

# Hydrologic Similarity Based on Width Function and Hypsometry: An Unsupervised Learning Approach

Prashanta Bajracharya (✉ [prashanta.bajracharya@maine.edu](mailto:prashanta.bajracharya@maine.edu))

University of Maine System <https://orcid.org/0000-0002-8262-4567>

Shaleen Jain

University of Maine System

---

## Research Article

**Keywords:** Width function, Hypsometric curve, Hydrologic response, Hydrologic similarity, Hierarchical clustering, Divergence measures

**Posted Date:** June 24th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-614652/v1>



**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Computers & Geosciences on March 1st, 2022. See the published version at <https://doi.org/10.1016/j.cageo.2022.105097>.

1 **Hydrologic similarity based on width function and**  
2 **hypsoetry: An unsupervised learning approach**

3 Prashanta Bajracharya  · Shaleen Jain 

4  
5 Received: June 10, 2021/ Accepted: date

6 **Abstract** In ungauged or data-scarce watersheds, systematic analyses of a  
7 set of proximate watersheds (for example, selected based on locational prox-  
8 imity or similarity in climate, morphometry, lithology, soils, and vegetation)  
9 have been shown to lend significant insights regarding hydrologic response and  
10 prediction. Current approaches often rely on: (a) statistical regression mod-  
11 els that use measurable watershed attributes, such as area, slope, and stream  
12 length; and (b) comparative hydrology that considers watershed characteris-  
13 tics to assess hydrologic similarity to select analogous gauged watersheds as  
14 proxies. Newer conceptions regarding hydrologic similarity focus on hydrologic  
15 response and therefore emphasize the use of dynamical measures of the stream  
16 network and watershed terrain. For example, the width function and hypso-  
17 metric curve can be readily estimated using the available global digital terrain  
18 datasets and represented as functional forms involving a small set of param-  
19 eters, thus achieving significant data reduction. In this study, a new approach to  
20 hydrological similarity in watersheds, one that utilizes these functional forms  
21 to identify dynamically similar watersheds, is presented. Dissimilarity matri-  
22 ces are created based on divergence measures, and watersheds are classified  
23 using hierarchical clustering. The joint analysis of watershed width functions  
24 and hypsometric curves allows for the classification of watersheds into a re-  
25 duced number of dynamically-similar groups. An illustrative case study for the  
26 Narmada River, with 72 sub-watersheds, is presented.

---

Prashanta Bajracharya  
Department of Civil and Environmental Engineering, University of Maine, Orono, Maine  
04469-5711 USA  
E-mail: prashanta.bajracharya@maine.edu

Shaleen Jain  
Department of Civil and Environmental Engineering, University of Maine, Orono, Maine  
04469-5711 USA  
Climate Change Institute, University of Maine, Orono, Maine 04469-5711 USA  
E-mail: shaleen.jain@maine.edu

1 **Keywords** Width function · Hypsometric curve · Hydrologic response ·  
2 Hydrologic similarity · Hierarchical clustering · Divergence measures

### 3 1 Introduction

4 Flow regimes and hydrologic response in watersheds are tied to the spatial  
5 pattern and distribution of a number of biogeophysical variables, including  
6 but not limited to topography, soil, vegetation, and built structures. In wa-  
7 tersheds where streamflow data is available, hydrologic models can be readily  
8 deployed and calibrated for the purposes of hydrologic prediction. However,  
9 in ungauged or data-scarce watersheds, current approaches to flow estima-  
10 tion rely on: (a) statistical regression models that use measurable watershed  
11 attributes, such as area, slope, and stream length; and (b) comparative hydro-  
12 logy that considers watershed characteristics to assess hydrologic similarity to  
13 select analogous gauged watersheds as proxies. Newer conceptions regarding  
14 hydrologic similarity focus on hydrologic response and therefore emphasize  
15 the use of dynamical measures of the stream network and watershed terrain  
16 (Bajracharya and Jain, 2020, 2021). Dynamical measures—width function and  
17 hypsometric curve—can be readily estimated using the available global digital  
18 terrain datasets. The computational burden, while significant, can be reduced  
19 by functional estimation and machine learning approaches (Bajracharya and  
20 Jain, 2020, 2021).

21 In hydrological sciences, machine learning has been used in applications  
22 such as precipitation analysis (Sun and Tang, 2020), rainfall-runoff processes  
23 (Hsu et al., 1995; Minns and Hall, 1996; Dawson and Wilby, 1998; Abrahart  
24 and See, 2000; Duan et al., 2020; Opiel and Mewes, 2020), ground water hy-  
25 drology (Karandish and Šimnek, 2016; Sahu et al., 2020), reservoir hydrology  
26 (Bai et al., 2016; Mital et al., 2020), hydraulic networks (Dibike et al., 1999),  
27 river basin management (Solomatine and Ostfeld, 2008), and flow mapping  
28 (Zhu and Guo, 2014). Applications to the problem of hydrologic prediction in  
29 ungauged or data-scarce environments presents an attractive opportunity to  
30 meld machine learning approaches with the knowledge of watershed dynamics.

31 In this study, we propose an approach that employs unsupervised classifi-  
32 cation to group similar basins based on distribution properties of hydrological  
33 basins. This provides a means for efficiently organizing a sea of data by sub-  
34 setting it into a smaller fraction of similar basins based on relevant physical  
35 characteristics that can then be further analyzed at a finer detail. We used  
36 the width function as a metric since it is a building block of the geomorpho-  
37 logical instantaneous unit hydrograph concept (Gupta and Waymire, 1983;  
38 Mesa and Mifflin, 1986; Bras, 1990), along with a hypsometric function to  
39 incorporate elevational information to complement the areal stream network  
40 topology encapsulated by the former. As a result, the clusters based on width  
41 functions and hypsometric curves on their own provide hydrological analogues  
42 based on unit rainfall response characteristics and elevation distribution, re-  
43 spectively, while a bivariate grouping can provide a synergistic combination of

1 the streamflow path characteristic alongside elevational profiles. This approach  
2 enables a computationally efficient means for finding hydrological analogues  
3 that can have large-scale applications, including at national and global scales,  
4 with minimal expert supervision.

5 In what follows, we first briefly review some common approaches to similar-  
6 ity assessment. Next, we discuss the study area and the dataset used. We then  
7 discuss the background information about hierarchical clustering, before pre-  
8 senting our methodology. Next we illustrate the results of the width function-  
9 and the hypsometric function-based clustering.

## 10 **2 Background**

### 11 2.1 Common approaches to hydrological similarity

12 Comparative hydrology is an approach to the prediction in ungauged basins  
13 (PUB) that examines a large number of catchments to distinguish patterns of  
14 hydrological behavior using common catchment and climatic characteristics.  
15 While there is no universal basis for hydrological classification of catchments  
16 (Blöschl et al., 2013), they are self-organizing systems whose hydraulic be-  
17 havior result from adaptive geomorphological processes (Sivapalan, 2006) and  
18 there are discernible patterns that form the foundations for understanding  
19 their hydrological nature. In general, catchments can be considered hydrologi-  
20 cally similar if they have similar response to climatic variability (Blöschl et al.,  
21 2013). Proximity is a commonly used, reliable metric for determining similar  
22 catchments, however this measure is limited in that it does not allow for the  
23 use of catchments are not closer to each other (Patil and Stieglitz, 2012). Since  
24 climate strongly impacts catchment characteristics and hydrological behavior,  
25 the hydro-climatic region where a catchment is located provides another basis  
26 for catchment classification (Budyko et al., 1974; L’vovich, 1979; Abrahams,  
27 1984; Milly, 1994; Sankarasubramanian and Vogel, 2002; Woods, 2006; Yadav  
28 et al., 2007). Similarly, readily observable spatial patterns in the catchment  
29 structure that affect the temporal response can be used as signatures to de-  
30 termine possible co-evolution of basin dynamics (Blöschl et al., 2013), and can  
31 be utilized to transfer hydrological information from data-rich catchments to  
32 ungauged basins to predict physical phenomenon such as hydrologic response  
33 (Burn and Boorman, 1993; Tung et al., 1997; Aryal et al., 2002; McIntyre et al.,  
34 2005; Wagener et al., 2007; Reichl et al., 2009; Archfield and Vogel, 2010; Oudin  
35 et al., 2010; Patil and Stieglitz, 2011, 2012; Razavi and Coulibaly, 2013; Athira  
36 et al., 2016; Brunner et al., 2018). The mostly commonly used technique in-  
37 volves the transfer of lumped characteristics such as catchment shape and size,  
38 Strahler ratios, drainage density, average slope, etc. that are used to explain hy-  
39 drogeomorphological characteristics (Horton, 1932, 1945; Strahler, 1957; Bras,  
40 1990; Rodríguez-Iturbe and Rinaldo, 2001). An issue with this is the possibil-  
41 ity of the loss of information in simplifying complex catchment properties into  
42 a single number (Wooldridge and Kalma, 2001; Wagener and Wheeler, 2006;

1 Tetzlaff et al., 2009; Chang et al., 2014). Alternatively, distribution curves can  
 2 be used to assess hydrological similarity. Examples of this include the use of  
 3 the distribution of topographic index, height above nearest drainage, reduced  
 4 dissipation per unit length index (Loritz et al., 2019), the distribution of ripar-  
 5 ian and hillslope effects on streams, the riparian-area change along the stream  
 6 network (McGlynn and Seibert, 2003), the hypsometric curve (Booij et al.,  
 7 2007; Ssegane et al., 2012; Hailegeorgis et al., 2015; Bajracharya and Jain,  
 8 2021), and the width function (Moussa, 2008; Bajracharya and Jain, 2020).  
 9 Furthermore, various mathematical models that link catchment structure to  
 10 hydrological response based on underlying physics or statistical relationships  
 11 have been used to explore catchment similarity and to develop similarity pa-  
 12 rameters (Hebson and Wood, 1982; Sivapalan et al., 1987; Larsen et al., 1994;  
 13 Milly, 1994; Reggiani et al., 2000; Aryal et al., 2002; Woods, 2003).

## 14 2.2 Dynamical representation of watershed morphometry

### 15 2.2.1 Width function

16 The width function represents the travel distance distribution of a stream  
 17 network (Lashermes and Foufoula-Georgiou, 2007). For a given drainage basin,  
 18 the width function,  $N(x)$ , denotes the areal extent between  $x$  and  $x+dx$ , where  
 19  $x$  represents the total distance along the flow path to the outlet (Veneziano  
 20 et al., 2000), termed here as the hydrological distance. As we do not distinguish  
 21 between the hillslope and channel network distance in this study, the width  
 22 function becomes synonymous with the area function. Under the assumption  
 23 of constant velocity, the width function represents the probability distribution  
 24 of travel times or the instantaneous unit hydrograph, reflecting the topological  
 25 features of a basin’s stream response (Lashermes and Foufoula-Georgiou, 2007;  
 26 Moussa, 2008). The width function is strongly linked to the peak and shape  
 27 of the hydrograph (Kirkby, 1976; Gupta and Waymire, 1983; Troutman and  
 28 Karlinger, 1984, 1989).

29 The width function is most commonly represented by a histogram with  
 30 the hydrological distance in the  $x$ -axis and the frequency or density of the  
 31 areal extent of streams in the  $y$ -axis (Figure 1). Bajracharya and Jain (2020)  
 32 demonstrated the use of a truncated skew-Normal ( $SN$ ) mixture model to  
 33 analytically represent the width function with the  $x$ -axis normalized by scal-  
 34 ing between 0 and 1, and demonstrated its utility in finding hydrologically  
 35 analogous drainage basins using divergence measures such as the  $L_2$  distance  
 36 (Tsybakov, 2008). The  $SN$  distribution is a three-parameter probability dis-  
 37 tribution formed by adding a skewness element to the Normal distribution.  
 38 For a continuous random variable,  $X$ , the  $SN$  distribution is represented as:

$$f(x; \xi, \omega^2, \alpha) = \frac{2}{\omega} \phi\left(\frac{x - \xi}{\omega}\right) \Phi\left(\alpha \frac{x - \xi}{\omega}\right), x \in (-\infty, \infty) \quad (1)$$

1 where  $\phi(x)$  denotes the standard Normal density function of  $x$ ,  $\Phi(x)$  denotes  
 2 the cumulative distribution function (*cdf*) of the standard Normal, and  $\xi$ ,  $\alpha$ ,  
 3 and  $\omega$  are the location, scale, and shape parameters, respectively. The domain  
 4 of the *SN* distribution is then truncated to  $[0, 1]$  using a correcting factor to  
 5 guarantee the validity of the normalization condition (Thomopoulos, 2017):

$$g(x) = \begin{cases} \frac{f(x)}{F(1)-F(0)}, & x \in [0, 1] \\ 0, & x \in (-\infty, 0) \cup (1, \infty) \end{cases} \quad (2)$$

6 where  $F(x)$  denotes the cumulative density function. Finally, a finite mixture  
 7 model of  $n$  truncated *SN* distributions is represented as:

$$h(x) = \sum_{i=1}^n w_i g(x; \xi_i, \omega_i^2, \alpha_i) \quad (3)$$

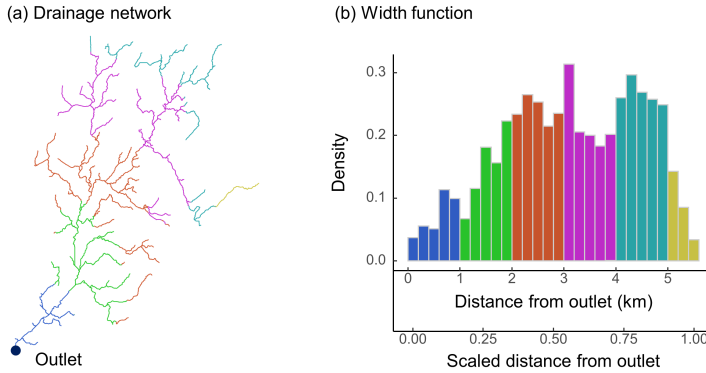
8 where  $w_i$  denote the non-negative mixing proportions that sum to one. Fur-  
 9 thermore, the  $L_2$  distance used by Bajracharya and Jain (2020) to measure  
 10 similarity between two width functions is computed as:

$$L_1 = \sqrt{\int (N_1 - N_2)^2 dx} \quad (4)$$

11 where  $N_1$  and  $N_2$  represent the two width functions. A value of zero indicates  
 12 identical width functions, while larger values reflect a larger difference.

### 13 2.2.2 Hypsometric function

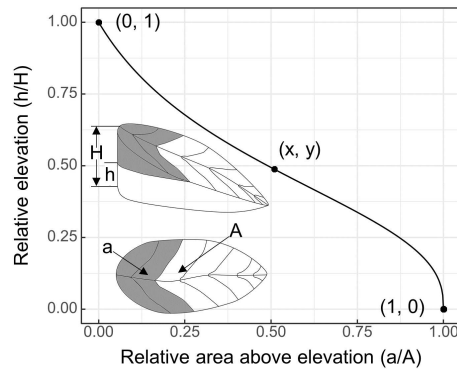
14 The hypsometric curve reflects the area-altitude distribution of a basin (Hor-  
 15 ton, 1932; Langbein, 1947) and reflects the distribution of landmass as a func-  
 16 tion of elevation (Harlin, 1984). The shape of the hypsometric curve indicates



**Fig. 1** (a) Drainage network, with color gradations based on flow path distances to the outlet denoting hydrological distances, and (b) width function with corresponding color gradation. The upper  $x$ -axis shows the hydrological distance in absolute units (km), while the lower  $x$ -axis presents the corresponding scaled hydrological distance.

1 the geomorphic maturity of catchments, with a concave up shape indicating  
 2 relatively mature basins with a high degree of erosive activity, and a concave  
 3 down shape indicating relatively young basins with a large proportion of  
 4 uneroded topography or creep-dominated hillslopes (Strahler, 1952; Moglen  
 5 and Bras, 1995; Pedrera et al., 2009; Willgoose, 2018). Furthermore, studies  
 6 have linked the hypsometric curve with various drainage basin features such  
 7 as the hydrograph time-to-peak, head-ward drainage development, regional  
 8 basin slopes (Harlin, 1984), average channel gradient (Howard, 1990), stream  
 9 network branching (Willgoose and Hancock, 1998), ground water interaction,  
 10 water table fluctuation (Marani et al., 2001), and surface and subsurface runoff  
 11 properties (Vivoni et al., 2008). Willgoose and Hancock (1998) further divided  
 12 the hypsometric curve into three regions: the 'head' (upper left-hand side),  
 13 the 'toe' (lower right-hand side), and the 'body' (between the head and the  
 14 toe), and linked the shape of the toe to stream branching characteristics of  
 15 the basin. Furthermore, hillslopes with active fluvial erosion and creep exhibit  
 16 concave down head and concave up tail (Willgoose, 2018). In long skinny catch-  
 17 ments and hillslopes with parallel flow lines, the hypsometric curve reflects the  
 18 hillslope long profile and can be used as an indirect test of the slope-area re-  
 19 lationship, while in more rounded catchments, the stream network branching  
 20 also affects the shape of the hypsometric curve (Willgoose, 2018).

21 The hypsometric curve can be plotted in absolute units, with elevation  
 22 in meters and area in square kilometers, or in relative units, with relative  
 23 elevation plotted against the relative area above said elevation (Figure 2). The  
 24 latter, termed as the percentage hypsometric curve, allows for the comparison  
 25 of basins of different altitudes and sizes (Strahler, 1952).



**Fig. 2** The scaled hypsometric curve showing the relative elevation plotted in relative to the proportion of area above this elevation.

26 Various functional forms have been developed to represent the hypsometric  
 27 curve (Strahler, 1952; Harlin, 1978; Sarkar and Patel, 2011; Vanderwaal and  
 28 Ssegane, 2013; Bajracharya and Jain, 2021). Bajracharya and Jain (2021) de-  
 29 veloped a three-parameter model named the Generalized Hypsometric function

1 by modifying the equation developed by Strahler (1952). The model places an  
 2 emphasis on the curvatures of the head, body, and the toe. The function is  
 3 defined as:

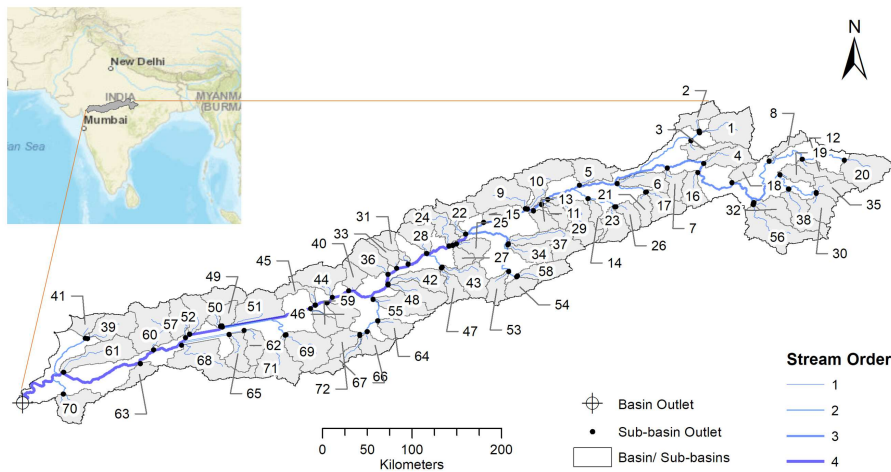
$$y = \left( \frac{1 - x^m}{1 + \beta x^m} \right)^z \quad (5)$$

4  
 5 where  $\beta$ ,  $z$ , and  $m$  denote the three parameters. Furthermore, Bajracharya  
 6 and Jain (2021) illustrated the use of hypsometry to find analogous basins  
 7 using the discordance index (DI), defined as the total absolute area between  
 8 two hypsometric curves.

### 9 3 Data and methods

#### 10 3.1 Case study

11 The Narmada River basin (NRB) is located in central India between latitudes  
 12  $21^{\circ}22' 0''$  N and  $23^{\circ}46' 30''$  N, and longitudes  $73^{\circ}4' 0''$  E and  $81^{\circ}45' 30''$  E.  
 13 The drainage area is  $95,000 \text{ km}^2$  (Figure 3). The elevation ranges from nearly  
 14 zero to over 1000 m above sea level, with an average slope of  $1.1^{\circ}$ . The basin  
 15 is bounded on the north, east, and south by hills, and on the west by the  
 16 Arabian sea. The lower middle reaches are comprised of fertile plain lands.  
 17 A number of reservoirs have been constructed in the basin for a variety of  
 18 purposes including water supply, irrigation, and hydropower generation. The  
 19 Narmada River passes through three states that face water shortages during  
 20 non-monsoon seasons (Ray and Goel, 2019).



**Fig. 3** Map of Narmada river basin and its location. Delineated sub-basins are shown along with their identifier ids.



1 The elevation data for the region was obtained from GTOPO30, a global  
2 digital elevation model (DEM) developed by the United States Geological  
3 Survey (USGS). It was derived from several raster and vector sources of to-  
4 pographic information (USGS, 1996). The dataset has a spatial resolution of  
5 30-arc seconds and a vertical accuracy of around 30 m. It is based on several  
6 sources of elevation information, including various vector and raster datasets,  
7 merged together, with a priority given to the data with a greater topographic  
8 detail and accuracy. With extensive accuracy checks, GTOPO30 data are suit-  
9 able for numerous regional and continental applications, including the extrac-  
10 tion of drainage features for hydrologic modeling (USGS, 1996).

11 The stream network was derived from the DEM in ESRI ArcGIS 10.5.1  
12 through standard Geographic Information System (GIS) procedures. First el-  
13 evation grids with undefined drainage directions, known as sinks, were filled;  
14 then the flow direction was determined based on the direction of steepest  
15 descent; followed by the computation of accumulated flow at each grid. A  
16 threshold of 396 km<sup>2</sup> was used to delineate the stream grids. This threshold  
17 was chosen to ensure a dense stream network, resulting in fourth order streams.  
18 This allowed for a delineation a considerable number of sub-basins to test the  
19 fits for diverse width function and hypsometric curve shapes. Finally, outlets  
20 were places at the confluences of first order and higher order streams to create  
21 72 non-overlapping sub-basins.

## 22 3.2 Methodology

23 Clustering is a descriptive unsupervised data mining technique for creating  
24 subsets by grouping similar data together based on some measure of similarity  
25 or dissimilarity (Veyssieres and Plant, 1998; Rokach and Maimon, 2005). The  
26 clustering structure is represented by a set of subsets,  $C = C_1, \dots, C_k$  of  $S$ ,  
27 such that  $S = \bigcup_{i=1}^k C_i$  and  $C_i \cap C_j = \emptyset$  for  $i \neq j$ . Hierarchical clustering is a  
28 clustering method that creates clusters by recursive partitioning, resulting in  
29 a dendrogram structure that represents the nested grouping of instances and  
30 similarity levels at which the groupings change. The recursive algorithm could  
31 be bottom-up, starting from every element in their individual cluster, with  
32 similar elements then grouped into a single cluster in each successive step  
33 (agglomerative clustering); or top-down, starting from all elements grouped  
34 in a single cluster, followed by the most dissimilar elements being separated  
35 into another cluster at each iteration (divisive hierarchical clustering). Vari-  
36 ous methods have been developed based on the manner in which the similarity  
37 measure is calculated and optimized, most of which are variants of single-link,  
38 complete-link, and minimum-variance algorithms (Jain et al., 1999). These al-  
39 gorithms consider the distance between two clusters to be equal to the shortest,  
40 longest, and average distance between a member of one cluster to a member  
41 of the other, respectively. Single-link methods are more versatile (Rokach and  
42 Maimon, 2005) but are susceptible to the "chaining effect", where a few points  
43 that form a bridge between two poorly separated, but distinct clusters lead to

1 them being merged at an early stage (Guha et al., 1998). On the other hand,  
 2 complete-link methods usually produce more compact clusters (Rokach and  
 3 Maimon, 2005). On the other hand, average-link clusters may cause the split-  
 4 ting of elongated clusters and the merging of portions of neighboring elongated  
 5 clusters (Guha et al., 1998).

6 In this study, we used the "agnes" function (Kaufman and Rousseeuw,  
 7 2009) from the "cluster" package (Maechler et al., 2021) in *R* programming  
 8 language (R Core Team, 2019) for the clustering analysis. This function pro-  
 9 vides the agglomerative coefficient (*ac*) which measure the amount of cluster-  
 10 ing structure. For a set of observations, *ac* is the average of  $1 - m(i)$ , where  
 11  $m(i)$  is the ratio of dissimilarity of each observation, *i*, to the first cluster it is  
 12 merged with to the dissimilarity of the final merger of the algorithm. *ac* varies  
 13 between zero and one, with larger values indicating more balanced cluster-  
 14 ing structures and values closer to zero indicating less well-formed structures.  
 15 For the given dataset, the Ward method (Ward, 1963), a type of minimum-  
 16 variance algorithm, was found to have a better *ac* value compared to the other  
 17 methods.

18 The width functions and hypsometric curves were first transformed to their  
 19 functional forms to facilitate efficient computation of dissimilarity matrices  
 20 (Figure 4). Width function clustering was done with the fitted *SN* functions,  
 21 using the  $L_2$  distance as the dissimilarity measure. This lead to width func-  
 22 tion analogues that share similarities in hydrological responses based on stream  
 23 network structures. Similarly, hypsometric clustering was done with the fitted  
 24 Generalized Hypsometric functions, using the *DI* as the dissimilarity measure.  
 25 These clusters are likely to share common hypsometric signatures in terms of  
 26 erosional/ depositional properties. While hypsometric curves are more closely  
 27 related to the erosional status of the basin, studies have indicated links between  
 28 hypsometric curves and hydrodynamic properties of basins (Harlin, 1984; Will-  
 29 goose and Hancock, 1998; Marani et al., 2001; Vivoni et al., 2008) due to the  
 30 topographic controls on stream generation and flow.

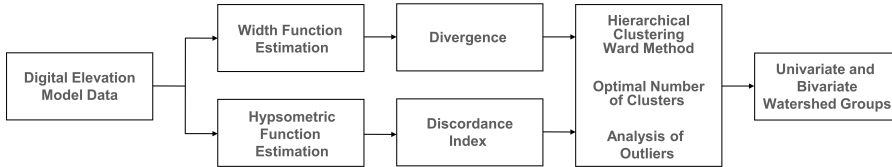


Fig. 4 Flowchart of the study methodology.

31 The gap statistic was used to determine the optimal number of clusters  
 32 (Tibshirani et al., 2001). For a dataset with  $k$  clusters based on distance mea-  
 33 sure  $d$ , the gap statistic is defined as

$$Gap_n(k) = E_n^*[\log(W_k)] - \log(W_k) \quad (6)$$

1  
2 where  $E_n^*$  represents the expected value for a sample size of  $n$  from the reference  
3 distribution and  $W_k$  is the pooled within-cluster sum of squares around the  
4 cluster means, defined as  $W_k = \sum_{r=1}^k \frac{1}{2n_r} \sum D_r$ . This statistic measures the  
5 deviation of the observed  $W_k$  from its expected value under the null hypothesis.  
6 The optimal number of clusters,  $\hat{k}$ , can be chosen based on various algorithms,  
7 including global maximum method, which maximizes  $Gap_n(k)$ , signifying the  
8 farthest deviation from uniform points distribution. Due to the lack of clear  
9 group demarcations in both width function and hypsometric curve shapes, we  
10 chose  $\hat{k}$  based on local maxima, where the increase in  $Gap_n(k)$  first tails off.  
11 There is a level of subjectivity in the choice of the number of clusters, with  
12 more groups leading to more homogeneity within the group members but a  
13 smaller number of members per group.

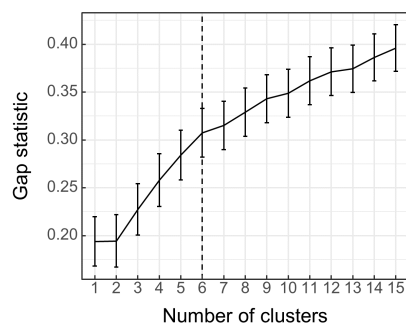
14 We also demonstrated the process of outlier detection to reduce intra-  
15 cluster variance with a simple algorithm based on similarity measures with  
16 the nearest neighbors. We used a minimum threshold approach where mem-  
17 bers exceeding a minimum similarity index with a selected number of nearest  
18 neighbors were classified as outliers and removed from the study. However,  
19 care was taken not to omit members with important and distinct physical  
20 characteristics. Finally, the sub-basins with common width function clusters  
21 and hypsometric function clusters were identified.

## 22 4 Watershed similarity

### 23 4.1 Width function clusters

#### 24 4.1.1 Hierarchical clustering

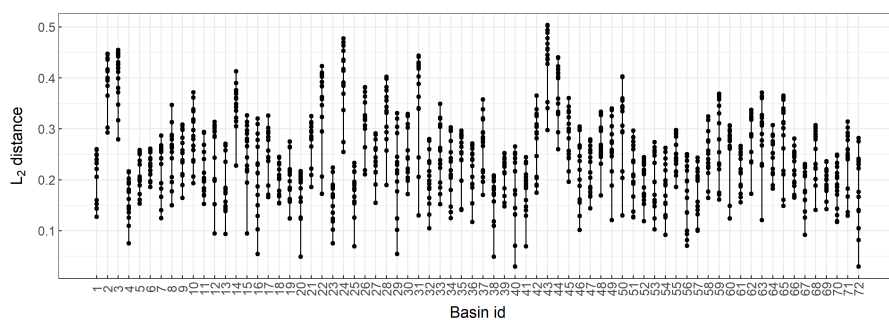
25 First, the optimal number of clusters was determined using the gap statistic.  
26 Figure 5 shows the gap statistic as a function of the number of clusters ( $k$ ). The  
27 graph shows that a larger number of clusters results in a higher gap statistic,  
28 and consequently, a better clustering. The continued increase in gap statistic  
29 with increasing number of clusters indicates that the different cluster regions  
30 are not sharply delineated. However, a large number of clusters impedes the  
31 interpretability of the width function shapes in each cluster. As such, the choice  
32 of optimal  $k$  involves some subjectivity. We based the choice on where the the  
33 rate of increase in the gap statistic first sharply decreases. The change in the  
34 gap statistic has a sharp decrease when  $k > 6$ , and as such, the optimal number  
35 of clusters for the width functions was chosen as six. The width functions  
36 in each cluster are shown in Figure S1. While there are some considerable  
37 variances in the width function shapes within each cluster, different clusters  
38 do exhibit noticeably different overall shapes.



**Fig. 5** Determination of the optimal number of width function clusters using gap statistic. The optimal number of clusters was chosen based on the change in the rate of increment of the gap statistic.

#### 1 4.1.2 Analysis of outliers

2 Outliers can cause chaining effects, leading to dissimilar objects being drawn  
 3 into the same cluster (Everitt et al., 2011). Removal of outliers can help reduce  
 4 intra-cluster variance. However, different outlier detection algorithms can lead  
 5 to different data points being classified as outliers. Moreover, outlier detection  
 6 can mistakenly classify small clusters as outliers and remove valuable information  
 7 from the data. Thus, outlier detection involves a degree of subjectivity.  
 8 Here we use a simple algorithm to analyze, detect, and remove outliers based  
 9 on similarity measures with nearest neighbors. Figure 6 shows the  $L_2$  distance  
 10 to fifteen closest neighbors for each width function. Based on this measure,  
 11 a threshold can be chosen subjectively to delineate outliers based on specific  
 12 goals. In this study, width functions with the  $L_2$  distance greater than 0.45  
 13 for up to 15 closest neighbors were marked as outliers. This led to only three  
 14 width functions being classified as outliers. Intra-cluster uniformity can be fur-  
 15 ther improved by lowering this threshold. While rigorous methods for removal  
 16 of outliers exist in the literature (Almeida et al., 2007; Fan et al., 2013; Krleža

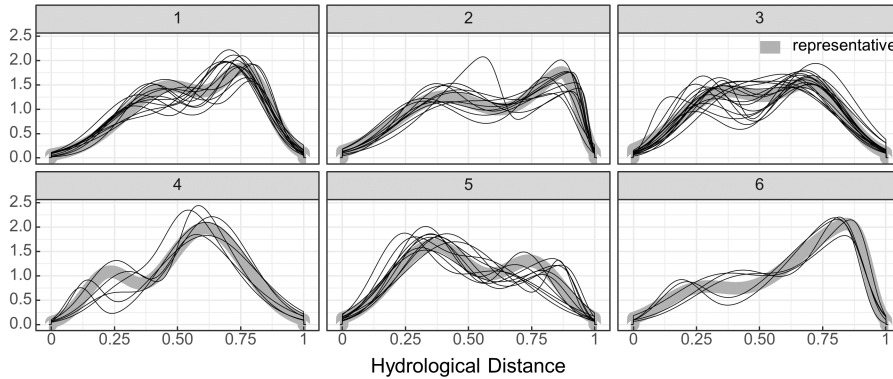


**Fig. 6**  $L_2$  distances between 15 closest neighbors for each width function.

1 et al., 2021), we employed this basic outlier detection algorithm as a proof of  
 2 concept, one that is easy to understand and can be readily applied.

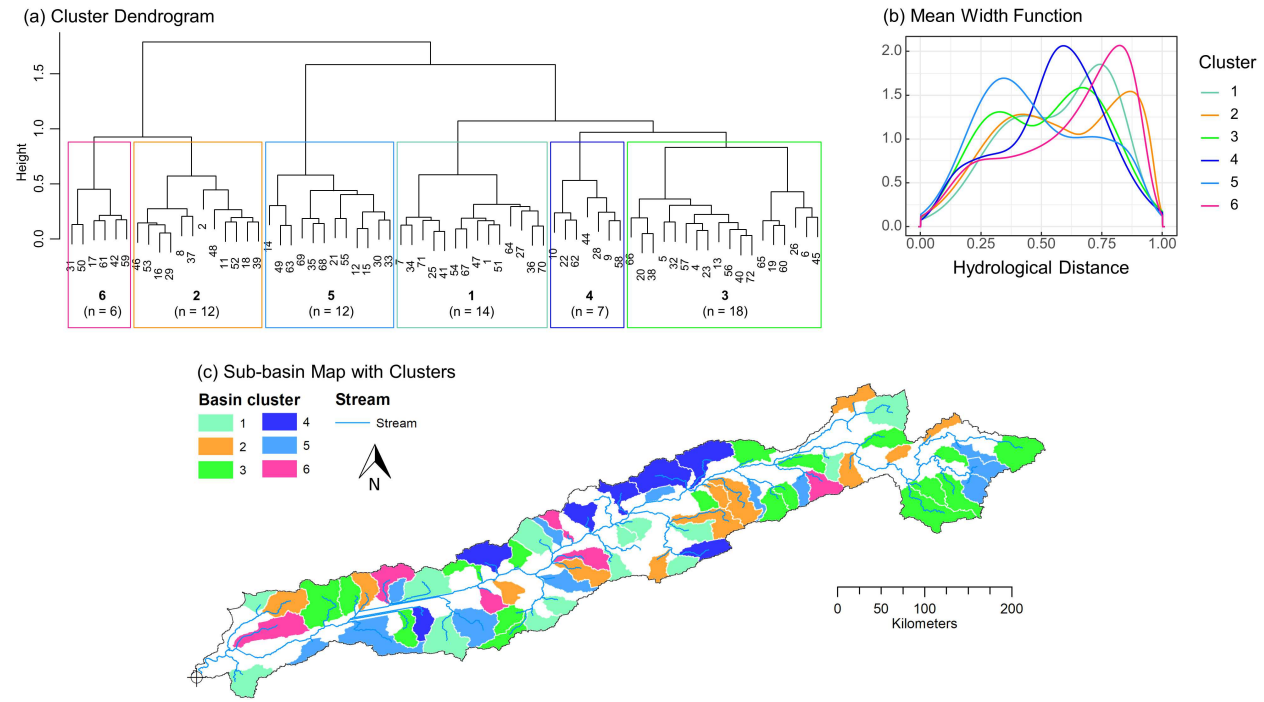
### 3 4.1.3 Analysis of clusters

4 After the removal of the outliers, the width functions were reclassified into  
 5 six clusters (Figure 7). With a removal of only three outliers, there is minimal  
 6 improvements in intra-cluster uniformity, as seen by the removal of two notable  
 7 outliers in cluster 3. To closely examine the properties of each cluster group,  
 8 representative width functions in each cluster have been highlighted in Figure  
 9 7. Representative width functions were chosen based on the lowest  $L_2$  distances  
 10 with the mean width functions within each cluster. Mean width functions were  
 11 calculated by averaging  $y$  values between all members of a given cluster at  
 12 each  $x$  value. Cluster 5 has a slightly higher peak in the first  $SN$  component,  
 13 while all other clusters have higher peaks in the second  $SN$  component, which  
 14 could indicate a difference in hydrograph peak locations. Among them, cluster  
 15 3 does not have a prominent peak, whereas cluster 6 has a prominent peak  
 16 towards the right end of the width function. Furthermore, the shape of the left  
 17 rising side and the right falling side of the curves differ between clusters. For  
 18 instance, the right side of the curves for clusters 2 and 6 are steeper compared  
 19 to other clusters. It should be noted that while the overall shape of the curves  
 20 are similar within clusters, there is still a considerable degree of heterogeneity  
 21 in the size and location of the peaks.



**Fig. 7** Width functions in each cluster after removing the outliers. The representative width functions for each cluster are shown as thick grey lines.

22 Hierarchical clustering can be best denoted using dendrograms. The den-  
 23 drogram notation of the width function clusters are shown in Figure 8, along  
 24 with the mean width functions and the location of the sub-basins. Figure 8  
 25 (b) further highlights the diversity in the shape of the width functions in



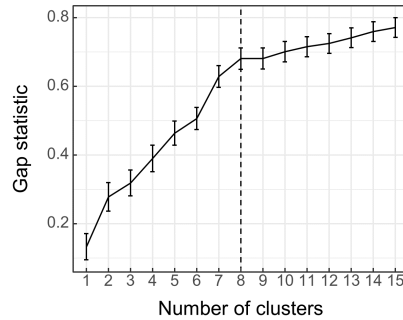
**Fig. 8** (a) Dendrogram of watershed width functions using hierarchical clustering using Ward's method. (b) Mean width functions for each cluster. (c) Map of sub-basins grouped by cluster.

1 each cluster. The width function shapes seem mostly independent of the loca-  
 2 tion of the sub-basin along the watershed as well as the sub-basin areas (Figure  
 3 8 (c)). Cluster 4 seems to be concentrated at the mid region of the watershed  
 4 and cluster 6 seems to be concentrated mostly in the bottom half, where as  
 5 all other clusters are spread across different regions. Interestingly, a number  
 6 of sub-basins within same cluster groups appear alongside each other.

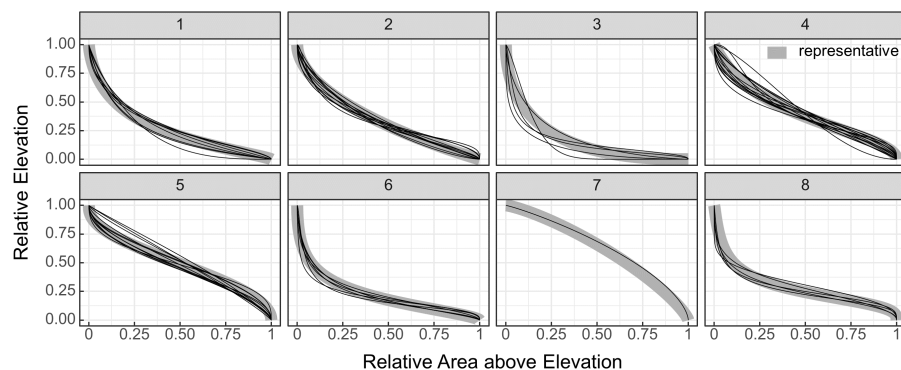
## 7 4.2 Hypsometric function clusters

### 8 4.2.1 Hierarchical clustering

9 Similar to width functions, the gap statistic was used to determine the optimal  
 10 number of hypsometric function clusters ( $k$ ) by evaluating the change in gap  
 11 statistic with the change in  $k$  (Figure 9). The change in the gap statistic sharply  
 12 decreases when  $k > 8$ , and hence, the optimal number of clusters was chosen  
 13 as eight. The classified hypsometric functions along with the corresponding  
 14 representative curves are shown in Figure 10. There is a clear distinction in  
 15 the shapes of the hypsometric curves in each cluster. Clusters 1, 2, 3, 6, and  
 16 8 comprise of concave up shapes, while cluster 7 comprises of concave down  
 17 shapes. The similarity in hypsometric curves could indicate similarity in ge-  
 18 omorphological characteristics within the clusters. Furthermore, hypsometric  
 19 curves in clusters 4, 5, and 8 have prominent tail regions following inflections  
 20 in the curve, where as other clusters lack prominent tail regions. However, it  
 21 is notable that there is some appreciable variability in the head and toe curva-  
 22 tures within each group. For instance, while the majority of curves in cluster  
 23 2 do not have an inflection point near the tail, there are a few curves with  
 24 prominent tail regions. On the other hand, some hypsometric curves with no  
 25 prominent tail curvatures are classified into clusters 4 and 5.



**Fig. 9** Determination of the optimal number of hypsometric function clusters using gap statistic. The optimal number of clusters was chosen based on the change in the rate of increment of the gap statistic.



**Fig. 10** Basin hypsometric curves. Corresponding representative curves for each cluster are shown as thick grey lines.

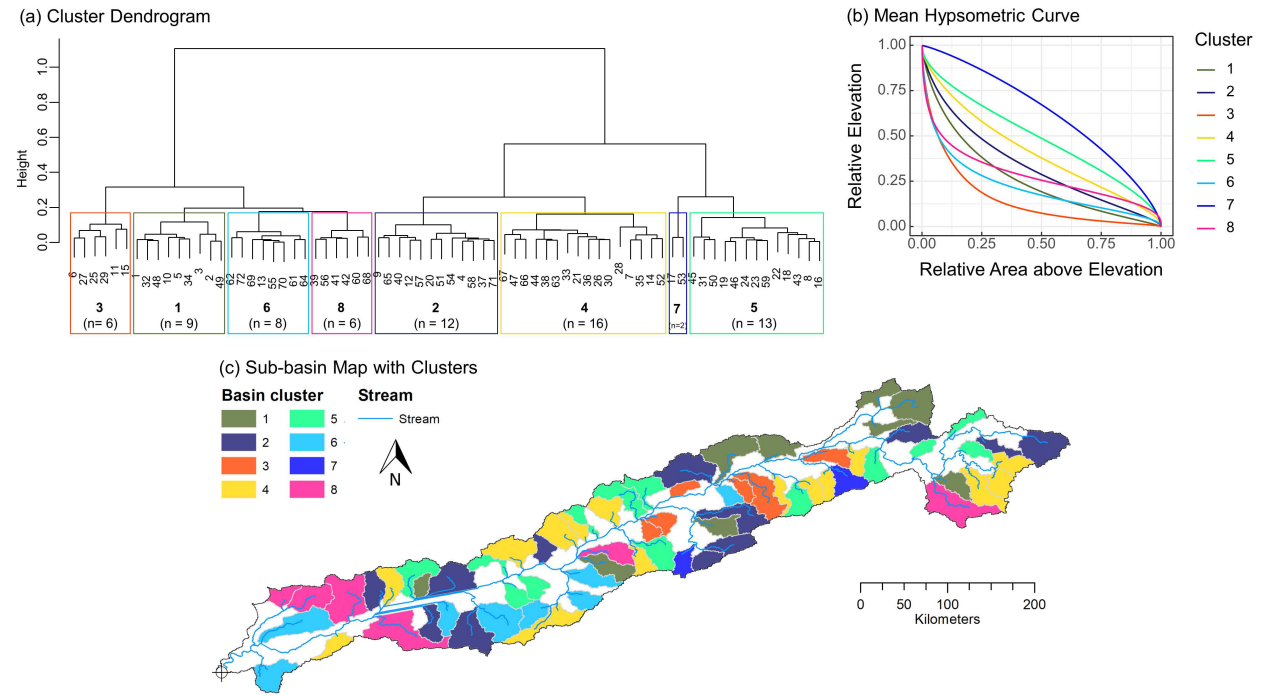
1 Similar outlier analysis algorithm was applied to these clusters, with a  
 2  $DI$  of 0.65 chosen as the threshold. However, this led to both sub-basins in  
 3 cluster 7 being classified as outliers. While this is computationally valid, cluster  
 4 7 is the only cluster comprised of concave down curves. Consequently, this  
 5 group carries an important geomorphological distinction as compared to other  
 6 clusters, and as such, should not be classified as an outlier or be removed  
 7 from the study. This indicates a shortcoming of the earlier outlier analysis  
 8 algorithm, and indicates that a degree of subjective choice may be necessary  
 9 in the outlier analysis so as to not omit important cluster groups. However,  
 10 lowering the number of nearest neighbors being considered to just one leads  
 11 to no member being classified as outliers. This matches visual inspection since  
 12 the intra-cluster variance in each group is already low. As a result of this, no  
 13 outlier was removed.

14 Cluster dendrograms are shown in Figure 11 (a), along with the mean hypso-  
 15 metric curves for each cluster group (Figure 11 (b)), and their locations (Figure  
 16 11 (c)). Mean hypsometric curves are computed by averaging the relative ele-  
 17 vations of each cluster member along the relative areas above the elevations.  
 18 The mean hypsometric curves indicate a gradual change from concave up to  
 19 concave down shapes along the clusters. There is no clear relationship between  
 20 the hypsometric curve shapes and the locations of the sub-basins along the wat-  
 21 ershed or the size of the sub-basins. Sub-basins in cluster 4 are concentrated  
 22 in the lower half of the watershed, while those in cluster 3 are concentrated  
 23 in the upper half. However sub-basins in other clusters are spread throughout  
 24 the watershed.

#### 25 4.3 Joint analysis of hierarchical clustering of width functions and 26 hypsometric curves

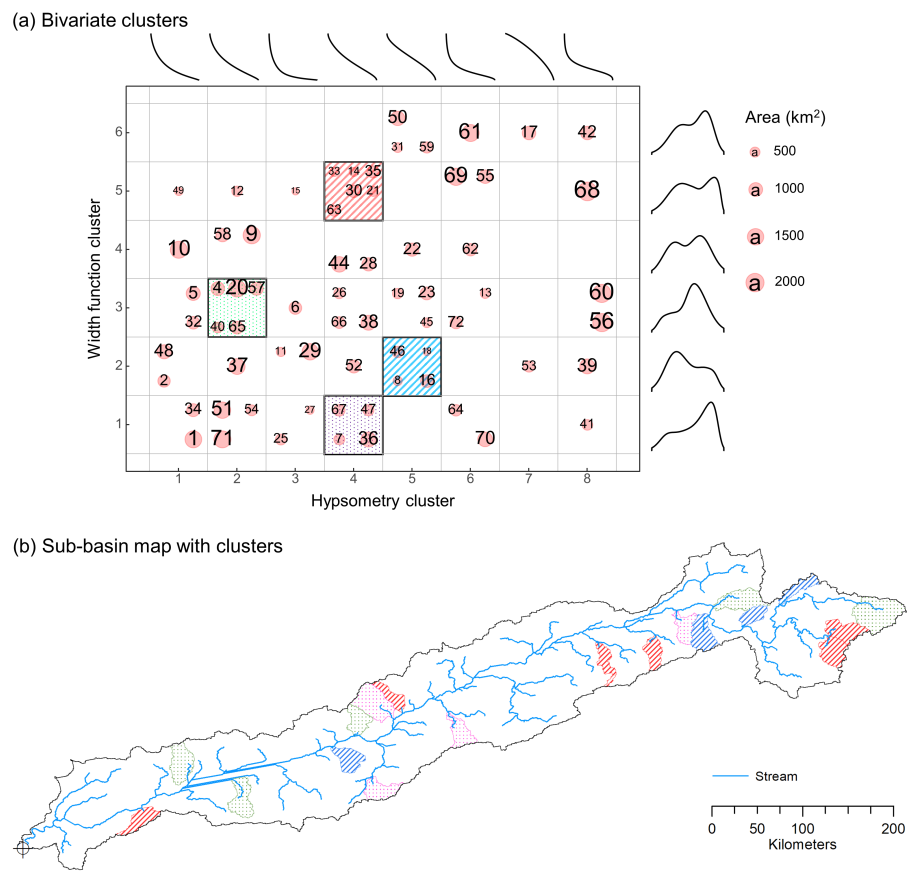
27 Next, we discuss the potential to combine the width function and the hyp-





**Fig. 11** (a) Dendrogram of basin hypsometric functions using hierarchical clustering using Ward's method. (b) Mean hypsometric functions for each cluster. (c) Map of sub-basins grouped by cluster.

1 sometric clustering to represent watershed analogs that take into account both,  
 2 the planar stream network geometry as well as the elevational characteristics  
 3 of the basin (Figure 12). This provides a framework for bivariate clustering  
 4 that incorporates multiple metrics that supplement each other. For instance,  
 5 sub-basins 14, 21, 30, 33, 35, and 60 fall in hypsometric cluster 4 and width  
 6 function cluster 5, with these members indicating mildly mature hypsometry  
 7 and width functions with the peak considerably skewed to the right. As  
 8 such, these sub-basins could potentially be analogues with similar hydrologi-  
 9 cal response properties. Sub-basins 17 and 53 have concave-down hypsometric  
 10 curves (hypsometric cluster 7), but have considerably different width function  
 11 shapes (width function clusters 2 and 5), indicating that the hydrological re-  
 12 sponse behaviour of these two sub-basins might be considerably different. As  
 13 such, width function and hypsometry can provide complementary properties,



**Fig. 12** (a) Bivariate cluster groups based on the width function and the hypsometric function. The mean curve for each cluster has been shown. (b) Map of sub-basins highlighting the bivariate groups with at least four members.

1 which results in a fuller description of basin processes. In Figure 12, we explore  
2 the spatial relationships between members in the bivariate groups. Group 5-2,  
3 with an early width function peak and a relatively linear hypsometric curve,  
4 is predominantly formed at the upstream region of the watershed. Group 4-5,  
5 with a highly steep falling limb of the width function and a relatively linear  
6 hypsometric curve, exhibited relatively smaller accumulation areas. However,  
7 in general, the spatial relationship within the highlighted bivariate groups was  
8 found to be weak.

## 9 **5 Discussion and concluding remarks**

10 New understandings and a sound physical basis for the prediction in ungauged  
11 basins has great theoretical and practical importance. To that end, this study  
12 provides an additional tool through the use of unsupervised learning and func-  
13 tional data reduction to derive dynamical measures of hydrologic response in  
14 watersheds. We demonstrated that the classification of basins through clus-  
15 tering when applied using dynamical measures of watershed behavior allows  
16 for the partitioning of watersheds into groups with consistent functional forms.  
17 We proposed a four-step approach for forming hydrologically similar analogues.  
18 This first step involves the functional estimation of two dynamic features, the  
19 width function and the hypsometric curve. Next, divergence measures are ap-  
20 plied across all basin pairs to form dissimilarity matrices, which are then used  
21 for hierarchical clustering. The clusters based on width functions and hypsomet-  
22 ries on their own provide groups of basins with similar drainage topology and  
23 elevation distribution, respectively. Finally, groups of basins with common  
24 width function and hypsometric function clusters serve as analogous basins  
25 with similar hydrological response characteristics. With the wide availability  
26 of terrain information, this method can be applied at large scales (national or  
27 global) to find a sizeable number of similar hydrological basins at low data and  
28 computational costs. This allows for a large number of catchments to be in-  
29 cluded in the donor pool and thus, provides a means for the statistical analysis  
30 of uncertainty in the hydrological signatures being transferred.

31 Our study illustrated this framework in the context of the Narmada River  
32 basin in India. The following observations and takeaways can be made about  
33 the hydrologically similar characteristics across the 72 selected sub-basins for  
34 the Narmada River:

- 35 1. The majority of width functions exhibit late peaks, with only one out  
36 of the six clusters exhibiting an early peak. Interestingly, the early peak  
37 cluster seems slightly concentrated near the outlet. Furthermore, a number  
38 of sub-basins that share cluster groups appear adjacent to each other.
- 39 2. The majority of sub-basins exhibited concave up hypsometric curves, with  
40 only two sub-basins showing concave down curves. This could indicate that  
41 these sub-basins are surface runoff-dominant and highly eroded. While the  
42 hierarchical clustering approach performed well in classifying the overall

1 concavity of the curves, it was slightly less effective in classifying the head  
2 and the tail curvatures.

- 3 3. There is a level of subjectivity in the choice of the number of clusters. The  
4 considerable degree of intra-cluster heterogeneity in the location of the  
5 peaks of width functions indicates the need for a relatively large number of  
6 clusters for width functions if a high degree of homogeneity is desired. On  
7 the other hand, a relatively lower number of hypsometric clusters might be  
8 sufficient due to the cumulative nature of the curve which tends to offer a  
9 lower variance.
- 10 4. Two bivariate groups with similar width functions and hypsometric func-  
11 tions were identified with at least four members, one was identified with  
12 five members, and one with six members from a total sample size of 72.  
13 These represent sub-basins with similar hydrological response character-  
14 istics. This can easily be scaled to thousands of watersheds around the  
15 world.

16 The lack of a definitive spatial pattern could indicate that spatial prox-  
17 imity alone might not be a strong predictor of basin hydrological response,  
18 especially at the chosen scale. The presence of pairs of sub-basins with similar  
19 width functions do indicate some spatial dependence. While spatial pattern is  
20 justifiably a good metric of hydrological similarity in most use cases, dynamic  
21 metrics such as the width function can serve as another strong measure in  
22 defining analogues.

23 Modern data collection techniques such as satellite hydrology and crowd-  
24 sourcing tools have led to an explosion in data volume. The future of water sci-  
25 ences hinges on our ability to harness this big data to understand hydrological  
26 phenomena based on smart, data-driven computational techniques (Peters-  
27 Lidard et al., 2017; Sit et al., 2020). Our approach focuses on the efficient  
28 use of large volumes of elevation data to find hydrological analogues through  
29 dynamical properties of terrains and facilitates large scale applications. This  
30 approach is consistent with the growing recognition in the hydrological com-  
31 munity regarding the use of explainable AI (XAI) techniques that build upon  
32 conceptual and machine learning models to explain hydrological phenomenon  
33 (Maksymiuk et al., 2020; Althoff et al., 2021). An application of hydrological  
34 similarity study is to assist in improving our understanding of hydrological  
35 processes in watersheds (Blöschl et al., 2013) and future works can build upon  
36 this study by integrating the width function and elevation-based slope and  
37 velocity distribution to create a robust dynamical metric for hydrological re-  
38 sponse quantification and similarity assessment.

### 39 **Declarations**

### 40 **Funding**

41 Not applicable

### 1 **Conflicts of interest/Competing interests**

2 The authors declare that they have no conflict of interest.

### 3 **Availability of data and material**

4 Not applicable

### 5 **Code availability**

6 Not applicable

### 7 **References**

- 8 Abrahams, A.D., 1984. Channel networks: a geomorphological perspective.  
9 *Water resources research* 20, 161–188.
- 10 Abrahart, R.J., See, L., 2000. Comparing neural network and autoregressive  
11 moving average techniques for the provision of continuous river flow forecasts  
12 in two contrasting catchments. *Hydrological processes* 14, 2157–2172.
- 13 Almeida, J., Barbosa, L., Pais, A., Formosinho, S., 2007. Improving hierar-  
14 chical cluster analysis: A new method with outlier detection and automatic  
15 clustering. *Chemometrics and Intelligent Laboratory Systems* 87, 208–217.
- 16 Althoff, D., Bazame, H.C., Nascimento, J.G., 2021. Untangling hybrid hydro-  
17 logical models with explainable artificial intelligence. *H2Open Journal* 4,  
18 13–28.
- 19 Archfield, S.A., Vogel, R.M., 2010. Map correlation method: Selection of a  
20 reference streamgage to estimate daily streamflow at ungauged catchments.  
21 *Water Resources Research* 46. doi:10.1029/2009WR008481.
- 22 Aryal, S.K., O’Loughlin, E.M., Mein, R.G., 2002. A similarity approach to  
23 predict landscape saturation in catchments. *Water Resources Research* 38,  
24 26–1.
- 25 Athira, P., Sudheer, K., Cibin, R., Chaubey, I., 2016. Predictions in ungauged  
26 basins: An approach for regionalization of hydrological models considering  
27 the probability distribution of model parameters. *Stochastic Environmental*  
28 *Research and Risk Assessment* 30, 1131–1149.
- 29 Bai, Y., Chen, Z., Xie, J., Li, C., 2016. Daily reservoir inflow forecasting using  
30 multiscale deep feature learning with hybrid models. *Journal of Hydrology*  
31 532, 193–206.
- 32 Bajracharya, P., Jain, S., 2020. Estimation of watershed width function: A  
33 statistical approach using LiDAR data. *Stochastic Environmental Research*  
34 *and Risk Assessment* , 1–15.
- 35 Bajracharya, P., Jain, S., 2021. Characterization of drainage basin hypsometry:  
36 A generalized approach. *Geomorphology* , 107645.

- 1 Blöschl, G., Bloschl, G., Sivapalan, M., Wagener, T., Savenije, H., Viglione,  
2 A., 2013. Runoff prediction in ungauged basins: synthesis across processes,  
3 places and scales. Cambridge University Press.
- 4 Booiij, M.J., Deckers, D., Rientjes, T.H., Krol, M.S., 2007. Regionalization for  
5 uncertainty reduction in flows in ungauged basins. IAHS Publication 313,  
6 329.
- 7 Bras, R.L., 1990. Hydrology: An Introduction to Hydrologic Science. Addison  
8 Wesley Publishing Company.
- 9 Brunner, M.I., Furrer, R., Sikorska, A.E., Viviroli, D., Seibert, J., Favre, A.C.,  
10 2018. Synthetic design hydrographs for ungauged catchments: A comparison  
11 of regionalization methods. Stochastic Environmental Research and Risk  
12 Assessment 32, 1993–2023.
- 13 Budyko, M.I., Miller, D.H., Miller, D.H., 1974. Climate and life. volume 508.  
14 Academic press New York.
- 15 Burn, D.H., Boorman, D.B., 1993. Estimation of hydrological parameters at  
16 ungauged catchments. Journal of Hydrology 143, 429–454. doi:10.1016/  
17 0022-1694(93)90203-L.
- 18 Chang, H., Johnson, G., Hinkley, T., Jung, I.W., 2014. Spatial analysis of  
19 annual runoff ratios and their variability across the contiguous US. Journal  
20 of Hydrology 511, 387–402. doi:10.1016/j.jhydro1.2014.01.066.
- 21 Dawson, C.W., Wilby, R., 1998. An artificial neural network approach to  
22 rainfall-runoff modelling. Hydrological Sciences Journal 43, 47–66.
- 23 Dibike, Y.B., Solomatine, D., Abbott, M.B., 1999. On the encapsulation of  
24 numerical-hydraulic models in artificial neural network. Journal of Hy-  
25 draulic research 37, 147–161.
- 26 Duan, S., Ullrich, P., Shu, L., 2020. Using convolutional neural networks for  
27 streamflow projection in California. Front. Water 2: 28 doi:10.3389/frwa.  
28 2020.00028.
- 29 Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Cluster analysis 5th ed.
- 30 Fan, S.K.S., Huang, H.K., Chang, Y.J., 2013. Robust multivariate control  
31 chart for outlier detection using hierarchical cluster tree in SW2. Quality  
32 and Reliability Engineering International 29, 971–985.
- 33 Guha, S., Rastogi, R., Shim, K., 1998. Cure: An efficient clustering algorithm  
34 for large databases. ACM Sigmod record 27, 73–84.
- 35 Gupta, V.K., Waymire, E., 1983. On the formulation of an analytical ap-  
36 proach to hydrologic response and similarity at the basin scale. Journal of  
37 Hydrology 65, 95–123. doi:10.1016/0022-1694(83)90212-3.
- 38 Hailegeorgis, T.T., Abdella, Y.S., Alfredsen, K., Kolberg, S., 2015. Evalua-  
39 tion of regionalization methods for hourly continuous streamflow simulation  
40 using distributed models in boreal catchments. Journal of Hydrologic Engi-  
41 neering 20, 04015028.
- 42 Harlin, J.M., 1978. Statistical moments of the hypsometric curve and its  
43 density function. Journal of the International Association for Mathematical  
44 Geology 10, 59–72.
- 45 Harlin, J.M., 1984. Watershed morphometry and time to hydrograph peak.  
46 Journal of Hydrology 67, 141–154.

- 1 Hebson, C., Wood, E.F., 1982. A derived flood frequency distribution using  
2 horton order ratios. *Water Resources Research* 18, 1509–1518.
- 3 Horton, R.E., 1932. Drainage-basin characteristics. *Eos, Transactions of the*  
4 *American Geophysical Union* 13, 350–361. doi:10.1029/TR013i001p00350.
- 5 Horton, R.E., 1945. Erosional development of streams and their drainage  
6 basins; Hydrophysical approach to quantitative morphology. *Geological So-*  
7 *ciety of America Bulletin* 56, 275–370.
- 8 Howard, A.D., 1990. Role of hypsometry and planform in basin hydrologic  
9 response. *Hydrological Processes* 4, 373–385.
- 10 Hsu, K.I., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network mod-  
11 eling of the rainfall-runoff process. *Water resources research* 31, 2517–2530.
- 12 Jain, A., Murty, M., Flynn, P., 1999. Data clustering: A review. *ACM Com-*  
13 *puting Surveys* 31, 264–323.
- 14 Karandish, F., Šimnek, J., 2016. A comparison of numerical and machine-  
15 learning modeling of soil water content with limited input data. *Journal of*  
16 *Hydrology* 543, 892–909.
- 17 Kaufman, L., Rousseeuw, P.J., 2009. Finding groups in data: an introduction  
18 to cluster analysis. volume 344. John Wiley & Sons.
- 19 Kirkby, M., 1976. Tests of the random network model, and its application to  
20 basin hydrology. *Earth Surface Processes* 1, 197–212.
- 21 Krleža, D., Vrdoljak, B., Brčić, M., 2021. Statistical hierarchical clustering  
22 algorithm for outlier detection in evolving data streams. *Machine Learning*  
23 110, 139–184.
- 24 Langbein, W.B., 1947. Topographic characteristics of drainage basins. Tech-  
25 nical Report Water Supply Paper 968- C. United States Geological Survey.  
26 doi:10.3133/wsp968C.
- 27 Larsen, J.E., Sivapalan, M., Coles, N.A., Linnet, P.E., 1994. Similarity analysis  
28 of runoff generation processes in real-world catchments. *Water Resources*  
29 *Research* 30, 1641–1652.
- 30 Lashermes, B., Foufoula-Georgiou, E., 2007. Area and width functions of river  
31 networks: New results on multifractal properties. *Water Resources Research*  
32 43.
- 33 Loritz, R., Kleidon, A., Jackisch, C., Westhoff, M., Ehret, U., Gupta, H.,  
34 Zehe, E., 2019. A topographic index explaining hydrological similarity by  
35 accounting for the joint controls of runoff formation. *Hydrology and Earth*  
36 *System Sciences* 23, 3807–3821.
- 37 L’vovich, M.I., 1979. World water resources and their future. *American Geo-*  
38 *physical Union*.
- 39 Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2021.  
40 *Cluster: Cluster analysis basics and extensions*. URL: [https://CRAN.](https://CRAN.R-project.org/package=cluster)  
41 [R-project.org/package=cluster](https://CRAN.R-project.org/package=cluster). r package version 2.1.1 — For new fea-  
42 tures, see the ‘Changelog’ file (in the package source).
- 43 Maksymiuk, S., Gosiewska, A., Biecek, P., 2020. Landscape of R packages for  
44 explainable Artificial Intelligence. arXiv preprint arXiv:2009.13248 .
- 45 Marani, M., Eltahir, E., Rinaldo, A., 2001. Geomorphic controls on re-  
46 gional base flow. *Water Resources Research* 37, 2619–2630. doi:10.1029/

- 1 2000WR000119.
- 2 McGlynn, B.L., Seibert, J., 2003. Distributed assessment of contributing area  
3 and riparian buffering along stream networks. *Water resources research* 39.
- 4 McIntyre, N., Lee, H., Wheeler, H., Young, A., Wagener, T., 2005. Ensemble  
5 predictions of runoff in ungauged catchments. *Water Resources Research*  
6 41. doi:10.1029/2005WR004289.
- 7 Mesa, O.J., Mifflin, E.R., 1986. On the relative role of hillslope and network  
8 geometry in hydrologic response, in: *Scale problems in hydrology*. Springer,  
9 pp. 1–17.
- 10 Milly, P., 1994. Climate, interseasonal storage of soil water, and the annual  
11 water balance. *Advances in Water Resources* 17, 19–24.
- 12 Minns, A., Hall, M., 1996. Artificial neural networks as rainfall-runoff models.  
13 *Hydrological sciences journal* 41, 399–417.
- 14 Mital, U., Dwivedi, D., Brown, J.B., Faybishenko, B., Painter, S.L., Steefel,  
15 C.I., 2020. Sequential imputation of missing spatio-temporal precipitation  
16 data using random forests. *Front. Water* 2: 20 doi:10.3389/frwa.2020.  
17 00020.
- 18 Moglen, G.E., Bras, R.L., 1995. The importance of spatially heterogeneous ero-  
19 sivity and the cumulative area distribution within a basin evolution model.  
20 *Geomorphology* 12, 173–185.
- 21 Moussa, R., 2008. What controls the width function shape, and can it be  
22 used for channel network comparison and regionalization? *Water Resources*  
23 *Research* 44. doi:10.1029/2007WR006118.
- 24 Oppel, H., Mewes, B., 2020. On the automation of flood event separation from  
25 continuous time series .
- 26 Oudin, L., Kay, A., Andréassian, V., Perrin, C., 2010. Are seemingly physically  
27 similar catchments truly hydrologically similar? *Water Resources Research*  
28 46. doi:10.1029/2009WR008887.
- 29 Patil, S., Stieglitz, M., 2011. Hydrologic similarity among catchments under  
30 variable flow conditions. *Hydrology and Earth System Sciences* 15, 989–997.  
31 doi:10.5194/hess-15-989-2011.
- 32 Patil, S., Stieglitz, M., 2012. Controls on hydrologic similarity: Role of nearby  
33 gauged catchments for prediction at an ungauged catchment. *Hydrology*  
34 *and Earth System Sciences* 16, 551–562. doi:10.5194/hess-16-551-2012.
- 35 Pedrera, A., Pérez-Peña, J.V., Galindo-Zaldívar, J., Azañón, J.M., Azor, A.,  
36 2009. Testing the sensitivity of geomorphic indices in areas of low-rate active  
37 folding (eastern Betic Cordillera, Spain). *Geomorphology* 105, 218–231.
- 38 Peters-Lidard, C.D., Clark, M., Samaniego, L., Verhoest, N.E., Van Emmerik,  
39 T., Uijlenhoet, R., Achieng, K., Franz, T.E., Woods, R., 2017. Scaling,  
40 similarity, and the fourth paradigm for hydrology. *Hydrology and Earth*  
41 *System Sciences* 21, 3701–3713. doi:10.5194/hess-21-3701-2017.
- 42 R Core Team, 2019. *R: A Language and Environment for Statistical Com-*  
43 *puting*. R Foundation for Statistical Computing. Vienna, Austria. URL:  
44 <https://www.R-project.org>.
- 45 Ray, L.K., Goel, N.K., 2019. Flood frequency analysis of Narmada River basin  
46 in India under nonstationary condition. *Journal of Hydrologic Engineering*



24, 05019018.

- Razavi, T., Coulibaly, P., 2013. Streamflow prediction in ungauged basins: Review of regionalization methods. *Journal of Hydrologic Engineering* 18, 958–975.
- Reggiani, P., Sivapalan, M., Hassanizadeh, S.M., 2000. Conservation equations governing hillslope responses: Exploring the physical basis of water balance. *Water Resources Research* 36, 1845–1863.
- Reichl, J.P.C., Western, A.W., McIntyre, N.R., Chiew, F.H.S., 2009. Optimization of a similarity measure for estimating ungauged streamflow. *Water Resources Research* 45. doi:10.1029/2008WR007248.
- Rodríguez-Iturbe, I., Rinaldo, A., 2001. *Fractal river basins: Chance and self-organization*. Cambridge University Press.
- Rokach, L., Maimon, O., 2005. Clustering methods, in: *Data mining and knowledge discovery handbook*. Springer, pp. 321–352.
- Sahu, R.K., Müller, J., Park, J., Varadharajan, C., Arora, B., Faybishenko, B., Agarwal, D., 2020. Impact of input feature selection on groundwater level prediction from a multi-layer perceptron neural network. *Frontiers in Water* 2, 46.
- Sankarasubramanian, A., Vogel, R.M., 2002. Annual hydroclimatology of the united states. *Water Resources Research* 38, 19–1.
- Sarkar, A., Patel, P., 2011. Topographic analysis of the Dulung River Basin. *The Indian Journal of Spatial Science* 2, 19.
- Sit, M., Demiray, B.Z., Xiang, Z., Ewing, G.J., Sermet, Y., Demir, I., 2020. A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology* 82, 2635–2670.
- Sivapalan, M., 2006. Pattern, process and function: elements of a unified theory of hydrology at the catchment scale. *Encyclopedia of hydrological sciences* .
- Sivapalan, M., Beven, K., Wood, E.F., 1987. On hydrologic similarity: 2. a scaled model of storm runoff production. *Water Resources Research* 23, 2266–2278.
- Solomatine, D.P., Ostfeld, A., 2008. Data-driven modelling: Some past experiences and new approaches. *Journal of hydroinformatics* 10, 3–22.
- Ssegane, H., Tollner, E., Mohamoud, Y., Rasmussen, T., Dowd, J., 2012. Advances in variable selection methods II: Effect of variable selection method on classification of hydrologically similar watersheds in three Mid-Atlantic ecoregions. *Journal of Hydrology* 438, 26–38.
- Strahler, A.N., 1952. Hypsometric (area–altitude) analysis of erosional topography. *Geological Society of America Bulletin* 63, 1117–1142.
- Strahler, A.N., 1957. Quantitative analysis of watershed geomorphology. *Eos, Transactions American Geophysical Union* 38, 913–920.
- Sun, A., Tang, G., 2020. Downscaling satellite and reanalysis precipitation products using attention-based deep convolutional neural nets. *Front. Water* 2: 536743 doi:10.3389/frwa.2020.536743.
- Tetzlaff, D., Seibert, J., Soulsby, C., 2009. Inter-catchment comparison to assess the influence of topography and soils on catchment transit times in

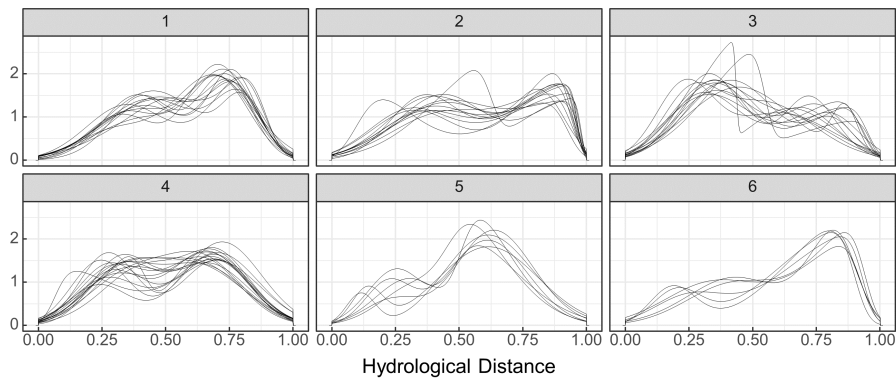
- 1 a geomorphic province; the Cairngorm mountains, Scotland. *Hydrological*  
2 *Processes* 23, 1874–1886. doi:10.1002/hyp.7318.
- 3 Thomopoulos, N.T., 2017. *Statistical distributions: Applications and Param-*  
4 *eter Estimates*. Springer.
- 5 Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clus-  
6 ters in a data set via the gap statistic. *Journal of the Royal Statistical*  
7 *Society: Series B (Statistical Methodology)* 63, 411–423.
- 8 Troutman, B.M., Karlinger, M.R., 1984. On the expected width function for  
9 topologically random channel networks. *Journal of Applied Probability* 21,  
10 836–849. doi:10.1017/S0021900200037530.
- 11 Troutman, B.M., Karlinger, M.R., 1989. Predictors of the peak width for  
12 networks with exponential links. *Stochastic Hydrology and Hydraulics* 3,  
13 1–16. doi:10.1007/BF01543424.
- 14 Tsybakov, A.B., 2008. *Introduction to nonparametric estimation*. Springer.
- 15 Tung, Y.K., Yeh, K.C., Yang, J.C., 1997. Regionalization of unit hydrograph  
16 parameters: 1. Comparison of regression analysis techniques. *Stochastic*  
17 *Hydrology and Hydraulics* 11, 145–171. doi:10.1007/BF02427913.
- 18 USGS, 1996. USGS EROS Archive – Digital Elevation – Global 30 Arc-Second  
19 Elevation (GTOPO30). <https://doi.org/10.5066/F7DF6PQS>. Accessed:  
20 2019-06-02.
- 21 Vanderwaal, J.A., Ssegane, H., 2013. Do polynomials adequately describe  
22 the hypsometry of Monadnock phase watersheds? *JAWRA Journal of the*  
23 *American Water Resources Association* 49, 1485–1495. doi:10.1111/jawr.  
24 12089.
- 25 Veneziano, D., Moglen, G.E., Furcolo, P., Iacobellis, V., 2000. Stochastic model  
26 of the width function. *Water Resources Research* 36, 1143–1157. doi:10.  
27 1029/2000WR900002.
- 28 Veysieres, M., Plant, R.E., 1998. Identification of vegetation state and transi-  
29 tion domains in California’s hardwood rangelands. University of California  
30 101.
- 31 Vivoni, E.R., Di Benedetto, F., Grimaldi, S., Eltahir, E.A., 2008. Hypsometric  
32 control on surface and subsurface runoff. *Water Resources Research* 44.
- 33 Wagener, T., Sivapalan, M., Troch, P., Woods, R., 2007. Catchment classifi-  
34 cation and hydrologic similarity. *Geography Compass* 1, 901–931. doi:10.  
35 1111/j.1749-8198.2007.00039.x.
- 36 Wagener, T., Wheeler, H.S., 2006. Parameter estimation and regionalization  
37 for continuous rainfall-runoff models including uncertainty. *Journal of hy-*  
38 *drology* 320, 132–154. doi:10.1016/j.jhydro.2005.07.015.
- 39 Ward, J.H., 1963. Hierarchical grouping to optimize an objective function.  
40 *Journal of the American Statistical Association* 58, 236–244.
- 41 Willgoose, G., 2018. *Principles of soilscape and landscape evolution*. Cam-  
42 bridge University Press.
- 43 Willgoose, G., Hancock, G., 1998. Revisiting the hypsometric curve as an  
44 indicator of form and process in transport-limited catchment. *Earth Sur-*  
45 *face Processes and Landforms: The Journal of the British Geomorphological*  
46 *Group* 23, 611–623.

- 
- 1 Woods, R., 2003. The relative roles of climate, soil, vegetation and topography  
2 in determining seasonal and long-term catchment dynamics. *Advances in*  
3 *Water Resources* 26, 295–309.
- 4 Woods, R., 2006. Global similarity indices for mean and seasonal hydrology  
5 of ungauged basins, in: Presentation at USA PUB Workshop.
- 6 Wooldridge, S.A., Kalma, J.D., 2001. Regional-scale hydrological modelling  
7 using multiple-parameter landscape zones and a quasi-distributed water  
8 balance model. *Hydrology and Earth System Sciences* 5. doi:10.5194/  
9 **hess-5-59-2001**.
- 10 Yadav, M., Wagener, T., Gupta, H., 2007. Regionalization of constraints on  
11 expected watershed response behavior for improved predictions in ungauged  
12 basins. *Advances in water resources* 30, 1756–1774.
- 13 Zhu, X., Guo, D., 2014. Mapping large spatial flow data with hierarchical  
14 clustering. *Transactions in GIS* 18, 421–435.

## 1 Supplementary materials

### 2 Original width function clusters

3  
4  
5 Figure S1 shows the width function clusters before the removal of outliers.  
6 Clusters 1, 5, and 6 have higher peaks in the right *SN* component while  
7 cluster 3 has a higher peak in the left *SN* component, potentially indicative of  
8 different location of peak flows in hydrographs. Furthermore, the high slopes  
9 on right sides of the curves for clusters 2 and 6 could be indicative of more  
10 rapidly falling recession limbs of hydrographs.



**Fig. S1** Width functions in each cluster.