Hydrological Interpretation of a Statistical Measure of Basin Complexity

Pande, Saket; Moayeri, Mehdi

DOI
10.1029/2018WR022675

Publication date
2018

Document Version
Final published version

Published in
Water Resources Research

Citation (APA)
Pande, S., & Moayeri, M. (2018). Hydrological Interpretation of a Statistical Measure of Basin Complexity. *Water Resources Research*, *54*(10), 7403-7416. https://doi.org/10.1029/2018WR022675

AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE

10.1029/2018WR022675

Key Points:

- Basin complexity is defined in terms of difficulty in learning statistical patterns of daily streamflow
- Relationship between complexity of MOPEX basins and hydrological characteristics is analyzed
- More complex basins that are drier have less seasonal rainfall, vegetation with more storage capacity and faster responsive soils

Correspondence to:
S. Pande,
s.pande@tudelft.nl

Citation:
Pande, S., & Moayeri, M. (2018). Hydrological interpretation of a statistical measure of basin complexity. *Water Resources Research*, *54*. https://doi.org/10.1029/2018WR022675

Received 30 JAN 2018
Accepted 6 SEP 2018
Accepted article online 21 SEP 2018

Hydrological Interpretation of a Statistical Measure of Basin Complexity

Saket Pande[1] **and Mehdi Moayeri**[2]

[1]Department of Water Management, Delft University of Technology, Delft, Netherlands, [2]Department of Water Engineering, University of Tabriz, Tabriz, Iran

Abstract This paper studies how streamflow predictability varies with basin characteristics. We introduce an index of basin complexity that is based on a model of least statistical complexity that is needed to reliably predict daily streamflow of the basin. We then relate it with climate, vegetation and soil characteristics of the basin. Daily streamflow is modeled using *k* nearest neighbor model of lagged streamflow that predicts next time step streamflow based on the occurrences of similar streamflow events from the past. In order to calculate basin complexity, we identify *difficult* streamflow events of the basin and then use Vapnik-Chervonenkis generalization theory, which trades off model performance with Vapnik-Chervonenkis dimension (i.e., a measure of model complexity), to find a *k* nearest neighbor model of appropriate complexity for predicting a difficult streamflow event of the basin. The average of selected model complexities corresponding to difficult events is then defined as the basin's complexity. Basin complexity of 412 Model Parameter Estimation Experiment basins from continental United States are then related with its six basin characteristics. All the characteristics have been derived from the Model Parameter Estimation Experiment database to represent climate, vegetation and soil characteristics of the basins in a concise manner. Results find that more complex basins that are drier have less seasonal rainfall, vegetation with more storage capacity (i.e., smaller 5-week Normalized Difference Vegetation Index gradient), and faster responsive soils. The results reaffirm prior observations that minimum complexity that is required to model a basin depends on its climate and landscape characteristics (e.g., complex models do not perform well in dry basins).

1. Introduction

One important lesson learned from the prediction in ungauged basins initiative of the International Association of Hydrological Sciences is that variability in streamflow is controlled by climate and landscape attributes of basins (Parajka et al., 2013; Viglione et al., 2013). Streamflow in ungauged basins are predicted based on the transfer of understanding of the transformation of rainfall to runoff, in terms of models of similar but gauged basins. Therefore, the ungauged basins where it is even more difficult to predict streamflow due to its attributes such as aridity will not only have poorer candidate models to tansfer but also will have poorer transfer of models itself.

Due to complex connections between processes and responses, streamflow predictability has been found to be influenced by basin characteristics such as area and aridity (Parajka et al., 2013; Viglione et al., 2013). Larger basins are expected to smooth out the effects of within basin process heterogeneities on its streamflow responses more than smaller basins. As a symptom of difficulty of predicting streamflow in such basins, often more complex models are needed to explain the response, which often is difficult when limited data are available. Thus, there has been a tendency to use simple models in dry basins and more complex model in humid basins, even though the latter are hypothesized to have more linear rainfall-runoff processes (Parajka et al., 2013). There is no doubt that models of different complexities work best only in certain environments (Fenicia et al., 2011), and therefore, those models should be selected that are appropriate for the climate, vegetation, and soil characteristics of a basin (van Werkhoven et al., 2009).

Basin characteristics have also been used to identify donor basins to transfer models to ungauged basins with the understanding that predictabilty of streamflow depends on such characteristics. Several similarity indices and transfer schemes have been proposed that rely on basin characteristics (Abdulla & Lettenmaier, 1997; Beck et al., 2016; Carrillo et al., 2011; Oudin et al., 2008; Parajka et al., 2005; Young, 2006; Zhang & Chiew, 2009). It is therefore not surprising that the performance of such transfer schemes itself depends on basin characteristics. This is because the performance of the models that they transfer is conditioned upon the latter (Viglione et al., 2013).

This paper explores the question of how predictability of streamflow varies with basins characteristics yet again, except that it studies how the former varies with multiple characteristics at the same time instead of with one characteristics at a time. For example, the interpretation that larger basins are easier to predict, based on one to one correlation, implicitly assumes that all other confounding characteristics, such as precipitation seasonality and drainage characteristics, are constant. What if bigger basins lie in regions of more seasonal rainfall? Can some wetter basins be more complex, in terms of its streamflow predictability, because its vegetation has more water storage capacity in a more seasonal climate?

We interpret streamflow predictability in terms of existing concepts of complexity and critical flow events. Several interpretations of complexity exist that either interpret complexity in terms of model parameter nonidentifiability (Beven, 2006), number of model parameters (e.g., Downer & Ogden, 2003; Jakeman & Hornberger, 1993; Keating et al., 2010; Young et al., 1996), number of reservoirs in a model structure (e.g., Bai et al., 2009; Buttsa et al., 2004; Farmer et al., 2003; Martinez & Gupta, 2011; Sivapalan et al., 2003; van der Linden & Woo, 2003), information content of data (e.g., Gupta et al., 2008, Marshall et al., 2005; Patil & Stieglitz, 2014; Ye et al., 2008), or richness of underlying nonlinear dynamics (e.g., Kumar, 2011; Puente & Sivakumar, 2007; Young et al., 1996). This paper interprets complexity in terms of difficulty in learning statistical patterns (Cucker & Smale, 2002) from streamflow sequences and thus in terms of difficulty in predicting historically similar streamflow events (Pande et al., 2009). Basins are then deemed more complex that generate streamflows that are more difficult to predict. In other words, if more complex model is required to predict hydrologic processes then this indicates that the basin is more complex. The paper relates basin complexity with its characteristics in order to understand how streamflow predictability varies as a function of climate, vegetation and soil characteristics of the basin. For this we comparatively analyze the characteristics and basin complexity of 412 MOPEX basins in the continental United States (Duan et al., 2006; Sawicz et al., 2011).

The paper is organized as follows. The first section on methodology presents the problem of learning from the statistical patterns of past streamflow events. MOPEX data set that is used and cluster analysis is then presented followed by the results section. These results are then discussed, followed by the conclusions drawn.

2. Methodology

The problem of streamflow prediction is framed in terms of *k* nearest neighbor modeling (Lall & Sharma, 1996; Sharma et al., 1997), which first identifies a *query* that is made up of lagged streamflows (e.g., streamflow sequence at the last ℓ time steps). This then queries the model to predict the next time step streamflow and makes a prediction based on finding lagged streamflow events from the past that are similar to the query.

A query is deemed more difficult if the error between what is predicted and what is observed in the next time step is larger in general. That is, the statistical pattern corresponding to the query is difficult to learn from past events and predict. This notion of difficulty, in particular the generality of errors (leave one out cross-validation error used for assessing streamflow predictability in ; Viglione et al., 2013, is one example of generalized error), of the next time step streamflow prediction is assessed based on Vapnik-Chervonenkis (VC) generalization theory (Vapnik, 1982; Vapnik & Chervonenkis, 1971).

VC generalization theory expresses the worst-case performance of a model based on the ratio of its performance of predicting streamflow in response to a given query and a measure of model complexity called VC dimension (Shao & Cherkasky, 2000; Vapnik, 1999). The same ratio has successfully been used to constrain average performance of models in high dimensional inverse problems (e.g., Bartlett & Kulkarni, 1998; Vapnik, 1999). The worst-case performance of a nearest neighbor model depends on the query it responds to. That is, a model will perform worse on a query that is more difficult to predict even if the best possible model is chosen for the purpose. Such a query may correspond to a rare streamflow event resulting from multiple underlying processes or as a consequence of crossing thresholds that are otherwise crossed less often, such as within basin connectivity of dry basins (Farmer et al., 2003). In order to identify such difficult to predict events, corresponding time indices are identified that lie on the boundary of streamflow data cloud in ℓ dimensional space (Singh & Bardossy, 2012).

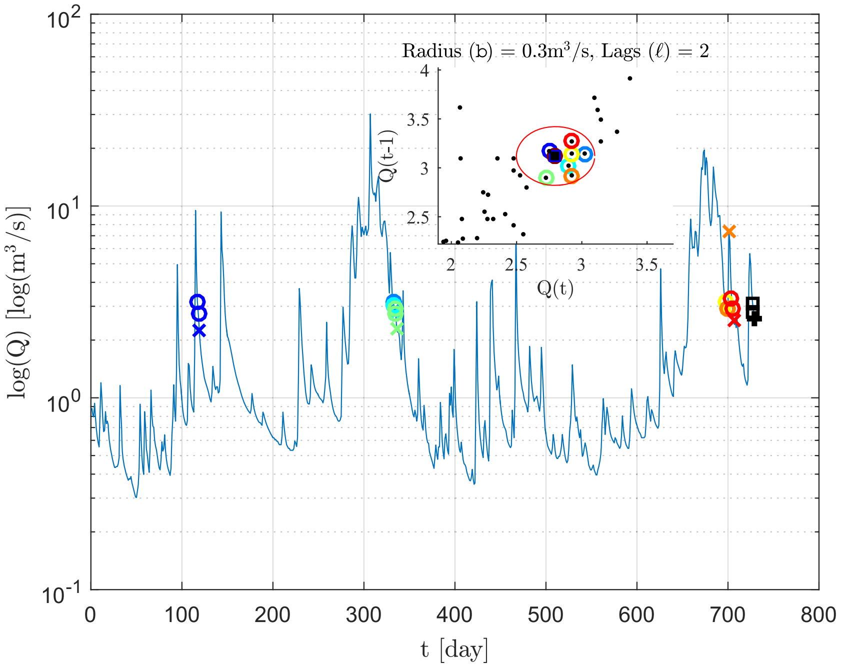

Figure 1. An illustration of *k* nearest neighbor streamflow prediction. Query, shown in black squares at the end of the time series, is made of two lagged streamflow variables and seeks to predict the next time step value that is observed to be +. For a given radius b = 0.3 m^3/s, the model seeks other lagged streamflow vectors that lie within the 0.3 m^3/s radius neighborhood of the query. This is shown in the inset by the black dot, and selected similar occurring vectors are shown within the circle with radius b. The prediction is the mean of next time step predictions of these neighboring streamflow vectors, that is, the mean of values shown by x.

More complex basins are then the ones which have higher worst-case performance of the *k* nearest neighbor models with least possible model complexity. That is, the difficulty in predicting streamflow of a basin is framed in terms of the performance of the best possible streamflow predictions on the most difficult streamflow events that the basin witnesses.

The following subsections further explain *k* nearest neighbor models and complexity regularized prediction that is used to find the best possible prediction for most difficult streamflow events.

2.1. The *k* Nearest Neighbor Streamflow Prediction

Nonparametric models, such as *k* nearest neighbor models, avoid strong assumptions about rainfall-runoff relationships (e.g., Karlsson & Yakowitz, 1987; Sharma et al., 1997; Yakowitz, 1993). The *k* nearest neighbor models (Lall & Sharma, 1996) first identify *k* time sequences from the past that are closest to a query of, for example, predicting the next time step streamflow. In case of hydrological time series, a sequence can be of lagged rainfall, lagged streamflow, etc. The predicted stream flow of the query is the average of next time step streamflow of *k* sets of similar sequences (i.e., *k* nearest neighbors).

Figure 1 is an illustration of the method of *k* nearest neighbor streamflow prediction. The input variables that drive the model to predict stream flow $Q(t+\Delta t)$ are lagged streamflow, $Q(t')$. The time indices, t', for the input variables lie between t and $t-\ell\Delta t$, that is, $t' \in \{t, t-\Delta t, \dots, t-\ell\Delta t\}$ for some $\ell \geq 0$ and $\Delta t > 0$ time step.

At any given time t, the nearest neighbor model predicts the streamflow for time $t+\Delta t$, $\hat{Q}(t+\Delta t)$. The model predicts $\hat{Q}(t+\Delta t)$ based on the following steps.

1. Query: Take a set of lagged streamflow of predetermined size ℓ, that is, $\{Q(t), Q(t-\Delta t), \dots, Q(t-\ell\Delta t)\}$. Call this a query,

$$\mathbf{x}_o = \{Q(t), Q(t-\Delta t), \dots, Q(t-\ell\Delta t)\}.$$

2. Nearest neighbors: For the query $\mathbf{x}_o$, find k nearest neighbors, $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ from the past time series of streamflow of the basin. This is done by arranging the time series in the same format as the query. For example, $\mathbf{x}_i$, for $i = 1, .., k$, is a set of lagged variable from time periods $t_i' < t$from the past, that is,

$$\mathbf{x}_i = \{Q(t_i'), Q(t_i' - \Delta t), \dots, Q(t_i' - \ell \Delta t)\}.$$

The k such vectors are picked that are closest to $\mathbf{x}_o$ in terms of Euclidean distance.

3. Prediction: The streamflow prediction $\hat{Q}(t + \Delta t)$ is given by the mean of one step ahead streamflow values of k nearest neighbors of the query $\mathbf{x}_o$. That is,

$$\hat{Q}(t + \Delta t) = \frac{\sum_{i=1}^{i=k} Q(t_i' + \Delta t)}{k}$$

We use a slight modification of the model as illustrated in Figure 1 (Pande et al., 2009). Instead of using k nearest neighbors of the query $\mathbf{x}_o$, we define a rectangular kernel centered at $\mathbf{x}_o$ that has a radius b. The kernel, $K(\mathbf{x}_j - \mathbf{x}_o; \mathrm{b})$, is such that whenever the Euclidean distance between $\mathbf{x}_o$ and any other point $\mathbf{x}_j$ is less than or equal to b it calls $\mathbf{x}_j$ as one of the nearest neighbors of $\mathbf{x}_o$. The value that the kernel takes for such a point is 1; otherwise, it is set to 0 (see also ; Pande et al., 2009). This is how $K_{\mathrm{b},\ell}$ neighbors of $\mathbf{x}_o$ are chosen for a given value of b and ℓ, which is by finding those $\mathbf{x}_j$ whose Euclidean distance from $\mathbf{x}_o$ is less than or equal to b.

If we represent the streamflow data set in lagged form by $\mathbf{D}_N = \{(Q(t_j + \Delta t), \mathbf{x}_j)\}_{j=1,\dots,N-\ell}$, then the nearest neighbor model $m(\mathbf{x}_o, \mathbf{D}_N; \mathrm{b}, \ell) : \mathbf{x}_o \rightarrow \hat{Q}$ makes a prediction $\hat{Q}(t + \Delta t)$ in response to the query $\mathbf{x}_o$ based on $K_{\mathrm{b},\ell}$ occurrences that are similar as

$$\hat{Q}(t + \Delta t) = \frac{\sum_{i=1}^{i=K_{\mathrm{b},\ell}} Q(t_i + \Delta t)}{K_{\mathrm{b},\ell}}, \tag{1}$$

where t_i is the time index of lagged streamflow that is one of the $K_{\mathrm{b},\ell}$ nearest neighbors of the query $\mathbf{x}_o$. It can be shown that the model prediction, $\hat{Q}$, in equation (1) is obtained by minimizing the following objective function with respect to y_p,

$$r_N(\mathbf{x}_o, \mathbf{D}_N; y_p, \mathrm{b}, \ell) = \frac{\sum_{t=\ell\Delta t}^{N} \left(Q(t + \Delta t) - y_p\right)^2 K(\mathbf{x}_i - \mathbf{x}_o; \mathrm{b})}{N}$$

Thus, the nearest neighbor model in equation (1) produces smallest weighted mean of residuals among all possible models for given (b, ℓ),

$$r_N(\mathbf{x}_o, \mathbf{D}_N; \mathrm{b}, \ell) = \min_{y_p} r_N(\mathbf{x}_o, \mathbf{D}_N; y_p, \mathrm{b}, \ell).$$

The nearest neighbor model seeks occurrences similar to a query in the past with the idea that similar streamflow is likely to follow next (i.e., corresponding streamflow pattern learned from the past will repeat). Therefore, similar occurrences but with larger lags (ℓ) may be needed to identify sufficient number of nearest neighbors to confidently predict the next time step streamflow in case of basins with heterogeneous landscape or with many processes involved in generating streamflow. The interlinkages between number of lags ℓ, radius of neighborhood b, and confidence in predicting next time step streamflow are discussed next.

2.2. Complexity Regularized Prediction

The predictions made by a nearest neighbor model depend on the chosen number of neighbors, and therefore on b and ℓ. Note that a prediction is made for each query, so a prediction is specific to a query. Therefore, (b, ℓ) can be locally specified for each query. A query about next time period streamflow resulting from multiple processes may require more lags of past streamflow in order to obtain more information. But this increases the dimensionality of the prediction problem and may require more nearest neighbors to make a confident prediction. This is because increased dimensionality increases model complexity, and therefore, more data (in terms of nearest neighbors) are needed for a confident prediction. Such a tradeoff relationship between data and model complexity is generally realized based on worst-case performance of models.

Pande et al. (2009) estimated a complexity index, called VC dimension $h(\mathrm{b}, \ell)$, for nearest neighbor models that is based on its worst-case performance (Shao & Cherkasky, 2000; Vapnik et al., 1994). In context of nearest neighbor modeling of streamflow, worst-case performance corresponds to difficult or critical queries that appear less often in stream flow time series. In this way, the estimated VC dimension describes the worst-case\vspace*{4pt} performance of a nearest neighbor model. Since model performance is never worse

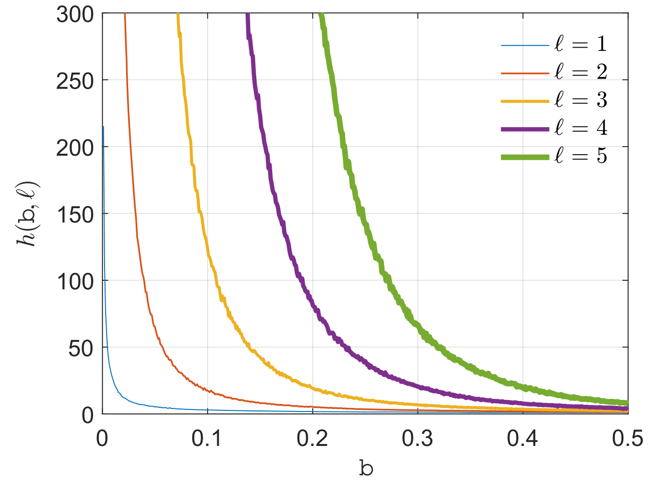

Figure 2. The complexity index called Vapnik-Chervonenkis dimension, $h(\mathrm{b},\ell)$, as a function of radius of neighborhood, b, and number of lags ℓ. Shown is $h(\mathrm{b},\ell)$ estimated on data normalized between 0 and 1. Hence, entire data are covered by a radius of b = 0.5. For a given number of lags, complexity increases with small radius of neighborhood, while for a given radius of neighborhood it increases with the number of lags.

than its worst-case performance, the complexity index defines an upper bound on model performance in general. Also note that different values of b and ℓ result in different models, different worst-case performances, and hence different complexity indices. Figure 2 shows the estimated VC dimension (h) as a function of b and ℓ, that is, $h(\mathrm{b},\ell)$.

If model performance of a nearest neighbor model is defined as the mean squared error between model prediction and next time step streamflow values of $\mathbf{x}_i$ vectors in the neighborhood of the query $\mathbf{x}_o$,

$$r_N(\mathbf{x}_o,\mathbf{D}_N;\mathrm{b},\ell)=\frac{\sum_{t=\ell\Delta t}^{N}\left(Q(t+\Delta t)-\hat{Q}(t+\Delta t)\right)^2 K(\mathbf{x}_i-\mathbf{x}_o;\mathrm{b})}{N}, \qquad (2)$$

then the general performance of the model in predicting stream flow corresponding to any given query $\mathbf{x}_o$ is given by the left-hand side of (Pande et al., 2009; Corani & Gatto, 2006)

$$r(\mathbf{x}_o;\mathrm{b},\ell)\le\frac{r_N(\mathbf{x}_o,\mathbf{D}_N;\mathrm{b},\ell)}{\mathrm{vm}(h(\mathrm{b},\ell),N)}. \qquad (3)$$

Here $r(\mathbf{x}_o;\mathrm{b},\ell)=\int r_N(\mathbf{x}_o,\mathbf{D}_N;\mathrm{b},\ell)\mathrm{P}(\mathbf{D}_N)\mathrm{dP}(\mathbf{D}_N)$ with $\mathrm{P}(\mathbf{D}_N)$ being the probability of sampling $\mathbf{D}_N$ data from the underlying streamflow-generating processes. The denominator, $\mathrm{vm}(h(\mathrm{b},\ell),N)$, on the right-hand side is called *Vapnik Measure* that is given by (Cherkasky & Mulier, 1998)

$$\mathrm{vm}(h(\mathrm{b},\ell),N)=\max 0,\left(1-\sqrt{p-p\log p+\frac{\log N}{N}}\right),$$

where $p=h(\mathrm{b},\ell)/N$. Note that $\mathrm{vm}(h(\mathrm{b},\ell),N)$ increases when model complexity $h(\mathrm{b},\ell)$ increases.

The best possible model for a given query $\mathbf{x}_o$ is obtained by selecting parameters (b^*,ℓ^*) such that the right-hand side of equation (3) is minimized, balancing the mean squared error with model complexity. This ensures that extremely poor performances are avoided on average, and the model is of complexity that is appropriate for the query at hand.

2.3. Critical Indices Using Data Depth Function

The worst-case performance of a model corresponds to the maximum of $r(\mathbf{x}_o;\mathrm{b},\ell)$ with respect to possible queries $\mathbf{x}_o$. This means that difficult queries bring out worst-case performances of the model. If $\tilde{\mathbf{x}}_o$ represent one such difficult query then the performance of its best possible k nearest neighbor prediction is given by

$$r(\tilde{\mathbf{x}}_o;\mathrm{b}^*,\ell^*)\approx\frac{r_N(\tilde{\mathbf{x}}_o,\mathbf{D}_N;\mathrm{b}^*,\ell^*)}{\mathrm{vm}(h(\mathrm{b}^*,\ell^*;\tilde{\mathbf{x}}_o),N)}. \qquad (4)$$

This is when inequality in equation (3) is tight (i.e., holds approximately with equality) and the worst-case performance of the model on the difficult query $\tilde{\mathbf{x}}_o$ in general (in general because it is over different realizations of $\mathbf{D}_N$), $r^*(\tilde{\mathbf{x}}_o)=r(\tilde{\mathbf{x}}_o;\mathrm{b}^*,\ell^*)$, can be obtained from the right-hand side of equation (4).

Here $h^*(\tilde{\mathbf{x}}_o)=h(\mathrm{b}^*,\ell^*;\tilde{\mathbf{x}}_o)$ corresponds to the best possible model that can be selected to predict the difficult streamflow event corresponding to the query $\tilde{\mathbf{x}}_o$. This essentially gives an indication of how difficult it is to learn from the statistical patterns of past streamflow events and to predict streamflow of the corresponding basin in general.

We use the concept of data depth function to identify critical time indices corresponding to difficult queries. The data depth concept orders multivariate data from the center of the multivariate data cloud in an outward direction by first establishing the *centrality* of observations. The aim is to find time indices corresponding to unusual streamflow events that lie on the boundary of the cloud, thereby yielding difficult queries at those indices (see Figure 3). We use Tukey half-space depth function (Tukey, 1975) for extracting unusual events based on ICE (identification of critical events) algorithm of Singh and Bardossy (2012). We operate ICE on data

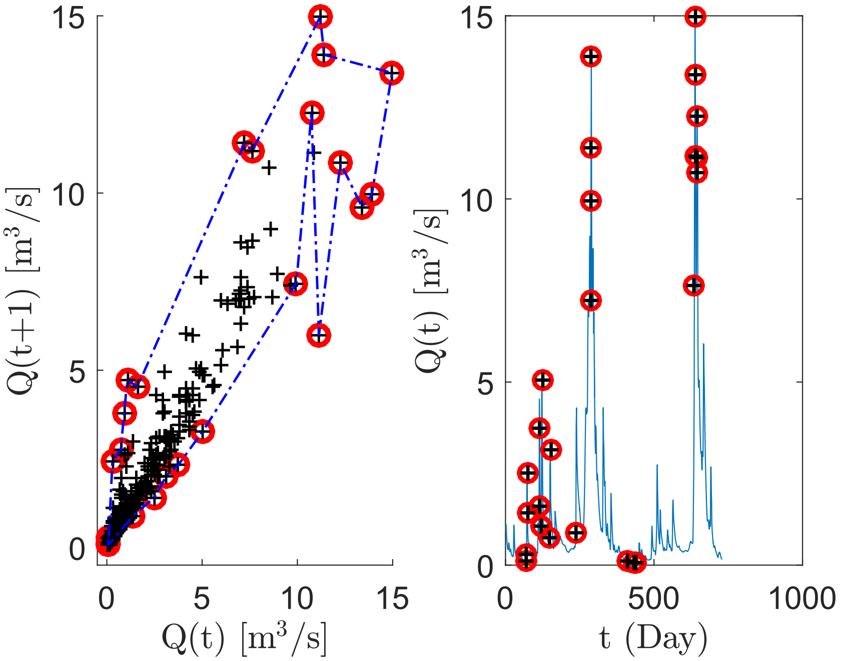

Figure 3. (left panel) Identification of critical streamflow events using Tukey half-space data depth function on the boundary of the data cloud. (right panel) Identification of time indices corresponding to critical streamflow events at the boundary of the data cloud. Queries are then constructed on the corresponding time indices.

cloud in three-dimensional space that is made up of lagged streamflow vectors as $\mathbf{q}(t) = \{Q(t), Q(t-\Delta t), Q(t-2\Delta t)\}$ for $t = 3, .., N$.

Figure 3 shows how the critical indices are obtained that identify difficult queries for data cloud in two-dimensional space. The queries constructed at these indices are then deemed difficult.

2.4. Basin Complexity

For a given difficult query $\tilde{\mathbf{x}}_o$ identified by the ICE algorithm, the complexity index value, $h^*(\tilde{\mathbf{x}}_o) = h(\mathrm{b}^*, \ell^*; \tilde{\mathbf{x}}_o)$ from equation (4) reflects the difficulty of finding occurrences similar to the streamflow event corresponding to the query. We define basin complexity as the average of log complexity, $\log h^*(\tilde{\mathbf{x}}_o^j)$, calculated on J difficult queries from $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}}_o^j \in \mathcal{X}, j = 1, \dots, J\}$ identified on $\mathbf{D}_N$ lagged streamflow data set of the basin. Here $\mathcal{X}$ is the set of all the queries. That is, the basin complexity is defined as

$$\mathcal{H} = \frac{\sum_{\tilde{\mathbf{x}}_o \in \tilde{\mathcal{X}}} \log h^*(\tilde{\mathbf{x}}_o)}{J} \tag{5}$$

where $h^*(\tilde{\mathbf{x}}_o) = h(\mathrm{b}^*, \ell^*; \tilde{\mathbf{x}}_o)$ corresponds to $(\mathrm{b}^*, \ell^*) \leftarrow \arg\min_{\mathrm{b},\ell} r(\tilde{\mathbf{x}}_o; \mathrm{b}, \ell)$.

3. Data Description and Analysis

3.1. MOPEX Basins and Its Hydrological Characteristics

The basin complexity described in equation (5) is calculated for 412 basins covering the continental United States. Daily time series of streamflow for the basins are obtained from the MOPEX data set (Duan et al., 2006; Sawicz et al., 2011), acknowledging that larger data set consisting of more diverse range of climate and geographical domain could also be used (Addor et al., 2017; Newman et al., 2017).

In total, six hydrological characteristics of the basins are selected and described in Table 1. These characteristics have been inspired by Sawicz et al. (2011) who clustered MOPEX basins from the eastern United States based on six streamflow-based signatures and found systematic variation in climate, vegetation, and soil characteristics of those clusters. The basin characteristics shown in Table 1 will be used to cluster the basins with the underlying assumption that variability of streamflow with basin characteristics does not functionally change within a cluster but changes across the clusters.

Table 1
Hydrological Characteristics of Selected 412 MOPEX Basins in the Eastern United States

Characteristics (units)	Explanation	Range
WRC (-)	Slope of the water retention curve (Clapp and Hornberger *b* parameter)	3.08 – 11.37
P/PE (-)	Annual precipitation/annual potential evaporation	0.22 – 4.32
GNDVI (-)	5-week gradient of NDVI before peak greeness	4.57 – 50.00
MeanSlope (%)	Mean slope	0.52 – 26.00
PSI (-)	Precipitation seasonality index	0.05 – 0.86
Porosity (-)	Porosity	0.35 – 0.48

Precipitation seasonality index (PSI) is calculated as $\sum_{n=1}^{12} |P_n - \bar{P}/12|/\bar{P}$, where P_n is the mean monthly precipitation (mm/month) and $\bar{P}$ is the mean annual precipitation (mm/year) of a basin. It is close to 0 for basins where monthly rainfall is uniformly distributed over a year and close to 1 for basins with highly seasonal monthly precipitation. The rest of the characteristics have been obtained from the MOPEX data set. Basins with higher water retention curve (WRC) are expected to drain faster, such basins also tend to have lower porosity. Drier basins have lower P/PE values. Basins with seasonal vegetation, such as prairies, are expected to have higher values of GNDVI.

3.2. Complexity Computation and Cluster-Specific Regression

The complexity $\mathcal{H}$ for each of the 412 basins are computed using equation (5). For this, a set of difficult queries $\tilde{\mathcal{X}}$ are generated over 10 years of daily streamflow data and complexity indices, $h^*(\tilde{\mathbf{x}}_o)$, are computed for each $\tilde{\mathbf{x}}_o \in \tilde{\mathcal{X}}$ by finding corresponding nearest neighbors from an independent set of 4 years of daily streamflow. In total, 14 years of daily streamflow data from 26 September 1950 to 22 September 1964 is selected (42 basins have a different period) and used for all the basins.

The computed basin complexity is then appended to the corresponding hydrological characteristics for all 412 characteristics. The basin characteristics and complexity are standardized individually, that is, by subtracting the mean of a variable (e.g., PSI or $\mathcal{H}$) calculated over 412 basins and dividing it by its standard deviation. We then explore whether basin complexities are significantly correlated with hydrological characteristics one at a time. Then the basins are clustered into groups of similar characteristics to geographically locate basins associated with certain complexity, assuming that the dependence of streamflow variability with basin characteristics is of similar nature within a cluster but is different across the clusters.

Agglomerative Hierarchical Clustering algorithm (Ward, 1963) is used for clustering the basins based on its characteristics (see Table 1). It is implemented in MATLAB using function *cluster*, which constructs clusters from the agglomerative hierarchical cluster tree as generated by the linkage function. The method used in the linkage function was *ward* (Release, M A T L A B, 2015). It starts with one basin in one cluster and then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying a similarity criteria (we use Ward method which minimizes the total within cluster variance), till all of the data are in one cluster. Then a subjective number of clusters is chosen that cuts off the hierarchy at a lower point. We choose the best number of clusters such that the mean of basin complexities within each cluster is sufficiently different from the means of other clusters.

Finally, we perform cluster-specific linear regressions between the characteristics and basin complexity $\mathcal{H}$, in order to explore multivariate linear relationships within each such cluster.

4. Results

Figure 4 shows how various basin characteristics correlate with basin complexity one at a time. All are significantly correlated with p value < 0.005. The climatic variables such as P/PE and PSI are most strongly correlated, followed by GNDVI and MeanSlope. Porosity and WRC characteristics appear to have weakest correlations. The basin complexity increases with aridity and seasonality. It is lower for basins that have steeper slopes and have faster changing vegetation (higher GNDVI). Finally, the basins that drain more easily, that is, associated with lower porosity and/or steeper WRC, have higher complexity though such correlations are relatively weak.

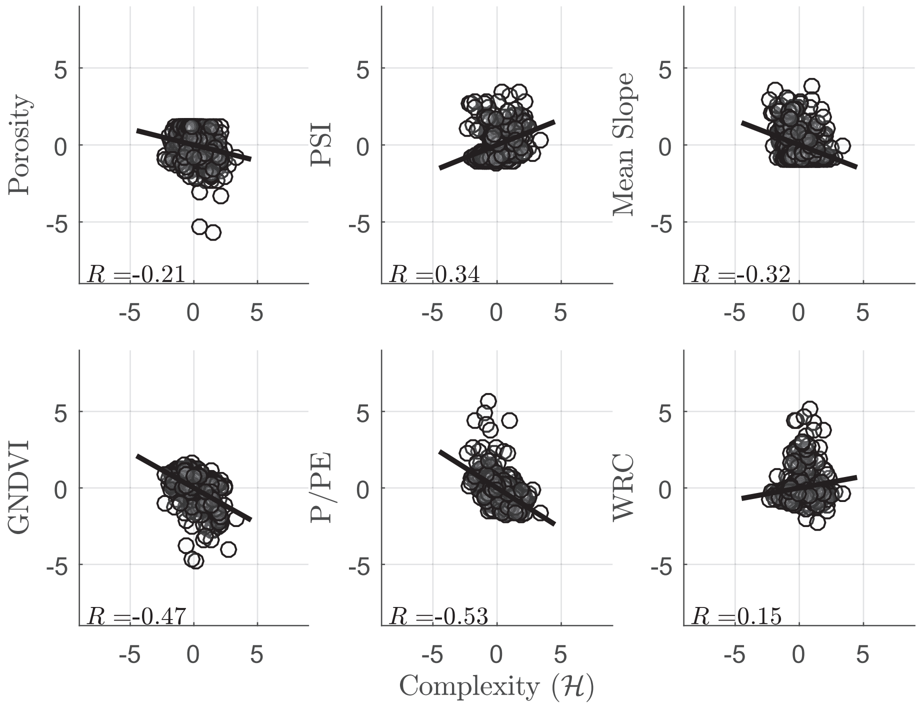

Figure 4. Pairwise correlations between hydrological characteristics and basin complexity $\mathcal{H}$. Characteristic and complexity values have been standardized by subtracting respective mean from it and dividing by corresponding standard deviation. The correlations, R, shown are statistically significant with $p < 0.005$. Also shown are the best fit lines based on pairwise linear regressions. PSI = precipitation seasonality index; WRC = water retention curve.

It is evident from Figure 4 that each of the pairwise scatters plots nonlinear relationships. Locally linear relationships in different parts of the plots are evident. These locally linear relationships appear to be multivariate as well; that is, the pairwise scatters are two-dimensional projections of higher dimensional covariation between complexity and the characteristics. This necessitates a local multivariate analysis, which is performed based on cluster analysis.

Only four clusters could be identified with statistically different means (p value < 0.001). However, the distribution of basins was quite unevenly distributed. The basins are therefore clustered into five clusters that have sufficiently different complexity means from each other. Figure 5 shows the variation of hydrological characteristics and basin complexity within the five clusters. While the covariation of basin characteristics with complexity is overall as observed in pairwise correlation plots in Figure 4, that is, WRC and PSI increase and other characteristics decrease with increasing basin complexity ($\mathcal{H}$), there are differences in how the characteristics covary within a cluster. For example, MeanSlope strongly varies within CL V, after gradually varying from CL I-CL IV. WRC gradually varies across all the clusters except CL IV, just as PSI except for its variation within CL I. P/PE, Porosity and GNDVI have visible variability in each of the five clusters.

Table 2, which shows cluster-specific linear regression results, corroborates some of the observations made in Figure 4. All but GNDVI and PSI characteristics are consistent in their signs of covariation with complexity.

Figure 6 maps the basin geographically according to their clusters. The lowest two complexity clusters lie in the northwest along the Pacific coastline and along the Appalachian mountain range in eastern United States. Most complex basins lie in the western United States.

Daily streamflow for each basin is predicted based on similar occurring events in the past. It is therefore expected that basins that generate streamflow as a result of overcoming multiple thresholds or are forced by less predictable rainfall events are more difficult to predict. Such basins are deemed as more complex. The signs of cluster-specific slopes suggest that flatter and faster draining basins have higher complexity. Also, basins with more seasonal rainfall or drier basins appear to be more complex. Farmer et al. (2003) corroborate this observation, who argued that drier catchment require more complex model structure because these catchments experience more disruption of within basin connectivity when compared to wetter catchments.

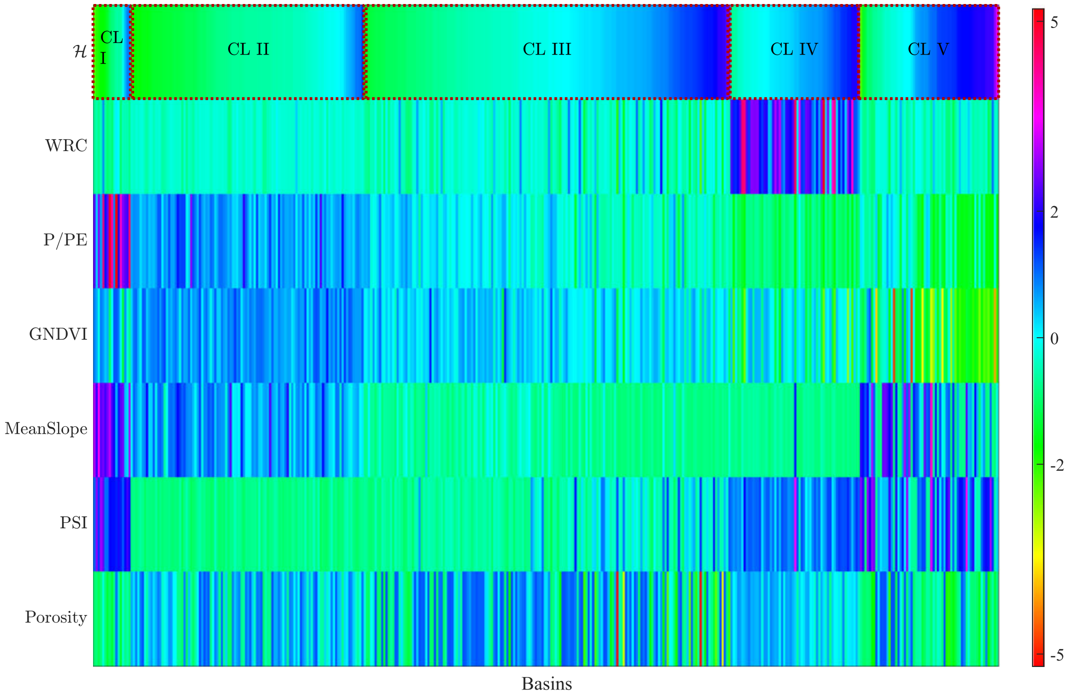

Figure 5. Heat map of basin characteristics for five identified clusters with mean complexities $\mu_1 = -0.880$ (CL I, $N = 17$), $\mu_2 = -0.666$ (CL II, $N = 106$), $\mu_3 = 0.050$ (CL III, $N = 167$), $\mu_4 = 0.511$ (CL IV, $N = 59$), and $\mu_5 = 0.747$ (CL V, $N = 63$). The clusters are identified such that means of basin complexity within clusters are sufficiently different from each other. PSI = precipitation seasonality index; WRC = water retention curve.

While drier basins are more complex to predict overall, basin complexity is positively related with basin dryness only within CL III and CL V complexity clusters. There is no significant relationship between dryness and complexity for the remaining clusters, contrary to the notion that drier basins are more complex to understand and predict (Parajka et al., 2013).

The case that flatter basins are more complex is only for lowest complexity basins that are also the wettest (CL I and CL II). These basins lie along the northwest coastline and along the Appalachian mountain range in eastern United States. Note that lowest complexity clusters also have basins with steep slope (see Figure 5). Therefore, flatter basins being more complex applies to the set of basins that are relatively steep. Here the vegetation characteristic (GNDVI) appears to bear inconsistent relationship with basin complexity across the two lowest complexity cluster. Further, CL II basins (along the Appalachian mountain range) witness its complexity

Table 2
Cluster-Specific Regression Between Hydrological Characteristics (as Independent Variable) With Basin Complexity (as the Dependent Variable)

	CL I		CL II		CL III		CL IV		CL V	
	mean	var	mean	var	mean	var	mean	var	mean	var
$\mathcal{H}$	−0.880	0.056	−0.666	0.004	0.050	0.005	0.511	0.008	0.747	0.021
	slope	*p*-val	slope	*p*-val	slope	*p*-val	slope	*p*-val	slope	*p*-val
WRC	—	—	1.769	<0.001	0.276	0.088	—	—	—	—
P/PE	—	—	—	—	−0.764	<0.001	—	—	−0.918	<0.001
GNDVI	−0.850	0.013	0.460	0.022	—	—	—	—	−0.218	0.047
Mean slope	−0.947	0.002	−0.224	0.066	—	—	—	—	—	—
PSI	—	—	—	—	—	—	—	—	−0.270	0.032
Porosity	—	—	−0.430	0.001	−0.251	<0.001	−0.675	0.011	−0.318	0.033

Note. The slopes of only those characteristics are shown that are significant with p value < 0.10. WRC = water retention curve; PSI = precipitation seasonality index.

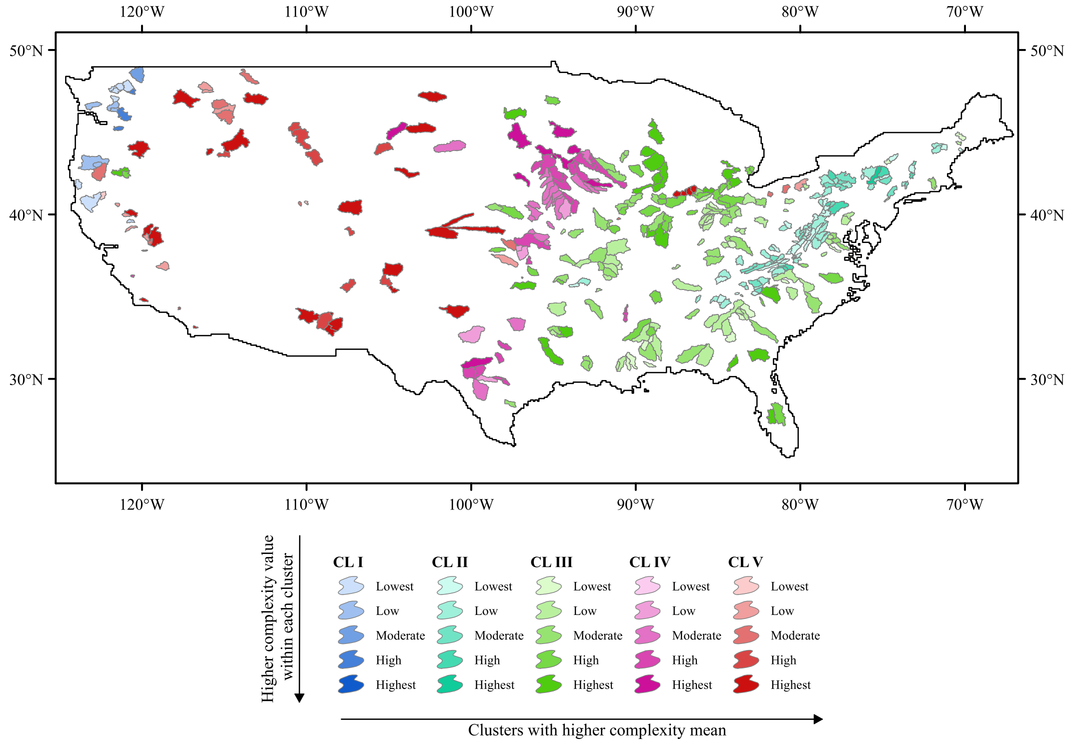

Figure 6. Spatial distribution of basin complexity. High complexity basins (CL IV and CL V) lie to the west, drier part of eastern United States. Low complexity basins (CL I and CL II) lie on the wet northwestern and northeastern part with steeper slopes.

vary strongly with soil characteristics in addition to mean slope and GNDVI. Therefore, wetter basins are more complex when they have more responsive vegetation and soils and are flatter (implying bigger basin areas; ; Willgoose et al., 1991). That is, bigger basins can be more complex if they have more responsive catchments (i.e., with more responsive vegetation and soils).

As move from lowest complexity cluster, CL I, to geographically disconnected second lowest complexity cluster CL II, the covariation of basin complexity with mean slope decreases. Meanwhile, the strong covariation with drainage properties of the basin appears. The vegetation, being deciduous, also responds in a fashion similar to evergreen vegetation of CL I. Vegetation with less rapidly changing greenness in response to similar rainfall and temperature variability has more storage capacity to buffer against variability in water availability. The vegetation therefore has lower variability in water stress as indicated by lower gradient in greenness as we see in CL I. Under strong drainage effect of soils, as we see in CL II, the buffer against variability in water availability is safeguarded by flowering of deciduous trees at the onset of warmer temperatures. This means that as the buffering capacity of vegetation increases, the variability of streamflow partitioned out of rainfall also increases resulting in higher basin complexity in terms of streamflow.

Meanwhile, highest complexity basins that are also the driest (see Figure 5) bear no relationship with average slope. Instead, basin complexity is higher for those basins that have slower changing vegetation (based on the gradient of NDVI, i.e., lower GNDVI) and drier climate. Surprisingly, the complexity of the basins belonging to the highest complexity cluster has strong negative relationship with seasonality in precipitation, while it remains insignificant for all other clusters. This is in spite of an overall positive correlation between basin complexity and seasonality. Also, complexity strongly increases with faster responding basins (negative with porosity) in this cluster. This thus means that faster responding basins that are drier but experience less seasonal rainfall are more difficult to predict.

Consider now the basins in the third cluster that are of medium complex (CL III). Here complexity increases with dryness and responsiveness of soils. The effect of dryness disappears while the effect of drainage property weakens as one moves to medium to high complexity cluster (CL IV) in the midwest. As one moves further west to most complex cluster of basins (i.e., CL V) dryness, seasonality, and vegetation begin to influence basin

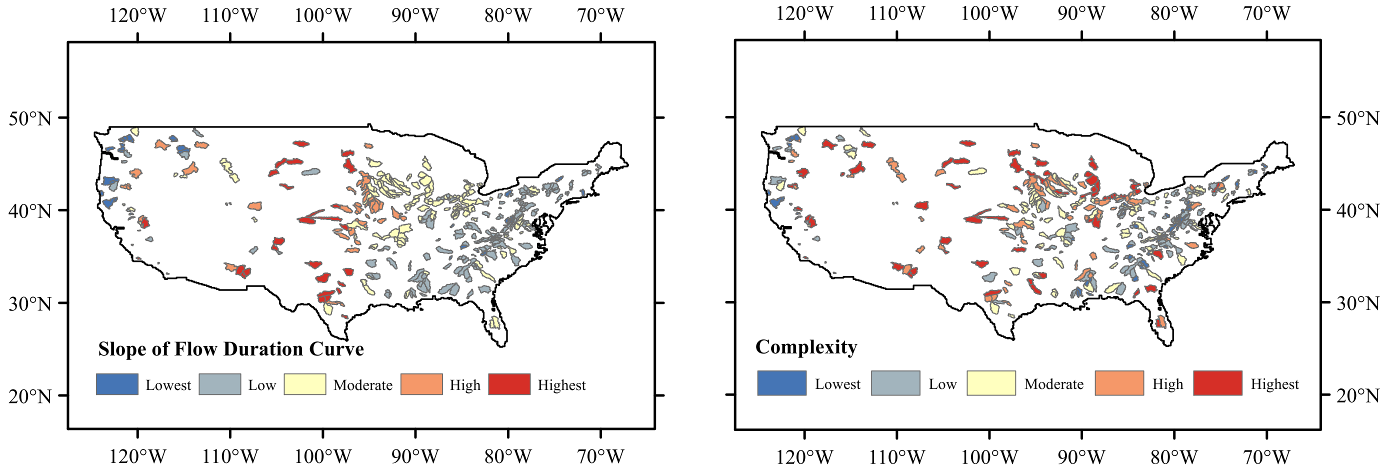

Figure 7. (left panel) Slope of the flow duration curve of the MOPEX basins. (right panel) Basin complexity, five clusters of which were shown in Figure 6. Overall similarities are evident; that is, more complex basins tend to have steeper FDC, except for northeastern and southeastern basins.

complexity again. The vegetation retains its effect as one surprisingly transitions further west from the cluster of most complex basins (CL V) to its neighboring cluster of lowest complexity basins (CL I).

5. Discussion

In summary, more complex basins that are drier have less seasonal rainfall, vegetation with more storage capacity (i.e., smaller 5-week NDVI gradient), and faster responsive soils. In other words, variability in streamflow reduces (predictability of streamflow increases) as basins get wetter and its vegetation respond faster. This may indicate that the vegetation is maximizing its use of water stored in the root zone (by absorbing any variability in rainfall). In other words, it adapts to use the water more efficiently as the climate dries by reducing its buffering capacity (which is less needed because less water is available overall and more is uniformly available throughout the year such as in basins of CL V) and quickly vaporizing any positive variation in rainfall (Troch et al., 2009).

This is explicit in Figure 7 that compares slopes of flow duration curve (FDC) with basin complexity. While Figure 6 shows clustered basins and the gradation of colors is intended to show how complexity changes within each cluster from low to high, Figure 7 (right panel) shows how complexity values vary across all the basins. Faster responding basins with lower storage capacity have steeper FDC (Yilmaz et al., 2008), resulting from more variable streamflow (Berghuijs et al., 2014). This may mean that steeper FDCs are associated with more complex basins and that corresponding basins are more difficult to predict. The dependence of slow and fast parts of FDC on timing of rainfall events and basin storage capacity respectively (Yokoo & Sivapalan, 2011) reiterates overall association of basin complexity with the slope of its FDC. Drier basins are found to be more complex (e.g., in CL IV; see Figure 6 and right panel of Figure 7) and to have intermediate slope of FDC (left panel of Figure 7) when rainfall seasonality does not play a role. Cluster V basins (most complex), which are drier but also bear intuitive relationships with seasonality, vegetation, and soil, also have similar patterns of basin complexity and slope of FDC. However, the association breaks down in the northeastern basins of cluster CL II (low complexity), the northern and southeastern basins of cluster CL III (intermediate complexity), and western basins of cluster V (high complexity).

The northern and northeastern basins of CL II and western basins of CL V have higher fraction of rain falling as snow, while the rainfall is mildly seasonal in the southeastern basins of CL III (Berghuijs et al., 2014). As a result, both such sets of basins are expected to have higher intra-annual variability in streamflow and hence higher complexity—which is what we observe in the map of basin complexity (right panel of Figure 7) but is not captured by the slope of FDC (left panel of Figure 7). This is because high intra-annual variability is captured by lower quantiles of FDC that are not considered when computing the slopes of FDC (Yaeger et al., 2012), but it is captured by the index of basin complexity presented here.

The statistical interpretation of the measure of complexity is that those basins that are more complex are more difficult to predict. It is therefore most difficult to predict in drier parts of western United States and the Great Plains (thereby requiring more complex models), while streamflow is easiest to predict along the northwest and northeastern coastline. If a complex hydrological model, that is, Variable Infiltration Capacity (Newman et al., 2015), is calibrated and used to predict stream flow in both such basins, one is likely to achieve unreliable predictions in the case of drier parts of western United States and the Great Plains (Mizukami et al., 2017; Newman et al., 2015). This is because the corresponding inverse problem (of estimating the parameters of the model by maximizing some measure of model performance on available data) is difficult and not posed well (Vapnik, 1982). One method to tame such difficult inverse problems is penalize the selection problem by a certain measure of model complexity (i.e., regularization of the inverse problem; ; Arkesteijn & Pande, 2013).

Often number of model parameters is considered a measure of corresponding model complexity (e.g., see ; Mo et al., 2006; Newman et al., 2015). This is indeed the case for models that are linear in its input, for example, if daily stream flow is modeled as a linear function of current and last *d* days of daily rainfall. However, in case of highly nonlinear models, complexity has also been found to depend on the magnitude of parameters that nonlinearly transform rainfall to runoff (Pande et al., 2012, 2015). The strong covariation of basin complexity with hydrological characteristics presented in Table 2 reinforces the observation. For example, the complexity of basins in cluster CL II increases as porosity decreases and slope of the WRC increases. Since higher porosity means higher water storage capacity and steeper slope of the WRC corresponds to a higher recession coefficient, this means that basin complexity increases with higher magnitude of recession coefficients (unit: T^{-1}) and lower magnitude of storage capacity. If one imagines that stream flow of the basin has been simulated by some underlying but unknown *universal* hydrological model, this would mean that corresponding model complexity increases with its recession coefficients and reduces with its soil (or vegetation) water storage capacity. Such conclusions have also been drawn elsewhere (Pande et al., 2015).

Basin characteristics have been used to regionalize model parameters in order to extrapolate streamflow predictions from gauged to ungauged locations (Abdulla & Lettenmaier, 1997; Young, 2006). The implied dependence of model parameters on hydrological characteristics can therefore be used to regularize a hydrological model of choice in order to predict a basin's streamflow. The regularization will depend on the cluster to which the basin belongs to. For example, the following steps may be taken in modeling the stream flows of basins in CL V: (i) identify parameters of the model that correspond to significant co-variates of basin complexity (i.e., GNDVI and porosity), (2) identify predictive equations of the parameters in terms of these covariates, and (3) then use a weighted sum of the absolute values of the predictive equations to regularize the calibration of the hydrological model. Since the measure of basin complexity also serves as a measure of similarity, the predictive equation identified in step 2 also identifies the sensitive parameters and how the parameters and the model can be extrapolated from gauged to ungauged locations in a robust manner. Such an approach complements other studies such as Archfield and Vogel (2010), Carrillo et al. (2011), and Patil and Stieglitz (2012) that have proposed methods to transfer models from gauged to similar but ungauged basins.

6. Conclusion

The *k* nearest neighbor model predicted streamflow based on the occurrence of similar events in the past for 412 MOPEX basins in continental United States. We used VC generalization theory and data depth function to estimate appropriate model complexity that is needed to predict critical streamflow events in all these basins. Basin complexity was then computed as the average model complexity needed to predict its critical events and regressed with characteristics related to climate, vegetation, and soil. Those basins were deemed more complex where it was more difficult to learn the statistical patterns of critical streamflow sequences. Statistically significant relationships were observed between model complexity and the characteristics, specific to three different clusters of the basins that were identified based solely on the characteristics.

Drier basins in general were found to be more complex; however, wetter basins that are flatter were also difficult to predict. Faster responding basins or those with lower storage capacity (either in vegetation or soil) were also difficult to predict, irrespective of its dryness. Spatial pattern of basin complexity was also found to be closely associated with that of the slope of corresponding FDC, suggesting that basins with steeper FDC are difficult to predict.

Even though the motivation of the paper was to interpret basin complexity in terms of learning from the statistical patterns of streamflow and how it relates with basin characteristics, the presented measure can also serve as an index of hydrological similarity in the light of its relationships with dryness, rainfall seasonality, drainage response, and storage capacity of the basins. This is because such relationships between streamflow and the characteristics can only emerge from the dynamics of climate, vegetation response, soil, and basin-scale streamflow response. Such emergent relationships can therefore be used to control the complexity of basin-specific hydrological models in order to predict steam flow.

Acknowledgments
The authors thank the Editors, three anonymous referees, and Wouter Berghuijs for their constructive comments on previous versions of the manuscript. The MOPEX data used were obtained from Duan et al. (2006), and the data underlying Figure 2 were obtained from Pande et al. (2009). All the codes and data sets generated in the paper are available from Pande and Moayeri (2018).

References

Abdulla, F. A., & Lettenmaier, D. P. (1997). Development of regional parameter estimation equations for a macroscale hydrologic model. *Journal of Hydrology*, *197*(1-4), 230–257. https://doi.org/10.1016/S0022-1694(96)03262-3

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences Discussions*, *2017*, 1–31. https://doi.org/10.5194/hess-2017-169

Archfield, S. A., & Vogel, R. M. (2010). Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments. *Water Resources Research*, *46*, W10513. https://doi.org/10.1029/2009WR008481

Arkesteijn, L., & Pande, S. (2013). On hydrological model complexity, its geometrical interpretations and prediction uncertainty. *Water Resources Research*, *49*, 7048–7063. https://doi.org/10.1002/wrcr.20529

Bai, Y., Wagener, T., & Reed, P. (2009). A top-down framework for watershed model evaluation and selection under uncertainty. *Environmental Modelling & Software*, *24*, 901–916.

Bartlett, P. L., & Kulkarni, S. R. (1998). The complexity of model classes, and smoothing noisy data. *Systems & Control Letters*, *34*, 133–140. https://doi.org/10.1016/S0167-6911(98)00008-5

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., et al. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, *52*, 3599–3622. https://doi.org/10.1002/2015wr018247

Berghuijs, W. R., Sivapalan, M., Woods, R. A., & Savenije, H. H. G. (2014). Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. *Water Resources Research*, *50*, 5638–5661. https://doi.org/10.1002/2014WR015692

Beven, K. J. (2006). A manifest for the equifinality thesis. *Journal of Hydrology*, *320*, 18–36.

Buttsa, M. B., Paynea, J. T., Kristensenb, M., & Madsen, H. (2004). An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *Journal of Hydrology*, *298*, 242–266.

Carrillo, G., Troch, P. A., Sivapalan, M., Wagener, T., Harman, C., & Sawicz, K. (2011). Catchment classification: Hydrological analysis of catchment behavior through process-based modeling along a climate gradient. *Hydrology and Earth System Sciences*, *15*(11), 3411–3430. https://doi.org/10.5194/hess-15-3411-2011

Cherkasky, V., & Mulier, F. M. (1998). *Learning from data*. New York: John Wiley.

Corani, G., & Gatto, M. (2006). VC-dimension and structural risk minimization for the analysis of nonlinear ecological models. *Applied Mathematics and Computation*, *176*, 166–176. https://doi.org/10.1016/j.amc.2005.09.050

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning/B. *American Mathematical Society*, *39*, 1–49.

Downer, C. W., & Ogden, F. L. (2003). Prediction of runoff and soil moistures at the watershed scale: Effects of model complexity and parameter assignment. *Water Resources Research*, *39*(3), 1045. https://doi.org/10.1029/2002WR001439

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

Farmer, D., Sivapalan, M., & Jothityangkoon, C. (2003). Climate, soil, and vegetation controls upon the variability of water balance in temperate and semiarid landscapes: Downward approach to water balance analysis. *Water Resources Research*, *39*(2), 1035. https://doi.org/10.1029/2001WR000328

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

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

Jakeman, A. J., & Hornberger, G. M. (1993). How much complexity is warranted in a rainfall–runoff model? *Water Resources Research*, *29*, 2637–2649.

Karlsson, M., & Yakowitz, S. (1987). Nearest neighbor methods for nonparametric rainfall-runoff forecasting. *Water Resources Research*, *23*(7), 1300–1308. https://doi.org/10.1029/WR023i007p01300

Keating, E. H., Doherty, J., Vrugt, J. A., & Kang, Q. (2010). Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. *Water Resources Research*, *46*, W10517. https://doi.org/10.1029/2009WR008584

Kumar, P. (2011). Typology of hydrologic predictability. *Water Resources Research*, *47*, W00H05. https://doi.org/10.1029/2010WR009769

Lall, U., & Sharma, A. (1996). A nearest neighbour bootstrap for resampling hydrologic time series. *Water Resources Research*, *32*(3), 679–693. https://doi.org/10.1029/95WR02966

Marshall, L., Nott, D., & Sharma, A. (2005). Hydrological model selection: A Bayesian alternative. *Water Resources Research*, *41*, W10422. https://doi.org/10.1029/2004WR003719

Martinez, G. F., & Gupta, H. V. (2011). Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. *Water Resources Research*, *47*, W12540. https://doi.org/10.1029/2011WR011229

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., et al. (2017). *Towards seamless large-domain parameter estimation for hydrologic models* (Vol. 53, pp. 8020–8040). https://doi.org/10.1002/2017WR020401

Mo, X., Pappenberger, F., Beven, K., Liu, S., de Roo, A., & Lin, Z (2006). Parameter conditioning and prediction uncertainties of the LISFLOOD-WB distributed hydrological model. *Hydrological Sciences Journal*, *51*(1), 45–65.

Newman, A. J., Clark, M. O., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, *18*(8), 2215–2225. https://doi.org/10.1175/JHM-D-16-0284.1

Oudin, L., Andreassian, V., Perrin, C., Michel, C., & LeMoine, N. (2008). Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. *Water Resources Research*, *44*, W03413. https://doi.org/10.1029/2007WR006240

Pande, S., Arkesteijn, L., Savenije, H., & Bastidas, L. A. (2015). Hydrological model parameter dimensionality is a weak measure of prediction uncertainty. *Hydrology and Earth System Sciences Discussions*, *12*, 3945–4004. https://doi.org/10.5194/hessd-12-3945-2015

Pande, S., Bastidas, L. A., Bhulai, S., & McKee, M. (2012). Parameter dependent convergence bounds and complexity measure for a class of conceptual hydrological models. *Journal of Hydroinformatics*, *14*, 443–463. https://doi.org/10.2166/hydro.2011.005

Pande, S., McKee, M., & Bastidas, L. A. (2009). Complexity-based robust hydrologic prediction. *Water Resources Research*, *45*, W10406. https://doi.org/10.1029/2008WR007524

Pande, S., & Moayeri, M (2018). Data and algorithms underlying "Hydrological interpretation of a statistical measure of basin complexity". 4TU. Centre for Research Data, Dataset. https://doi.org/10.4121/uuid:08608567-6970-46a0-b14c-3365732ece6b

Parajka, J., Merz, R., & Blöschl, G. (2005). A comparison of regionalisation methods for catchment model parameters. *Hydrology and Earth System Sciences*, *9*(3), 157–171. https://doi.org/10.5194/hess-9-157-2005

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins—Part 1: Runoff–hydrograph studies. *Hydrology and Earth System Sciences*, *17*, 1783–1795. https://doi.org/10.5194/hess-17-1783-2013

Patil, S., & Stieglitz, M. (2012). Controls on hydrologic similarity: Role of nearby gauged catchments for prediction at an ungauged catchment. *Hydrology and Earth System Sciences*, *16*(2), 551–562.

Patil, S., & Stieglitz, M. (2014). Modelling daily streamflow at ungauged catchments: What information is necessary? *Hydrological Processes*, *28*(3), 1159–1169.

Puente, C. E., & Sivakumar, B. (2007). Modeling geophysical complexity: A case for geometric determinism. *Hydrology and Earth System Sciences*, *11*, 721–724.

Release, M A T L A B (2015). *The MathWorks, Inc. Natick*. Massachusetts: United States.

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

Shao, X., & Cherkasky, V. (2000). Measuring the VC-dimension using optimized experimental design. *Neural Comput.*, *12*(8), 1969–1986. https://doi.org/10.1162/089976600300015222

Sharma, A., Tarboton, D., & Lall, U. (1997). Streamflow simulation: A non-parametric approach. *Water Resources Research*, *33*(2), 291–308. https://doi.org/10.1029/96WR02839

Singh, S. K., & Bardossy, A. (2012). Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, *38*, 81–91.

Sivapalan, M., Blöschl, G., Zhang, L., & Vertessy, R (2003). Downward approach to hydrological prediction. *Hydrological Processes*, *17*, 2101–2111. https://doi.org/10.1002/hyp.1425

Troch, P. A., Martinez, G. F., Pauwels, V. R. N., Durcik, M., Sivapalan, M., Harman, C., et al. (2009). Climate and vegetation water use efficiency at catchment scales. *Hydrological Processes*, *23*, 2409–2414. https://doi.org/10.1002/hyp.7358

Tukey, J. (1975). Mathematics and the picturing of data. In Proc. 1975 Inter. Cong. Math., Vancouver 523-531 Montreal: Canad. Math. Congress.

van Werkhoven, K., Wagener, T., Reed, P., & Tang, Y. (2009). Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. *Advances in Water Resources*, *32*, 1154–1169.

van der Linden, S., & Woo, M.-K. (2003). Application of hydrological models with increasing complexity to subarctic catchments. *Journal of Hydrology*, *270*, 145–157. https://doi.org/10.1016/S0022-1694(02)00291-3

Vapnik, V (1982). *Estimation of dependencies based on empirical data*. New York: Springer Verlag.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. https://doi.org/10.1109/72.788640

Vapnik, V., & Chervonenkis, A. (1971). On uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, *16*(2), 264–280.

Vapnik, V., Levin, E., & LeCun, Y. (1994). Measuring the VC dimension of a learning machine. *Neural Computation*, *6*(5), 851–876.

Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., & Bloäschl, G. (2013). Comparative assessment of predictions in ungauged basins—Part 3: Runoff signatures in Austria. *Hydrology and Earth System Sciences*, *17*, 2263–2279. https://doi.org/10.5194/hess-17-2263-2013

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–44.

Willgoose, G., Bras, R. L., & Rodriguez-Iturbe, I. (1991). A physical explanation of an observed link area-slope relationship. *Water Resources Research*, *27*(2), 1697–1702.

Yaeger, M., Coopersmith, E., Ye, S., Cheng, L., Viglione, A., & Sivapalan, M. (2012). Exploring the physical controls of regional patterns of flow duration curves- Part 4: A synthesis of empirical analysis, process modeling and catchment classification. *Hydrology and Earth System Sciences*, *16*, 4483–4498.

Yakowitz, S. (1993). Nearest neighbor regression estimation for null recurrent Markov time series. *Stochastic Processes and Their Applications*, *48*, 311–318.

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

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

Yokoo, Y., & Sivapalan, M. (2011). Towards reconstruction of the flow duration curve: Development of a conceptual framework with a physical basis. *Hydrology and Earth System Sciences*, *15*, 2805–2819. https://doi.org/10.5194/hess-15-2805-2011

Young, A. R. (2006). Stream flow simulation within UK ungauged catchments using a daily rainfall-runoff model. *Journal of Hydrology*, *320*(1–2), 155–172.

Young, P., Parkinson, S., & Lees, M. J. (1996). Simplicity out of complexity in environmental modelling: Occam's razor revisited. *Journal of Applied Statistics*, *23*(2–3), 165–210.

Zhang, Y. Q., & Chiew, F. H. S. (2009). Relative merits of different methods for runoff predictions in ungauged catchments. *Water Resources Research*, *45*, W07412. https://doi.org/10.1029/2008WR007504