Supplement of Hydrol. Earth Syst. Sci., 26, 429–444, 2022
https://doi.org/10.5194/hess-26-429-2022-supplement

Hydrology and Earth System Sciences
Open Access

Supplement of

Watershed zonation through hillslope clustering for tractably quantifying above- and below-ground watershed heterogeneity and functions

Haruko M. Wainwright et al.

Correspondence to: Haruko M. Wainwright (hmwainw@mit.edu)

Texts S1. Airborne Data Descriptions

The National Ecological Observation Network (NEON) team collected airborne hyperspectral and LiDAR data in June 2018, which corresponds roughly to the peak growing season in these catchments. The pre-campaign planning, acquisition procedures, and postprocessing are described in detail by Chadwick et al. (2020). The onboard instrumentation includes Airborne Visible-Infrared Imaging Spectrometer - Next Generation (AVIRIS-NG), which measures surface reflectance over the range of 380-2510 nm with the spectral resolution of 5 nm (426 spectral bands). After the standard postprocessing, we used data products including aboveground biomass, Leaf Area Index (LAI), Normalized Difference Vegetation Index (NDVI), and Normalized Difference Water Index (NDWI).

NASA Airborne Snow Observatory (ASO) data were collected in April 2018, which corresponds to the peak snow timing. ASO couples an imaging spectrometer and scanning LiDAR, as well as distributed snow modeling, to estimate snow depth and snow-water-equivalent (SWE) over the watersheds (Painter et al., 2016). (These datasets are publicly available at https://nsidc.org/data/aso.) From these data products, we incorporated the peak SWE map at 50 m resolution (https://www.jpl.nasa.gov/missions/airborne-snow-observatory-aso/).

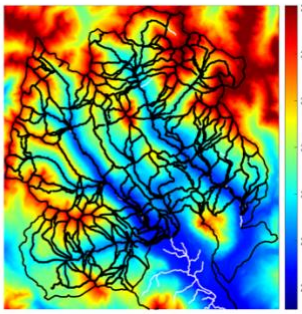
An airborne electromagnetic (AEM) survey was conducted in fall 2017. The detailed descriptions are available in Zamudio et al. (2020) and Uhlemann et al. (submitted). We used the electrical resistivity averaged over the top 20 meters from the land surface as near-surface electrical resistivity, which is considered as a proxy for bedrock properties, including fracture density, weathering, and hydrothermal alternation. Since the data does not cover the entire domain, we gap-filled the datasets by assigning the average value of the same geology at each pixel.

Texts 2. Characteristics of cluster methods

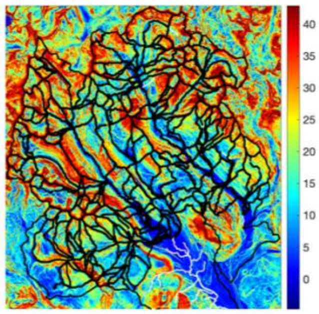
- K-means (KM) is the most commonly used clustering method. The K-means algorithm identifies clusters by minimizing an objective function called inertia, which is the total average intra-cluster distance between samples assigned to a cluster and its centroid. The K-means algorithm starts from a random set of starting points. A sufficient number of random sets are used so that clusters converge.
- Hierarchical clustering (HC) is a decision-tree-based method that divides data points based on a series of binary splits. It is agglomerative clustering, which begins by considering each data point as an individual cluster, then progressively merges clusters at increasing levels of dissimilarity until all data points form one large cluster (Kassambara, 2017). To define the linkage (or the distance) between any two clusters, we used the Ward method, which has been successfully used for ecosystem zonation at the hillslope scale (Devadoss et al., 2020). Ward's method computes the variance within each cluster, measuring the distance between each observation and the cluster's mean, and then taking the sum of the distances' squared.
- Gaussian mixture models (GMM) are often called soft clustering, which allows us to determine a probability or membership degree for each sample to belong to each cluster, instead of deterministic clusters. The distribution of datasets in the multidimensional space is approximated by the weighted sum of multivariate Gaussian distributions. The expectation maximization algorithm is used to maximize the likelihood of parameters that define a Gaussian distribution for each cluster. In addition, GMM evaluates the number of clusters based on the Bayesian Information Criterion (BIC), which is a function of the maximized likelihood and the number of parameters (i.e., the number of clusters). Geometric features (shape, volume, orientation) of each cluster can be specified or optimized through BIC as EII, VII, EEI, VEI, EVI, VVI, EEE, EEV, VEV, and VVV. E stands for "equal", V for "variable" and I for "coordinate axes" in the order of volume, shape and orientation.

Table S1. The list of 16 hillslope features.

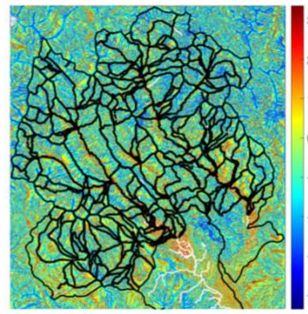
	Variable	Computation in each hillslope
1	Elevation	Hillslope average of elevation (LiDAR)
2	Slope	Hillslope average of slope (LiDAR)
3	Relief	The difference between the minimum and maximum elevation (LiDAR)
4	TWI	Hillslope average of TWI (LiDAR)
5	Radiation	Hillslope average of annual net potential radiation (LiDAR)
6	Resistivity	Hillslope average of top-20-m bedrock resistivity (AEM)
7	NDVI	Hillslope average of peak NDVI (NEON hyperspectral + LiDAR) in 2018
8	NDWI	Hillslope average of NDWI (NEON hyperspectral + LiDAR) corresponding to the peak phenology timing in 2018.
9	Biomass	Hillslope average of peak aboveground biomass (NEON hyperspectral + LiDAR) in 2018
10	SWE	Hillslope average of peak SWE (NASA ASO) in 2018
11	%Igneous	Percent coverage of igneous rock
12	%Sandstone	Percent coverage of sandstone
13	%Evergreen	Percent coverage of evergreen forest
14	%Aspen	Percent coverage of aspen
15	%Barren	Percent coverage of bare ground
16	%Clay	Hillslope average of %Clay
17	%Sand	Hillslope average of %Sand



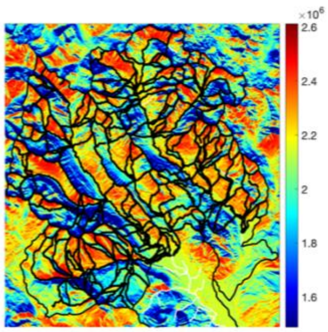
(a)



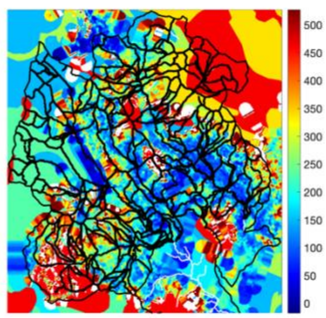
(b)



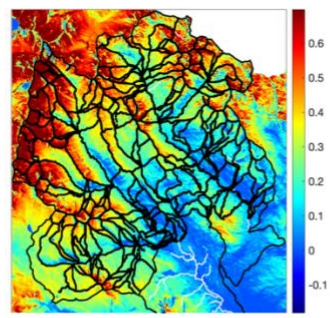
(c)



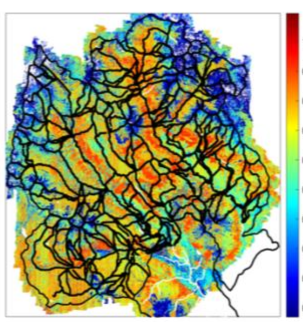
(d)



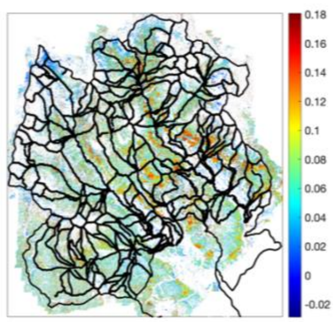
(e)



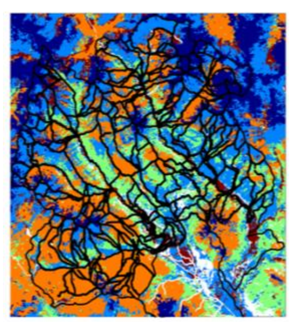
(f)



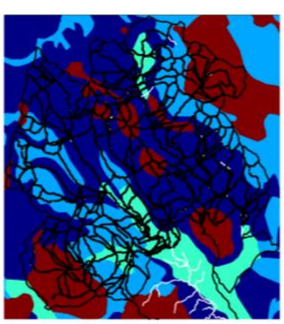
(g)



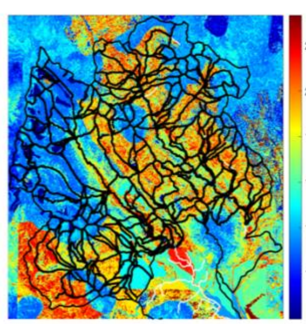
(h)



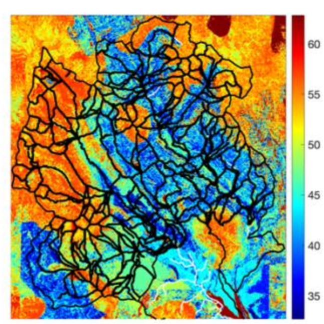
(i)



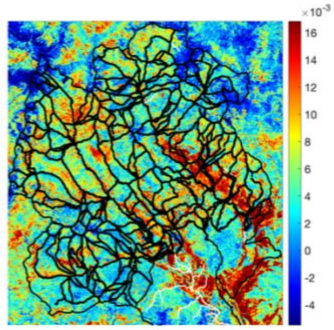
(j)



(k)



(l)



(m)

Figure S1. Spatial data layers: (a) elevation in meter, (b) slope in degrees, (c) log10-TWI, (d) net potential radiation in W/m^2, (e) top 20-m electrical resistivity in Ohm. m, (f) 2018 peak SWE in meters, (g) NDVI in July2018, and (h) NDWI in July 2018, (i) 2011 NLCD landcover map, and (j) geological map, (k) %clay, (l) %sand, and (m) foresummer drought sensitivity from Wainwright et al. (2020). In (i), the dark blue region is barren, the right blue is grassland, the green is aspen forest and the orange is evergreen forest, and the red is a riparian region. In (j), the dark blue region is shale, the light blue is sandstone, the red is igneous rock, and the white-green is glacial deposit.

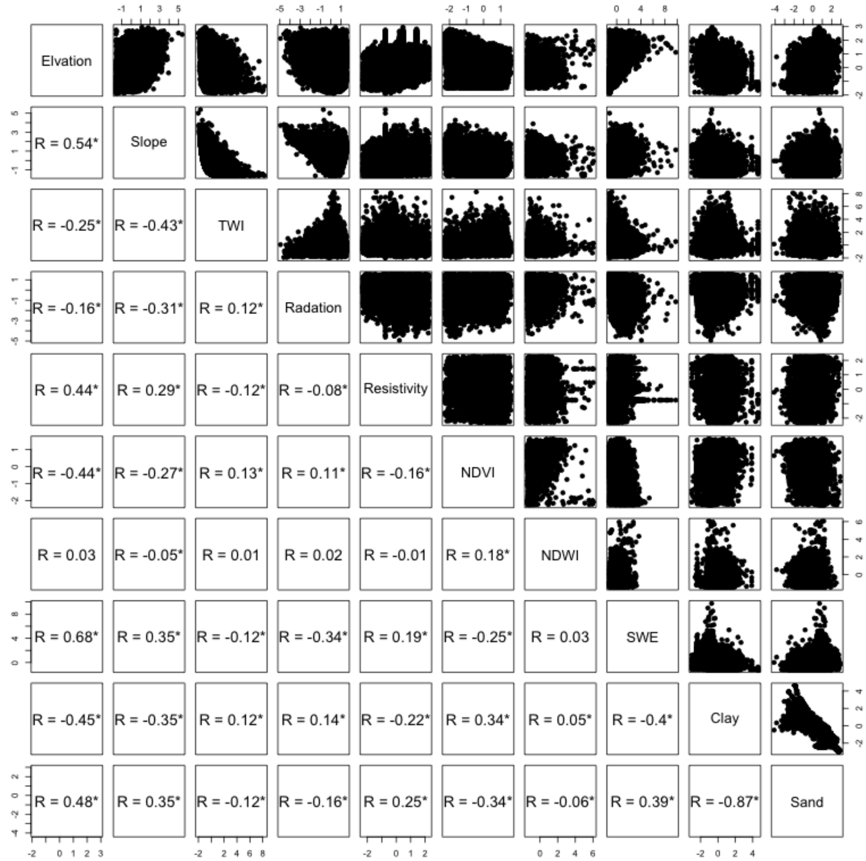


Figure S2(a). Pixel-by-pixel correlation (the grid resolution of 9 m) and scatter plots (Pearson's correlation coefficients) among selected metrics. The * sign represents that the p-values are smaller than 0.01.

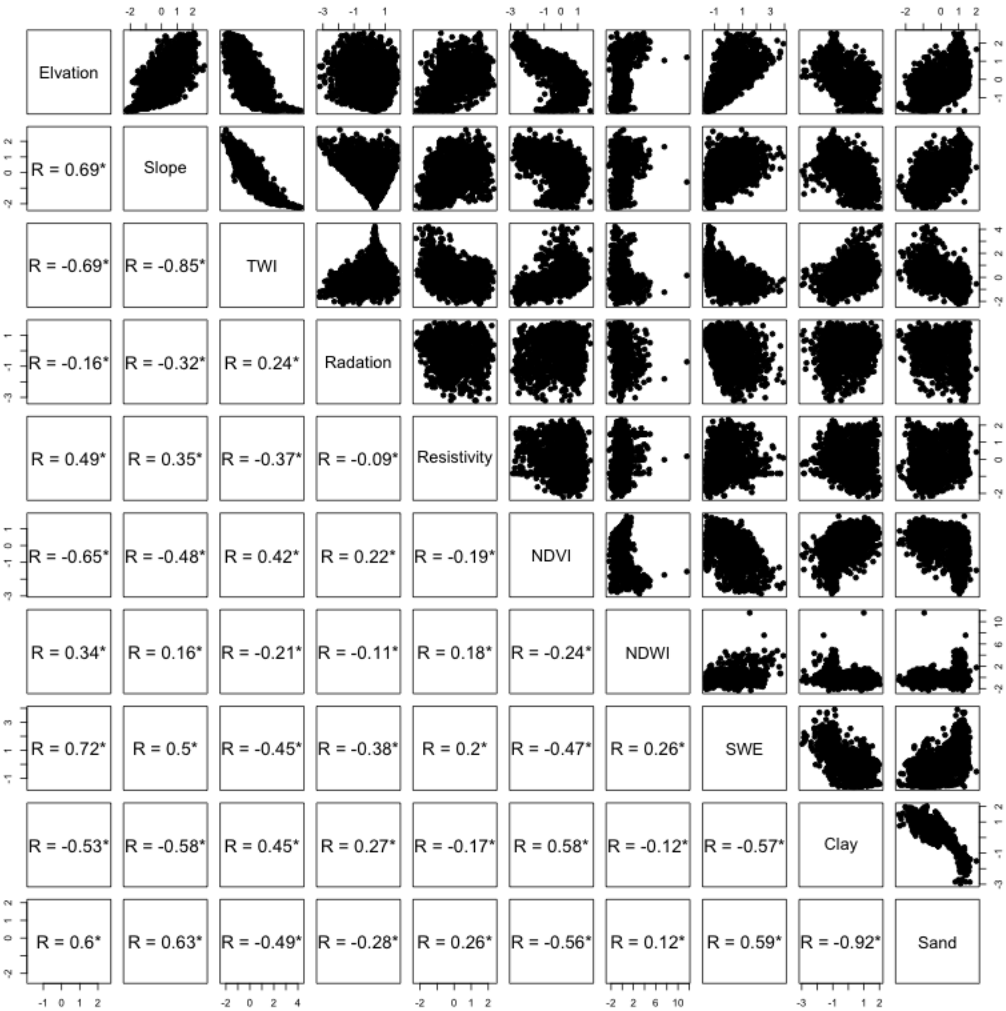
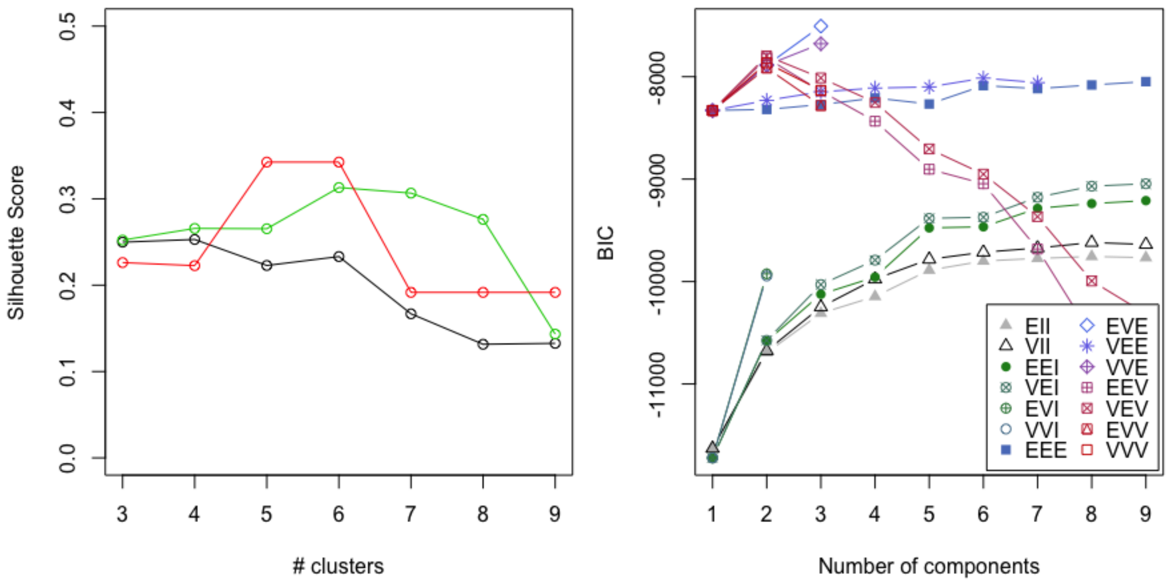
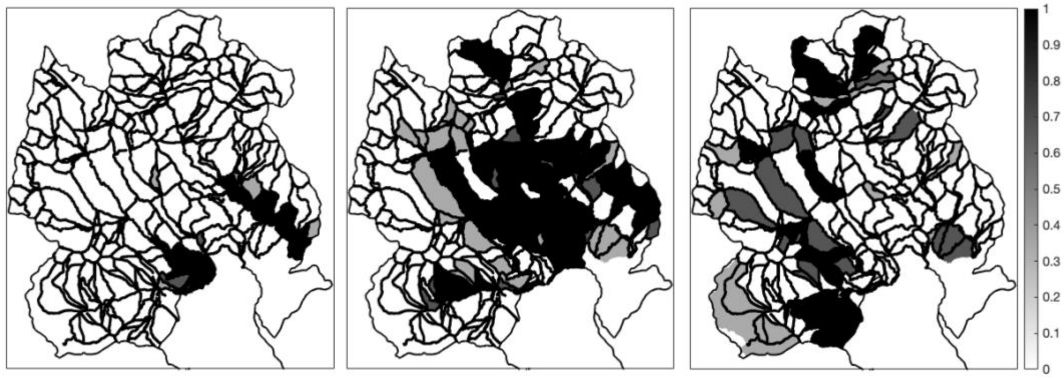


Figure S2(b). Upscaled pixel-by-pixel correlation and scatter plots at the grid resolution of 100 m (Pearson's correlation coefficients) among selected metrics. The * sign represents that the p-values are smaller than 0.01.

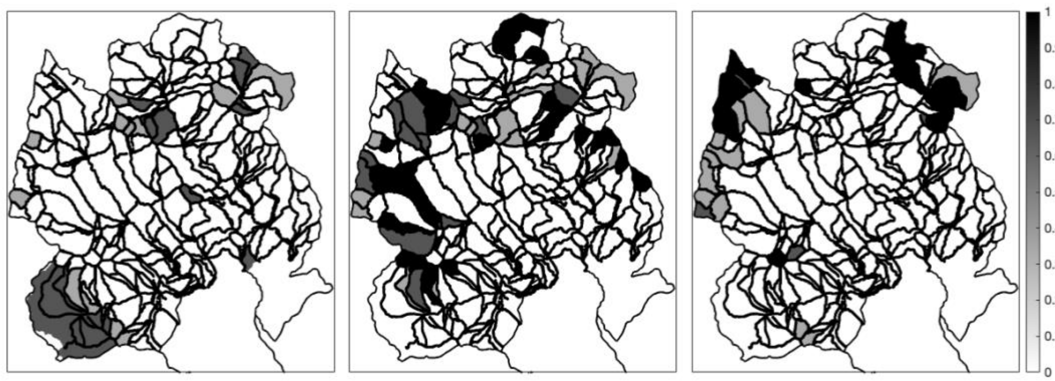


(a) (b)

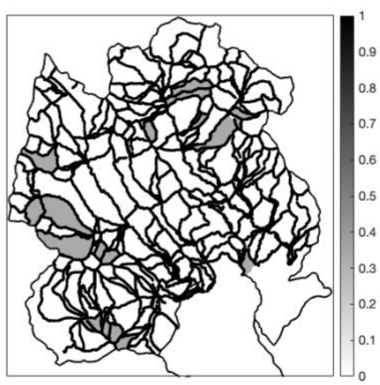
Figure S3. Cluster metrics: (a) silhouette score, and (b) Bayesian Information Criteria (BIC) as a function of the number of clusters. In (a), the black line is GMM, the red line is KM, and the green line is HC. In (a), the silhouette score is defined in the range [−1, +1]; a higher value means that data points are closer in its own clusters than with the ones in other clusters. In (b), geometric features (shape, volume, orientation) of each cluster can be specified or optimized through BIC as EII, VII, EEI, VEI, EVI, VVI, EEE, EEV, VEV, and VVV. E stands for “equal”, V for “variable” and I for “coordinate axes” in the order of volume, shape and orientation.



(a) (b) (c)



(d) (e) (f)



(g)

Figure S4. Frequency of each hillslope being categorized into each zone across the three clustering methods: (a) Zone 1, (b) Zone 2, (c) Zone 3, (d) Zone 4, (e) Zone 5, (f) Zone 6, and (g) Zone 7. The frequency is normalized to one such that it is one when the three methods agree with each other.

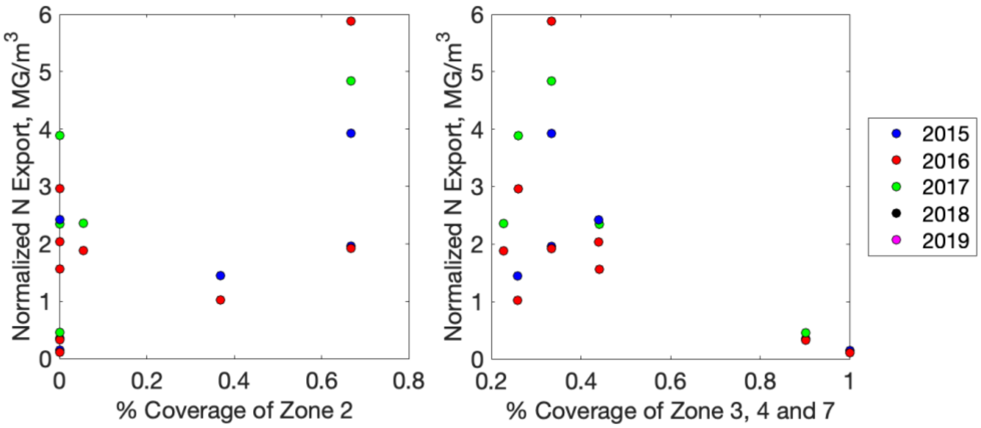


Figure S5. Normalized N export (N export divided by annual discharge) as a function of the percent coverage of (a) Zone 2 and (b) Zone 3, 4 and 7(the conifer dominated zone).