Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions

Regionalization with Hierarchical Hydrologic Similarity and Ex-situ Data for the Estimation of Mean Annual Groundwater Recharge at Ungauged Watersheds

Ching-Fu Chang[1] and Yoram Rubin[1]

[1]Department of Civil and Environmental Engineering, University of California, Berkeley, U.S.A.

Correspondence: Yoram Rubin (rubin@ce.berkeley.edu)

Abstract. There are various methods available for annual groundwater recharge estimation with in-situ observations. However, a great number of watersheds around the world still remain ungauged, i.e., without in-situ observations of hydrologic responses. One approach for making estimates at ungauged watersheds is through regionalization, namely, transferring information obtained at gauged watersheds to ungauged ones. The reliability of regionalization depends on (1) the underlying system of hydrologic similarity, i.e., the similarity in how watersheds respond to precipitation input, as well as (2) the approach by which information is transferred.

In this paper we present a set of ready-to-use tools for obtaining informative estimates of hydrologic responses at ungauged watersheds, using a nested tree-based modeling approach to condition the estimates on ex-situ data. It invokes a two-leveled hierarchical hydrologic similarity, where the higher level determines the relative importance of various watershed characteristics under different conditions, and the lower level performs the regionalization and estimation of hydrologic responses based on the watershed characteristics of the ungauged watershed of interest.

We apply the nested tree-based modeling approach to investigate the complicated relationship between mean annual groundwater recharge and watershed characteristics, and to test the applicability and usefulness of the hierarchical hydrologic similarity. Our findings reveal the decisive role of soil available water content in hydrologic similarity at regional and annual scales, as well as certain conditions under which it is risky to resort to climate variables for determining hydrologic similarity. These findings contribute to the understanding of the physical principles governing robust information transfer.

1 Introduction

Groundwater resources supply approximately 50% of the drinking water and roughly 40% of the irrigation water worldwide (National Ground Water Association, 2016). Yet, the groundwater has increasingly been depleted since the late 20th century (Wada et al., 2010). Therefore, groundwater recharge, here broadly defined as the replenishing of water to a groundwater reservoir, plays a critical role in sustainable water resources management (de Vries and Simmers, 2002). Several studies have reviewed and compared multiple methods for recharge estimation at a wide spectrum of temporal and spatial scales, including lysimeter tests, seepage tests, water table fluctuation, chemical and heat tracers, baseflow analysis, water budget, and numerical modeling (Scanlon et al., 2002; Healy, 2010; Heppner et al., 2007). However, the aforementioned methods rely on in-situ data,

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions

while many watersheds worldwide still remain effectively ungauged (i.e., ungauged, poorly gauged, or previously gauged) (Loukas and Vasiliades, 2014).

This fact leads us to a critical question: How can one estimate hydrologic responses without in-situ data? Studying ungauged watersheds has been a popular research topic for more than a decade, especially since The Prediction in Ungauged Basins (PUB) initiative by the International Association of Hydrological Sciences (IAHS) (Sivapalan et al., 2003). Facing the lack of in-situ data, studies have attempted transferring ex-situ information from gauged watersheds to ungauged ones; this data transfer is also termed "regionalization". Regionalization has been applied to constrain the estimates of the parameters of hydrologic models (especially rainfall-runoff models), which could then be used to make predictions at ungauged watersheds (Kuczera, 1982; Singh et al., 2014; Razavi and Coulibaly, 2017; Wagener and Montanari, 2011; Blöschl et al., 2013). Such constraining is expected to lead to more accurate and precise estimates, and could be in the form of (1) relationships between model parameters and watershed characteristics, (2) subsets of the parameter space, or (3) plausible parameter values from models built for other hydrologically similar watersheds (Singh et al., 2014).

However, the application of regionalization is not without challenges. One of the key factors of predictive uncertainty identified by the PUB initiative is the unsuitability of information transfer techniques, due to a lack of comparative studies across watersheds and a lack of understanding of the physical principles governing robust regionalization (Hrachowitz et al., 2013). Different regionalization techniques have been applied in different cases with different assumptions. For example, Li et al. (2018) attempted a simple form of regionalization, where kernel density estimation was applied on recharge values obtained from various hydrologically similar sites, in order to build an ex-situ prior distribution (i.e., a prior distribution conditioned on ex-situ data). However, one limitation in Li et al. (2018) was that hydrologic similarity was treated as a Boolean variable, and therefore, there was no way to systematically distinguish a highly similar site from a slightly similar site. To pursue this further in this study, we must ask the following question: How can we tell that two watersheds are hydrologically similar? Sawicz et al. (2011) applied Bayesian mixture clustering to watersheds across the eastern U.S. They found that spatial proximity was a valuable first indicator of hydrological similarity because it reflected strong climatic control in their study area. Oudin et al. (2008) reported similar findings based on 913 French watersheds, despite acknowledging the lack of some key physical descriptors in their data set. However, Smith et al. (2014) attempted regionalization of hydrologic model parameters in eastern Australia, and suggested that spatial proximity was an unreliable metric of hydrological similarity. For their part, Tague et al. (2013) presented successful regionalization of hydrologic parameters based on geologic similarity at watersheds in the U.S. Oregon Cascades, a mountain range that features geological heterogeneity. Although not directly shown, their findings also went against the use of applying spatial proximity, for they discussed the sharp contrasts in hydrology at proximal watersheds based primarily on geological differences. The indication from these findings is that, although spatial proximity is of practical importance due to its common usage, its simplicity, and its demonstrated effectiveness in specific areas (Smith et al., 2014), it is not the true controlling factor, but rather a confounding factor.

One can resort to other physical characteristics of watersheds for the determination of hydrologic similarity. However, what those characteristics are may be a complicated question. Razavi and Coulibaly (2017) tested the effect of combinations of neural-network-based classification techniques and regionalization techniques in Canada, and found that classifying water-

sheds before regionalization improves regionalization for streamflow, baseflow, and peak flow predictions, but also discovered that the best combination of techniques varied from one watershed to another. Singh et al. (2014) applied classification and regression tree (CART) to determine the relationship between catchment similarity and regionalization in the U.S., finding that the dominant controls of successful regionalization vary significantly with the spatial scale, with the region of interest, and with the objective function used. Similarly, Kuentz et al. (2017) found that different physiographic variables controlled various flow characteristics across Europe, showing how different descriptors could account for different dominant hydrologic processes and flow characteristics. These studies indicate an important challenge, that the factors determining hydrologic similarity may vary under different conditions, and a universal system of hydrologic similarity still remains unavailable. Loritz et al. (2018) suggested an interesting perspective describing a dynamic hydrologic similarity system, where similarity and uniqueness are not mutually exclusive; rather, they suggested that hydrologic systems operate by gradually changing to different levels of organization in which their behaviors are partly unique and partly similar.

In this study, we would like to integrate the perspective in Loritz et al. (2018), that similarity and uniqueness are not mutually exclusive, into our regionalization framework for groundwater recharge estimation at ungauged watersheds. It is thus critical to identify a number of plausible controlling factors. Although few studies have directly identified the controlling factors, some insights can be learned from previous studies. For example, the effective recharge (i.e., the net source term in the groundwater flow equation) in a steady, depth-integrated, and unbounded groundwater flow was found to be correlated with the spatial distributions of transmissivity and hydraulic head (Rubin and Dagan, 1987a, b). From a recharge-mechanism-based perspective, previous studies have also found a list of plausible controlling factors of recharge via recharge potential mapping (Yeh et al., 2016, 2009; Naghibi et al., 2015; Rahmati et al., 2016). These variables include watershed topography, land cover, soil properties, and geology. At regional scale, climate variables have been found to be among the primary controlling factors of groundwater table depth (Fan et al., 2013), mean annual groundwater recharge (Nolan et al., 2007), and mean annual baseflow (Rumsey et al., 2015), the latter of which is often used as a surrogate of recharge under the steady state assumption. Other examples include Xie et al. (2017), who showed that evapotranspiration data provided more conditioning power and more uncertainty reduction than soil moisture data in long-term mean recharge estimation, and Hartmann et al. (2017), who reported variations of the sensitivity of annual groundwater recharge to annual precipitation with aridity. Although these studies did not apply regionalization explicitly and did not target ungauged watersheds directly, their findings provide guidance for us to identify some watershed characteristics—especially climate variables—that might play an important role in the regionalization process for recharge estimation.

Given a set of watershed characteristics, the next important question is how the regionalization is carried out. Gibbs et al. (2012) provided a generic framework of regression regionalization, which involves a multi-objective optimization for calibration, a sensitivity analysis to determine the most important model parameters, and a final step relating watershed characteristics with model parameters. However, the framework does not include a straightforward quantification of uncertainties in calibration and in regionalization. On the other hand, Smith et al. (2014) applied a hierarchical Bayesian model (which they termed Bayes empirical Bayes in their study) for the statistical pooling of information from multiple donor watersheds. It was able to transfer

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

parametric distributions rather than plausible parameter sets, thus allowing for full Bayesian quantification of uncertainty, but it required a set of donor watersheds determined *a priori*.

The objectives of this study are twofold. First, to address the aforementioned challenges in regionalization technique, we propose a data-driven, Bayesian, and non-linear regression approach, which features simultaneous full Bayesian quantification of uncertainty and non-linear regression to model the predictor-response relationship. Second, we augment the approach with a classification-tree-based model comparison component, and propose a hypothesis of hierarchical hydrologic similarity. The augmented approach and the hypothesis are applied to a case study to estimate mean annual groundwater recharge at ungauged watersheds, with the goal of revealing the key controlling factors of a dynamic hydrologic similarity system, which could ultimately contribute to robust information transfer.

The remainder of this paper is organized as follows. The details of the methodology are provided in Sect. 2. The data, the study area, and the application of the approach in the case study are explained in Sect. 3. Sect. 4 presents and compares the results, while Sect. 5 provides further discussion. Finally, we conclude in Sect. 6.

2 Methodology

The data-driven, Bayesian, and non-linear regression approach is powered by Bayesian Additive Regression Tree (BART). The details of BART, including the establishment of prior distribution (which we term prior), the calculation of likelihoods, and the posterior inference statistics are well documented in Chipman et al. (2010) and in Kapelner and Bleich (2016). Here, we provide a brief conceptual introduction to the implementation and advantages of BART, as well as how BART it augmented in this study.

2.1 BART

Consider a fundamental problem of making inference about an unknown function that estimates a response variable of interest using a set of predictor variables available at hand. The general form of this problem can be expressed as follows:

$$R = \hat{R} + \epsilon = f(\boldsymbol{\theta}, \mathbf{x}) + \epsilon, \tag{1}$$

where R is the response variable, $f(\cdot)$ is a model that outputs the estimate of the response variable, $\hat{R}$ is the estimate, $\boldsymbol{\theta}$ is the vector of model parameters, $\mathbf{x}$ is the vector of predictors, and ϵ is a Gaussian white noise with finite variance, i.e., $\epsilon \sim N(0, \sigma^2)$. The observation of R is denoted by r.

BART solves this problem by applying a Bayesian version of the additive ensemble tree model. The additive ensemble tree model is the sum of J individual tree models (Fig. 1, **(a)** and **(b)**), each of which consists of a tree structure ($T_j, j = 1, ..., J$) and a set of terminal node (or leaf node) values ($\mathbf{M}_j, j = 1, ..., J$), shown as follows:

$$\hat{R} = f(\boldsymbol{\theta}, \mathbf{x}) = \sum_{j=1}^{J} g(T_j, \mathbf{M}_j, \mathbf{x}). \tag{2}$$

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions

where $\boldsymbol{\theta} = \{T_1, \mathbf{M}_1, ..., T_J, \mathbf{M}_J\}$. Each tree model, $g(\cdot)$, functions similarly as a CART model: T_j recursively applies binary partitioning to the predictor space, and $\mathbf{M}_j$ provides a set of different outputs corresponding to the partitioned predictor subspaces (Breiman, 1984; Chipman et al., 1998).

BART defines the following joint prior of all the tree structures, all the leaf node value sets, and the white noise variance:

$$p\left(T_1, \mathbf{M}_1, ..., T_j, \mathbf{M}_j, \sigma^2\right) = p(\sigma^2) \prod_{j=1}^{J} p(T_j) P(\mathbf{M}_j | T_j). \tag{3}$$

BART then applies a tailored version of backfitting Markov Chain Monte Carlo (MCMC) simulation algorithm to condition the prior on r, where backfitting means the jth tree model is iteratively updated with its partial residual. The stationary distribution toward which the MCMC simulations converge is then used to approximate the true posterior distribution (which we term posterior):

$$p\left(T_1, \mathbf{M}_1, ..., T_j, \mathbf{M}_j, \sigma^2 | r\right). \tag{4}$$

A schematic diagram of the MCMC simulation iteration procedure is shown in Fig. 1 **(c)**. For each MCMC simulation, both T_j and $\mathbf{M}_j$ for each tree in the ensemble tree model are iteratively simulated using a Metropolis-within-Gibbs sampler (the loop in the blue circle in Fig. 1 **(c)**). After simulating all the trees, the error variance (σ^2) is simulated with a Gaussian-Gamma-conjugate Gibbs sampler. Together, this process completes one MCMC simulation. We can see by the loop in the red square in Fig. 1 **(c)**, the MCMC simulation is continuous, until the simulated values converge to a stationary distribution. These post-convergence simulated values approximate realizations from Eq. 4, and thus we approximate the true posterior in Eq. 4 by the stationary distribution obtained by MCMC simulation. At this point, we have reached a BART model that is conditioned on data r, because all the BART parameters (tree structures, leaf node values, and the white noise variance) have been conditioned on r.

Given the aforementioned conditioned BART model, we now turn our attention to estimating recharge at an ungauged watershed that was not included in the data on which the BART model was conditioned. Here, we let $\tilde{\mathbf{x}}$ denote the predictor vector at an ungauged watershed. We wish to transfer and apply the information we learned with the conditioned model to this ungauged watershed, and obtain an informative predictive distribution of recharge. Firstly, Eq. 1 can be rewritten as:

$$R \sim N\left(\hat{R}, \sigma^2\right). \tag{5}$$

Both the mean and the variance in Eq. 5 are uncertain, and have their respective posteriors. By combining Eqs. 2 and 5, and after plugging in the post-convergence MCMC simulated values and $\tilde{\mathbf{x}}$, we obtain a plausible realization (indexed by the superscript l, $l = 1, ..., L$) of predictive distribution as follows:

$$\begin{aligned} &N\left(\hat{R}^{(l)}, (\sigma^2)^{(l)}\right) = \\ &N\left(f\left(\boldsymbol{\theta}^{(l)}, \tilde{\mathbf{x}}\right), (\sigma^2)^{(l)}\right) = \\ &N\left(\sum_{j=1}^{J} g\left(T_j^{(l)}, \mathbf{M}_j^{(l)}, \tilde{\mathbf{x}}\right), (\sigma^2)^{(l)}\right). \end{aligned} \tag{6}$$

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

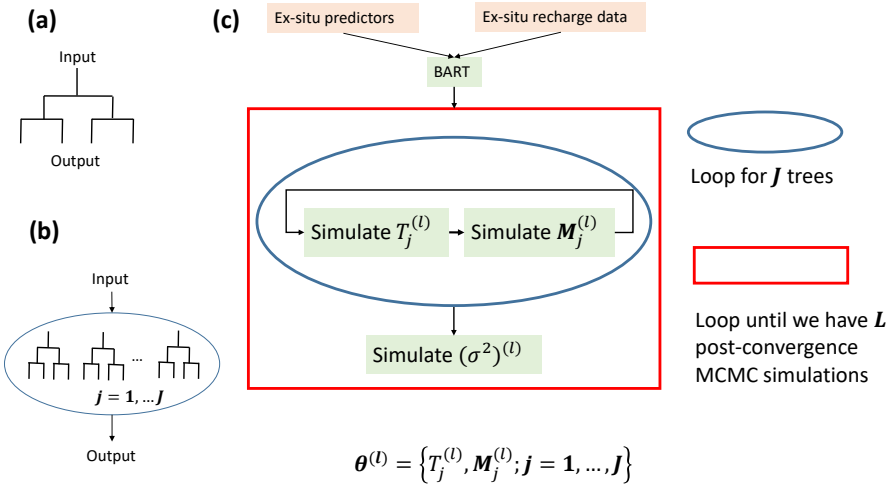

Figure 1. Schematic diagrams of **(a)** a regression tree model, **(b)** an ensemble tree model which consists of J additive regression tree models, and **(c)** the loops structure that BART uses to draw MCMC simulations, consisting of an inner loop for J additive regression tree models and an outer loop that continues until we have a total of L MCMC simulations after convergence toward a stationary distribution.

The collection of many plausible realizations yields an approximated posterior of predictive distributions. Thus, at the ungauged watershed of interest we have now obtained a fully Bayesian Gaussian predictive model, where the mean and the variance have their respective posteriors, achieved by transferring the information gained from conditioning the BART model on r and $\mathbf{x}$ to the ungauged watershed of interest.

2.2 Advantages of BART

The key advantage of BART is that it combines the non-linear regression for the predictor-response relationship with Bayesian inference, allowing for the determination of a full Bayesian posterior of predictive distribution, rather than one or a few estimates/predictions.

The estimation and the regionalization processes are data-driven. Prior knowledge of the physics is only minimally accounted for in terms of the composition of the predictor sets and the user-defined prior of the splitting rules (which are embedded in the tree structure variable, T_j). The underlying physics is inferred from the ex-situ data via obtaining conditional simulations of the tree structures and the terminal nodes (similar to the calibration stage), and thus, is implicitly embedded rather than explicitly defined. Therefore, the extent to which physics could be inferred is restricted by the training data —here, the ex-situ data, which is a common limitation of data-driven approaches.

However, in compensation, we avoid two disadvantages of the application of physically based models in the case of ungauged watersheds. First of all, the limited information at the ungauged watershed comes in as is, and it is unrealistic to expect that certain watershed characteristics should be known. Data availability could hinder the implementation of powerful hydro-

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

logic models (Razavi and Coulibaly, 2017) because some of the required model inputs may be unavailable at the ungauged watersheds (Xie et al., 2017; Gemitzi et al., 2017). It is possible to treat missing inputs as part of the parameters, and run simulations to impute them or apply stochastic methods to estimate them. Nonetheless, the corresponding computational demand grows in power law with the number and the plausible range of the missing inputs, which is of great practical importance when evaluating the pros and cons of an approach. Second, at ungauged watersheds, the conceptual model uncertainty due to misconceptions of physically based models is often compensated for, and thus disguised by, parameter uncertainty. The application of BART allows for estimates conditioned on and only on the available data at hand, without requiring specific predictors/inputs. Conceptual model uncertainty can be either directly accounted for, which will be explained in Sect. 2.3, or indirectly represented, which will be explained in Sect. 2.4. These advantages make our approach practically feasible at almost every ungauged watershed.

Note that in this study there is no intention to show the superiority of either the data-driven or the physically based approaches. As Wagener and Montanari (2011) pointed out, the ultimate goal is not to define parameters of a model, but rather, to understand what behavior we should expect at the ungauged watersheds of interest. We have simply shown why our approach is suitable in this study, and we will show how it helps us understand the behavior at ungauged watersheds in Sect. 2.4.

2.3 Bayesian model averaging

This subsection shows how one can account for conceptual model uncertainty with Bayesian model averaging. Suppose that one establishes K different BART models, denoted as $B_k, k = 1, ..., K$, to estimate recharge at an ungauged watershed with watershed characteristics (i.e., the predictors) $\tilde{\mathbf{x}}$. To do so, one would revisit Eq. 5, in which the posteriors of the mean and the variance depend on the model. That dependence would now need to be explicitly shown and addressed since one would now like to account for model uncertainty. This can be done via Bayesian model averaging, where we average the posterior of the estimate over the conditional probability mass function of the K models:

$$p\left(\hat{R}|\tilde{\mathbf{x}}, r\right) = \sum_{k=1}^{K} p\left(\hat{R}|\tilde{\mathbf{x}}, r, B_k\right) p\left(B_k|r\right), \tag{7}$$

where $p\left(\hat{R}|\tilde{\mathbf{x}}, r, B_k\right)$ can be approximated with MCMC simulations in the same way as that shown in Eq. 6, except that the previously omitted dependence on the kth model is now explicitly shown. The conditional probability mass function of the models, $p(B_k|r)$, can be obtained by invoking Bayes rule and the Total Probability rule:

$$p(B_k|r) = \frac{p(r|B_k)\,p(B_k)}{\sum_{k=1}^{K} p(r|B_k)\,p(B_k)}. \tag{8}$$

The integrated likelihood, $p(r|B_k)$, can be obtained by integration over the parameter space of B_k:

$$p(r|B_k) = \int\int p\left(r|B_k, \boldsymbol{\theta}_k, \sigma_k^2\right) p\left(\boldsymbol{\theta}_k, \sigma_k^2|B_k,\right) d\boldsymbol{\theta}_k d\sigma_k^2, \tag{9}$$

which can be approximated with a total of L MCMC simulations:

$$p(r|B_k) \approx \frac{1}{L}\sum_{l=1}^{L} p\left(r|B_k, \boldsymbol{\theta}_k^{(l)}, (\sigma_k^2)^{(l)}\right). \tag{10}$$

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

The last missing piece is $p(B_k)$, signifying the prior probability mass function of the models. As opposed to a uniform distribution $p(B_k) = 1/K$, one could adjust the prior based on the characteristics of the ungauged watershed. This is where physical knowledge as well as the knowledge about hydrologic similarity come into play, which will be elaborated on in Sect. 2.4. With $p(B_k)$ defined, one can now follow Eqs. 7 through 10 to obtain an estimate with the model uncertainty accounted for.

2.4 Nested tree-based modeling and the hierarchical similarity hypothesis

The second main objective of this study is revealing the key controlling factors of a dynamic hydrologic similarity system for mean annual groundwater recharge, which could contribute to a better determination of $p(B_k)$ in future applications. Here the details of how it works are provided.

First, we propose a hypothesis of hierarchical similarity. We hypothesize that hydrologic similarity is controlled by a hierarchy that follows two levels:

- The lower level is the predictor similarity, meaning that if two watersheds have some similar predictor values, then their hydrologic responses will be similar.
- The higher level is the regionalization similarity, meaning that if two watersheds share regionalization similarity, then their predictor similarities will be governed by similar predictors.

Put simply, regionalization similarity determines the predictor-predictor relationship and tells us which predictors to extract information from, while predictor similarity determines the predictor-response relationship that actually estimates the hydrologic response. This could explain why the controlling factors of hydrologic similarity change under different conditions.

To test this hypothesis, we define a nested tree-based modeling approach by nesting multiple BART models under a CART model for classification, and apply it to a case study. The details of the case study are provided in Sect. 3, while a brief general introduction is given here.

First, we apply the holdout method to divide a set of gauged watersheds into two subsets: the **training** watersheds and the **testing** watersheds. We represent conceptual model uncertainty indirectly by building multiple BART models using various plausible predictor sets, and fit the models to the data at the training watersheds. These data are the ex-situ data with respect to the testing watersheds. After model fitting, at each testing watershed, we evaluate the performance of the BART models, by comparing the data at the testing watershed with the predictive distributions from the BART models. Then, a label is given to each testing watershed, indicating which BART model has the highest predictive accuracy. Finally, we use a CART model to classify the testing watersheds based on their labels.

With this setup, we use the BART models to explore predictor similarity with different predictor sets, and use the CART model to explore regionalization similarity. The latter indicates under what condition does a certain BART model stand out in terms of predictive power, thus showing how the dominant factors of hydrologic similarity change under different conditions.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

3 Case study

In this case study, we are going to apply the methodology described in Sects. 2.1 and 2.4 to investigate the predictor similarity and the regionalization similarity in the study and to test the hypothesis of hierarchical similarity, with the goal of better understanding of the physical principles governing robust regionalization and better determination of $p(B_k)$ in future applications. The Bayesian model averaging in Sect. 2.3 is not demonstrated in this case study, as it is not necessary for achieving the aforementioned goal.

This Sect. provides the details about the case study setup, including the watersheds, the recharge data, the watershed characteristics data, the partitioning of data, and the evaluation metrics.

3.1 Watersheds and recharge estimates

The conterminous United States can be divided into eight major river basins (MRBs), each of which consists of thousands of watersheds (The United States Geological Survey, 2005; Brakebill and Terziotti, 2011). At each and every watershed, watershed-average annual recharge estimate and watershed characteristics data are retrieved from publicly available databases, and will be described in the following subsections. In our work, the recharge estimates are used as the target response while the characteristics are used as predictors in the regionalization process.

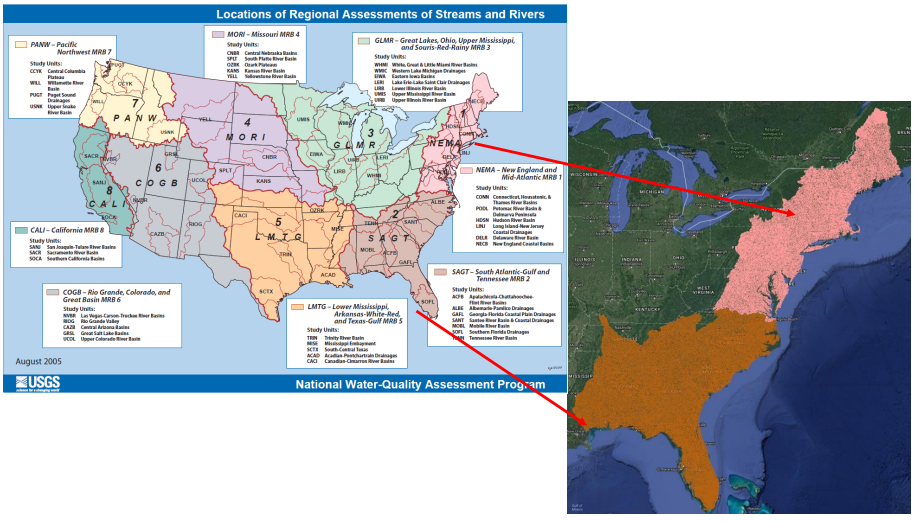

Figure 2. Annotated map of the coverage of the MRBs (left hand side), modified from The United States Geological Survey (2005), and a zoomed-in view of the study area (right hand side), which includes MRB 1 and MRB 2.

It is important to note that the predictors considered in this case study are not supposed to constitute a comprehensive list of controlling factors of recharge, nor are their respective spatiotemporal heterogeneities and uncertainties well accounted for. Rather, we provide them as an example of what could be available at ungauged watersheds, where one tries to condition recharge estimates with only a limited amount of information.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions

In 2002, annual groundwater recharge at each watershed was estimated via baseflow analyses by the U.S. Geological Survey (USGS) (Wieczorek and LaMotte, 2010h; Wolock, 2003). The reliability of baseflow analyses for recharge estimations depends on the spatiotemporal homogenization of recharge. Given the long-term steady-state assumption embedded in baseflow analysis, we made a working assumption in this study that the recharge mechanism remains steady at annual (or larger) temporal scale, so the inter-annual variability in recharge is solely due to the inter-annual variability of watershed characteristics.

The more arid U.S. Midwest may have more pronounced localized recharge (de Vries and Simmers, 2002), which cannot be effectively captured by baseflow analysis (Scanlon et al., 2002). This, then, does not fit well with our working assumption. Therefore, following the suggestion of Nolan et al. (2007), our study area includes only the relatively humid eastern parts of the U.S., namely MRB 1 and 2 (Fig. 2). After excluding watersheds with less desirable data coverage, we consider a total of 3609 watersheds in MRB 1 and 7413 watersheds in MRB 2. The distributions of the recharge data from all the watersheds in the study area are shown in Fig. 3 **(a)**.

3.2 Climate

As discussed in Sect. 1, climate predictors are found to be among the most important factors to control recharge at regional scale. At each watershed included in the study the following data are retrieved from publicly available databases: the long-term average annual precipitation ($\bar{P}$) averaged from 1970 to 2000 (Wieczorek and LaMotte, 2010a), the annual precipitation in the year 2002 (P) (Wieczorek and LaMotte, 2010b), and the long-term average annual potential evapotranspiration (E_p) averaged from 1960 to 1990 (Title and Bemmels, 2017). Given the precipitation and evapotranspiration, we obtained two additional climate variables: the long-term aridity index, estimated as $\bar{\phi} = E_p/\bar{P}$, and the 2002 aridity index, estimated as $\phi = E_p/P$. Given that the recharge data are based on baseflow analysis for the year 2002, P and ϕ represent the climate controls of that same year, while $\bar{P}$, E_p, and $\bar{\phi}$ represent climate controls over the long-term.

The distributions of P are shown in Fig. 3 **(b)**. The annual recharge data (in volume of water per unit watershed area) can be normalized by P (also in volume of water per unit watershed area), as in Fig. 3 **(c)**. This stems from the concept of water budgets and has been commonly used in hydrological studies worldwide (e.g., Magruder et al., 2009; Rangarajan and Athavale, 2000; Obuobie et al., 2012; Heppner et al., 2007; Takagi, 2013; Yang et al., 2009). Here, we apply logit transformation, which is common for proportions or probabilities (Gelman et al., 2014), to that normalized recharge, relaxing the physical bounds (0 and 1) of the values of the target variable (Fig. 3 **(d)**). This step is advantageous as it opens the opportunity to estimate recharge with parametric statistical models without special accommodations for the bounds. Therefore, in this case study the logit normalized recharge is used as the target variable.

3.3 Non-climate watershed characteristics

We also consider various non-climate watershed characteristics in this study, including topography, land cover, soil properties, and geology. The land cover is based on data published in 2001, which we feel is close enough to 2002 to provide the appropriate information. The other characteristics are based on raw data obtained in different years before 2002; it is assumed

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

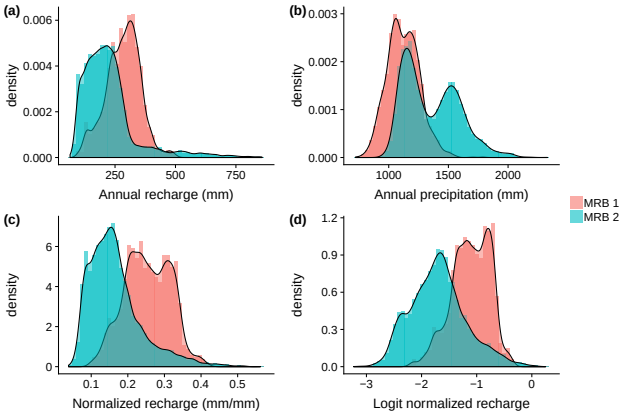

Figure 3. Distributions of **(a)** annual recharge in 2002, **(b)** annual precipitation in 2002, **(c)** normalized recharge, and **(d)** logit normalized recharge at all the watersheds in MRB 1 and 2.

that they remain steady at sub-century time scales. We provide the details of these watershed characteristics in the following subsections.

3.3.1 Topography and land cover

The topographic predictors are taken from publicly available databases (Wieczorek and LaMotte, 2010g); they are summarized in Table 1. The land cover variables are the percentages of watershed area corresponding to each land cover class (Wieczorek and LaMotte, 2010f); these are summarized in Table 2. The land cover classes are based on the 2001 National Land Cover Database (NLCD2001), the categories of which include water, developed land, barren land, forest, shrubland, herbaceous land, cultivated land, and wetland, with each having its own sub-classes. The details of NLCD2001 can be found in Homer et al. (2007).

3.3.2 Soil property

The soil property predictors include watershed scale statistics (e.g., average, upper bound, and lower bound) of soil properties (Wieczorek and LaMotte, 2010e); these are summarized in Table 3. The spatial statistics of the soil properties within each watershed were obtained over gridded source data values from the State Soil Geographic database (STATSGO) (Schwarz and Alexander, 1995), which were depth-averaged over all soil layers (Wolock, 1997).

3.3.3 Geology

The geology predictors used in this study were retrieved from publicly available databases (Wieczorek and LaMotte, 2010c, d) and they can be classified into two subcategories: surficial geology (surface sediment) and bedrock geology. As the predictors, we used fractions of the watershed area corresponding to each of the 45 surficial geology types (Wieczorek and LaMotte,

2010d; Clawges and Price, 1999) and each of the 162 bedrock geology types (Wieczorek and LaMotte, 2010c; Schruben et al., 1994). Details regarding each geology type can be found in Wieczorek and LaMotte (2010c) and Wieczorek and LaMotte (2010d). Note that in geological terminology, rock type or rock composition data are referred to as lithology data. Compared to lithology, structural geology data might be more informative for groundwater studies (e.g., orientation, fracture properties, discontinuity, etc.). However, structural geology information usually requires in-situ investigation, which cannot be expected at ungauged watersheds. Therefore, we consider only lithology data in this study.

3.4 Data partitioning

This Sect. explains the setup of the holdout method (i.e., the partitioning of data into two mutually exclusive subsets) for the watersheds, as well as the partitioning of the predictors into various subsets in order to evaluate the effects of different predictors.

3.4.1 Watershed partitioning

Because we cannot evaluate predictive accuracy at real ungauged watersheds due to the lack of in-situ recharge observations, we follow the holdout method described in Sect. 2.4 to partition the watersheds: the watersheds in MRB 1 are the **testing** watersheds and the watersheds in MRB 2 are the **training** watersheds. The ex-situ data (i.e., data in MRB 2) are used to fit multiple BART models (to be explained shortly), which are then used to obtain predictive distributions of recharge at all the testing watersheds.

There are two reasons for this MRB-based data partitioning:

- For reasons touched on in Sect. 1, we do not consider spatial proximity as a predictor in this study. Separating the two MRBs partly ensures the exclusion of the confounding effect of spatial proximity, and thus the regionalization is solely based on the watershed characteristics.
- Considering the logit normalized recharge (Fig. 3 **(d)**), the range of values in MRB 2 fully covers the range of values in MRB 1. However, the reverse is not true. It is thus advantageous to train the models with MRB 2 to avoid poor model fitting due to lack of data coverage.

After partitioning the watersheds, we now turn our attention to the partitioning of predictors.

3.4.2 Predictor partitioning

As mentioned in Sect. 1, climate variables are among the most important factors in hydrologic similarity at regional scale, but there might be other controlling factors to consider as well, and the dominance of climate variables may not be always present. To investigate the various effects of different predictors, we define a total of six different predictor sets to build six unique BART models, which are indexed by k, $k = 1, 2, ...6$ (Table 4). Note that the determination of the six predictor sets is guided by the idea of testing the relative importance of different categories of predictors under different conditions, instead of aiming

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions

for high accuracy and precision. Therefore, by no means is Table 4 an exhaustive list of all possible sets, nor does it necessarily include the "best" set. The design of the six predictor sets simply facilitates the investigation of the effects of various categories of predictors on predictive accuracy and uncertainty.

In addition to the six BART models, we also build a simple model by using the estimated distribution of logit normalized recharge at the training watersheds (via kernel density estimation) as the predictive distribution for the testing watersheds, without considering any predictor. This is a model that ignores hydrologic similarity altogether, and it can be considered as an extreme case of the ex-situ prior in Li et al. (2018), with a lot more watersheds and much less stringent criteria of similarity. From this point forward, we refer to this model as the benchmark model, for it is used as a benchmark against which the BART models are compared.

3.5 Evaluation of predictive distributions

As mentioned in Sect. 2.4, we label each testing watershed by the best-performing model. Thus, the metric with which we evaluate predictive distributions matters.

In this study, two different accuracy metrics are adopted. The first is the root mean squared error (RMSE), defined as

$$E_{i,k} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\hat{R}_{i,k}^{(l)} - \tilde{r}_i\right)^2} \tag{11}$$

where $\tilde{r}_i$ is the recharge data at the ith testing watershed, and $E_{i,k}$ is the RMSE of the kth model at the ith testing watershed. Note that $\hat{R}_{i,k}^{(l)}$ is obtained by following Eq. 6, but now subscripts are added to indicate that we plug in the predictors from the ith testing watershed to the kth model . This metric evaluates the predictive performance in an estimation problem, where we wish to obtain a "best estimate" of recharge with minimal expected error.

The second metric is the median log predictive probability density (LPD) at the value of recharge observation, defined as

$$L_{i,k} = \text{median}_{l=1,\ldots,L}\left\{\ln\left[p\left(R = \tilde{r}_i | \hat{R}_{i,k}^{(l)}, (\sigma^2)_k^{(l)}\right)\right]\right\} \tag{12}$$

where $L_{i,k}$ is the LPD of the kth model at the ith testing watershed. The subscript of $(\sigma^2)_k^{(l)}$ indicates the kth model. This metric evaluates the predictive performance in a simulation problem, where we wish the realizations from the predictive distributions are likely to be the same as the observation.

In addition to accuracy, we also quantify the predictive uncertainty. This is done by first recognizing the two components of uncertainty for the kth model at the ith testing watershed:

1. σ_k^2, which we refer to as the predictive variance, and is approximated as the sample median of $(\sigma^2)_k^{(l)}$ over $l = 1, ..., L$, and

2. the posterior variance of $\hat{R}_{i,k}$, which we refer to as the estimate variance, and is approximated as the sample variance of $\hat{R}_{i,k}^{(l)}$ over $l = 1, ..., L$.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

The predictive variance indicates how informative the inferred predictor-response relationship is, while the estimate variance indicates how certain a BART model can infer that relationship. In this case study we weigh the two components equally, as we wish to obtain an informative relationship with certainty. To that end, we define the total predictive variance as the summation of the two components, and use it as the metric of predictive uncertainty in this study.

4 Results

As discussed above, we built six BART models (Table 4) with ex-situ data. In-situ predictors were then fed into the models to yield posterior realizations of predictive distributions (Eq. 6). With the metrics of accuracy and uncertainty defined, we are then able to quantify the predictive performance of the BART models, and classify them based on either the RMSE-based labels or the LPD-based labels with the nested tree-based modeling approach. This allows for the investigation of the effects of various predictors under different conditions, which will be presented in this Sect.

4.1 Evaluation of predictive distributions

The following subsections present the effects of different predictor sets on predictive accuracy and uncertainty.

4.1.1 Predictive uncertainty

The effect of regionalization with the different predictor sets on predictive uncertainty is shown in Fig. 4. The estimate variance (Fig. 4 **(a)**) represents how well the BART models capture the predictor-response relationships. We see that the geology predictors lead to the lowest estimate variance, probably because of the significantly larger number of predictors used (see Table 4). Yet, there is a surprise in Fig. 4 **(a)**. First, at $k = 1$ and $k = 2$ the estimate variances are generally quite low, despite the low number of predictors. However, at $k = 3$, the estimate variances increase significantly. Intuitively, since aridity is the ratio of evapotranspiration to precipitation, one would expect that the variances at $k = 3$ would be similar to, if not lower than, those at $k = 1$ and $k = 2$. One plausible explanation here is that although aridity indices and precipitation/evapotranspiration carry ample information to be extracted and conditioned upon, the respective predictor-response relationships we get might be significantly different. When used together, the BART models were not able to formulate a universal relationship. This will be revisited in Sect. 5.2.

The predictive variance (Fig. 4 **(b)**) represents how informative the predictor-response relationships are, which is a different aspect of uncertainty compared to the estimate variance. One could obtain a predictor-response relationship fairly confidently (low estimate variance), but the relationship is less informative (high predictive variance), like that found at $k = 6$. The opposite case is that one could not confidently obtain a predictor-response relationship, but once that relationship is obtained it is quite informative, like that found at $k = 5$.

The total predictive variance (Fig. 4 **(c)**) provides an overall metric that considers the above two sources of uncertainties. While the medians are rather similar, the spread of the box plots does vary significantly with k. The condensed box plots (e.g., $k = 1$ and $k = 6$) indicate that the total predictive variances are essentially constant throughout all testing watersheds, while the

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

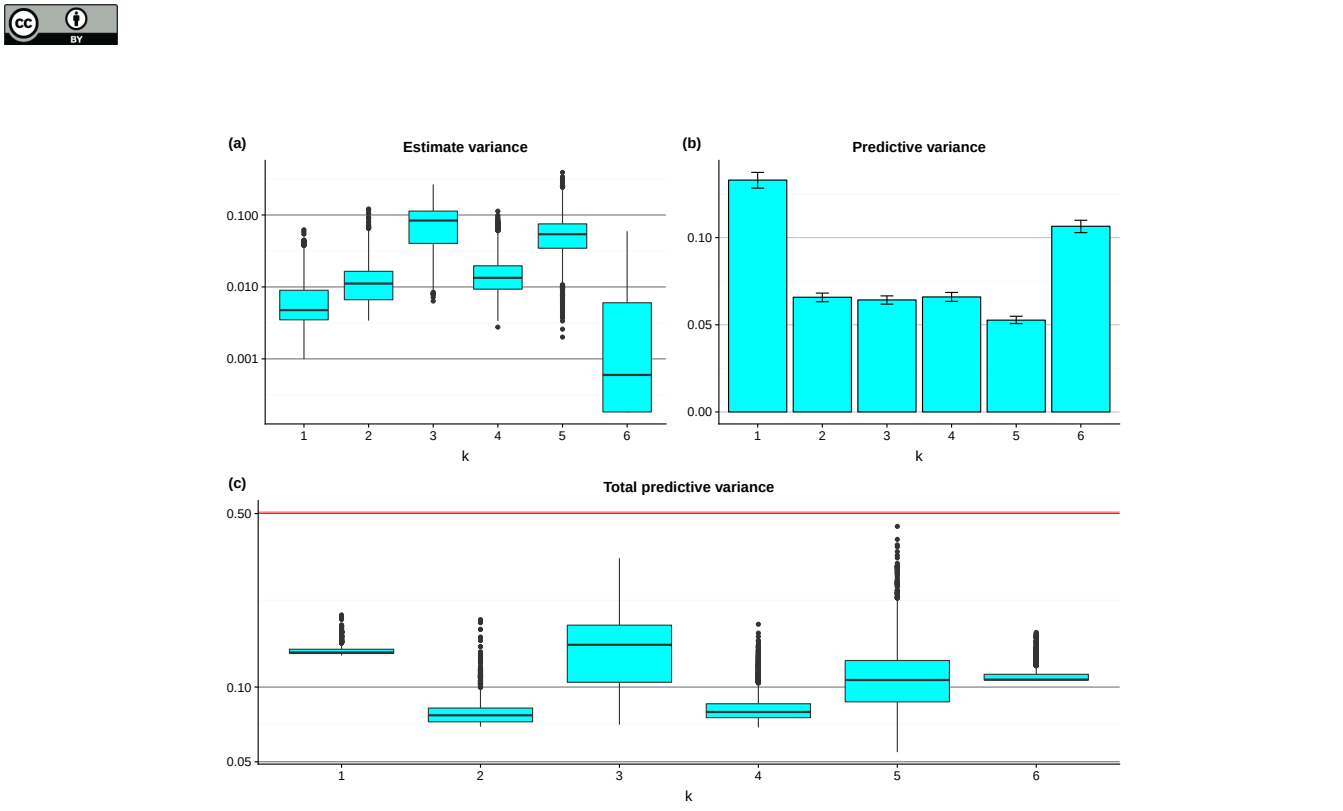

Figure 4. The box plots of the estimate variances at the testing watersheds **(a)**, the bar plot of the predictive variances with 95% intervals shown by the error bars **(b)**, and the box plots of the total predictive variances at the testing watersheds **(c)**. The red line indicates the variance of the benchmark model for comparison.

spread-out box plots (e.g., $k = 5$) indicate that the effect of the predictors may vary significantly from one testing watershed to another. This indicates that there might not be one single predictor set that always leads to the lowest uncertainty, and thus the effects of predictors on predictive uncertainty may vary from one condition to another. That said, regardless of the testing watersheds and predictor sets, the total predictive variance is always lower than the variance of the benchmark model, which clearly shows that regionalization using watershed characteristics definitely improves predictive precision.

4.1.2 Predictive accuracy

The effect of regionalization with the different predictor sets on RMSE is shown in Fig. 5. The RMSE of the benchmark model (Fig. 5 **(a)**) at each testing watershed is simply the difference between the sample mean of the ex-situ recharge data and the in-situ recharge observation. For the BART models (Fig. 5 **(b)**), it is calculated by the root of the average squared errors over post-convergence MCMC simulations.

Regardless of k, we see that, compared with the benchmark model, RMSE is reduced at least at half of the testing watersheds. Surprisingly, the largest overall RMSE reduction is observed when only the aridity indices are used for regionalization, indicating that at most of the watersheds tested in this study, aridity similarity implies recharge similarity at regional and annual scales to a high degree. On the other hand, we observe some outliers that have high RMSE reduction at $k = 4$ through $k = 6$,

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

indicating that topography, land cover, soil properties, and geology may not have an overall effect that is as strong, but under certain circumstances, they could still be important factors.

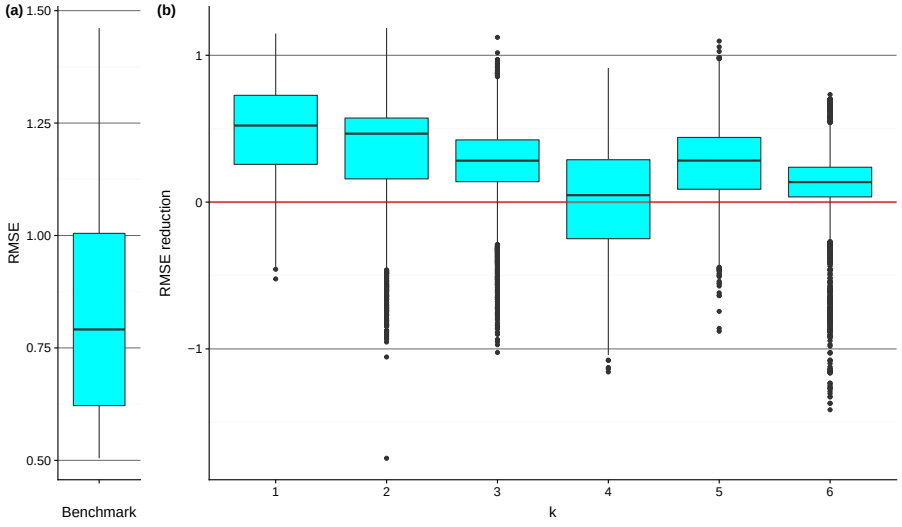

Figure 5. The box plot of the RMSE of the benchmark model at the testing watersheds **(a)**, and the box plots of the RMSE reduction introduced by applying the BART models at the testing watersheds **(b)**. The red line indicates zero RMSE reduction for comparison.

The effect of regionalization with different predictor sets on LPD is shown in Fig. 6. It is immediately clear that the accuracy improvement is not as prominent as that in Fig. 5. Only when $k = 1$ is LPD increased at most of the watersheds . We also find that all of the distributions of LPD are heavily negatively skewed with a lot of outliers.

Looking at Figs. 4 through 6 together, one can observe the different effects of the predictor sets on predictive accuracy, stemming from the different natures of an estimation and a simulation problem. From the point of view of the overall effect, for $k = 2$ through $k = 5$ (i.e., the predictors other than aridity indices) RMSE is reduced at more than half of the testing watersheds, but LPD does not increase to the same extent. This suggests that the predictive distributions are centered closer to the in-situ observations due to regionalization, but that the conditioning also significantly reduces the predictive variances, causing the predictive distribution to be too narrow. Therefore, compared to a relatively flat, spread-out, and uninformative or weakly informative distribution, the predictive density decays too quickly when deviating from the predictive mean, resulting in low LPD. This might be a sign of over-conditioning, or the disproportional reduction of predictive uncertainty, as exemplified in Fig. 7. The cyan curve is an example of an over-conditioned distribution. Although its mean is close to the true value, the small variance causes rapid decay of probability density; therefore, at the true value the predictive density is lower than that of the weakly informative distribution, and is essentially the same as that of the uninformative uniform distribution. The only predictor set that improves both RMSE and LPD at most of the testing watersheds is $k = 1$, the aridity indices, and one could expect the corresponding predictive distributions to be more similar to the case of the dark blue curve in Fig. 7.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions

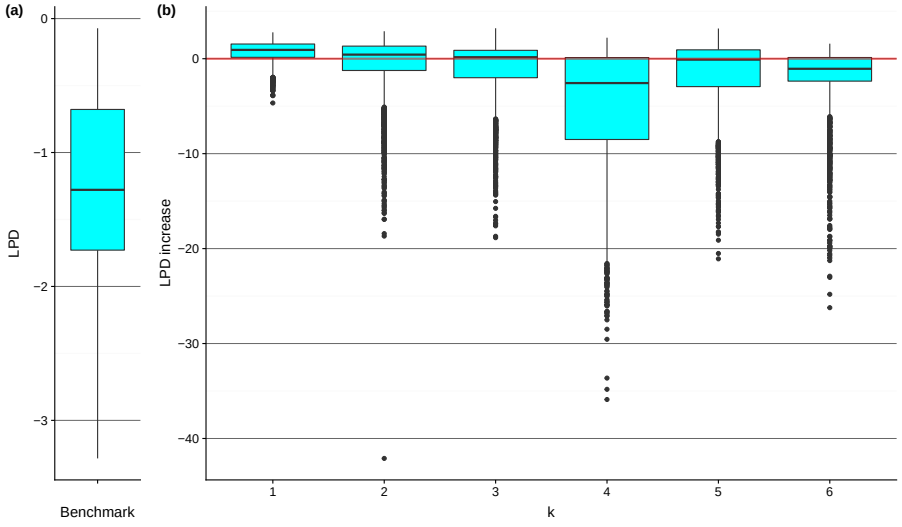

Figure 6. The box plot of the LPD of the benchmark model at the testing watersheds **(a)**, and the box plots of the LPD increase introduced by applying the BART models at the testing watersheds **(b)**. The red line indicates zero LPD increase, used for comparison.

Over-conditioning can occur when model fitting or model calibration leads to well-constrained parameters that are, in fact, subject to different forms of model uncertainty (Hutton et al., 2014; Beven et al., 2008). In this study, it could be that the uncertainty regarding the predictor-recharge relationship at the testing watersheds is characterized differently compared to the uncertainty at the training watersheds. For the sake of comparing the relative importance of the different predictor sets, instead of accounting for model uncertainty, we evaluated and compared the models directly. However, in another application where the estimates are to be refined, model uncertainty should be and can be considered (as shown in Sect. 2.3).

4.2 Regionalization similarity

The box plots in Fig. 4 through 6 showed different distributions of the predictive performance metrics for the different predictor sets. An interesting follow-up question here would be how model performance varies with watershed characteristics. It was shown that, in consistency with previous studies, aridity is indeed the most important controlling factor at regional and annual scales on average, but there are few cases where this aridity dominance is replaced. In other words, how might we identify the conditions under which a specific predictor set could be more informative than others?

To investigate this further, we give each testing watershed two labels: the model with the lowest RMSE, and the model with the highest LPD; we refer to these labels as the RMSE labels and the LPD labels, respectively. The possible values of each label include $k = 1$ through $k = 6$ and $benchmark$, representing the six BART models and the benchmark model, respectively. Then, using all the available predictors, we built two CART models to classify watersheds based on the RMSE labels (Fig. 8), and the LPD labels (Fig. 9).

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

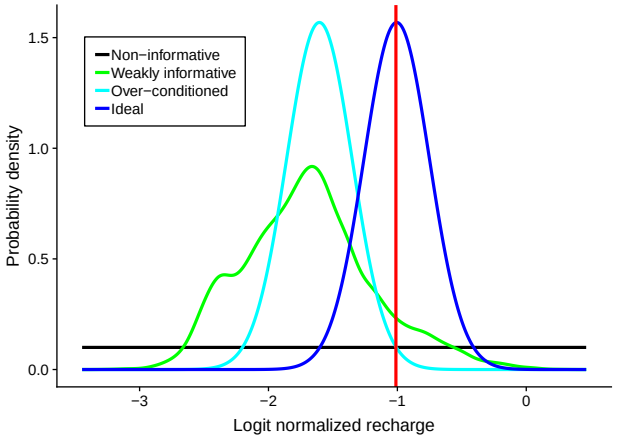

Figure 7. An example of over-conditioning: the probability density at the true value (indicated by the red vertical line) of the over-conditioned distribution is not higher than that of the non-informative distribution or that of the weakly informative distribution, not because the conditioning does not work, but because of the disproportional reduction of the variance of the distribution.

4.2.1 RMSE labels

The available water content (AWC) is the first indicator of regionalization similarity (Fig. 8 node 1): at watersheds with high AWC, aridity stands out as the dominant factor, consistent with various previous studies cited in Sect. 1. However, there is a potential risk if one uses aridity as the primary indicator of hydrologic similarity regardless of AWC. In previous studies, AWC was found to be an important predictor correlated with surface runoff, baseflow, and groundwater recharge (Arnold et al., 2000), and it was among the most important parameters to which water balance models are sensitive (Finch, 1998). In the current study, we are not claiming that AWC cannot be a predictor, but rather, we are suggesting a hierarchical structure in which AWC is placed —together with other predictors —to help estimate recharge at ungauged watersheds. Since AWC is governed by field capacity and wilting point, it is an indicator of the storage capacity of the soil for usable/consumable water: the larger the storage capacity, the higher the degree to which the system is supply-limited, thus pointing to aridity. If the storage capacity is low, on the other hand, the more complicated interplay among various predictors needs to be considered, and one cannot simply assume that aridity is the primary indicator of hydrologic similarity.

Further down the classification tree, watersheds with lower AWC are classified roughly as arid or humid watersheds by the long-term aridity index. For the more humid watersheds (Fig. 8, nodes 4 through 14), regionalization similarity is controlled by different predictors, but the dominant predictors for recharge estimation are almost always the climate variables (nodes 6, 8, 11 and 12, which contain 1576 watersheds in total). Only at a handful of watersheds (nodes 13 and 14, which contain only 185 watersheds in total) are aridity indices not dominant. However, some interesting conjectures can be made by taking a closer look at these two nodes.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions

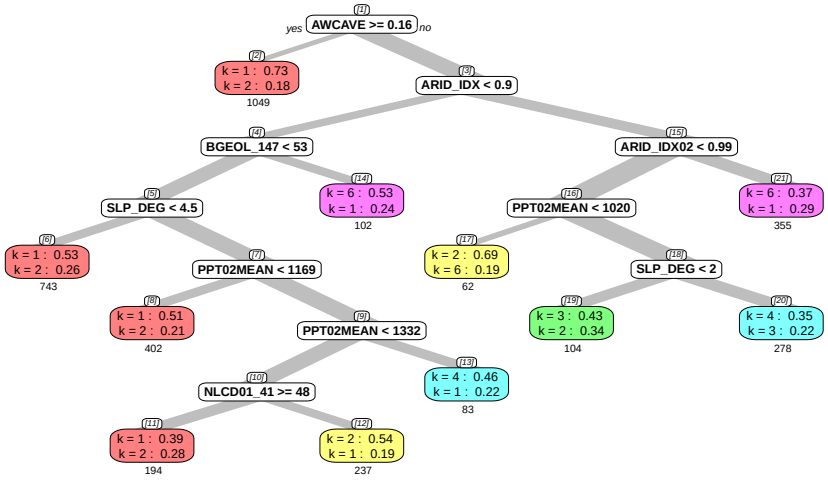

Figure 8. CART model classifying the RMSE labels of the testing watersheds. Splitting rules are shown in white nodes, while terminal nodes are colored based on the classification results. On top of every node, in brackets, is the node number, provided for convenient referencing. The predictors in the splitting rules are expressed in code names for convenience; a reference list is found in Table 5. The width of the tree branch (grey line) is proportional to the impurity of the node that the branch leads to, where impurity is defined as the probability that two randomly chosen watersheds within the node have different labels. For each terminal node, the class of the highest multinomial probability is shown first, which is the classification result, followed by the class of the second highest probability to indicate how impure the node is. Underneath each terminal node box is the number of watersheds belonging to the node.

Node 14 is a small but unique cluster, featuring watersheds that have low AWC, are humid, and have relatively homogeneous paragneiss and/or schist bedrock. Both of these bedrock types belong to the category of crystalline rock, and often feature layering in a particular orientation. The groundwater movement in such rock formation often depends on foliation, i.e., rock breaks along approximately parallel surfaces, which affect the direction of the regional groundwater flow (Singhal and Gupta, 2010). Hence we observe a condition where the ample water supply cannot be substantially held by the soil due to low AWC, and the regional groundwater movement might be controlled by bedrock layering and foliation. Low AWC is an indication of less clayey soils, and implies that infiltration/percolation through the soil layer might be facilitated by relatively higher permeability. Water could thus easily enter the bedrock layer, which is rather horizontally homogeneous. To that end, those predictor sets other than $k = 6$ become less informative, while the predictor set $k = 6$ becomes relatively more informative. In fact, these watersheds are mostly the positive outliers at $k = 6$ in Fig. 5 **(b)**, where the predictive power of the geology predictors is at its best.

Node 13 features watersheds that have low AWC, are humid, are not dominated by homogeneous paragneiss and/or schist, have a relatively steep average slope, and have large amount of annual precipitation. The low aridity is primarily driven by precipitation rather than evapotranspiration. In fact, these watersheds are mostly outliers featuring extremely low aridity index (below 0.65) due to ample precipitation. Under such condition, evapotranspiration is expected to operate to its full potential, i.e.,

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

it is shifting from water-limited state to energy-limited and canopy-controlled state. Because the evapotranspiration demand of the canopy can be met, the land cover type now starts to play a dominant role in hydrologic similarity. It is noteworthy to point out node 20 here. Node 20 features watersheds that are relatively humid among the arid watersheds ($\bar{\phi}$ in the range from 0.9 to 0.99) and have ample precipitation. The similarity of node 20 with node 13 supports our conjecture that the dominance of land cover predictors is due to the precipitation-driven humid environment that is relatively more capable of catering to the evapotranspiration water demand.

On the other side of the tree (Fig. 8, node 15 through 21), the resulting classification is quite diverse, and the purity of each node is relatively lower. Aridity no longer plays the dominant role, and the hierarchical similarity structure becomes complicated that it is difficult to make straightforward physical interpretations. The most important message we get is the significant risk one would face if one considers aridity, or any climate variable in general, as the primary indicator of hydrologic similarity when AWC is low and aridity index is high. In summary, although climate predictors are still the most important ones on average, within the context of the hierarchical similarity we have identified certain conditions under which either non-climate predictors become dominant or no dominant predictor set can be straightforwardly identified, all of which contribute to the understanding of the dynamic hydrologic similarity.

4.2.2 LPD labels

The classification of the LPD labels is shown in Fig. 9. In general, the root part of the classification tree (node 1 through 3) is quite similar to that found in Fig. 8, where AWC and long-term aridity define two sequential overarching separations of watersheds. However, further down the tree the leaf part is significantly different. The classification essentially leads to only three big clusters (Fig. 9, nodes 2, 7, and 9), and the other terminal nodes only contain a few watersheds. Node 9 features arid watersheds with low AWC, where we end up with a highly impure terminal node, and even the highest multinomial probability is only 0.27. No further splitting rule could significantly reduce classification error. This is supportive towards our previous argument that when aridity index is high and AWC is low, it is risky to resort to climate variables for hydrologic similarity, as shown here that it is difficult to even identify a dominant predictor set. As mentioned in Sect. 4.1.2, underestimation of the predictive variance (σ_k^2) leads to low LPD, and thus it is difficult to make physical interpretation our of the results in Fig. 9, except for node 1 through 3, which are quite similar to their counterparts in Fig. 8. Therefore, with the LPD labels we are only able to identify the overarching regionalization similarity controlled by AWC and long-term aridity.

RMSE and LPD represent views of predictive accuracy in an estimation problem and a simulation problem, respectively. Intuitively, if one only considers unimodal predictive distribution with limited skewness, a high predictive density at a value directly implies a closeness of the distribution central tendency to that value. However, the reverse is not necessarily true: either over- or underestimation of variance might possibly lead to low predictive density, even if the mean is close to the target value (e.g., Fig. 7). Based on whether RMSE or LPD is used as the accuracy metric —which implies the scope of recharge estimation —we can observe some common features as well as some distinctions of the structure of the hypothesized hierarchical similarity.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions

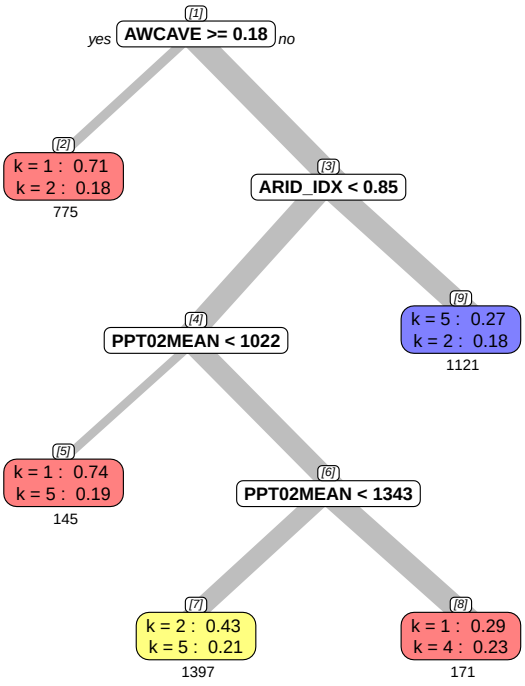

Figure 9. Same as Fig. 8, except that here the classification is done using the LPD labels. The predictors in the splitting rules are expressed in code names for convenience; a reference list is found in Table 5.

Fortunately, regardless of the metric of predictive accuracy, in both Figs. 8 and 9 the first three nodes are remarkably consistent, and the effect of the metric of predictive accuracy is only manifested at watersheds with low AWC. This supports the suggestion that AWC plays a pivotal role in hydrologic similarity for mean annual groundwater recharge estimation.

5 Discussion

In this section, we discuss the key findings as well as the limitations of the case study.

5.1 The hierarchical similarity hypothesis and the shift in dominant physical processes

With BART's ability to simultaneously model non-linear and/or interaction effects and present uncertainty in a fully Bayesian fashion, we are able to show how the controlling factors of hydrologic similarity vary among different watersheds, among different conditions, and among different accuracy metrics. These are all manifested in the case study under the context of the hierarchical similarity hypothesis.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Climate variables have been identified as the dominant factors in previous studies (see Sect. 1), and they are indeed on average the most dominant factors in our case study. However, the hierarchical similarity shows potential risks if one resorts to climate variables to define hydrologic similarity without considering other physical watershed characteristics, especially the soil available water content.

The details of the hierarchical similarity are inferred from the data in the fashion of supervised machine learning, using a nested application of tree-based modeling approach, consisting of six BART models and one benchmark model nested under one CART model. It is of great importance to have two levels in such a system, as it allows for identification of the shifts of dominant factors under different conditions. These shifts indicate shifts in dominant physical processes, as exemplified by node 13 and 20 in Fig. 8 where we observed the shift from water-limited evapotranspiration to energy-limited evapotranspiration. Therefore, we conjecture that it is the shift in dominant physical processes that is driving, and thus is reflecting, the shift in the controlling factors of hydrologic similarity under different conditions.

5.2 Limitations of the case study

Here, we provide discussions about the limitations of the case study from the aspects of the target response and the partitioning of watersheds.

5.2.1 Scale of the target response

A major limitation of the case study is that the target hydrologic response is the logit normalized watershed-averaged annual groundwater recharge. This is a large-scale spatiotemporally homogenized response, and in this study, the data were based on baseflow analyses. Streamflow-based estimation of recharge, such as baseflow analysis, is commonly used in humid regions. As put forward by Healy (2010), there are three key questions that should be carefully checked before applying baseflow analysis: (1) Is all recharging water eventually discharged into the stream where the baseflow is measured? (2) Do low flows consist entirely of groundwater discharge? (3) Does the contributing area of the aquifer differ significantly from that of the watershed? At an ungauged watershed, it is unlikely that one would have enough data to verify the answers to these three questions. To that end, a working assumption about the reliability of the baseflow analysis was made without a rigorous proof. The findings of the case study are all under the context of this working assumption, and thus, they should not be applied to recharge at other spatiotemporal scales or other hydrologic responses without careful considerations. Fortunately, from a post hoc check, the recharge estimates fall within the typical scales at which baseflow analysis is more suitable: a recharge scale from hundreds to thousands mm per year, a spatial scale of hundreds of m^2 to hundreds of km^2, and temporal scales from months to decades (Scanlon et al., 2002).

5.2.2 Artifact due to the partitioning of watersheds: ϕ versus P and E_p

Intuitively, since aridity index is the ratio of potential evapotranspiration to precipitation ($\phi = E_p/P$), one might be surprised by the differences among the cases of $k = 1$, $k = 2$, and $k = 3$ in the results. The main reason is revealed in Fig. 10. The E_p

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

values at the training and testing watersheds are so distinct that, essentially, all the testing watersheds are outliers from the point of view of a BART model trained at the training watersheds. On the other hand, the ϕ values at the training and testing watersheds share the range from about 0.6 to 1.2, and only differ at the two extreme ends. In other words, the predictor-response relationships inferred by using ϕ can be transferred due to the overlapping range (Fig. 10 **(c)**), but the relationships inferred using $E_p > 1000mm$ cannot be effectively transferred to watersheds with $E_p < 1000mm$ (Fig. 10 **(b)**). Although it is not shown, a similar case can be found by comparing $\bar{\phi}$ with E_p.

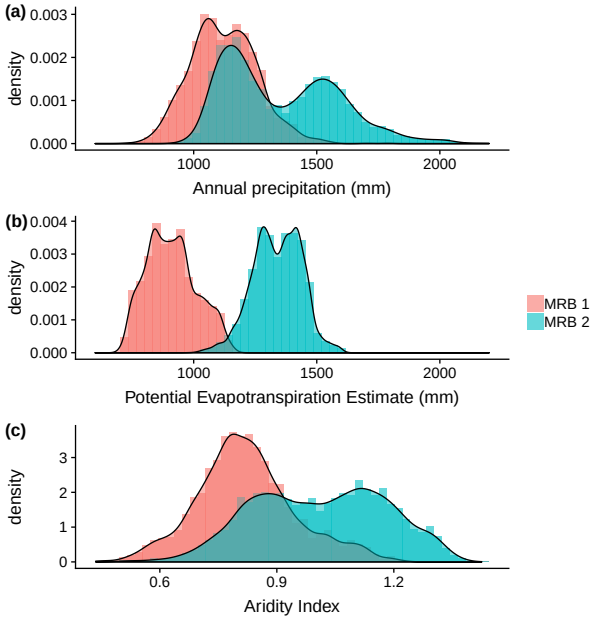

Figure 10. Distributions of **(a)** P, **(b)** E_p, and **(c)** ϕ, at watersheds in MRB 1 (the testing watersheds) and MRB 2 (the training watersheds).

Although this might have been avoidable by using a more sophisticated design of cross-validation, we kept the MRB-based holdout method on purpose. In addition to the reasons that were explained in Sect. 3.4.1, another motivation is that, in reality, the data at hand come in as is. This means there is no guarantee that the measurements will cover a particular range or that the watershed characteristics of the ungauged watersheds of interest are within a desirable range. The prevailing superiority of ϕ and $\bar{\phi}$ over P, $\bar{P}$, and E_p found in our results shows an important advantage of dimensionless predictors, that they tend to be more transferable from one site to another, and hence, they may be more suitable for studies targeting ungauged watersheds.

6 Conclusions

In this work, we proposed a nested tree-based modeling approach with three key features: (1) data-driven and non-linear regression for regionalization and estimation, (2) full Bayesian representation of the predictive uncertainty, and (3) CART-based model comparison, and an additional potential feature of accounting for conceptual model uncertainty via Bayesian

model averaging. We applied the nested tree-based modeling approach to obtain recharge estimates conditioned on ex-situ data at ungauged watersheds in a case study in the eastern U.S. We hypothesized a hierarchical similarity structure to account for the dynamic hydrologic similarity underlying the regionalization.

The findings of this study contribute to the understanding of one aspect of the key factors of predictive uncertainty identified in the PUB initiative: the physical principles governing robust regionalization among watersheds. Firstly, in consistency with previous studies, we found that the climate variables are on average the most important controlling factors of hydrologic similarity at regional and annual scales, which means a climate-based regionalization technique is on average more likely to result in better estimates. However, with our hierarchical similarity hypothesis we revealed certain conditions under which non-climate variables become more dominant than climate variables. In particular, we demonstrated how soil available water content stood out to be the pivotal indicator of the variable importance of aridity in hydrologic similarity. Moreover, we showed that with hierarchical similarity one could identify shifts in dominant physical processes that are reflecting shifts in the controlling factors of hydrologic similarity under different conditions, such as water-limited evapotranspiration versus energy-limited evapotranspiration, or homogeneous and foliated bedrock versus heterogeneous bedrock. As the controlling factors change from one condition to another, the suitable regionalization technique also changes. We demonstrated how the hierarchical similarity hypothesis could indicate mechanisms by which available water content, aridity, and other watershed characteristics dynamically affect hydrologic similarity. The nested tree-based modeling approach can be applied to identify plausible sets of watershed characteristics to be considered in the regionalization process.

The contributions of this study may be viewed differently depending on individual cases. In a situation where groundwater recharge is the ultimate target variable at ungauged watersheds, the nested tree-based modeling approach offers a systematic way to obtain informative predictive distributions that are conditioned on ex-situ data. In a difference case, where recharge estimation at ungauged watersheds is but one component of a greater project, the aforementioned informative predictive distributions can be treated as informative ex-situ priors, which could be further updated and/or integrated into simulation-based stochastic analyses where recharge is an input/component of other models/functions. At ungauged watersheds that will become gauged in the foreseeable future, the informative predictive distributions again serve as informative ex-situ priors that could guide the design of the sampling campaign, as different recharge flux magnitudes require different quantifying techniques (Scanlon et al., 2002; Healy, 2010). The hierarchical similarity hypothesis offers one plausible explanation of the dynamic nature of hydrologic similarity, which affects the application of regionalization. Lastly, it should be pointed out that the nested tree-based modeling approach is independent of the target response and the predictors of interest, so it could be integrated into future studies within or beyond the field of hydrology in search of a hierarchical predictor-response relationship.

Competing interests. The authors declare that they have no conflict of interest.

References

Arnold, J. G., Muttiah, R. S., Srinivasan, R., and Allen, P. M.: Regional estimation of base flow and groundwater recharge in the Upper Mississippi river basin, Journal of Hydrology, 227, 21–40, 2000.

Beven, K. J., Smith, P. J., and Freer, J. E.: So just why would a modeller choose to be incoherent?, Journal of Hydrology, 354, 15–32, 2008.

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

Brakebill, J. W. and Terziotti, S. E.: A Digital Hydrologic Network Supporting NAWQA MRB SPARROW Modeling – MRB_E2RF1WS, Report, 2011.

Breiman, L.: Classification and regression trees, Routledge, 1 edn., 1984.

Chipman, H. A., George, E. I., and McCulloch, R. E.: Bayesian CART Model Search, Journal of the American Statistical Association, 93, 935–948, 1998.

Chipman, H. A., George, E. I., and McCulloch, R. E.: BART: Bayesian additive regression trees, Ann. Appl. Stat., 4, 266–298, 2010.

Clawges, R. M. and Price, C. V.: Digital data set describing surficial geology in the conterminous U.S., US Geological Survey Open-File Report, 99, 77, 1999.

de Vries, J. J. and Simmers, I.: Groundwater recharge: an overview of processes and challenges, Hydrogeology Journal, 10, 5–17, 2002.

Fan, Y., Li, H., and Miguez-Macho, G.: Global Patterns of Groundwater Table Depth, Science, 339, 940–943, 2013.

Finch, J. W.: Estimating direct groundwater recharge using a simple water balance model – sensitivity to land surface parameters, Journal of Hydrology, 211, 112–125, 1998.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.: Bayesian data analysis, vol. 2, Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

Gemitzi, A., Ajami, H., and Richnow, H.-H.: Developing empirical monthly groundwater recharge equations based on modeling and remote sensing data – Modeling future groundwater recharge to predict potential climate change impacts, Journal of Hydrology, 546, 1–13, 2017.

Gibbs, M. S., Maier, H. R., and Dandy, G. C.: A generic framework for regression regionalization in ungauged catchments, Environmental Modelling & Software, 27–28, 1–14, 2012.

Hartmann, A., Gleeson, T., Wada, Y., and Wagener, T.: Enhanced groundwater recharge rates and altered recharge sensitivity to climate variability through subsurface heterogeneity, Proceedings of the National Academy of Sciences, 114, 2842–2847, 2017.

Healy, R. W.: Estimating groundwater recharge, Cambridge University Press, 2010.

Heppner, C. S., Nimmo, J. R., Folmar, G. J., Gburek, W. J., and Risser, D. W.: Multiple-methods investigation of recharge at a humid-region fractured rock site, Pennsylvania, USA, Hydrogeology Journal, 15, 915–927, 2007.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J. N., and Wickham, J.: Completion of the 2001 national land cover database for the counterminous United States, Photogrammetric engineering and remote sensing, 73, 337, 2007.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58, 1198–1255, 2013.

Hutton, C. J., Kapelan, Z., Vamvakeridou-Lyroudia, L., and Savić, D.: Application of Formal and Informal Bayesian Methods for Water Distribution Hydraulic Model Calibration, Journal of Water Resources Planning and Management, 140, 04014 030, 2014.

Kapelner, A. and Bleich, J.: bartMachine: Machine Learning with Bayesian Additive Regression Trees, 2016, 70, 40, 2016.

Kuczera, G.: Combining site-specific and regional information: An empirical Bayes Approach, Water Resources Research, 18, 306–314, 1982.

Kuentz, A., Arheimer, B., Hundecha, Y., and Wagener, T.: Understanding hydrologic variability across Europe through catchment classification, Hydrol. Earth Syst. Sci., 21, 2863–2879, 2017.

Li, X., Li, Y., Chang, C.-F., Tan, B., Chen, Z., Sege, J., Wang, C., and Rubin, Y.: Stochastic, goal-oriented rapid impact modeling of uncertainty and environmental impacts in poorly-sampled sites using ex-situ priors, Advances in Water Resources, 111, 174–191, 2018.

Loritz, R., Gupta, H., Jackisch, C., Westhoff, M., Kleidon, A., Ehret, U., and Zehe, E.: On the dynamic nature of hydrological similarity, Hydrology and Earth System Sciences Discussions, 2018, 1–38, 2018.

Loukas, A. and Vasiliades, L.: Streamflow simulation methods for ungauged and poorly gauged watersheds, Nat. Hazards Earth Syst. Sci., 14, 1641–1661, 2014.

Magruder, I. A., Woessner, W. W., and Running, S. W.: Ecohydrologic process modeling of mountain block groundwater recharge, Groundwater, 47, 2009.

Naghibi, S. A., Pourghasemi, H. R., and Dixon, B.: GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran, Environmental Monitoring and Assessment, 188, 44, 2015.

National Ground Water Association: Facts About Global Groundwater Usage, 2016.

Nolan, B. T., Healy, R. W., Taber, P. E., Perkins, K., Hitt, K. J., and Wolock, D. M.: Factors influencing ground-water recharge in the eastern United States, Journal of Hydrology, 332, 187–205, 2007.

Obuobie, E., Diekkrueger, B., Agyekum, W., and Agodzo, S.: Groundwater level monitoring and recharge estimation in the White Volta River basin of Ghana, Journal of African Earth Sciences, 71–72, 80–86, 2012.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., and Le Moine, N.: Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, Water Resources Research, 44, n/a–n/a, 2008.

Rahmati, O., Pourghasemi, H. R., and Melesse, A. M.: Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran, CATENA, 137, 360–372, 2016.

Rangarajan, R. and Athavale, R. N.: Annual replenishable ground water potential of India—an estimate based on injected tritium studies, Journal of Hydrology, 234, 38–53, 2000.

Razavi, T. and Coulibaly, P.: An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 42, 2–20, 2017.

Rubin, Y. and Dagan, G.: Stochastic identification of transmissivity and effective recharge in steady groundwater flow: 1. Theory, Water Resources Research, 23, 1185–1192, 1987a.

Rubin, Y. and Dagan, G.: Stochastic identification of transmissivity and effective recharge in steady groundwater flow: 2. Case study, Water Resources Research, 23, 1193–1200, 1987b.

Rumsey, C. A., Miller, M. P., Susong, D. D., Tillman, F. D., and Anning, D. W.: Regional scale estimates of baseflow and factors influencing baseflow in the Upper Colorado River Basin, Journal of Hydrology: Regional Studies, 4, 91–107, 2015.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G.: Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrol. Earth Syst. Sci., 15, 2895–2911, 2011.

Scanlon, B. R., Healy, R. W., and Cook, P. G.: Choosing appropriate techniques for quantifying groundwater recharge, Hydrogeology Journal, 10, 2002.

Schruben, P. G. A., Bawiec, R. E., King, W. J., Beikman, P. B., and Helen, M.: Geology of the Conterminous United States at 1: 2,500,000 Scale–A Digital Representation of the 1974 PB King and HM Beikman Map, 1994.

Schwarz, G. E. and Alexander, R.: State soil geographic (STATSGO) data base for the conterminous United States, Report 2331-1258, 1995.

Singh, R., Archfield, S. A., and Wagener, T.: Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments – A comparative hydrology approach, Journal of Hydrology, 517, 985–996, 2014.

Singhal, B. B. S. and Gupta, R. P.: Applied hydrogeology of fractured rocks, Springer Science & Business Media, 2010.

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J. J., Mendiondo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., and Zehe, E.: IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences, Hydrological Sciences Journal, 48, 857–880, 2003.

Smith, T., Marshall, L., and Sharma, A.: Predicting hydrologic response through a hierarchical catchment knowledgebase: A Bayes empirical Bayes approach, Water Resources Research, 50, 1189–1204, 2014.

Tague, C. L., Choate, J. S., and Grant, G.: Parameterizing sub-surface drainage with geology to improve modeling streamflow responses to climate in data limited environments, Hydrol. Earth Syst. Sci., 17, 341–354, 2013.

Takagi, M.: Evapotranspiration and deep percolation of a small catchment with a mature Japanese cypress plantation, Journal of Forest Research, 18, 73–81, 2013.

The United States Geological Survey: Locations of Regional Assessments of Streams and Rivers, 2005.

Title, P. O. and Bemmels, J. B.: ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling, Ecography, pp. n/a–n/a, 2017.

Wada, Y., van Beek, L. P. H., van Kempen, C. M., Reckman, J. W. T. M., Vasak, S., and Bierkens, M. F. P.: Global depletion of groundwater resources, Geophysical Research Letters, 37, 2010.

Wagener, T. and Montanari, A.: Convergence of approaches toward reducing uncertainty in predictions in ungauged basins, Water Resources Research, 47, n/a–n/a, 2011.

Wieczorek, M. E. and LaMotte, A. E.: Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: 30-Year Average Annual Precipitation, 1971-2000, Report, U.S. Geological Survey, 2010a.

Wieczorek, M. E. and LaMotte, A. E.: Attributes for MRB_E2RF1 Catchments by Major Rivers Basins in the Conterminous United States: Total Precipitation, 2002, Report, U.S. Geological Survey, 2010b.

Wieczorek, M. E. and LaMotte, A. E.: Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: Bedrock Geology, Report, U.S. Geological Survey, 2010c.

Wieczorek, M. E. and LaMotte, A. E.: Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: Surficial Geology, Report, U.S. Geological Survey, 2010d.

Wieczorek, M. E. and LaMotte, A. E.: Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: STATSGO Soil Characteristics, Report, U.S. Geological Survey, 2010e.

Wieczorek, M. E. and LaMotte, A. E.: Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: NLCD 2001 Land Use and Land Cover, Report, U.S. Geological Survey, 2010f.

Wieczorek, M. E. and LaMotte, A. E.: Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: Basin Characteristics, 2002, Report, U.S. Geological Survey, 2010g.

Wieczorek, M. E. and LaMotte, A. E.: Attributes for MRB_E2RF1 Catchments by Major River Basins in the Conterminous United States: Estimated Mean Annual Natural Groundwater Recharge, 2002, Report, U.S. Geological Survey, 2010h.

Wolock, D. M.: STATSGO soil characteristics for the conterminous United States, Report 2331-1258, US Geological Survey, 1997.

Wolock, D. M.: Estimated mean annual natural ground-water recharge in the conterminous United States, Report, U.S. Geological Survey, 2003.

Xie, Y., Cook, P. G., Simmons, C. T., Partington, D., Crosbie, R., and Batelaan, O.: Uncertainty of groundwater recharge estimated from a water and energy balance model, Journal of Hydrology, 2017.

Yang, F.-R., Lee, C.-H., Kung, W.-J., and Yeh, H.-F.: The impact of tunneling construction on the hydrogeological environment of "Tseng-Wen Reservoir Transbasin Diversion Project" in Taiwan, Engineering Geology, 103, 39–58, 2009.

Yeh, H.-F., Lee, C.-H., Hsu, K.-C., and Chang, P.-H.: GIS for the assessment of the groundwater recharge potential zone, Environmental Geology, 58, 185–195, 2009.

Yeh, H.-F., Cheng, Y.-S., Lin, H.-I., and Lee, C.-H.: Mapping groundwater recharge potential zone using a GIS approach in Hualian River, Taiwan, Sustainable Environment Research, 26, 33–43, 2016.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Table 1. Watershed topography predictors.

Variable	Explanation
Basin index	Watershed area divided by watershed perimeter squared (dimensionless).
Stream density	Reach length divided by watershed area (m^{-1}).
Sinuosity	Reach length divided by the length of the straight line connecting the beginning and the ending of the reach (dimensionless).
Slope	Mean watershed slope calculated from digital elevation data (degree).

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Table 2. Land cover classification by NLCD2001.

Class	Subclass
Water	Open water
	Perennial ice
Developed	Open space
	Low intensity
	Medium intensity
	High intensity
Barren	Barren land
Forest	Deciduous
	Evergreen
	Mixed
Shrubland	Dwarf shrub
	Shrub/scrub
Herbaceous	Grassland
	Sedge
	Lichens
	Moss
Cultivated	Pasture/hay
	Crops
Wetlands	Woody wetland
	Emergent herbaceous wetland

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Table 3. Soil property predictors.

Soil property	Unit	Statistics[a]
Calcium carbonate equivalent	%	Lower/higher bounds
Cation exchange capacity	cmolc kg^{-1}	Lower/higher bounds
Depth to the seasonally high water table	m	Average and Lower/higher bounds
Soil thickness	m	Lower/higher bounds
Hydrologic soil group classification	%	Average
Soil erodibility factor	dimensionless	Average
Permeability	m h^{-1}	Average and Lower/higher bounds
Available water content	fraction	Average and Lower/higher bounds
Bulk density	g cm^{-3}	Average and Lower/higher bounds
Organic matter content	%	Average and Lower/higher bounds
Clay soil content	%	Average and Lower/higher bounds
Silt soil content	%	Average
Sand soil content	%	Average
Percent finer than nos.4, 10, and 200 sieve	%	Average and Lower/higher bounds

[a]: Spatial statistics calculated across the watershed.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Table 4. Table of the six different predictor sets.

k	predictors included	Number of predictors
1	$\bar{\phi}$ and ϕ	2
2	$\bar{P}$, P, and E_p	3
3	All climate predictors: $\bar{P}$, P, E_p, $\bar{\phi}$ and ϕ	5
4	Topography and land cover predictors	20
5	Soil predictors	48
6	Geology predictors	206

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-561
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 January 2019

Hydrology and Earth System Sciences
Discussions

Table 5. Reference list of the splitting variables in Fig. 8 and Fig. 9.

Fig. 8		Fig. 9	
Node number	**Splitting variable[a]**	**Node Number**	**Splitting variable[a]**
1	Average available water content (AWCAVE)	1	Average available water content (AWCAVE)
3	Long term average aridity index (ARID_IDX)	3	Long term average aridity index (ARID_IDX)
4	% area of Paragneiss and Schist bedrock (BGEOL_147)	4	Precipitation in 2002 (PPT02MEAN)
5	Average slope (SLP_DEG)	6	Precipitation in 2002 (PPT02MEAN)
7	Precipitation in 2002 (PPT02MEAN)		
9	Precipitation in 2002 (PPT02MEAN)		
10	% area of Deciduous Forest (NLCD01_41)		
15	Aridity index in 2002 (ARID_IDX02)		
17	Precipitation in 2002 (PPT02MEAN)		
19	Average slope (SLP_DEG)		

[a]: Underneath each predictor, in parentheses, is the predictor code name.