.2. Analysis 2. Evaluation of the Probabilistic RF Regionalization Model

This analysis comprises an exploration of the probability distribution fitted to the residual errors of the regionalization model. The following aspects are considered:

1. The first aspect is regionalization model bias, quantified by the average of the normalized mean parameter of the residual errors of the regionalization model. Two normalization approaches are considered: (i) by the range of observed flow index PCs across all catchments and (ii) by the range of simulated flow index PCs in the given catchment, obtained using the hydrological model with parameter sets sampled from the prior $p(\theta)$. A normalized mean of 0 corresponds to an unbiased regionalization.
2. The second aspect is regionalization model precision, quantified by the standard deviation parameter of the residual errors, using the two normalizations described above. The average of the normalized standard deviation quantifies the relative errors of the regionalization model with respect to the range of flow indices PCs. A narrow spread indicates lower uncertainty (higher precision).

In addition to the parameter analysis above, the fitted residual error distribution is plotted for one of the catchments.

### 3.4.3. Analysis 3. Evaluation of the Correspondence Between Model Adequacy in the Flow Index PC Space and Model Performance in the Flow Time Series Space

This analysis investigates the behavior of the adequacy tests DistanceTest and InfoTest, which are applied in the flow index PC space yet are intended to provide at least an indirect indication of model ability to predict flow time series. The following investigations are carried out over the 16 “ungauged” catchments:

1. Apply the adequacy tests under five scenarios of model quality and data availability (section 3.5).
2. Evaluate flow predictions using the probabilistic Nash-Sutcliffe efficiency (probabilistic NSE) $\Phi_{NSP}$ and the 95% posterior probability limits $\gamma_{95\%}$ (section 3.6). Results are then briefly discussed in the context of operational predictions in the same geographical area.
3. Carry out an analysis of variance (ANOVA) to quantify whether improvements in the hydrological and/or regionalization models make a statistically significant difference on the quality of the predictions as assessed using the $\Phi_{NSP}$ and $\gamma_{95\%}$ performance metrics.

### 3.4.4. Analysis 4. Illustration of Model Performance in Specific Catchments

This analysis reports observed and simulated flow time series for two selected catchments:

1. Leizarán River Basin (code c8z1 in Figure 2), which is located in the Basque Country and has one of the highest data quality of the available catchments. This catchment is typical of north of Spain, with a humid climate and no snow; PDM is expected to perform well in this catchment.

2. Ara River Basin (code X9040 in Figure 2), which is located in the Pyrenees and has a glacial valley at its higher elevations. This catchment experiences snow melt in April–July; PDM is expected to perform poorly in this catchment.

### 3.5. Description of Scenarios in Analysis 3

The studies in Analysis 3 are carried out for the following five scenarios, intended to represent different levels of model quality and available data.

- *Scenario 0*. This scenario assumes that $\mathbf{z}^{\text{obs}}$ is available, which allows a direct assessment of the adequacy of the regionalization and hydrological models but in real practice will not be possible in ungauged catchments. The regionalization model is tested by estimating $\mathbf{z}^{\text{reg}}$ from the catchment descriptors in the "ungauged" catchment and applying DistanceTest and InfoTest with $\mathbf{z} = \mathbf{z}^{\text{obs}}$. The hydrological model is tested by applying the adequacy tests with $\mathbf{z} = \mathbf{z}^{\text{obs}}$.
- *Scenario 1*. This scenario represents the intended usage of the proposed framework for flow prediction in ungauged catchments, where observation-based flow index PCs are not available. Flow index PCs for an ungauged catchment are estimated using the regionalization model as described in section 2.2, and regionalized flow index PCs are used to condition hydrological model parameters/simulations via Bayes equation (13). Adequacy tests are performed only for the hydrological model, with $\mathbf{z} = \mathbf{z}^{\text{reg}}$.
- *Scenario 2*. This scenario is devised such that the regionalization model is accurate and has low noise (i.e., near exact). Specifically, with reference to equation (4), the deterministic term is set equal to the observed values of flow index PCs in the catchment treated as "ungauged" (rather than estimated using RF regression), and the random noise term is set to a Gaussian distribution with zero mean and a standard deviation equal to 5% of the full range of observed flow index PCs across the 92 catchments treated as "gauged." Though other values of the standard deviation might be used, a standard deviation equal to 5% is used in this paper. The purpose of using this reduced noise is to explore the impact of improvements in the regionalization model on the adequacy tests and predictive performance. The adequacy tests are applied only to the hydrological model, because the regionalization model is adequate by construction ($p$ value = 1 and BF > 1).
- *Scenario 3*. In this scenario the hydrological model is replaced by the observed time series (and hence the hydrological model reproduces the flow index PCs exactly) in the ungauged catchment, but the regionalization model has a "realistic" error (based on scenario 1). The flow index PCs in a given ungauged catchment are generated synthetically, as follows. First, a "reference" synthetic flow time series $\mathbf{q}^{\text{ref}} = \mathbf{H}(\theta)$ is generated using the hydrological model with a known "reference" parameter set $\theta_{\text{ref}}$. This reference parameter set is obtained by minimizing the normalized distance between the $N_\mathbf{z}$ dominant PCs calculated for the observed and simulated flows in the ungauged catchment, $\sqrt{\sum_{i=1}^{N\mathbf{z}} \left(\frac{\mathbf{z}_i^{\text{sim}} - \mathbf{z}^{\text{obs}}}{\text{sdev}\left[\mathbf{z}_i^{\text{sim}}\right]}\right)^2}$, where $\text{sdev}\left[\mathbf{z}_i^{\text{sim}}\right]$ denotes the standard deviation of the $i$th component computed from 1,000 model simulations with parameters sampled from the prior $p(\theta)$. Second, the flow index PCs of the synthetic flow time series are computed, and treated as synthetic $\mathbf{z}^{\text{obs}}$. Third, these flow index PCs are corrupted with the same error values as incurred by the regionalization model in scenarios 0 and 1 and set to represent a synthetic deterministic term $\mathbf{r}$ in equation (4). Fourth, the random term in equation (4) is set to have the same characteristics as estimated in scenario 1. Given this synthetic model setup, adequacy tests of the regionalization model produce the same results as in scenario 0. However, as will be discussed in section 4.3.1, an "exact" hydrological model might be not "adequate" in simulating flow index PCs if the latter are biased and/or noisy.
- *Scenario 4*. In this scenario the hydrological model is replaced by the observed time series and the regionalization model is near perfect (accurate and with little noise in the space of flow index PCs). Synthetic $\mathbf{z}^{\text{obs}}$ are generated as in scenario 3, and then the regionalization model is constructed as in scenario 2.

### 3.6. Model Performance Metrics for Flow Time Series

In Analyses 3 and 4, model performance in the flow time series space is appraised in terms of three attributes of a probabilistic prediction: accuracy, precision, and reliability. Accuracy quantifies the distance between the central values of a predictive distribution (e.g., its expectation) and the observed values; precision characterizes the spread of the predictive distribution; and reliability quantifies the degree of statistical consistency of the predictive distribution with the observed data.

Accuracy and precision are quantified jointly using a probabilistic extension of the original (deterministic) NSE (Nash & Sutcliffe, 1970). In this metric, denoted as $\Phi_{\mathrm{NSP}}$, accuracy is penalized by (poor) precision as follows (Bulygina et al., 2009):

$$\Phi_{\mathrm{NSP}} = \text{accuracy}-\text{precision} = \left[1-\frac{\sum_{t=1}^{T}\left(\mathrm{E}[\mathbf{q}_t]-q_t^{\mathrm{obs}}\right)^2}{\sum_{t=1}^{T}\left(q_t^{\mathrm{obs}}-\mathrm{E}\left[q_t^{\mathrm{obs}}\right]\right)^2}\right]-\frac{\sum_{t=1}^{T}\left(\mathrm{var}[\mathbf{q}_t]\right)}{\sum_{t=1}^{T}\left(q_t^{\mathrm{obs}}-\mathrm{E}\left[q_t^{\mathrm{obs}}\right]\right)^2} \tag{14}$$

where $\mathbf{q}_t$ is the set of simulated flows at time step t generated using the set of sampled hydrological model parameters, $q_t^{\mathrm{obs}}$ is the observed flow, E[.] denotes the expected value, var[.] denotes the variance, and $T$ is the total number of time steps. The accuracy term corresponds to the deterministic NSE of the expected values of simulated flows. The precision term is given by the sum of variances of predicted flows, scaled by the sum of squared residuals (higher variance corresponds to lower precision). When $\Phi_{\mathrm{NSP}} = 1$, the predictions provide a perfect match to the data (i.e., both perfectly accurate and precise).

Note that except for the case of predictions with no uncertainty (not relevant here), $\Phi_{\mathrm{NSP}}$ always has lower values than has the deterministic NSE, due to the (nonnegative) precision penalty term in equation (14). Therefore, the probabilistic NSE is a more stringent metric than is the deterministic one. It is also a more complete metric because it accounts for the width of probability limits (precision).

Predictive reliability is quantified using the 95% coverage interval $\gamma_{95\%}$, defined as the percentage of observations that fall into the 95% posterior probability limits (Yadav et al., 2007). A reliable prediction is characterized by $\gamma_{95\%}$ values close to 95%; larger and smaller values indicate, respectively, underestimation and overestimation of predictive uncertainty. A more comprehensive measure of reliability is given by the predictive QQ plot (Renard et al., 2010), but is not carried out in this work.

# 4. Results

## 4.1. Analysis 1: Selection of Flow Index PCs

From the complete set of 103 flow indices $\boldsymbol{w}^{\mathrm{obs}}$ in each of the 92 catchments, the application of PCA and the broken-stick method identifies four flow index PCs $\mathbf{z}^{\mathrm{obs}}$ that collectively explain 87% of the variability in the flow indices (individual contributions of 63%, 12%, 7%, and 5%, respectively). These dominant PCs appear to be made up by combination of multiple flow indices, with no single index having much larger weight than the others.

## 4.2. Analysis 2: Evaluation of the Probabilistic Regionalization Model

The four-dimensional likelihood function $p\left(\mathbf{z}^{\mathrm{reg}}|\mathbf{z}_{\theta}^{\mathrm{sim}}\right)$ in equation (13) is estimated by fitting a parametric pdf to the distribution of residuals $\mathbf{z}^{\mathrm{reg}}-\mathbf{z}_{\theta}^{\mathrm{sim}}$ of the RF regionalization model (section 2.2). Pearson's linear correlation test indicates a significant correlation ($p$ value of 0.03) between the residuals of $z_2$ and $z_3$, and no significant correlation (at the 0.05 significance level) between other pairs of residuals. Hence, the likelihood function is constructed as a product of marginal probability distributions of regionalization residual errors for $z_1$ and $z_4$ and joint probability distribution for $\mathrm{z}_2$ and $\mathrm{z}_3$, as follows:

$$p\left(\mathbf{z}^{\mathrm{reg}}|\mathbf{z}_{\theta}^{\mathrm{sim}}\right) = p\left(\mathbf{z}^{\mathrm{reg}}-\mathbf{z}_{\theta}^{\mathrm{sim}}\right) = p\left(z_1^{\mathrm{reg}}-z_{1,\theta}^{\mathrm{sim}}\right)\times p\left(z_2^{\mathrm{reg}}-z_{2,\theta}^{\mathrm{sim}}, z_3^{\mathrm{reg}}-z_{3,\theta}^{\mathrm{sim}}\right)\times p\left(z_4^{\mathrm{reg}}-z_{4,\theta}^{\mathrm{sim}}\right) \tag{15}$$

where subscripts index the components of the dominant flow index PCs.

The $\chi^2$ test (Pearson, 1900) on the PC residuals suggests (at 0.05 significance) that (1) PC1 residual distribution $p\left(z_1^{\mathrm{reg}}-z_{1,\theta}^{\mathrm{sim}}\right)$ can be approximated using the Extreme Value Type 1 (Gumbel) distribution and (2) PC4 residual distribution $p\left(z_4^{\mathrm{reg}}-z_{4,\theta}^{\mathrm{sim}}\right)$ can be approximated using a Gaussian distribution. In addition, $\chi^2$ tests (Pearson, 1900) and Mardia tests (Mardia, 1970) applied at the 0.05 significance level suggest that the joint residual distribution $p\left(z_2^{\mathrm{reg}}-z_{2,\theta}^{\mathrm{sim}}, z_3^{\mathrm{reg}}-z_{3,\theta}^{\mathrm{sim}}\right)$ is approximately Gaussian.

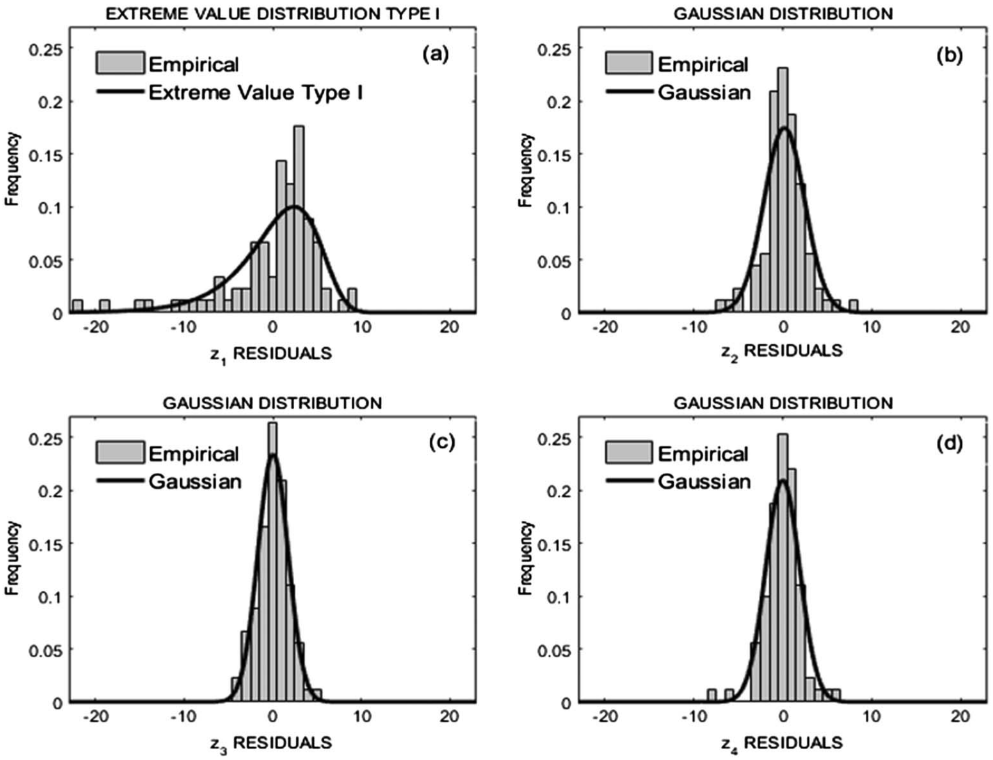


**Figure 4.** Empirical and fitted distributions of residual errors of the regionalization model in catchment c8z1 (Leizarán). The regionalization model was constructed using 91 catchments, excluding catchment c8z1.

Figure 4 illustrates the marginal and joint distributions of residual errors of the regionalization model estimated from 91 catchments (excluding the Leizarán River Basin c8z1 treated as ungauged). Table 4 shows the ranges of the mean and standard deviation of the residuals distribution of the regionalization model (columns 2 and 3 in Table 4) and averaged normalized mean and standard deviation of the residual errors (columns 4 to 7 in Table 4) across all 92 catchments. The mean and standard deviation are normalized by the range of flow index PCs from observations (columns 4 and 5 in Table 4) or from hydrological model simulations (columns 6 and 7 in Table 4).

When normalized by the range of observed PCs, the regionalization model residuals have a bias (mean) of 0–12% and a spread (standard deviation) of 12–16%. The regionalization of $z_2$, $z_3$, and $z_4$ appears relatively unbiased (mean of residual errors close to 0), whereas the regionalization of $z_1$ has a (normalized) bias of 12%. The regionalization of $z_1$ and $z_4$ is the most precise (narrowest spread), while the regionalization of $z_3$ is the least precise (widest relative spread).

**Table 4**
*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 | Range of values of the standard deviation | Average values normalized by range of $\mathbf{z}^{obs}$ | | Average values normalized by range of $\mathbf{z}^{sim}$ | |
|---|---|---|---|---|---|---|
| | | | Mean | Standard deviation | Mean | Standard deviation |
| $z_1$ | 4.13–4.72 | 4.51–4.86 | 0.12 | 0.12 | 0.40 | 0.42 |
| $z_2$ | 0.14–0.20 | 1.34–2.37 | 0.01 | 0.13 | 0.01 | 0.09 |
| $z_3$ | 0.02–0.05 | 1.65–2.71 | 0.00 | 0.16 | 0.00 | 0.10 |
| $z_4$ | 3.52–3.79 | 1.84–1.93 | 0.00 | 0.12 | 0.00 | 0.17 |

*Note.* Normalizations with respect to the observation-based PC range (based on 92 catchments) and catchment-specific hydrological model simulation-based PC range are also reported. PC = principal component.

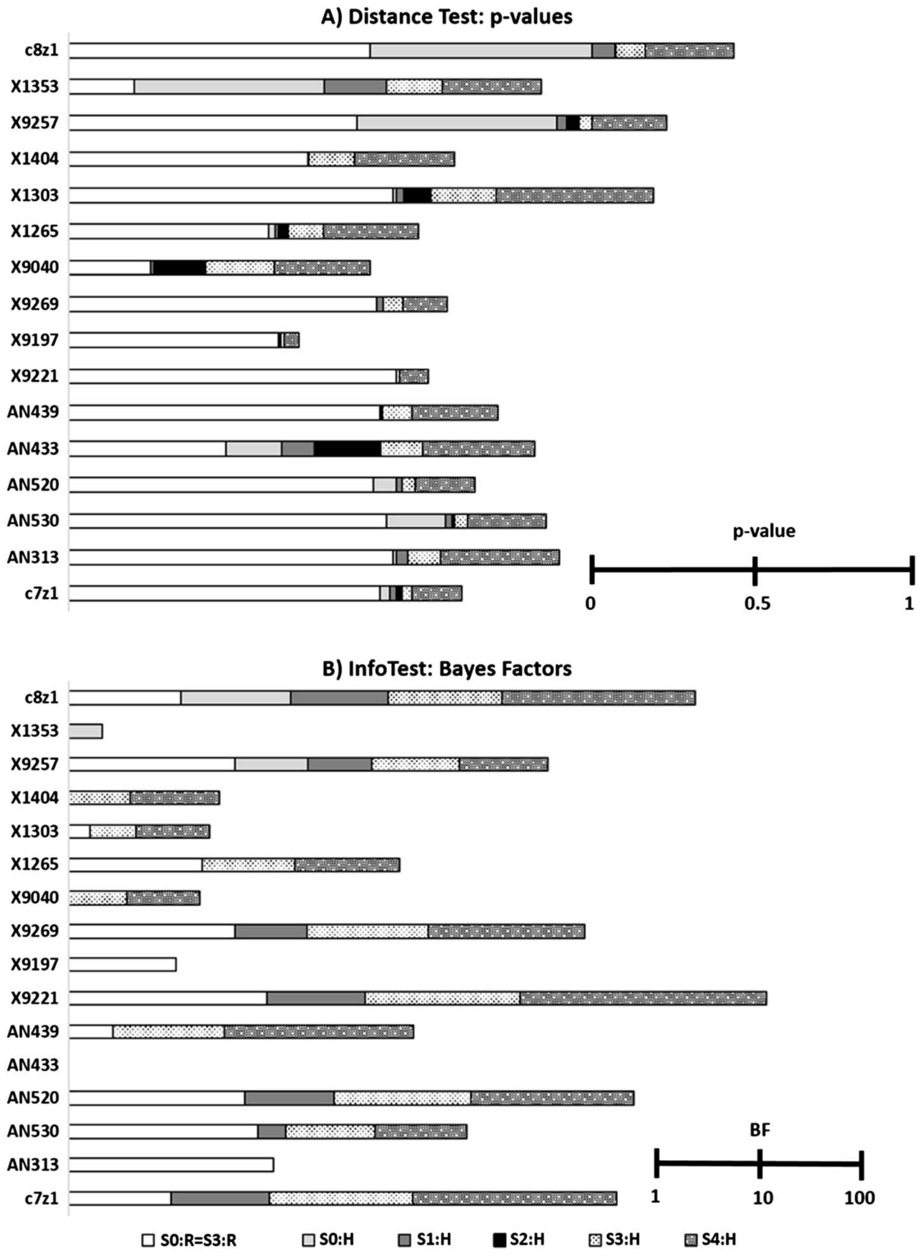


**Figure 5.** Model adequacy tests DistanceTest and InfoTest in Analysis 3, scenarios 0–4. “H” denotes metrics applied to the hydrological model, and “R” denotes metrics applied to the regionalization model. The *y*-axis lists the catchments; the *p* values and Bayes factors (BFs) are given by the length of the bars, with BF shown in log scale and only for scenarios where BF > 1. See supporting information Tables S1 and S2 for the numerical values of the metrics.

When the parameters of the regionalization error model are normalized by the range of simulated PCs, the apparent bias and spread increase, due to a tighter normalization range. The biggest effect is seen for the $z_1$ residuals—normalized bias and spread increase to 40% and 42%, respectively.

## 4.3. Analysis 3: Understanding the Properties of Adequacy Tests

### 4.3.1. Adequacy Test Results

Figure 5 shows DistanceTest and InfoTest results for the five scenarios of Analysis 3 in each of the 16 catchments treated as ungauged. The scenarios are “S0:R,” “S0:H,” “S1:H,” “S2:H,” “S3:R,” “S3:H,” and “S4:H,” where “R” and “H” refer, respectively, to whether the regionalization or hydrological model is being tested. Note that the regionalization model can only be tested in scenarios 0 and 3. The complete set of numerical values is reported in the supporting information. The results are summarized below for each scenario.

- *Scenario 0*: $\mathbf{z}^{obs}$ available and used to check the hydrological and regionalization models individually

Figure 5 (S0:R) shows that the RF regionalization model passes DistanceTest in all 16 catchments but passes InfoTest in only 12 catchments. In other words, the regionalization model is able to adequately reproduce $\mathbf{z}^{obs}$ in 12 of 16 catchments (75% of the catchments). Figure 5 (S0:H) shows that the hydrological model PDM passes DistanceTest in 6 catchments and passes InfoTests in 3 catchments, so that PDM is able to adequately reproduce $\mathbf{z}^{obs}$ in only 3 of the 16 catchments (c8z1, X1353, and X9257). The large number of catchments where the regionalization model is adequate, but the hydrological model is not, suggests that the hydrological model tends to be the dominant source of error, at least in the case study area.

- *Scenario 1*: Real operating conditions where the hydrological model is conditioned on regionalized flow index PCs

Figure 5 (S1:H) shows that the combined regionalization/hydrological model is adequate in only a single catchment, c8z1, where it reproduces regionalized flow index PCs with high probability and adds information over the prior knowledge about the flow index PCs.

- *Scenario 2*: Condition an "inexact" hydrological model on flow index PCs with small regionalization error

In this scenario, the regionalization model is devised to be adequate (section 3.5); hence, its adequacy tests results are not shown in Figure 5. The hydrological model is inadequate in representing the regionalized information in all catchments (Figure 5, S2:H). More specifically, the hydrological model passes DistanceTest in 3 of 16 catchments but fails InfoTest in all of them (BF < 0.5).

- *Scenario 3*: Condition an "exact" hydrological model on flow index PCs with "realistic" regionalization errors.

Figure 5 (S3:H) shows that even an "exact" hydrological model conditioned on inaccurate and noisy $\mathbf{z}^{reg}$ fails DistanceTest and InfoTests in 9 of 16 catchments.

- *Scenario 4*: Condition an "exact" hydrological model on flow index PCs with small regionalization error

In this scenario, the regionalization model is devised to be adequate (section 3.5), the same as in scenario 2 (results hence not shown in Figure 5). The focus is hence on the hydrological model, which in this synthetic scenario is "exact" (section 3.5). Interestingly, Figure 5 shows that the hydrological model fails DistanceTest in 1 of 16 catchments and fails InfoTest in 4 of 16 catchments (in different catchments than in scenario 0). The single DistanceTest failure is not unexpected: As the *p* value threshold to pass DistanceTest is set to 0.05 (5%), even an adequate model is expected to fail DistanceTest on average once for every 20 catchments.

#### 4.3.2. Performance of Flow Time Series Predictions in the Ungauged Catchments

This section reports model performance in terms of flow time series predictions and relates it to model performance in terms of the adequacy tests. Figure 6 reports the $\Phi_{NSP}$ and $\gamma_{95\%}$ performance metrics (section 3.6) in all 16 "ungauged" catchments.

First, consider changes in model performance from scenario 1 to scenario 2. Figures 6a and 6c show that in the single catchment where the hydrological model is found adequate in representing regionalized information (i.e., catchment c8z1; see section 4.3.1, scenario 1), improving the regionalization model results in improved flow predictions: $\Phi_{NSP}$ improves slightly from 0.68 to 0.71, and $\gamma_{95\%}$ improves from 70% to 74%. In contrast, in the 15 catchments where the hydrological model is found inadequate, improving the regionalization model as achieved "synthetically" in scenario 2 does not systematically improve prediction quality. For example, in catchment AN313 $\Phi_{NSP}$ worsens from 0.3 to 0.25 and $\gamma_{95\%}$ worsens from 77% to 75%, but in catchment X1353 $\Phi_{NSP}$ improves from 0.6 to 0.7 and $\gamma_{95\%}$ improves from 45% to 51%.

Second, consider changes in model performance from scenario 3 to scenario 4. Figure 6 shows improvements in $\Phi_{NSP}$ and $\gamma_{95\%}$ in all catchments. For example, in catchment AN313 $\Phi_{NSP}$ improves from 0.77 to 0.99 and $\gamma_{95\%}$ remains at 100% (indicating an overestimation of predictive uncertainty). This result

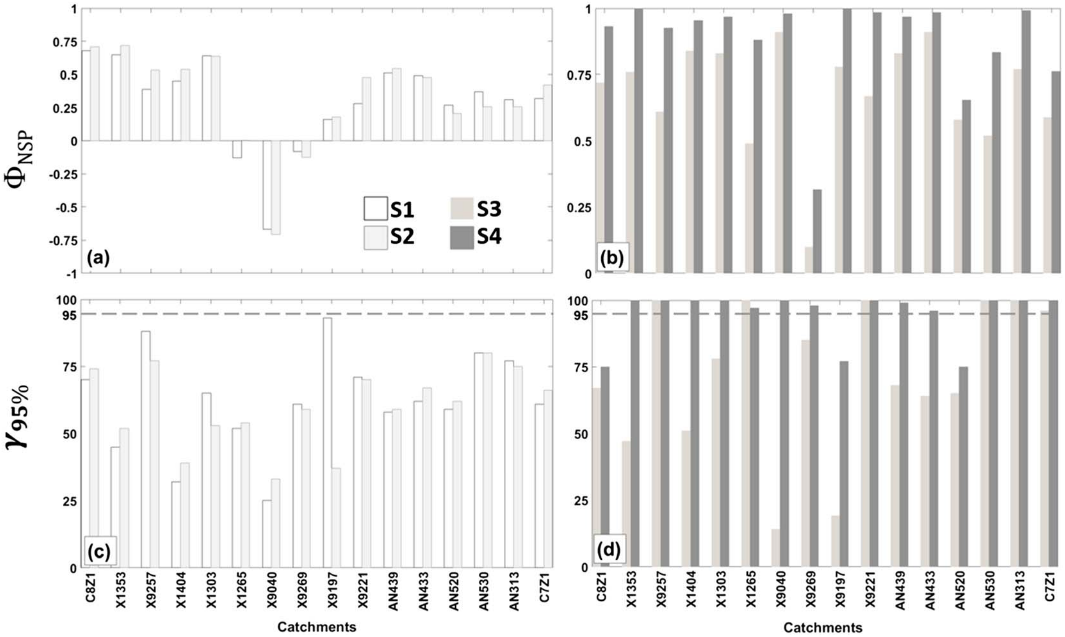


**Figure 6.** Model performance in flow time series space. (a) Probabilistic Nash-Sutcliffe efficiency, $\Phi_{NSP}$, for scenarios 1 and 2; (b) $\Phi_{NSP}$ for scenarios 3 and 4; (c) 95% coverage interval, $\gamma_{95\%}$, for scenarios 1 and 2; and (d) $\gamma_{95\%}$ for scenarios 3 and 4. In scenarios 1 and 2 model predictions are compared against observed data; in scenarios 3 and 4 model predictions are compared against synthetic data (see section 3.5).

suggests that when the hydrological model is exact, improving the regionalization model leads to better flow predictions. This finding is also supported by the ANOVA, which rejects the null hypothesis that "when the hydrological model is exact, improving regionalization has no effect on $\Phi_{NSP}$ and $\gamma_{95\%}$."

### 4.4. Analysis 4: Detailed Illustration of Model Performance in Selected Catchments

Figures 7a–7d show the predicted flow time series (95% probability limits) for scenarios 1–4 in catchment c8z1, where PDM is adequate. In scenarios where regionalization quality is representative of real conditions (scenarios 1 and 3), the adequate hydrological model provides predicted flow time series of similar quality to the exact hydrological model. Further, improving the regionalization model improves the quality of predicted flow time series generated by the adequate hydrological model—both when switching from scenario 1 to scenario 2 and when switching from scenario 3 to scenario 4.

Figures 7e–7h show the predicted flow time series (95% probability limits) for scenarios 1 to 4 in catchment X9040, where neither the hydrological nor regionalization models are adequate. Note that catchment X9040 is affected by snow melt, which is not represented by PDM (April–July period in scenarios 1 and 2 in Figures 7e and 7f). Results in scenario 2 illustrate that when the hydrological model is inadequate, its error precludes the reproduction of flow time series even when the regionalization model is adequate. Moreover, results in scenario 3 show the importance of regionalization model adequacy: Even when the hydrological model is exact, conditioning on poor-quality flow index PCs yields low-quality flow predictions; for example, the 95% probability limits envelop just 14% of observed flows ($\gamma_{95\%} = 14\%$).

## 5. Discussion

### 5.1. Selection of Flow Indices PCs

Section 4.1 shows that the first 4 flow index PCs collectively explain 87% of the variability in the flow indices. This indicates that most of the information contained in the hydrological indices can be captured by just a few quantities, which is expected to be useful in ungauged catchment applications (Wagener & Montanari, 2011). Further, section 4.3 shows that in some catchments, the most significant PCs regionalized by RF are sufficient to obtain good performance of the hydrological model predictions. Note that the

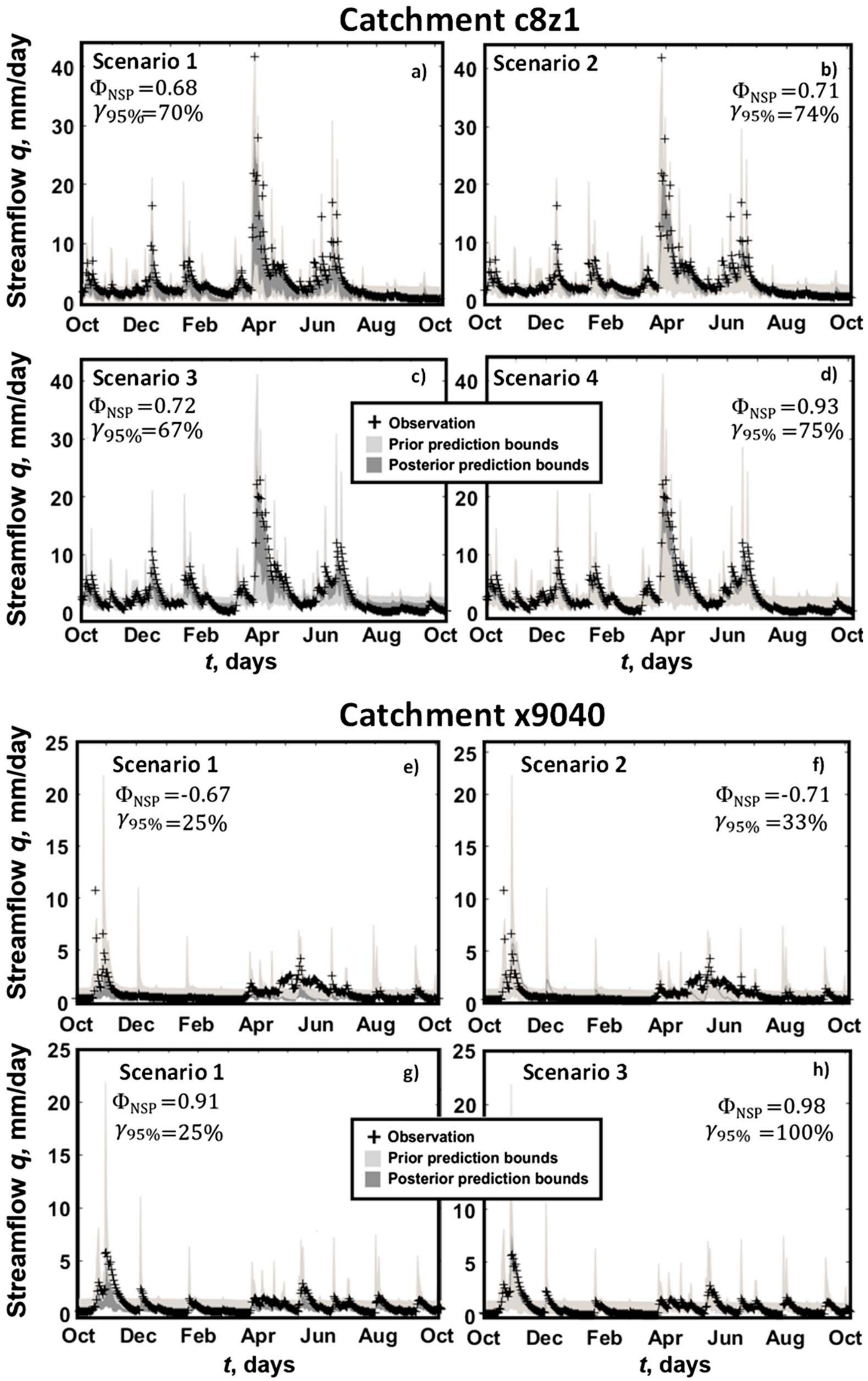


**Figure 7.** (a–d) Prior and posterior 95% probability limits and posterior performance metrics for flow predictions in catchment c8z1 (Leizarán), scenarios 1–4. Both the hydrological and regionalization models are found adequate in this catchment. (e–h) Prior and posterior 95% probability limits and posterior performance metrics for flow predictions in catchment X9040 (Ara), scenarios 1–4. Neither the hydrological nor regionalization models are found adequate in this catchment.

available information (here 103 hydrological indices) is still a small portion of all possible existing information.

### 5.2. Evaluation of the Probabilistic Regionalization Model

Section 4.2 shows that when the parameters of the regionalization model are normalized by the range of most significant PCs computed from observed data, the spread is relatively narrow. The spread is comparable with the case in which the normalization range is based on the range of most significant PCs computed from model simulations. The only exception is $z_1$, for which the regionalized value has the highest bias and spread when normalized by the range of simulated PCs (see Table 4). This behavior indicates that the hydrological model provides only a limited coverage of $z_1$ value range generated by the regionalization model and suggests that the regionalization error is large relative to the range of $z_1$ values simulated by the hydrological model. Therefore, $z_1$ might be of limited utility for conditioning the selected hydrological model parameters despite being the dominant (most informative) flow index PC from the complete set of flow indices $\boldsymbol{w}$, as conditioning on $z_1$ might not be efficient in reducing PDM parametric uncertainty or generalized $z_1$ might be outside of $z_1$ range produced by PDM.

Nonparametric distributions (e.g., mixtures of Gaussians) might be advantageous when working with larger data sets (e.g., flow indices from more than the 90 catchments used in this study). For shorter data sets, nonparametric distributions could be difficult to fit parsimoniously. For example, to model the four-dimensional vector of flow index PCs considered in the paper using a single joint mixture of Gaussians distribution, one needs 4 parameters to define a Gaussian kernel mean, 10 parameters to define the covariance matrix, and 1 kernel weight coefficient, hence 15 parameters in total. If the mixture distribution were to comprise several such kernels, say, 6 kernels, the number of parameters ($6 \times 15 - 1 = 90 - 1 = 89$) would approach the number of data points (90 sets of PCs in our case). This might result in poor identifiability and overfitting.

### 5.3. Adequacy Tests in Analysis 3

Scenario 0 in section 4.3.1 shows that catchments where the regionalization model fails InfoTest are harder to model using the combined regionalization/hydrological model, because the information available to condition the hydrological model is less reliable. In addition, results in scenario 0 indicate that there is a large number of catchments where the regionalization model is adequate, but the hydrological model is not (see scenario 0 in Figure 5). In other words, in the case study area, the hydrological model tends to be the dominant source of error in comparison with the regionalization model.

Moving from the adequacy test for the hydrological model in scenario 0 (S0:H in Figure 5) to scenario 1 (S1:H in Figure 5), the number of catchments where the hydrological model is adequate reduces from 3 to 1 (see section 4.3.1). This finding can be attributed to the uncertainty due to the regionalization model, especially as the uncertainty in $\mathbf{z}^{\text{reg}}$ can often be expected to exceed the uncertainty in $\mathbf{z}^{\text{obs}}$. More generally, the combined regionalization/hydrological model could be inadequate for several potential reasons, including model structure deficiencies and data errors. For example, for the hydrological model, the following potential sources of error are noted:

1. The PDM model structure does not represent snow melt (e.g., catchment X9040 in Figure 7), or Hortonian runoff process/intermittent rivers (e.g., catchment X9257), or deep aquifers (e.g., catchment AN439—Larraún Irutzun River Basin; J. L. Navarra, personal communication, March 15, 2018).
2. The rainfall data are uncertain due to a limited number of rain gauges and their nonuniform coverage, while rainfall variability is expected to be high due to orographic precipitation in catchments draining into the Cantabrian sea and convective precipitation in catchments draining into the Mediterranean sea. Daily PET values are calculated for each month and distributed uniformly to produce daily values, which might be inaccurate for dry catchments.
3. The flow data (and hence flow index PCs) are expected to be affected by discharge gauging errors, rating curves extrapolation, flow regime hysteresis (Westerberg et al., 2011), and channel hydraulic property changes due to the lack of a control section in the majority of the catchments (Kuczera, 1996; Renard et al., 2011).

Potential sources of error in the regionalization model have been noted in section 2.2.2.

Results in scenario 2 are generally consistent with results in scenario 0: In both scenarios the hydrological model struggled to match the observed flow index PCs. Scenario 2 adds small noise to the observed flow index PCs, and the hydrological model adequacy deteriorates further, as compared with scenario 0.

Results of the adequacy test for the hydrological model in scenario 3 are reassuring as they imply that the hydrological model will not reproduce erroneous flow index PCs. However, it is not clear if this finding is specific to PDM, nor is it clear if a more heavily parameterized model might behave differently because of its increased flexibility in matching conditioning data.

Finally, there are some catchments that fail to pass InfoTest in scenario 4. This might be due to the 5% noise in the regionalization model. Note that InfoTest is only indicative of whether the model generates $\mathbf{z}^{\mathrm{sim}}$ close to $\mathbf{z}^{\mathrm{reg}}$ more frequently than the prior (see section 2.3.2), and achieving a high frequency of such occurrences is not guaranteed for the hydrological model parameter set used in this scenario. Hence, unless the hydrological model is conditioned to additional data (e.g., regionalized flow index PCs beyond those identified as dominant by the broken stick method; see section 4.3.2), the frequency at which the model will reproduce the exact flow index PCs will be relatively low. This question links to information quality and quantity of information, that is, what additional data/information are needed for those catchments that fail in scenario 4 and what is the size of error that can be propagated into the hydrological model. These research questions are recommended for future work.

### 5.4. Performance of Flow Time Series Predictions

The analysis of flow time series predictions in ungauged catchments (section 4.3.2) showed that moving from scenario 1 to scenario 2—that is, improving only the regionalization model while the hydrological model is kept as is—does not systematically translate into an improvement in the flow hydrograph (little improvement is seen in $\Phi_{\mathrm{NSP}}$ and $\gamma_{95\%}$). In other words, the error in the hydrological model (which includes its structure, inputs and parameters) is precluding an improvement in flow predictions even if the regionalization model improves (see scenario 2).

However, when the hydrological model is exact, improving the regionalization model (i.e., moving from scenario 3 to scenario 4) leads to better flow predictions. This finding is supported by the ANOVA, which rejects the null hypothesis that "when the hydrological model is exact, improving regionalization has no effect on $\Phi_{\mathrm{NSP}}$ and $\gamma_{95\%}$." Consequently, an improvement in the regionalization leads to improved $\Phi_{\mathrm{NSP}}$ and $\gamma_{95\%}$ over most catchments only if the hydrological model error is eliminated/reduced first (see scenario 4). This finding is consistent with the results in section 4.3.1 (scenarios 2 and 4).

Therefore, given their definition based on flow index PCs, the model adequacy tests proposed in this work can be expected to become better indicators of flow time series performance if the following enhancements are undertaken in the regionalization process: (i) more diverse and representative sets of hydrological (flow) indices $\boldsymbol{w}$ and catchment descriptors $\mathbf{d}^{\mathrm{obs}}$ are used and (ii) more PCs are included in $\mathbf{z}^{\mathrm{obs}}$.

### 5.5. Benefits and Limitations of the Adequacy Metrics

This section compares the adequacy metrics proposed in section 2.3 with traditional metrics of hydrological model performance.

First, and most importantly in the context of prediction in ungauged catchments, the adequacy metrics are applied *a priori* using regionalized information. In contrast, conventional application of NSE-type metrics (deterministic or probabilistic) and prediction interval coverage to predicted flow time series represent *a posteriori* diagnostic metrics and cannot be applied in ungauged applications (where observed flow data are unavailable).

Second, the adequacy metrics are applied to flow index PCs, which allows quantifying the degree to which the model reproduces dominant flow characteristics (Fenicia et al., 2018; Westerberg & McMillan, 2015; Yilmaz et al., 2008). This study uses indices characterizing annual and monthly flows, high and low flows, hydrograph timing, and so on (section 3.2), and the modelers could refine their selection if guided by particular operational priorities. In contrast, conventional application of NSE and prediction interval coverage to streamflow time series yields highly aggregated metrics (e.g., Clark et al., 2011; Gupta et al., 2008; Seibert, 2001). For example, NSE and similar metrics are most sensitive to high flow values, which often results in poor prediction of low flows and is undesirable when modeling low-yield ephemeral catchments (e.g., Ye et al., 1998). In the context of this work, $\Phi_{\mathrm{NSP}}$ and $\gamma_{95\%}$ do not provide any indication of the added value of using a model to predict the dominant flow characteristics (as represented by the flow index PCs).

Third, the adequacy metrics are probabilistic in nature, whereas many traditional hydrological metrics, most notably the original NSE (Nash & Sutcliffe, 1970), cater solely to deterministic predictions. In this respect, the adequacy metrics contribute toward better probabilistic prediction assessment, which currently includes the probabilistic NSE (Bulygina et al., 2009), the continuous rank probability score (Hersbach, 2000), reliability and precision metrics (McInerney et al., 2017), and other metrics.

However, like any other diagnostic metric, the adequacy tests have their limitations. As highlighted in section 2.2.3, passing the adequacy tests for the flow indices does not guarantee that the model is able to reproduce the flow time series.

Overall, these considerations make the adequacy metrics highly attractive for assessing model performance, especially in ungauged applications.

### 5.6. Illustration of Model Performance in Selected Catchments

Figure 7 illustrates model performance in two representative cases, exemplified by catchments c8z1 and x9040. Figures 7a–7d show that in catchment c8z1, where the hydrological model is adequate, improving the regionalization model improves the quality of flow predictions. However, in catchment X9040 (Figures 7e–7h), where the hydrological and regionalization models are inadequate, the hydrological model error precludes the reproduction of flow time series even when the regionalization model is adequate. In this catchment, an improvement in the regionalization model will lead to better predictions only if the hydrological model is improved first.

### 5.7. Model Performance in the Context of Operational Predictions

The hydrological prediction results obtained in this study can be put in the context of operational work in the same geographical area. In scenario 1, where real data/models are used, the *probabilistic* NSE of predictions in catchments (treated as *ungauged*) was in the range [−0.65, 0.68], as seen in Figure 6. In operational studies in the north of Spain, representative ranges of *deterministic* NSE of predictions in *gauged* catchments were found to be [0.25, 0.72] for a study in Cantabria (García et al., 2008; Gobierno de Cantabria, 2005) and [−0.4, 0.9] for a study in the Basque Country (Agencia Vasca del Agua, 2003). Considering that this research study calibrated to regionalized flow indices, whereas the operational studies cited above calibrated to observed flow time series data, and that the probabilistic NSE is lower than the deterministic NSE (e.g., by 0.02–0.2 units in scenario 1), it can be seen that comparable or somewhat better performance is achieved in this research study. According to the classification suggested by Foglia et al. (2009) to interpret deterministic NSE values of flow prediction in gauged catchments, the probabilistic NSE values in ungauged catchments achieved in this work are in the range of “sufficient” to “very good” in 13 of the 16 catchments. While this classification is obviously subjective, taken together with the quantitative results cited above, it highlights the promise offered by the prediction in ungauged catchments methods developed in the present paper.

### 5.8. Future Research

First, our application assumes that hydrological model errors are small/negligible compared with regionalization model errors. This assumption follows published work on the conditioning of hydrological parameters to streamflow statistics (e.g., Almeida et al., 2016; Bulygina et al., 2009, 2012; Yadav et al., 2007). However, this assumption is questionable (see the conclusions in Almeida et al., 2016; Bulygina et al., 2009, 2011), and it is of interest to investigate ways to relax it, for example, by considering an ensemble of (sufficiently different) hydrological models. Multiple models can yield insight into hydrological model errors (e.g., Clark et al., 2008; Clark et al., 2015; Fenicia et al., 2011; Wagener et al., 2004; Wagener & Montanari, 2011), and in some cases model ensembles have been shown to reduce predictive errors (e.g., Georgakakos et al., 2004; Hsu et al., 2009; Shamseldin et al., 1997). Future work could consider the use of adequacy metrics on model ensembles.

Second, the error in the regionalized estimates of **z** could be estimated by expanding the number of residuals draws and looking into individual tree predictive errors (not the entire forest average). This approach would reduce the computational cost of the estimation at the expense of larger regionalization errors.

Computational cost is not a major aspect in this work (less than a second per catchment) but could be important in some other applications.

Third, the sensitivity of the predictions of the flow dynamics (hydrograph) to the quantity and quality of information in **z** warrants further investigation. For example, the relationship between the variance represented by **z** and the information needed to reproduce the hydrograph is of interest, as well as understanding how this relationship depends on the type of catchment.

Finally, the analyses presented in this paper could be extended to a larger set of catchments, and the model results verified more comprehensively using cross-validation over more than the 16 catchments used here. This was not possible in this study due to the lack of synchronized data.

## 6. Conclusions

This study offers two advances to flow prediction in ungauged catchments: (1) combination of a regionalization method, implemented using the machine learning technique RF augmented with a probabilistic residual error model, with a Bayesian inference formulated for regionalized PCs of a set of flow indices, and (2) development of model adequacy tests, namely, DistanceTest and InfoTest, computed using the regionalized flow index PCs, to provide an a priori indication of the ability of a hydrological model to predict flow time series in an ungauged catchment. More specifically, in a given ungauged catchment, DistanceTest quantifies whether a model (hydrological and/or regionalization) is likely to reproduce regionalized flow index PCs, and InfoTest quantifies whether a model adds information about flow index PCs beyond what is already known from prior knowledge (here the ranges of flow index PCs in gauged catchments).

An empirical case study based on 92 catchments in northern Spain is presented, where the proposed methods are tested on 16 catchments treated as ungauged. The following findings are obtained:

1. The high-dimensional space of flow indices is reduced substantially via PC analysis and the broken-stick method, from 103 indices to just 4 dominant PCs that explain 87% of the variance in the indices.
2. The errors in the regionalized flow index PCs estimated using RF regression are of the order of 12–16% of the observed values. Regionalization via RF regression is adequate in 12 of the 16 catchments (75% of catchments).
3. The first four PCs regionalized via RFs provide sufficient information for predictions of flow time series in many (though not all) of the case study catchments (treated as ungauged), with probabilistic NSE values ranging from −0.65 to 0.68. A preliminary comparison with selected operational water resources studies on gauged catchments in the same geographic area suggests broadly comparable performance, with opportunities for further improvement.
4. The adequacy tests in the flow index PC space are indicative of model performance in the flow time series space. The following insights are attained:
   a. When a hydrological model is adequate (i.e., passes the adequacy tests), improvements in the regionalization model translate into improvements in flow predictions.
   b. When a hydrological model is inadequate, improvements in the regionalization model do not systematically propagate into improved flow predictions. The model adequacy tests can be considered a prerequisite for a hydrological model to attain meaningful and high-quality flow time series predictions in ungauged catchments.
5. The adequacy tests yield useful insights that can help identify dominant sources of predictive error: hydrological model, regionalization model, or both. This error attribution can help prioritize future studies and developments.

An important limitation identified using model adequacy tests and flow time series metrics is the poor performance of the hydrological model PDM forced with observed rainfall and estimated PET in the case study catchments. In these catchments, PDM struggles to reproduce the hydrological characteristics given by the observed flow index PCs. Note that this study has not attempted to separate total flow error into individual contributions due to model structural errors versus the effects of observational data errors. Overall, current results suggest that priority should be given to improving the hydrological model structure and its inputs (including a better characterization of predictive uncertainty) and then to improving the regionalization model.

## Appendix A

InfoTest equation (10) is derived by applying Bayes' law and using statistical independence between $\mathbf{z}^{\text{reg}}$, $\mathbf{H}$, and $\mathbf{x}$ to estimate how much information is added by the combination of the hydrological model, its inputs, and the regionalization over the regionalization alone:

$$\text{BF}_{\mathbf{H}} = \frac{p(\mathbf{z}|\mathbf{z}^{\text{reg}},\mathbf{H},\mathbf{x})}{p(\mathbf{z}|\mathbf{z}^{\text{reg}})} = \frac{p(\mathbf{z}^{\text{reg}}|\mathbf{z},\mathbf{H},\mathbf{x})p(\mathbf{z}|\mathbf{H},\mathbf{x})}{p(\mathbf{z}^{\text{reg}}|\mathbf{H},\mathbf{x})p(\mathbf{z}|\mathbf{z}^{\text{reg}})} = \frac{p(\mathbf{z}^{\text{reg}}|\mathbf{z})p(\mathbf{z}|\mathbf{H},\mathbf{x})}{p(\mathbf{z}^{\text{reg}})p(\mathbf{z}|\mathbf{z}^{\text{reg}})} = \frac{p(\mathbf{z}|\mathbf{H},\mathbf{x})}{p(\mathbf{z})} \tag{A1}$$

Similarly, InfoTest equation (11) is derived by considering how much information is added by the combination of the hydrological model, its inputs, and the regionalization model over the hydrological model alone:

$$\text{BF}_{\mathbf{R}} = \frac{p(\mathbf{z}|\mathbf{z}^{\text{reg}},\mathbf{H},\mathbf{x})}{p(\mathbf{z}|\mathbf{H},\mathbf{x})} = \frac{p\left(\mathbf{z}_{\theta}^{\text{sim}}|\mathbf{z},\mathbf{z}^{\text{reg}}\right)p(\mathbf{z}|\mathbf{z}^{\text{reg}})}{p\left(\mathbf{z}_{\theta}^{\text{sim}}|\mathbf{z}^{\text{reg}}\right)p\left(\mathbf{z}|\mathbf{z}_{\theta}^{\text{sim}}\right)} = \frac{p\left(\mathbf{z}_{\theta}^{\text{sim}}|\mathbf{z}\right)p(\mathbf{z}|\mathbf{z}^{\text{reg}})}{p\left(\mathbf{z}_{\theta}^{\text{sim}}\right)p\left(\mathbf{z}|\mathbf{z}_{\theta}^{\text{sim}}\right)} = \frac{p(\mathbf{z}|\mathbf{z}^{\text{reg}})}{p(\mathbf{z})} \tag{A2}$$

**Acknowledgments**
We thank Editor Martyn Clark, the anonymous Associate Editor, and two reviewers for handling our manuscript and their insightful and constructive comments, which helped us substantially improve the paper. We also thank the Spanish Meteorological Agency (AEMET) for providing the climatological data and Francisco Jesús Peñas Silva for sharing the hydrological indices and catchments descriptors developed in his PhD that have been used in this paper.

## References

Agencia Vasca del Agua (2003). Estudio de evaluación de los recursos hídricos totales en el ámbito de la CAPV. *Report prepared by INTECSA-INARSA*. Retrieved from http://www.uragentzia.euskadi.eus/contenidos/documentacion/evaluacion_recursos_hidricos/es_rh/adjuntos/memoria.pdf, 02/02/2019.

Almeida, S., Bulygina, N., McIntyre, N., Wagener, T., & Buytaert, W. (2013). Improving parameter priors for data-scarce estimation problems. *Water Resources Research*, *49*, 6090–6095. https://doi.org/10.1002/wrcr.20437

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

Almeida, S. L., Bulygina, N., McIntyre, N., Wagener, T. & Buytaert, W. (2012). Predicting flows in ungauged catchments using correlated information sources. In: *British Hydrological Society's Eleventh National Hydrology Symposium, Hydrology for a Changing World*, Dundee 2012

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

Arnell, N. W. (1999). A simple water balance model for the simulation of streamflow over a large geographic domain. *Journal of Hydrology*, *217*(3-4), 314–335. https://doi.org/10.1016/S0022-1694(99)00023-2

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

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

Beven, K. J., & Smith, P. J. (2015). Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *ASCE Journal of Hydrologic Engineering*, *20*(1). https://doi.org/10.1061/(ASCE)HE.1943-5584.0000991

Blöschl, G., Sivapalan, M., Savenije, H., Wagener, T., & Viglione, A. (Eds.) (2013). *Runoff prediction in ungauged basins: Synthesis across processes, places and scales*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139235761

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

Booker, D. (2013). Spatial and temporal patterns in the frequency of events exceeding three times the median flow (FRE3) across New Zealand. *Journal of Hydrology. New Zealand*, *52*(1), 15–39.

Booker, D. J., & Snelder, T. H. (2012). Comparing methods for estimating flow duration curves at ungauged sites. *Journal of Hydrology*, *434*, 78–94.

Booker, D. J., & Woods, R. A. (2014). Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. *Journal of Hydrology*, *508*, 227–239. https://doi.org/10.1016/j.jhydrol.2013.11.007

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, Mass: Addison-Wesley Pub. Co.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., FriedmanJ, H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

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

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

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

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*, W02503. https://doi.org/10.1029/2010WR009240

Cabus, P. (2008). River flow prediction through rainfall–runoff modelling with a probability-distributed model (PDM) in Flanders, Belgium. *Agricultural Water Management*, *95*(7), 859–868. https://doi.org/10.1016/j.agwat.2008.02.013

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

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

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

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

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, *10*(3), 197–208. https://doi.org/10.1023/a:1008935410038

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

European Commission (2000). Directive 2000/60/EC of the European Parliament and of the Council, of 23 October 2000, establishing a framework for Community action in the field of water policial. *Official Journal of the European Economics L*, *327*(1), 22–12.

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

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

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

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

Foglia, L., Hill, M. C., Mehl, S. W., & Burlando, P. (2009). Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. *Water Resources Research*, *45*, W06427. https://doi.org/10.1029/2008WR007255

Gallego, G., Cuevas, C., Mohedano, R., & Garcia, N. (2013). On the Mahalanobis distance classification criterion for multidimensional normal distributions. *IEEE Transactions on Signal Processing*, *61*(17), 4387–4396. https://doi.org/10.1109/TSP.2013.2269047

García, A., Sainz, A., Revilla, J. A., Álvarez, C., Juanes, J. A., & Puente, A. (2008). Surface water resources assessment in scarcely gauged basins in the north of Spain. *Journal of Hydrology*, *356*(3–4), 312–326. https://doi.org/10.1016/j.jhydrol.2008.04.019

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed., p. 675). New York: Chapman and Hall/CRC.

Geological and Mining Institute of Spain (2013). Retrieved from http://info.igme.es/cartografiadigital/geologica/Magna50.aspx, 01/12/2013

Georgakakos, K. P., Seo, D. J., Gupta, H., Schake, J., & Butts, M. B. (2004). Characterizing streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology*, *298*(1-4), 222–241. https://doi.org/10.1016/j.jhydrol.2004.03.037

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *69*(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Gobierno de Cantabria (2005). Estudio de los Recursos Hídricos de los Ríos de la Vertiente Norte de Cantabria. Consejería de Medio Ambiente, Gobierno de Cantabria, Santander, Spain (in Spanish).

Gottschalk, L. (1985). Hydrological regionalization of Sweden. *Hydrological Sciences Journal*, *30*(1), 65–83. https://doi.org/10.1080/02626668509490972

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

He, Y., Bárdossy, A., & Zehe, E. (2011). A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences*, *15*(11), 3539–3553. https://doi.org/10.5194/hess-15-3539-2011

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

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

Hsu, K. I., Moradkhani, H., & Sorooshian, S. (2009). A sequential Bayesian approach for hydrologic model selection and prediction. *Water Resources Research*, *45*, W00B12. https://doi.org/10.1029/2008WR006824

Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, *74*(8), 2204–2214. https://doi.org/10.2307/1939574

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, *55*(2), 163–172. https://doi.org/10.2307/1403192

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Clarendon Press.

Kavetski, D., Franks, S. W., & Kuczera, G. (2002). Confronting input uncertainty in environmental modelling. In Q. Duan, et al. (Eds.), *Calibration of watershed models*, *Water Sci. Appl.* (Vol. 6, pp. 49–68). Washington, DC: American Geophysical Union.

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modelling: 1. Theory. *Water Resources Research*, *42*, W03407. https://doi.org/10.1029/2005WR004368

Kollat, J. B., Reed, P. M., & Wagener, T. (2012). When are multiobjective calibration trade-offs in hydrologic models meaningful? *Water Resources Research*, *48*, W03520. https://doi.org/10.1029/2011WR011534

Koren, V., Smith, M., & Duan, Q. (2003). Use of a priori parameter estimates in the derivation of spatially consistent parameter sets of rainfall runoff models. In Q. Duan, et al. (Eds.), *Calibration of Watershed Models*, *Water Sci. Appl.* (Vol. 6, pp. 239–254). Washington, DC: American Geophysical Union.

Kuczera, G. (1996). Correlated rating curve error in flood frequency inference. *Water Resources Research*, *32*(7), 2119–2127. https://doi.org/10.1029/96WR00804

Kuczera, G., & Parent, P. (1998). Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm. *Journal of Hydrology*, *211*(1-4), 69–85. https://doi.org/10.1016/S0022-1694(98)00198-X

Lee, H., McIntyre, N., Wheater, H., & Young, A. (2005). Selection of conceptual models for regionalization of the rainfall-runoff relationship. *Journal of Hydrology*, *312*(1-4), 125–147. https://doi.org/10.1016/j.jhydrol.2005.02.016

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*, *2*(3), 18–22.

Lilliefors, H. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399–402. https://doi.org/10.1080/01621459.1967.10482916

Lilliefors, H. (1969). On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, *64*(325), 387–389. https://doi.org/10.1080/01621459.1969.10500983

MacKellar, N. C., Dadson, S. J., New, M., & Wolski, P. (2013). Evaluation of the JULES land surface model in simulating catchment hydrology in Southern Africa. *Hydrology and Earth System Sciences Discussions*, *10*(8), 11,093–11,128. https://doi.org/10.5194/hessd-10-11093-2013

Mardia, K. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530. https://doi.org/10.2307/2334770

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, *53*, 2199–2239. https://doi.org/10.1002/2016WR019168

McIntyre, N., Lee, H., Wheater, H., Young, A., & Wagener, T. (2005). Ensemble predictions of runoff in ungauged catchments. *Water Resources Research*, *41*, W12434. https://doi.org/10.1029/2005WR004289

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111. https://doi.org/10.1002/hyp.9384

Ministry of Agriculture, Food and Environment (2013). Spain, Retrieved from http://servicios2.marm.es/sia/visualizacion/descargas/dma.jsp, 01/12/2013

Moore, R. J. (2007). The PDM rainfall-runoff model. *Hydrology and Earth System Sciences*, *11*(1), 483–499. https://doi.org/10.5194/hess-11-483-2007

Muller, P., Erkanli, A., & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, *83*(1), 67–79. https://doi.org/10.1093/biomet/83.1.67

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I. A discussion of principles. *Journal of Hydrology*, *222*(1), 1–9. https://doi.org/10.1016/j.bbr.2011.03.031

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

Oudin, L., Kay, A., Andréassian, V., & Perrin, C. (2010). Are seemingly physically similar catchments truly hydrologically similar? *Water Resources Research*, *46*, W11558. https://doi.org/10.1029/2009WR008887

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series*, *5*(302), 157–175. https://doi.org/10.1080/14786440009463897

Pechlivanidis, I. G., McIntyre, N. R., & Wheater, H. S. (2010). Calibration of the semi-distributed PDM rainfall–runoff model in the Upper Lee catchment, UK. *Journal of Hydrology*, *386*(1-4), 198–209. https://doi.org/10.1016/j.jhydrol.2010.03.022

Peñas, F. J. (2013). Classification of the natural flow regime and prediction of hydroecological characteristics in the northern third of the Iberian Peninsula (in Spanish *Clasificación Del Régimen Hidrológico Natural Y Predicción De Características Hidroecológicas En El Tercio Norte De La Península Ibérica*), (PhD thesis). University of Cantabria, Spain.

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

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, *49*(4), 974–997. https://doi.org/10.1016/j.csda.2004.06.015

Razavi, T., & Coulibaly, P. (2013). Streamflow prediction in ungauged basins: Review of regionalization methods. *Journal of Hydrologic Engineering*, *18*(8), 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

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

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., & Franks, S. W. (2011). Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resources Research*, *47*, W11516. https://doi.org/10.1029/2011WR010643

Ribeiro, M. I. (2004). Gaussian probability density functions: Properties and error characterization. Institute for Systems and Robotics. Retrieved from http://users.isr.ist.utl.pt/~mir/pub/probability.pdf

Riggs, H. C. (1973). *Regional analysis of streamflow characteristics*. Washington, DC: US Geological Survey Techniques of Water Resources, United States Government Printing Office.

Samaniego, L., Bárdossy, A., & Kumar, R. (2010). Streamflow prediction in ungauged catchments using copula-based dissimilarity measures. *Water Resources Research*, *46*, W02506. https://doi.org/10.1029/2008WR007695

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carrillo, G. (2011). Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences*, *15*(9), 2895–2911. https://doi.org/10.5194/hess-15-2895-2011

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*, W10531. https://doi.org/10.1029/2009WR008933

Seibert, J. (2001). On the need for benchmarks in hydrological modelling. *Hydrological Processes*, *15*(6), 1063–1064. https://doi.org/10.1002/hyp.446

Shamseldin, A. Y., O'Connor, K. M., & Liang, G. C. (1997). Methods for combining the outputs of different rainfall–runoff models. *Journal of Hydrology*, *197*(1–4), 203–229. https://doi.org/10.1016/S0022-1694(96)03259-3

Singh, R. (2013). An uncertainty framework for hydrologic projections in gauged and ungauged basins under non-stationary climate conditions, (PhD thesis). Pennsylvania State University.

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

Smith, T., Marshall, L., & Sharma, A. (2014). Predicting hydrologic response through a hierarchical catchment knowledgebase: A Bayes' empirical Bayes' approach. *Water Resources Research*, *50*, 1189–1204. https://doi.org/10.1002/2013WR015079

Snelder, T. H., Barquin Ortiz, J., Booker, D. J., Lamouroux, N., Pella, H., & Shankar, U. (2012). Can bottom-up procedures improve the performance of stream classifications? *Aquatic Sciences*, *74*(1), 45–59. https://doi.org/10.1007/s00027-011-0194-7

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

Snelder, T. H., Lamouroux, N., Leathwick, J. R., Pella, H., Sauquet, E., & Shankar, U. (2009). Predictive mapping of the natural flow regimes of France. *Journal of Hydrology*, *373*(1-2), 57–67. https://doi.org/10.1016/j.jhydrol.2009.04.011

Srikanthan, R., Amirthanathan, G. E., & Kuczera, G. (2007). Real-time flood forecasting using ensemble Kalman filter. In MODSIM 2007 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand (pp. 1789–1795).

Thompson, J. R., Green, A. J., Kingston, D. G., & Gosling, S. N. (2013). Assessment of uncertainty in river flow projections for the Mekong River using multiple GCMs and hydrological models. *Journal of Hydrology*, *486*, 1–30. https://doi.org/10.1016/j.jhydrol.2013.01.029

Trujillo-Ortiz, A., & Hernandez-Walls R. (2003). Obrientest: O'Brien's test for homogeneity of variances. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=3335&objectType=FILE, accessed November 2013.

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

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

Wagener, T., & Wheater, H. S. (2006). Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. *Journal of Hydrology*, *320*(1-2), 132–154. https://doi.org/10.1016/j.jhydrol.2005.07.015

Wagener, T., Wheater, H. S., & Gupta, H. V. (2004). *Rainfall-runoff modelling in gauged and ungauged catchments*. London, UK: Imperial College Press. https://doi.org/10.1142/p335

Westerberg, I. K., & Birkel, C. (2015). Observational uncertainties in hypothesis testing: Investigating the hydrological functioning of a tropical catchment. *Hydrological Processes*, *16*(6), 1135–4879. https://doi.org/10.1002/hyp.1053

Westerberg, I. K., Gong, L., Beven, K. J., Seibert, J., Semedo, A., Xu, C.-Y., & Halldin, S. (2014). Regional water balance modelling using flow-duration curves with observational uncertainties. *Hydrology and Earth System Sciences*, *18*(8), 2993–3013. https://doi.org/10.5194/hess-18-2993-2014

Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., et al. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, *15*(7), 2205–2227. https://doi.org/10.5194/hess-15-2205-2011

Westerberg, I. K., & McMillan, H. K. (2015). Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences*, *19*(9), 3951–3968. https://doi.org/10.5194/hess-19-3951-2015

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

Willems, P., Mora, D., Vansteenkiste, T., Taye, M. T., & Van Steenbergen, N. (2014). Parsimonious rainfall-runoff model construction supported by time series processing and validation of hydrological extremes—Part 2: Intercomparison of models and calibration approaches. *Journal of Hydrology*, *510*, 591–609. https://doi.org/10.1016/j.jhydrol.2014.01.028

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, W., Jakeman, A. J., & Young, P. C. (1998). Identification of improved rainfall-runoff models for an ephemeral low-yielding Australian catchment. *Environmental Modelling & Software*, *13*(1), 59–74. https://doi.org/10.1016/S1364-8152(98)00004-8

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

Young, A. R. (2006). Stream flow simulation within UK ungauged catchments using a daily rainfall-runoff model. *Journal of Hydrology*, *320*(1-2), 155–172. https://doi.org/10.1016/j.jhydrol.2005.07.017

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*, W00B04. https://doi.org/10.1029/2008WR006833