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

Prashanta Bajracharya (prashanta.bajracharya@maine.edu)
University of Maine System

Shaleen Jain
University of Maine System

Research Article

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

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
Hydrologic similarity based on width function and hypsometry: An unsupervised learning approach

Prashanta Bajracharya · Shaleen Jain

Abstract In ungauged or data-scarce watersheds, systematic analyses of a set of proximate watersheds (for example, selected based on locational proximity or similarity in climate, morphometry, lithology, soils, and vegetation) have been shown to lend significant insights regarding hydrologic response and prediction. Current approaches often rely on: (a) statistical regression models that use measurable watershed attributes, such as area, slope, and stream length; and (b) comparative hydrology that considers watershed characteristics to assess hydrologic similarity to select analogous gauged watersheds as proxies. Newer conceptions regarding hydrologic similarity focus on hydrologic response and therefore emphasize the use of dynamical measures of the stream network and watershed terrain. For example, the width function and hypsometric curve can be readily estimated using the available global digital terrain datasets and represented as functional forms involving a small set of parameters, thus achieving significant data reduction. In this study, a new approach to hydrological similarity in watersheds, one that utilizes these functional forms to identify dynamically similar watersheds, is presented. Dissimilarity matrices are created based on divergence measures, and watersheds are classified using hierarchical clustering. The joint analysis of watershed width functions and hypsometric curves allows for the classification of watersheds into a reduced number of dynamically-similar groups. An illustrative case study for the Narmada River, with 72 sub-watersheds, is presented.
Keywords Width function · Hypsometric curve · Hydrologic response · Hydrologic similarity · Hierarchical clustering · Divergence measures

1 Introduction

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

In hydrological sciences, machine learning has been used in applications such as precipitation analysis [Sun and Tang, 2020], rainfall-runoff processes [Hsu et al., 1995; Minns and Hall, 1996; Dawson and Wilby, 1998; Abrahart and Sec, 2000; Duan et al., 2020; Oppel and Mewes, 2020], groundwater hydrology [Karandish and Simnek, 2016; Sahu et al., 2020], reservoir hydrology [Bai et al., 2016; Mital et al., 2020], hydraulic networks [Dibike et al., 1999], river basin management [Solomatine and Ostfeld, 2008], and flow mapping [Zhu and Guo, 2014]. Applications to the problem of hydrologic prediction in ungauged or data-scarce environments presents an attractive opportunity to meld machine learning approaches with the knowledge of watershed dynamics.

In this study, we propose an approach that employs unsupervised classification to group similar basins based on distribution properties of hydrological basins. This provides a means for efficiently organizing a sea of data by sub-setting it into a smaller fraction of similar basins based on relevant physical characteristics that can then be further analyzed at a finer detail. We used the width function as a metric since it is a building block of the geomorphological instantaneous unit hydrograph concept [Gupta and Waymire, 1983; Mesa and Mifflin, 1986; Bras, 1990], along with a hypsometric function to incorporate elevational information to complement the areal stream network topology encapsulated by the former. As a result, the clusters based on width functions and hypsometric curves on their own provide hydrological analogues based on unit rainfall response characteristics and elevation distribution, respectively, while a bivariate grouping can provide a synergistic combination of
the streamflow path characteristic alongside elevational profiles. This approach enables a computationally efficient means for finding hydrological analogues that can have large-scale applications, including at national and global scales, with minimal expert supervision.

In what follows, we first briefly review some common approaches to similarity assessment. Next, we discuss the study area and the dataset used. We then discuss the background information about hierarchical clustering, before presenting our methodology. Next we illustrate the results of the width function- and the hypsometric function-based clustering.

2 Background

2.1 Common approaches to hydrological similarity

Comparative hydrology is an approach to the prediction in ungauged basins (PUB) that examines a large number of catchments to distinguish patterns of hydrological behavior using common catchment and climatic characteristics. While there is no universal basis for hydrological classification of catchments (Bloschl et al., 2013), they are self-organizing systems whose hydraulic behavior result from adaptive geomorphological processes (Sivapalan, 2006) and there are discernible patterns that form the foundations for understanding their hydrological nature. In general, catchments can be considered hydrologically similar if they have similar response to climatic variability (Bloschl et al., 2013). Proximity is a commonly used, reliable metric for determining similar catchments, however this measure is limited in that it does not allow for the use of catchments not closer to each other (Patil and Stieglitz, 2012). Since climate strongly impacts catchment characteristics and hydrological behavior, the hydro-climatic region where a catchment is located provides another basis for catchment classification (Budyko et al., 1974; L’vovich, 1979; Abrahams, 1984; Milly, 1994; Sankarasubramanian and Vogel, 2002; Woods, 2006; Yadav et al., 2007). Similarly, readily observable spatial patterns in the catchment structure that affect the temporal response can be used as signatures to determine possible co-evolution of basin dynamics (Bloschl et al., 2013), and can be utilized to transfer hydrological information from data-rich catchments to ungauged basins to predict physical phenomenon such as hydrologic response (Burn and Boorman, 1993; Tung et al., 1997; Aryal et al., 2002; McIntyre et al., 2005; Wagener et al., 2007; Reichl et al., 2009; Archfield and Vogel, 2010; Oudin et al., 2010; Patil and Stieglitz, 2011; 2012; Razavi and Coulthart, 2013; Athira et al., 2016; Brunner et al., 2018). The mostly commonly used technique involves the transfer of lumped characteristics such as catchment shape and size, Strahler ratios, drainage density, average slope, etc. that are used to explain hydrogeomorphological characteristics (Horton, 1932; 1945; Strahler, 1957; Bras, 1990; Rodriguez-Iturbe and Rinaldo, 2001). An issue with this is the possibility of the loss of information in simplifying complex catchment properties into a single number (Wooldridge and Kalma, 2001; Wagener and Wheater, 2006).
Alternatively, distribution curves can be used to assess hydrological similarity. Examples of this include the use of the distribution of topographic index, height above nearest drainage, reduced dissipation per unit length index (Loritz et al., 2019), the distribution of riparian and hillslope effects on streams, the riparian-area change along the stream network (McGlynn and Seibert, 2003), the hypsometric curve (Booij et al., 2007; Ssegane et al., 2012; Hailegeorgis et al., 2015; Bajracharya and Jain, 2021), and the width function (Moussa, 2008; Bajracharya and Jain, 2020). Furthermore, various mathematical models that link catchment structure to hydrological response based on underlying physics or statistical relationships have been used to explore catchment similarity and to develop similarity parameters (Hebson and Wood, 1982; Sivapalan et al., 1987; Larsen et al., 1994; Milly, 1994; Reggiani et al., 2000; Aryal et al., 2002; Woods, 2003).

2.2 Dynamical representation of watershed morphometry

2.2.1 Width function

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

The width function is most commonly represented by a histogram with the hydrological distance in the \( x \)-axis and the frequency or density of the areal extent of streams in the \( y \)-axis (Figure 1). Bajracharya and Jain (2020) demonstrated the use of a truncated skew-Normal (SN) mixture model to analytically represent the width function with the \( x \)-axis normalized by scaling between 0 and 1, and demonstrated its utility in finding hydrologically analogous drainage basins using divergence measures such as the \( L_2 \) distance (Tsybakov, 2008). The SN distribution is a three-parameter probability distribution formed by adding a skewness element to the Normal distribution.

For a continuous random variable, \( X \), the SN distribution is represented as:

\[
f(x; \xi, \omega^2, \alpha) = \frac{2}{\omega} \phi \left( \frac{x - \xi}{\omega} \right) \Phi \left( \alpha \frac{x - \xi}{\omega} \right), \quad x \in (-\infty, \infty) \quad (1)
\]
where $\phi(x)$ denotes the standard Normal density function of $x$, $\Phi(x)$ denotes the cumulative distribution function (cdf) of the standard Normal, and $\xi$, $\alpha$, and $\omega$ are the location, scale, and shape parameters, respectively. The domain of the $SN$ distribution is then truncated to $[0,1]$ using a correcting factor to guarantee the validity of the normalization condition (Thomopoulos, 2017):

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

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

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

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

$$L_2 = \sqrt{\int (N_1 - N_2)^2 dx}$$

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

2.2.2 Hypsometric function

The hypsometric curve reflects the area-altitude distribution of a basin (Horton, 1932; Langbein, 1947) and reflects the distribution of landmass as a function 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.
the geomorphic maturity of catchments, with a concave up shape indicating
relatively mature basins with a high degree of erosive activity, and a con-
cave down shape indicating relatively young basins with a large proportion of
uneroded topography or creep-dominated hillslopes (Strahler 1952; Moglen
and Bras 1995; Pedrera et al. 2009; Willgoose 2018). Furthermore, studies
have linked the hypsometric curve with various drainage basin features such
as the hydrograph time-to-peak, head-ward drainage development, regional
basin slopes (Harlin 1984), average channel gradient (Howard 1990), stream
network branching (Willgoose and Hancock 1998), ground water interaction,
water table fluctuation (Marani et al. 2001), and surface and subsurface runoff
properties (Vivoni et al. 2008). Willgoose and Hancock (1998) further divided
the hypsometric curve into three regions: the ’head’ (upper left-hand side),
the ’toe’ (lower right-hand side), and the ’body’ (between the head and the
toe), and linked the shape of the toe to stream branching characteristics of
the basin. Furthermore, hillslopes with active fluvial erosion and creep exhibit
concave down head and concave up tail (Willgoose 2018). In long skinny catch-
ments and hillslopes with parallel flow lines, the hypsometric curve reflects the
hillslope long profile and can be used as an indirect test of the slope-area re-
lationship, while in more rounded catchments, the stream network branching
also affects the shape of the hypsometric curve (Willgoose 2018).

The hypsometric curve can be plotted in absolute units, with elevation
in meters and area in square kilometers, or in relative units, with relative
elevation plotted against the relative area above said elevation (Figure 2). The
latter, termed as the percentage hypsometric curve, allows for the comparison
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.

Various functional forms have been developed to represent the hypsometric
curve (Strahler 1952; Harlin 1978; Sarkar and Patel 2011; Vanderwaal and
Ssegane 2013; Bajracharya and Jain 2021). Bajracharya and Jain (2021) de-
veloped a three-parameter model named the Generalized Hypsometric function
by modifying the equation developed by Strahler (1952). The model places an emphasis on the curvatures of the head, body, and the toe. The function is defined as:

\[ y = \left( \frac{1 - x^m}{1 + \beta x^m} \right)^z \]  

(5)

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

3 Data and methods

3.1 Case study

The Narmada River basin (NRB) is located in central India between latitudes 21°22' 0" N and 23°46' 30" N, and longitudes 73°4' 0" E and 81°45' 30" E. The drainage area is 95,000 km\(^2\) (Figure 3). The elevation ranges from nearly zero to over 1000 m above sea level, with an average slope of 1.1°. The basin is bounded on the north, east, and south by hills, and on the west by the Arabian sea. The lower middle reaches are comprised of fertile plain lands. A number of reservoirs have been constructed in the basin for a variety of purposes including water supply, irrigation, and hydropower generation. The Narmada River passes through three states that face water shortages during 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.
The elevation data for the region was obtained from GTOPO30, a global digital elevation model (DEM) developed by the United States Geological Survey (USGS). It was derived from several raster and vector sources of topographic information [USGS 1996]. The dataset has a spatial resolution of 30-arc seconds and a vertical accuracy of around 30 m. It is based on several sources of elevation information, including various vector and raster datasets, merged together, with a priority given to the data with a greater topographic detail and accuracy. With extensive accuracy checks, GTOPO30 data are suitable for numerous regional and continental applications, including the extraction of drainage features for hydrologic modeling [USGS 1996].

The stream network was derived from the DEM in ESRI ArcGIS 10.5.1 through standard Geographic Information System (GIS) procedures. First elevation grids with undefined drainage directions, known as sinks, were filled; then the flow direction was determined based on the direction of steepest descent; followed by the computation of accumulated flow at each grid. A threshold of 396 km² was used to delineate the stream grids. This threshold was chosen to ensure a dense stream network, resulting in fourth order streams. This allowed for a delineation a considerable number of sub-basins to test the fits for diverse width function and hypsometric curve shapes. Finally, outlets were places at the confluences of first order and higher order streams to create 72 non-overlapping sub-basins.

3.2 Methodology

Clustering is a descriptive unsupervised data mining technique for creating subsets by grouping similar data together based on some measure of similarity or dissimilarity [Veyssieres and Plant 1998; Rokach and Maimon 2005]. The clustering structure is represented by a set of subsets, \( C = C_1, ..., C_k \) of \( S \), such that \( S = \bigcup_{i=1}^{k} C_i \) and \( C_i \cap C_j = \emptyset \) for \( i \neq j \). Hierarchical clustering is a clustering method that creates clusters by recursive partitioning, resulting in a dendrogram structure that represents the nested grouping of instances and similarity levels at which the groupings change. The recursive algorithm could be bottom-up, starting from every element in their individual cluster, with similar elements then grouped into a single cluster in each successive step (agglomerative clustering); or top-down, starting from all elements grouped in a single cluster, followed by the most dissimilar elements being separated into another cluster at each iteration (divisive hierarchical clustering). Various methods have been developed based on the manner in which the similarity measure is calculated and optimized, most of which are variants of single-link, complete-link, and minimum-variance algorithms [Jain et al. 1999]. These algorithms consider the distance between two clusters to be equal to the shortest, longest, and average distance between a member of one cluster to a member of the other, respectively. Single-link methods are more versatile [Rokach and Maimon 2005] but are susceptible to the “chaining effect”, where a few points that form a bridge between two poorly separated, but distinct clusters lead to
them being merged at an early stage (Guha et al., 1998). On the other hand, complete-link methods usually produce more compact clusters (Rokach and Maimon, 2005). On the other hand, average-link clusters may cause the splitting of elongated clusters and the merging of portions of neighboring elongated clusters (Guha et al., 1998).

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

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

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

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

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

4 Watershed similarity

4.1 Width function clusters

4.1.1 Hierarchical clustering

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

Outliers can cause chaining effects, leading to dissimilar objects being drawn into the same cluster (Everitt et al., 2011). Removal of outliers can help reduce intra-cluster variance. However, different outlier detection algorithms can lead to different data points being classified as outliers. Moreover, outlier detection can mistakenly classify small clusters as outliers and remove valuable information from the data. Thus, outlier detection involves a degree of subjectivity. Here we use a simple algorithm to analyze, detect, and remove outliers based on similarity measures with nearest neighbors. Figure 6 shows the $L_2$ distance to fifteen closest neighbors for each width function. Based on this measure, a threshold can be chosen subjectively to delineate outliers based on specific goals. In this study, width functions with the $L_2$ distance greater than 0.45 for up to 15 closest neighbors were marked as outliers. This lead to only three width functions being classified as outliers. Intra-cluster uniformity can be further improved by lowering this threshold. While rigorous methods for removal of outliers exist in the literature (Almeida et al., 2007; Fan et al., 2013; Krelaža...
et al. [2021], we employed this basic outlier detection algorithm as a proof of concept, one that is easy to understand and can be readily applied.

4.1.3 Analysis of clusters

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

Hierarchical clustering can be best denoted using dendrograms. The dendrogram notation of the width function clusters are shown in Figure 8, along with the mean width functions and the location of the sub-basins. Figure 8 (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.
each cluster. The width function shapes seem mostly independent of the location of the sub-basin along the watershed as well as the sub-basin areas (Figure 8 (c)). Cluster 4 seems to be concentrated at the mid region of the watershed and cluster 6 seems to be concentrated mostly in the bottom half, where as all other clusters are spread across different regions. Interestingly, a number of sub-basins within same cluster groups appear alongside each other.

4.2 Hypsometric function clusters

4.2.1 Hierarchical clustering

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

![Fig. 9](image.png)

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

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

4.3 Joint analysis of hierarchical clustering of width functions and
hypsometric curves

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.
Hydrologic similarity based on width function and hypsometry

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

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

5 Discussion and concluding remarks

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

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

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

3. There is a level of subjectivity in the choice of the number of clusters. The considerable degree of intra-cluster heterogeneity in the location of the peaks of width functions indicates the need for a relatively large number of clusters for width functions if a high degree of homogeneity is desired. On the other hand, a relatively lower number of hypsometric clusters might be sufficient due to the cumulative nature of the curve which tends to offer a lower variance.

4. Two bivariate groups with similar width functions and hypsometric functions were identified with at least four members, one was identified with five members, and one with six members from a total sample size of 72. These represent sub-basins with similar hydrological response characteristics. This can easily be scaled to thousands of watersheds around the world.

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

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

Declarations

Funding

Not applicable
Conflicts of interest/Competing interests
The authors declare that they have no conflict of interest.

Availability of data and material
Not applicable

Code availability
Not applicable

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

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.l., 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.
41 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/
Hydrologic similarity based on width function and hypsometry

2000WR000119

McGlynn, B.L., Seibert, J., 2003. Distributed assessment of contributing area and riparian buffering along stream networks. Water resources research 39.

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

Mesa, O.J., Mifflin, E.R., 1986. On the relative role of hillslope and network geometry in hydrologic response, in: Scale problems in hydrology. Springer, pp. 1–17.

Milly, P., 1994. Climate, interseasonal storage of soil water, and the annual water balance. Advances in Water Resources 17, 19–24.

Minns, A., Hall, M., 1996. Artificial neural networks as rainfall-runoff models. Hydrological sciences journal 41, 399–417.

Mital, U., Dwivedi, D., Brown, J.B., Faybishenko, B., Painter, S.L., Steefel, C.I., 2020. Sequential imputation of missing spatio-temporal precipitation data using random forests. Front. Water 2: 20 doi:10.3389/frwa.2020.00020.

Moglen, G.E., Bras, R.L., 1995. The importance of spatially heterogeneous erosion and the cumulative area distribution within a basin evolution model. Geomorphology 12, 173–185.

Moussa, R., 2008. What controls the width function shape, and can it be used for channel network comparison and regionalization? Water Resources Research 44. doi:10.1029/2007WR006118

Oppel, H., Mewes, B., 2020. On the automation of flood event separation from continuous time series.

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

Patil, S., Stieglitz, M., 2011. Hydrologic similarity among catchments under variable flow conditions. Hydrology and Earth System Sciences 15, 989–997. doi:10.5194/hess-15-989-2011

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, 551–562. doi:10.5194/hess-16-551-2012

Pedrera, A., Pérez-Peñia, J.V., Galindo-Zaldívar, J., Azañón, J.M., Azor, A., 2009. Testing the sensitivity of geomorphic indices in areas of low-rate active folding (eastern Betic Cordillera, Spain). Geomorphology 105, 218–231.

Peters-Lidard, C.D., Clark, M., Samaniego, L., Verhoest, N.E., Van Emmerik, T., Uijlenhoet, R., Achieng, K., Franz, T.E., Woods, R., 2017. Scaling, similarity, and the fourth paradigm for hydrology. Hydrology and Earth System Sciences 21, 3701–3713. doi:10.5194/hess-21-3701-2017

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org

Ray, L.K., Goel, N.K., 2019. Flood frequency analysis of Narmada River basin in India under nonstationary condition. Journal of Hydrologic Engineering
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.

Rodriguez-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., Sernet, 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., Tolnner, 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
Hydrologic similarity based on width function and hypsometry

a geomorphic province: the Cairngorm mountains, Scotland. Hydrological Processes 23, 1874–1886. doi:10.1002/hyp.7318

Thomopoulos, N.T., 2017. Statistical distributions: Applications and Parameter Estimates. Springer.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63, 411–423.

Troutman, B.M., Karlinger, M.R., 1984. On the expected width function for topologically random channel networks. Journal of Applied Probability 21, 836–849. doi:10.1017/S0021900200037530

Troutman, B.M., Karlinger, M.R., 1989. Predictors of the peak width for networks with exponential links. Stochastic Hydrology and Hydraulics 3, 1–16. doi:10.1007/BF01543424

Tsybakov, A.B., 2008. Introduction to nonparametric estimation. Springer.

Tung, Y.K., Yeh, K.C., Yang, J.C., 1997. Regionalization of unit hydrograph parameters: 1. Comparison of regression analysis techniques. Stochastic Hydrology and Hydraulics 11, 145–171. doi:10.1007/BF02427913.

USGS, 1996. USGS EROS Archive – Digital Elevation – Global 30 Arc–Second Elevation (GTOPO30). https://doi.org/10.5066/F7DF6PQS Accessed: 2019-06-02.

Vanderwaal, J.A., Ssegane, H., 2013. Do polynomials adequately describe the hypsometry of Monadnock phase watersheds? JAWRA Journal of the American Water Resources Association 49, 1485–1495. doi:10.1111/jawr.12089.

Veneziano, D., Moglen, G.E., Furcolo, P., Iacobellis, V., 2000. Stochastic model of the width function. Water Resources Research 36, 1143–1157. doi:10.1029/2000WR900002

Veyssieres, M., Plant, R.E., 1998. Identification of vegetation state and transition domains in California’s hardwood rangelands. University of California 101.

Vivoni, E.R., Di Benedetto, F., Grimaldi, S., Eltahir, E.A., 2008. Hypsometric control on surface and subsurface runoff. Water Resources Research 44.

Wagener, T., Sivapalan, M., Troch, P., Woods, R., 2007. Catchment classification and hydrologic similarity. Geography Compass 1, 901–931. doi:10.1111/j.1749-8198.2007.00039.x

Wagener, T., Wheeler, H.S., 2006. Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. Journal of Hydrology 320, 132–154. doi:10.1016/j.jhydrol.2005.07.015

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244.

Willgoose, G., 2018. Principles of soilscape and landscape evolution. Cambridge University Press.

Willgoose, G., Hancock, G., 1998. Revisiting the hypsometric curve as an indicator of form and process in transport-limited catchment. Earth Surface Processes and Landforms: The Journal of the British Geomorphological Group 23, 611–623.
Woods, R., 2003. The relative roles of climate, soil, vegetation and topography in determining seasonal and long-term catchment dynamics. Advances in Water Resources 26, 295–309.

Woods, R., 2006. Global similarity indices for mean and seasonal hydrology of ungauged basins, in: Presentation at USA PUB Workshop.

Wooldridge, S.A., Kalma, J.D., 2001. Regional-scale hydrological modelling using multiple-parameter landscape zones and a quasi-distributed water balance model. Hydrology and Earth System Sciences 5. doi:10.5194/hess-5-59-2001.

Yadav, M., Wagener, T., Gupta, H., 2007. Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. Advances in Water Resources 30, 1756–1774.

Zhu, X., Guo, D., 2014. Mapping large spatial flow data with hierarchical clustering. Transactions in GIS 18, 421–435.
Supplementary materials

Original width function clusters

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

Fig. S1 Width functions in each cluster.