Manuscript title
A Network Approach for Multiscale Catchment Classification using Traits
Authors
Fabio Ciulla and Charuleka Varadharajan

Response to Reviewer 1 Comments:

Major comment

COMMENT #1: The authors introduce a novel method to cluster catchments that is based on traits. The dataset is impressive and the network-based classification is, to my understanding, a relevant and innovative approach in this case. Methods and results are well presented.

AUTHOR RESPONSE #1 - We thank the reviewer for the positive comments. The reviewer brings up important concerns, and we have provided detailed responses below for each concern that was raised. We have made substantial changes that address these comments in the revised manuscript. We also tried to condense some of the original text and merged Figures 6 and 7 of the original manuscript to make room for some of the major additions.

COMMENT #2 - My main concern with such unsupervised classification is how we can use it for practical hydrological studies.

AUTHOR RESPONSE #2: We agree with the comments from both reviewers that we could do more to demonstrate the utility of our methodology for hydrological applications. In addition to responding to comments from reviewer 1 below, we note that we have done additional analysis to show how we can gain hydrological insights from our approach in our response to comments from reviewer 2.

COMMENT #3: From the introduction and discussion, it appears one aim of clustering is application to ungauged basins. In this sense, the results of the paper are discouraging, because the clustering technique does not succeed in relating ‘traits’ clusters to hydrological behaviors, except for some specific hydrological traits. This part is essential, in my opinion, for switching from a mere clustering exercise to something which could actually be useful in hydrological practice. I do not know how the method can be tuned to improve the overlap between the geographical and hydrological clusters, but my wish is that the authors tackle this issue in the paper. I realize that this implies a significant change in the paper.

AUTHOR RESPONSE #3: As pointed out, one of the applications of our methodology is for predictions in unmonitored basins (PUBs). We had originally demonstrated that our approach to trait-based clustering (which the reviewer refers to as geographical clustering) results in statistically distinct hydrological behavior (based on signatures) relative to other

catchments within the CONUS. We assume the reviewer thinks the results are discouraging, based on the boxplots shown in Figure 13 in the original manuscript, where it appears as though some of the distributions overlap for the two indices shown.

However, we would like to point out that our current approach does result in distinguishing the streamflow behavior for most of the indices, indicating that there already is significant overlap between the trait-based clustering and the hydrological groupings. We made the following changes to address this comment.

First, we added the following text in Section 2.10, line numbers 364-373

"We conducted two additional statistical tests to further examine whether the hydrological indices of catchment clusters are significantly distinct. The first is a nonparametric 1-sample Kolmogorov-Smirnov (K-S) Test that compares a sample distribution to a reference one for each hydrological index. This expands on the Kruskall-Wallis test, but allows us to determine the number of clusters that are statistically distinct from the entire catchment dataset. Here, each sample is constituted by the indices of one cluster and the reference distribution is based on all the catchments within the CONUS. The null hypothesis is that samples are drawn from the reference distribution when using 0.05 as threshold for the p-value.

The second is a 2-sample K-S test comparing the distributions of indices for pairs of catchment clusters, which allows us to determine how different the clusters are from each other. Here, each sample pair is constituted by the distribution of indices for the clusters being compared. Similar to the one-sample test, the null hypothesis is that samples are drawn from the same distribution when using 0.05 as threshold for the p-value..

We then added the following text to results, section 3.5, lines 461-466

"The results of the statistical tests are summarized in appendix in Table G1 for all 34 streamflow indices. When averaged for all the indices, 83% of the clusters for the 1-sample K-S test and 79% for the 2-sample K-S test reject the null hypothesis, meaning that the distribution of their streamflow indices is mostly distinct from the overall distribution at the continental scale. These results show that the trait-based clustering approach results in distinct signature classification. See Sect. 4.5 for further discussion about the distinct hydrological behavior across the catchment clusters."

We also modified the text in discussion, section 4.1, lines 555-508

"We have demonstrated that our network-based workflow outperforms the K-means or the Ward's hierarchical algorithms for two different metrics (see Sect. 3.6), particularly with the parameters we chose for dimensionality reduction and the disparity filter (black dashed lines

in Fig. 8 corresponding to k = 20 dimensions retained). We conclude that our method produces more distinct clusters, and is hence a better choice than traditionally used classification approaches."

Hydro-logical index	Description	K-W test	K–S test one-sample	K–S test two-samples
ma41	Mean annual flow divided by catchment area (m3/s/km2)	True	97.06	95.19
dh13	Mean annual of 30-day maximum divided by median flow	True	94.12	90.55
mh14	Median of the highest annual daily flow divided by the median annual daily flow	True	97.06	90.37
ma5	Skewness in daily flow	True	91.18	90.20
mh16	Mean of the 10th percentile from the flow duration curve divided by median daily flow across all years	True	91.18	88.95
ma3	Coefficient of variation in daily flows	True	94.12	88.95
fh7	Mean number of high flow events per year using an upper threshold of 7 times the median flow over all years	True	94.12	87.88
ra6	Median of difference between natural logarithm of flows between two consecutive days with increasing flow	True	88.24	86.27
fh6	Mean number of high flow events per year using an upper threshold of 3 times the median flow over all years	True	100.00	85.92
th3	Maximum proportion of the year (number of days/365) during which no floods have ever occurred over the period of record	True	85.29	85.20
fl3	Total number of low flow spells (threshold equal to 5% of mean daily flow)	True	82.35	83.60
fh3	High flood pulse count (high flood: at least 3 times median of daily flows)	True	85.29	83.07
ml21	Coefficient of variation in annual minimum flows averaged across all years	True	85.29	81.46
ml17	Seven-day minimum flow divided by mean annual daily flows averaged across all years	True	94.12	81.28
dh16	Coefficient of variation of high flood pulse (high flood: at least 75th percentile of daily flows)	True	91.18	80.93
dl13	Mean annual of 30-day minimum divided by median flow	True	79.41	80.39
fh2	Coefficient of variation of high flood pulse count (high flood: at least 75th percentile of daily flows)	True	88.24	78.97
ml18	Coefficient of variation of seven-day minimum flow divided by mean annual daily flows averaged across all years	True	76.47	78.25
fl2	Coefficient of variation of low flood pulse count	True	94.12	77.36
ta1	Constancy	True	79.41	77.36
ml4	Mean minimum monthly flow for the months of April (m3/s)	True	79.41	77.36
dh20	Mean duration of high flood pulse (high flood: at least 25th percentile of median flows) (days)	True	85.29	77.18
mh10	Mean maximum flows for the months of October (m3/s)	True	94.12	76.47
tl2	Variability in Julian date of annual minimum (days)	True	82.35	76.11
mh8	Mean maximum flows for the months of August (m3/s)	True	85.29	75.94
ma11	Spread in 75th-25th percentile range on decimal logarithm transformed daily flows	True	76.47	75.58
dl17	Coefficient of variation of low flood pulse count	True	79.41	75.22
dh15	Mean duration of high flood pulse (high flood: at least 75th percentile of daily flows) (days)	True	79.41	74.15
ra8	Number of negative and positive changes in water conditions from one day to the next	True	70.59	73.98
ra5	Ratio of days where the flow is higher than the previous day	True	67.65	67.02
ra9	Coefficient of variation of the number of negative and positive changes in water conditions from one day to the next	True	64.71	62.75
fl1	Low flood pulse count (low flood: below 25th percentile of daily flows)	True	61.76	59.36
dl18	Number of zero-flow days [days]	True	70.59	56.33
dl16	Mean duration of low flood pulse [days]	True	47.06	50.98

Table G1:Table showing the results from statistical tests comparing the distributions of streamflow indices of catchment clusters resulting from our network-based methodology. The first two columns list the streamflow indices used in this study as alphanumeric codes with brief descriptions as in Olden and Poff (2003). The last 3 columns show the result of the Kruskall-Wallis (K-W) test indicating that not all the samples have the same distribution, 1-sample and 2-sample Kolmogorov-Smirnov (K-S) tests. The indices are sorted according to the 2-sample Kolmogorov-Smirnov Test in descending order."

As mentioned in our earlier response, Specifically regarding the statements *"the clustering technique does not succeed in relating 'traits' clusters to hydrological behaviors, except for some specific hydrological traits. This part is essential, in my opinion, for switching from a mere clustering exercise to something which could actually be useful in hydrological practice."* , we note that since the dataset of traits is large, and encompasses a broad set of categories (climate, geology, land use, human activities etc.), it is expected that not all traits (or trait clusters) are going to be related to hydrologic behavior. The question about which traits are most relevant for a particular hydrologic function of interest, such as streamflows, is still largely unresolved. In general, prior large sample studies such as (Eng et al., 2017) and (Addor et al., 2018) have not had much success in using traits to predict hydrologic signatures with statistical or classical machine learning approaches. Tackling the issue of relating trait clusters to hydrologic behavior is a different study that is out of scope for this paper. This is the subject of multiple follow-on studies and papers that we are working on, which require building models to show the relationships between the trait clusters and hydrologic signatures. Please see our response #4 below where we elaborate more on the practical uses for our approach.

COMMENT #4: In the case the authors stick to unsupervised clustering, I guess that the paper might be of interest, but in my opinion, the authors should:

- introduce in more details the practical implications of such clustering, and

AUTHOR RESPONSE #4:

We have added a significant amount of text to describe the practical implications of our workflow as described below. Also see responses #12 and #13 for additional analysis that was added as per reviewer 2's suggestions.

First, we added a new Section 4.7, lines 677-686 with the following text:

"Our methodology can be used as initial steps of typical workflows used for predictions for unmonitored basins, which is (1) to classify catchments into groups for regionalization, and (2) to select a subset of traits from a large predictor space, a common challenge in many large sample studies. For the former, we choose to classify catchments using traits, since geospatial datasets are now available with a lot of trait information that allows us to do catchment classification at large spatial scales including for unmonitored catchments. For the latter, we find that we can condense a very large dataset containing hundreds of traits into 25 trait categories with the network approach, due to the redundancy in the traits. Thus our paired catchment-trait networks approach not only classifies catchments into clusters, but enables the reduction of a large dataset of traits into an interpretable set of trait categories by eliminating their redundancy. This provides the ability to identify distinct trait

categories that are over- or under-expressed in catchment clusters to streamflow behaviors. The parallel analysis of cluster and traits data as networks is an important characteristic that distinguishes our method from other typical unsupervised clustering workflows."

We added the following text to Section 4.2, lines 532-538

"To illustrate how our approach helps with reducing trait redundancy, we computed the spearman correlation coefficients (ρ) between streamflow indices and catchment traits, which account for non linearities in the data. We find that traits belonging to the same trait categories have similar correlations with the streamflow indices (Fig. I1 in appendix). We find that the spearman correlation coefficients between the streamflow indices and individual catchment traits can be effectively represented as an aggregated median value for the trait categories generated in our method (Fig. 10), which indicates that our reduced set of trait categories are sufficient to determine the general relationships between traits and hydrological signatures. "

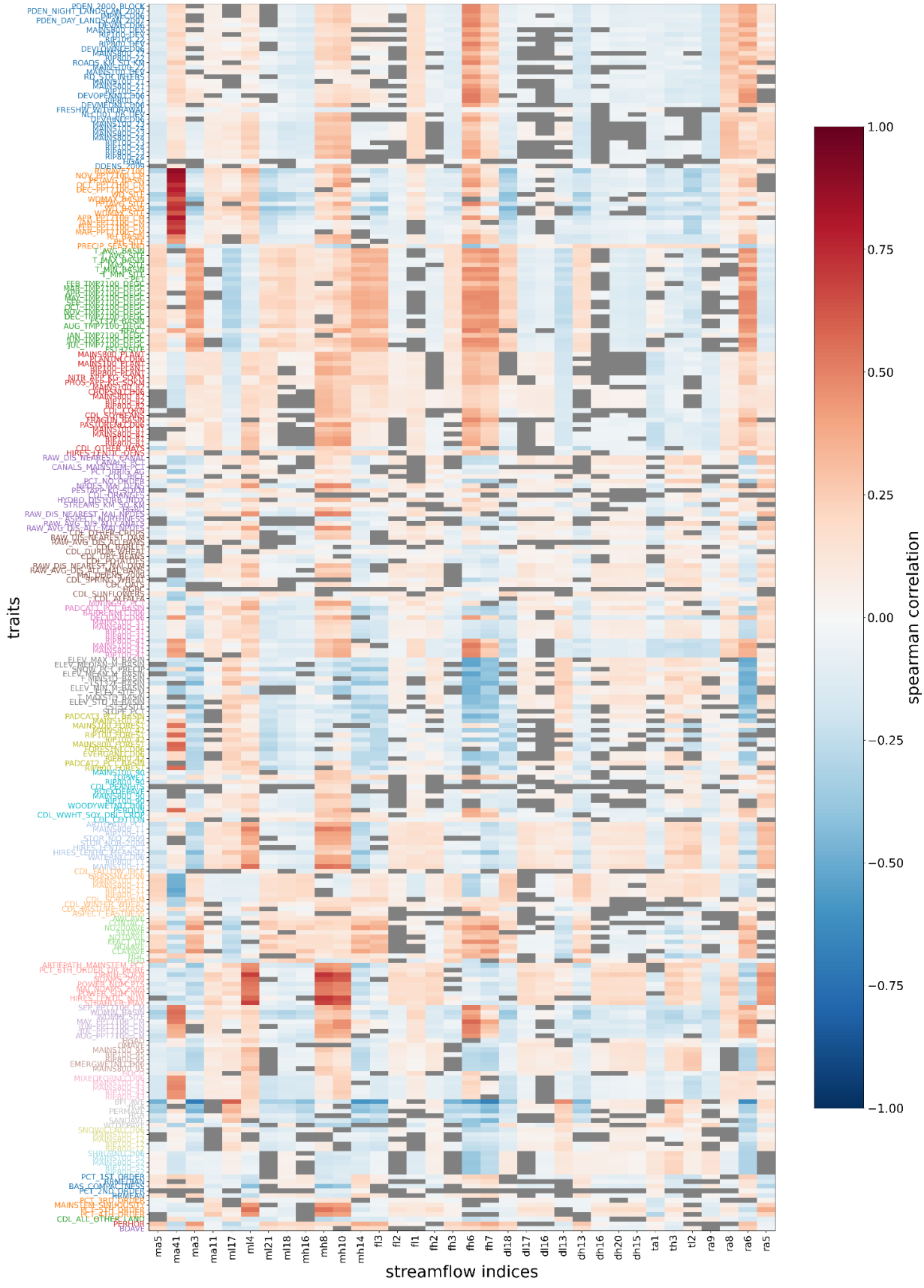
traits
streamflow indices
spearman correlation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
ma5
ma41
ma3
ma11
ml17
ml4
ml21
ml18
mh16
mh8
mh10
mh14
fl3
fl2
fl1
fh2
fh3
fh6
fh7
dl18
dl17
dl16
dl13
dh13
dh16
dh20
dh15
ta1
th3
tl2
ra9
ra8
ra6
ra5

Figure I1. Heatmap showing spearman correlation coefficients between streamflow indices and the traits used in the study. The intensity of the colors show the degree of correlation or anticorrelation as indicated in the color bar. Traits on the horizontal axis are ordered and colored according to the traits categories they belong to. Gray boxes indicate correlations that are above the significance level of p>0.05. The primary purpose of this plot is to provide a visual representation of the redundancy in the correlations between the traits within the same category and streamflow indices.

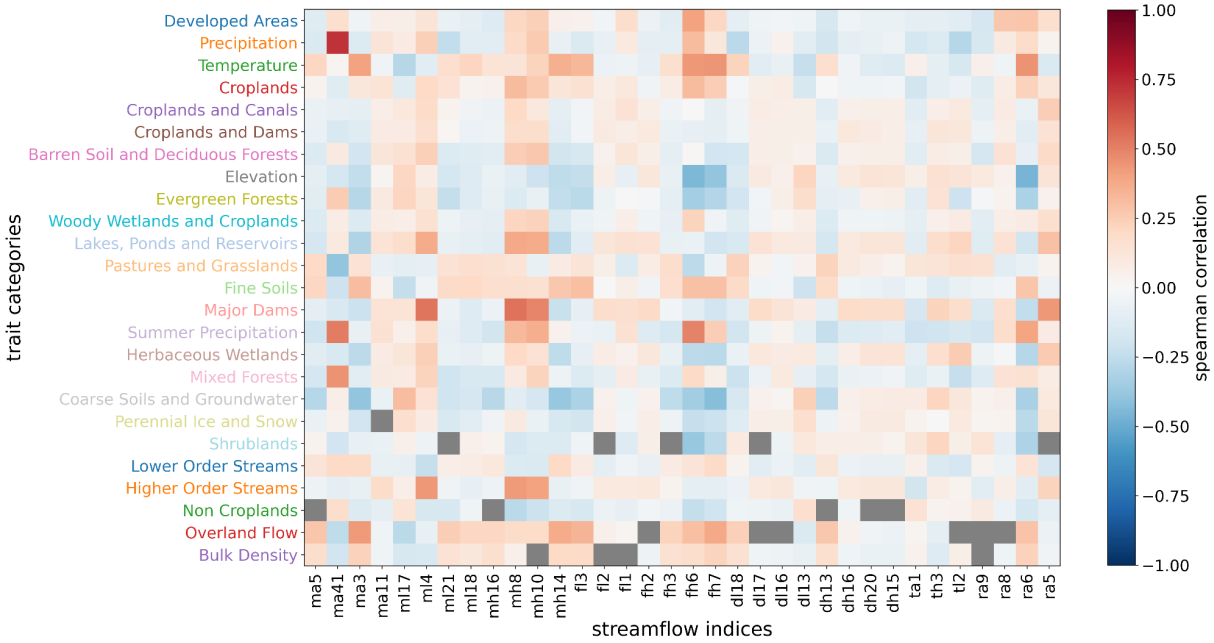

Figure 10. Heatmap showing spearman correlation coefficients between streamflow indices and the traits categories generated in the study. The intensity of the colors show the degree of correlation or anticorrelation as indicated in the color bar. Gray boxes indicate correlations that are above the significance level of p>0.05.

We also added an additional example for the use of our methodology in Section 4.5, lines 620-637:

“Additionally, we utilize the ability to determine correlations between stream flow indices and a reduced set of interpretable trait categories (Sect. 4.2). We find that across the 9067 catchments, mean annual runoff (ma41) is not just positively correlated with traits related to precipitation, but also with the presence of mixed forests (ρ = 0.45) and to a lesser degree evergreen forests (ρ = 0.25). The ma41 index is also negatively correlated with the ’pastures and grasslands’ trait category (ρ = −.40). This highlights the role of vegetation in mediating flows, and is somewhat counterintuitive given that in the absence of management, forested catchments with higher evapotranspiration would be expected to have lower flows compared to grasslands. Similarly, the fh6 index is not just positively correlated with precipitation traits, but also with traits related to developed areas (ρ = 0.40),

croplands (ρ = 0.32) and temperature (ρ = 0.43). It is also inversely correlated with elevation (ρ = −0.45), presence of shrublands (ρ = −0.38), evergreen forests (ρ = −0.28), coarse soils and groundwater (ρ = −0.35). These relationships are consistent across other flood indices. For example, the fh7 index showing the propensity for heavy floods (above 7 times median flows) similarly has a moderate positive correlation with temperature (ρ = 0.44) and overland flow (ρ = 0.38), and a moderate negative correlation with elevation (ρ = −0.39) and coarse soils/groundwater (ρ = −0.43). This indicates how flooding is affected by the complex relationships between land use, vegetation, soil infiltration capacity and base flows.

These examples highlight the use of our methodology to demonstrate how specific hydrological behaviors can be connected to catchment traits such as their climatic conditions, topography, land use or anthropogenic influence. The ability to link catchment traits to specific hydrological behaviors, enables further analysis of the factors that influence different stream flow characteristics (e.g. high versus low flows). In particular, the distinction between anthropogenically-influenced trait categories and natural traits (see Table 1) enables further analysis of human activities on hydrologic behavior."

Hence we argue that the trait categories generated using our method are interpretable in a manner that is harder to do with a dimensionality reduction approach using eigenvectors, where the contributions of traits can be distributed across many principal components. For additional applications of our unsupervised clustering approach, we refer the reviewer to our response #13 to reviewer 2 comments. Here we conducted additional analysis to identify relationships between hydrologic signatures and the catchment clusters, as well as the predominant traits of the catchments in those clusters. This analysis has produced new insights that on its own can be used to generate hypotheses about processes that influence hydrological behavior.

COMMENT #5:

- compare the obtained classification with a benchmark clustering approach.

AUTHOR RESPONSE #5: We acknowledge that the manuscript we submitted does not provide sufficient evidence that our proposed method based on networks and cosine similarity performs better than traditional unsupervised clustering algorithms. We thank the reviewer for this suggestion and in response performed a more comprehensive analysis comparing the performance of our method against benchmark hierarchical clustering and k-means approaches. To address this comment, we made the following changes.

First, we added the following text to methods as a new section 2.11, lines 375-400

"We compare the clusters obtained from our methodology that uses network theory and the cosine distance metric, with the ones resulting from the traditional K-means and hierarchical clustering approaches. Since these are all unsupervised methods, we cannot use any target

variable to compute the accuracy of the classification to compare performance. For this reason, we identified two metrics to evaluate the performance of the different methods.

The first metric, which we refer to as "cluster similarity", reflects the similarity between traits of the catchment clusters, which are represented by the average trait z-scores aggregated across the catchments in each cluster as described in Sect. 2.8. Here, each catchment cluster is compared to the others by calculating their pairwise cosine similarity. The highest value of the cosine distance within each catchment cluster is used as a conservative measure of inter-cluster similarity, to assess how far

apart the catchment clusters are from each other. The median value of the inter-cluster similarities represents how distinct the clusters produced by each algorithm are. We aim to minimize this metric, since a good classification algorithm should produce more distinct clusters.

The second metric is the silhouette score (Rousseeuw, 1987), which is a measure of intra-cluster similarity. It represents how similar each element (i.e., a catchment) is to other elements within its cluster relative to elements in other clusters. The values of this metric range between -1 and 1, with higher values denoting that an element is well placed in its cluster compared to other clusters. The silhouette values are averaged for all items in the dataset. A good clustering algorithm would produce higher values of the silhouette score.

We use these two metrics to compare our clustering approach with the hierarchical clustering (in its common implementation using the Ward criterion; Ward (1963) and the k-means clustering algorithm (MacQueen, 1967). Additionally, to determine the effects of the distance metric, we compare the results from our workflow that uses the cosine distance with a version where the pairwise similarity between nodes is computed using the Euclidean distance. Finally, to show the robustness of our approach, we examine various choices for the two free parameters in our workflow, namely the number of reduced dimensions after the PCA (k) and the cluster granularity, which is governed by the disparity filter parameter α used to tune the removal of network edges during the backbone extraction step (Sect. 2.5). Three different values of k are investigated; k=6 corresponding to 50% of information retained after PCA, k=20 (our choice in the study) corresponding to 72% of retained information, and k=90 corresponding to 95% of retained information. For each value of k, we generate clusters with different α values, with the number of clusters covering 95% of the dataset ranging between 20 and 120."

We added the following text to results, section 3.6, lines 468-474

"The two metrics, cluster similarity and silhouette score, indicate that our workflow performs better than traditional unsupervised methods of classification (Fig. 8). We also find that the network-based clustering (red and green points) is considerably superior to both k-means (yellow points) and hierarchical clustering (blue points) across the different values of k and cluster granularity (Fig. H1 in appendix). This is evident from the consistently low values of

the median cluster similarity and higher values of silhouette scores for our methodology. Also, the network generated using the cosine distance as a similarity metric (red points) performs better than its counterpart that uses the Euclidean distance (green points). This confirms that the cosine similarity should be preferred as a distance metric in high dimensions and the directionality of the data can contain valuable information."

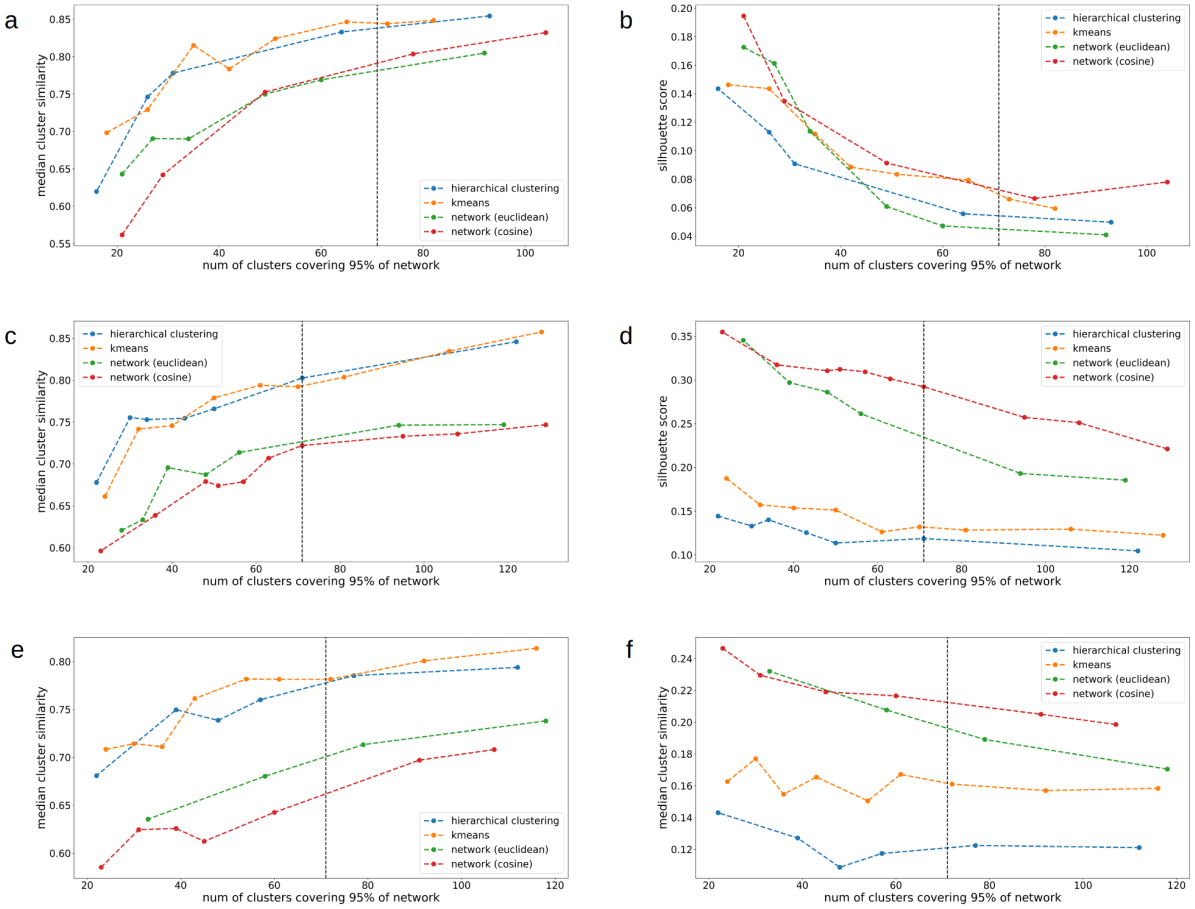

Figure H1. Median cluster similarity values (a, c, e) and silhouette scores (b, d, f) for different clustering methods and similarity measures used in the network analysis. The number of reduced dimensions after PCA is equal to (a, b) 6, (c, d) 20 (used in the study) and (e,f) 90 corresponding to 50%, 72% and 95% of retained information respectively. The vertical black dashed line refers to the cluster granularity used in the paper. The colored dashed lines are shown for visualization of trends. Lower values of the median cluster similarity metric (a, c, e) correspond to better clustering performance. Higher values of the silhouette scores (b, d, f) correspond to better clustering performance.

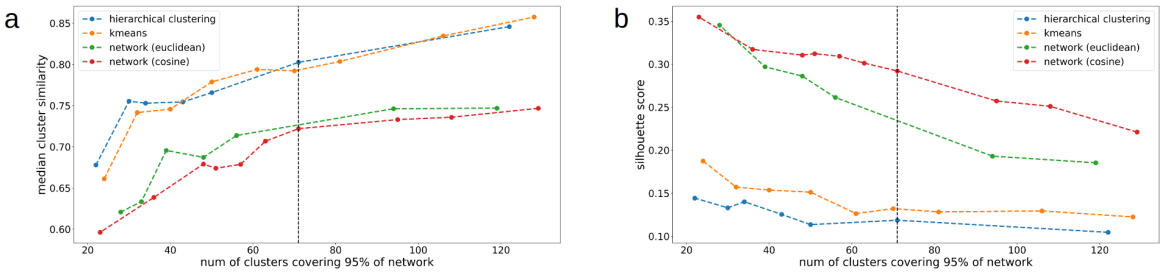

Figure 8. Median cluster similarity values (a) and silhouette scores (b) for different clustering methods and similarity measures used in the network analysis. The number of reduced dimensions after PCA is equal to 20, corresponding to 72% of retained information. The vertical black dashed line refers to the cluster granularity used in the paper. The colored dashed lines are shown for visualization of trends. Lower values of the median cluster similarity metric (a) correspond to better clustering performance. Higher values of the silhouette scores (b) correspond to better clustering performance.

Minor comments

COMMENT #6: l.5: please clarify the term “subject to degradation”

AUTHOR RESPONSE #6: The term “degradation” is often used in computational literature dealing with metrics in high dimensions and refers to the property of a distance metric to perform worse as the number of dimensions grows. This concept of “degradation” is related to the “curse of dimensionality”. It can be understood by recognizing that, counter to our intuition, what applies in three dimensions does not necessarily hold in higher dimensions. For example, in high dimensions, most of the mass of the points distributed according to a well behaved Gaussian distribution does not lie around the mean but becomes increasingly distant from it. Most of the mass migrates toward the surface of the domain leaving the bulk of the inner space empty. One of the consequences of this increased sparsity in high dimensional space is that the ratio of the distances of the nearest and farthest neighbors to a given point is almost 1, namely the points become uniformly distant from each other (Beyer et al., 1999).

To clarify, we added the following to the manuscript in Section 4.1, lines 481-484

“In particular, the distance metrics perform worse (referred to as "degradation") as the number of dimensions grows (Aggarwal et al., 2001). One of the consequences of this phenomenon is that the ratio of the distances of the nearest and farthest neighbors to a given point in high dimensions approaches 1, meaning that the points become uniformly distant from each other (Beyer et al., 1999).”

COMMENT #7: l.43, l.48 and in many other places: problems with in-line referencing.

AUTHOR RESPONSE #7: Thanks for pointing that out. We have corrected these references.

COMMENT #8: Section 2.3: I understand that traits values are standardized, but are their distributions normal? I guess no and I wonder how this may affect PCA and low dimensional vectors extracted from PCA.

AUTHOR RESPONSE #8: The referee's intuition about the non-normality behavior of the traits distribution is correct. The Shapiro-Wilk test, which checks if the data is drawn from a normal distribution, reveals that none of the traits are normally distributed when testing with a p-value of 0.05. However, the PCA method does not require normality in the input data. However, there are other factors that have to be considered when choosing PCA as the dimensionality reduction approach, which includes whether there are non-linear relationships between the variables, and whether the dataset contains outliers. To address these concerns, we added the following text in Section 2.3, lines 184-192:

*“The traits in the GAGES-II dataset contain significant redundancies, with 84% of pair-wise Pearson correlation coefficients (Pearson, 1895) and 92% of pair-wise Spearman coefficients, which accounts for non-linear relationships (Spearman, 1987), have a significant p-value of 0.05. The coefficient of determination between these two metrics is equal to 0.76, which indicates that although nonlinear relationships among the traits are present, they are not so dominant to prevent the use of a linear dimensionality reduction method such as PCA. Another factor that can affect the PCA algorithm’s performance is the presence of outliers. We determined that the PCA is a reasonable choice for the GAGES-II dataset, since only 8.1% of the traits lies outside their “inner fence”, a common threshold for outliers, defined as the range between Q1 - 1.5 * IQR and Q3 + 1.5 * IQR for each trait, where Q1 and Q3 are the first and third quartiles respectively, and IQR = Q3 - Q1 is the interquartile range. “*

COMMENT #9: l.473-475: Please clarify the added values of the network-based approach compared to other clustering techniques. Many of them address already the problem of dimensionality by working on Eigen-vectors.

AUTHOR RESPONSE #9: In response #4, we explain some of the benefits of our clustering approach. As shown in response #5, we have demonstrated that the network-approach outperforms other clustering techniques. Specific to the comment regarding the use of Eigen-vectors, we point out that although working with eigenvectors reduces the dimensionality of the problem, often the reduced vector space is still high dimensional. In our case, retaining 72% of the information - using the variance explained by the SVD singular values matrix - from the dimensionality reduction, still leads to a 20 dimension vector space. Although there is no universally accepted threshold for “high-dimensional” data, we argue that 20 constitutes a high number of dimensions where distance calculations are impacted.

Using a network approach allows one to choose the similarity metric and not to rely on euclidean distance, a metric needed in most of the traditional unsupervised clustering methods like k-means and hierarchical clustering. Thus we are able to use the cosine similarity metric that is less affected by issues of high dimensionality and include directionality information in the data. This is explained in the Method section 2.11, Result section 3.6, and Discussion section 4.1.

The resulting clusters from our network approach are computed using the information from both the transformed matrix and the principal components. As illustrated in response #4, we believe that our workflow produces more interpretable results than the ones that would be obtained using only the Eigenvectors obtained from a PCA. This is because the network approach allows to separate traits or catchments into distinct clusters, produced using network connections statistics validated by the disparity filter introduced in line 231 in the original manuscript. Conversely the contributions of traits or catchments is generally distributed among multiple elements of the transformed matrix and the principal components in the PCA, making it difficult to produce clear categories or connect groups of traits to specific hydrological behaviors.

COMMENT #10 Figure 13: what is the unit of MA41?

AUTHOR RESPONSE #10: The streamflow index identified by the code MA41 refers to the "mean annual flow divided by catchment area" and its dimensions currently are $m^3.s^{-1}.km^{-2}$. We acknowledge that this is an uncommon choice and have changed into the more commonly used mm/day. Units for MA41 are explicitly indicated in Figure 14 axis and caption of the revised manuscript.

Response to Reviewer 2 Comments:

COMMENT #11 - This article describes the application of a post-PCA clustering algorithm for classification, in this case for catchments. There is no strong argument that the technique is much better than other methods in this particular application, but the breadth, quality and density of the GAGES-II dataset make it an attractive test bed.

The authors do not apply any effort in showing the improvement their technique makes over others. For example, the justification for their network-based approach is a single paragraph and three numbers. In a more structured analysis, the differences between PCA only, and each of the three post-PCA clustering techniques, would be outlined and their differences tabulated with relevant measures (with an equivalent of Figure 3 for each). There would also be a baseline measure, the PCA or one clustering technique with a minimum number of clusters, and some limited exploration of the number of clusters (or the two free parameters mentioned).

AUTHOR RESPONSE #11: We acknowledge that the manuscript we submitted does not provide sufficient evidence that our proposed method based on networks and cosine similarity performs better than traditional unsupervised clustering algorithms. We thank the reviewer for this suggestion. See response #5 where we performed a more comprehensive analysis comparing the performance of our method against benchmark hierarchical clustering and k-means approaches, as well as exploration of the free parameters.

COMMENT #12 - It is not remarkable (line 579) that a classification method using indices and data from a database (of over 300 measures on over 9000 catchments) specifically designed to described gauged catchments for evaluating streamflow would result in a classification that was related to streamflow measures. It will be no surprise to hydrologists that high rainfall, high elevation, forested catchments behave hydrologically differently to flatter, lower rainfall, cropland areas, or that higher rainfall catchments with lots of urban areas get more flooding. What the results might show however is the bidirectionality such that starting from the stream flow indices we get catchment clusters, and that starting from catchment traits we can get groups of catchments with distinct flow behavior.

AUTHOR RESPONSE #12: Thanks for this comment. Our original write up was intended to highlight that the approach produces intuitive results that are immediately obvious to all readers. Based on this comment, and those from reviewer #1, we have expanded the analysis to include some additional hydrological insights that can be gained with this methodology. In particular we expanded our discussion Section 4.2 and Section 4.5 (see response #4 for changes made to the manuscript). We also removed the text referring to our results as remarkable.

The issue of bidirectionality is interesting but beyond the scope of this paper. We are working on building models to predict hydrological indices using trait clusters, and understanding the traits of signature-based classification as part of multiple follow-on studies.

COMMENT #13: What would also have been of interest is the places where the flow indices and clusters do not match well. For example, if there are two areas that are low slope, low elevation cropland that have distinctly different baseflow regime, one may be influenced by groundwater discharge or a factor not yet captured, and this would be useful additional data to know or require to be collected.

AUTHOR RESPONSE #13: Thanks for the suggestion and we agree that it is interesting to investigate subsets of catchments within a cluster where the flow indices do not match well. To investigate this aspect, we performed a new analysis that focuses on anomalies in the hydrologic indices within catchment clusters. We have added the following text to the discussion as a new Section 4.6 “Examining diversity of hydrologic behavior within catchment clusters” in lines 639-689:

“We can also use our methodology to gain insights into the traits that may result in diversity of hydrologic behavior within similar catchments. For this purpose we selected catchment subsets that are considered outliers - i.e. where the index is either above the 90th percentile or below the 10th percentile of all indices in the cluster, for each streamflow index and for each cluster of catchments. We compare the z-scores of the traits associated with the catchment subsets relative to the entire catchment cluster to evaluate whether there are differences in traits that would explain the anomalous hydrological behavior. As an example, we look at catchments within a cluster that have distinct baseflow regimes, based on a baseflow index (ml17) in Olden and Poff (2003) that represents the 7-day minimum flows divided by mean annual daily flows. The results for anomalous catchments have higher than normal (>90th percentile) baseflow are shown in Fig. 15, where the size and color of the bubbles are the relative z-scores of the trait categories.

We focus on a crop-dominated catchment cluster such as the one generally encompassing the Ohio Valley region (cluster 2), displayed in the third row of the bubble plot. This cluster is characterized by relatively low elevation, presence of croplands and fine soil as indicated by the higher z-scores of these trait categories relative to the rest of the CONUS catchments (Fig. 16a). Using our approach, we can identify the over and under expressed traits of the catchments with anomalously high baseflows in cluster 2 that generally has low elevation croplands. In Fig. 15, we find there is a positive association of high baseflows with coarse soils (Z=0.98) and a negative one with fine soils (Z=-0.51), which is not surprising. In addition, there is an association of high baseflows with the “Non-cropland” trait category (third last column with green label, Z=0.85), which aggregates all non-agricultural land use such as urban areas and forests. This indicates that within the context of a cropland-dominated cluster, the catchments that have relatively lower areas of croplands have higher baseflows. Interestingly, there is also a strong positive association of high baseflows with shrubland (Z=1.12) and a moderate negative association with temperature (Z=-0.65). One possible explanation for these results is that pumping groundwater for agriculture decreases the groundwater input into streams resulting in lower baseflows. This depletion of groundwater discharge into streams does not occur in shrublands or other areas without croplands.

Another catchment cluster with a strong agricultural presence is cluster 14, generally located in North and South Dakota, which are characterized by low temperatures, herbaceous wetlands and croplands (Fig. 16b). Similar to cluster 2, there is a positive association of anomalously high baseflows with coarse soils (Z=0.76) and non-croplands (Z=1.10), and a negative association with fine soils (Z=-0.84). However, in comparison to cluster 2, several other factors have a positive association with high baseflows including precipitation/summer precipitation (Z=0.90, Z=1.04 respectively), the presence of lakes, ponds and reservoirs (Z=1.19), herbaceous wetland areas (Z=0.76), evergreen/mixed/deciduous forests (Z=0.58, Z=0.74, Z=0.88 respectively), and developed areas (Z=0.66). There is also a negative association with overland flows (Z=-0.82). This

reveals that, in catchment cluster 14, anomalously high baseflows are more likely in the presence of surface water bodies such as lakes and wetlands that have the potential for increased surface-groundwater exchange. High baseflows also occur in forested areas of these agricultural catchments, potentially indicating that the partitioning of precipitation is weighted towards infiltration and recharge over evapotranspiration in these catchments.

Overall, averaged z-scores for all catchments in the CONUS (shown in the last row of Fig. 15) indicates there is a moderate positive association of anomalously high base flows with the presence of lakes, ponds and reservoirs (11th column in light blue, Z=0.29), and with coarse soils and groundwater trait categories (18th column in gray, Z=0.40). Conversely, there is a negative link to fine soils (13th column in green, Z=-0.25). This indicates the potential for surface-groundwater exchange in regions where water bodies are present, and not surprisingly the importance of soil texture in mediating baseflow through infiltration and recharge."

Figure 15. Bubble plot showing the z-scores of catchments where the baseflow is above the 90th percentile relative to the entire catchment cluster. The baseflow index is computed as the seven-day minimum flow divided by mean annual daily flows (averaged across all years). Bubble size is proportional to the absolute value of the z-score. Colors separate positive from negative values as indicated by the colorbar. Catchment clusters are displayed on the vertical axis using an identifier consistent with the one used in the original paper, a name describing their main characteristics, their approximate geographical area (if applicable), and the number of anomalous catchments above

the 90th percentile shown in parenthesis in parenthesis. Only clusters with an anomalous set of catchments larger than 10 are included. Traits categories are displayed on the horizontal axis and are sorted in descending order, according to their size in terms of number of nodes in the traits network, and colored consistently with the trait clusters in said network. The last row of each plot refers to the average value of the trait z-scores of the clusters displayed in the plot and provides an idea of how much a trait category is over or under expressed across different clusters with different characteristics.

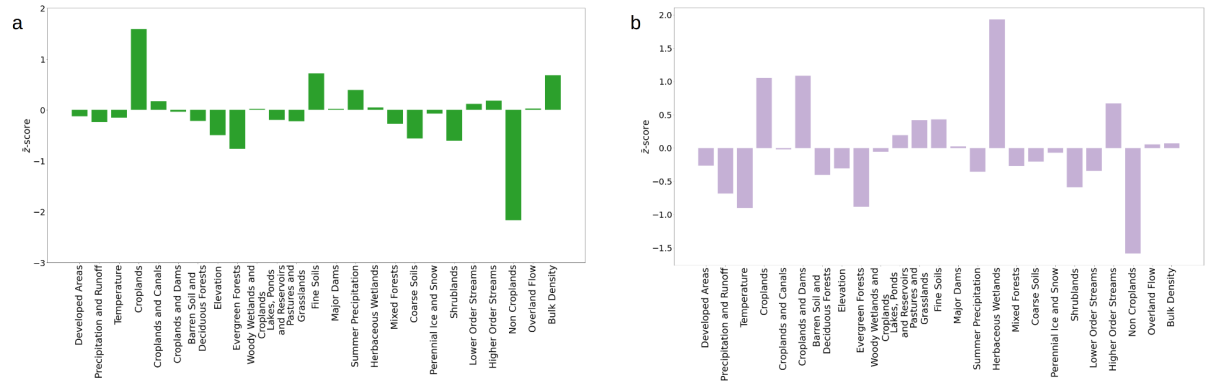

Figure 16. (a) Bar chart of traits z-scores of for the catchment cluster 2, characterized by croplands and fine soils. The catchments in this cluster are generally located in the Ohio Valley region. (b) Bar chart of traits z-scores of for the catchment cluster 14 , characterized by low temperatures, croplands and wetlands. The catchments in this cluster are generally located in North and South Dakota.

COMMENT #14: The citing of references within the text is inconsistent and non-standard, while many of the listed references do not use capital letters where appropriate in journal names or proceedings.

AUTHOR RESPONSE #14: Thanks for pointing that out. We have corrected the references.

REFERENCES IN OUR RESPONSES:

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P., 2018. A ranking of hydrological signatures based on their predictability in space. Water Resources Research, 54, pp.8792–8812.

Beyer K., Goldstein J., Ramakrishnan R., Shaft U., 1999. When is Nearest Neighbors Meaningful? ICDT Conference Proceedings

Eng, K., Grantham, T.E., Carlisle, D.M. and Wolock, D.M., 2017. Predictability and selection of hydrologic metrics in riverine ecohydrology. Freshwater Science, 36(4), pp.915-926.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, pp.53-65.

Olden, J.D. and Poff, N.L., 2003. Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River research and applications, 19(2), pp.101-121.