Clustering approaches for analysing similarity in ungauged catchments:
input variable selection for hydrological predictions

Nilay Dogulu[1], Inci Batmaz[2], and Elcin Kentel[1]

1 Middle East Technical University, Department of Civil Engineering, Water Resources Laboratory, Ankara, Turkey
2 Middle East Technical University, Department of Statistics, Ankara, Turkey

1 Background

Catchments are hydrological units that exhibit unique but distinct features that greatly contribute to **heterogeneity and complexity** of rainfall-runoff processes.

While the lure of understanding such diversity has underpinned the focus of many research efforts in hydrology, including **predictions in ungauged basins**, there is still room for improving our ability to benefit from this diversity in the context of **data-driven hydrologic regionalization**.

An outstanding issue in this line of research concerns enhanced utilization of knowledge on dominant factors affecting catchments' hydrologic response behaviour under different types of streamflow.

Research Objectives:

To explore the potential value of different clustering methods in identifying similar groups of catchments

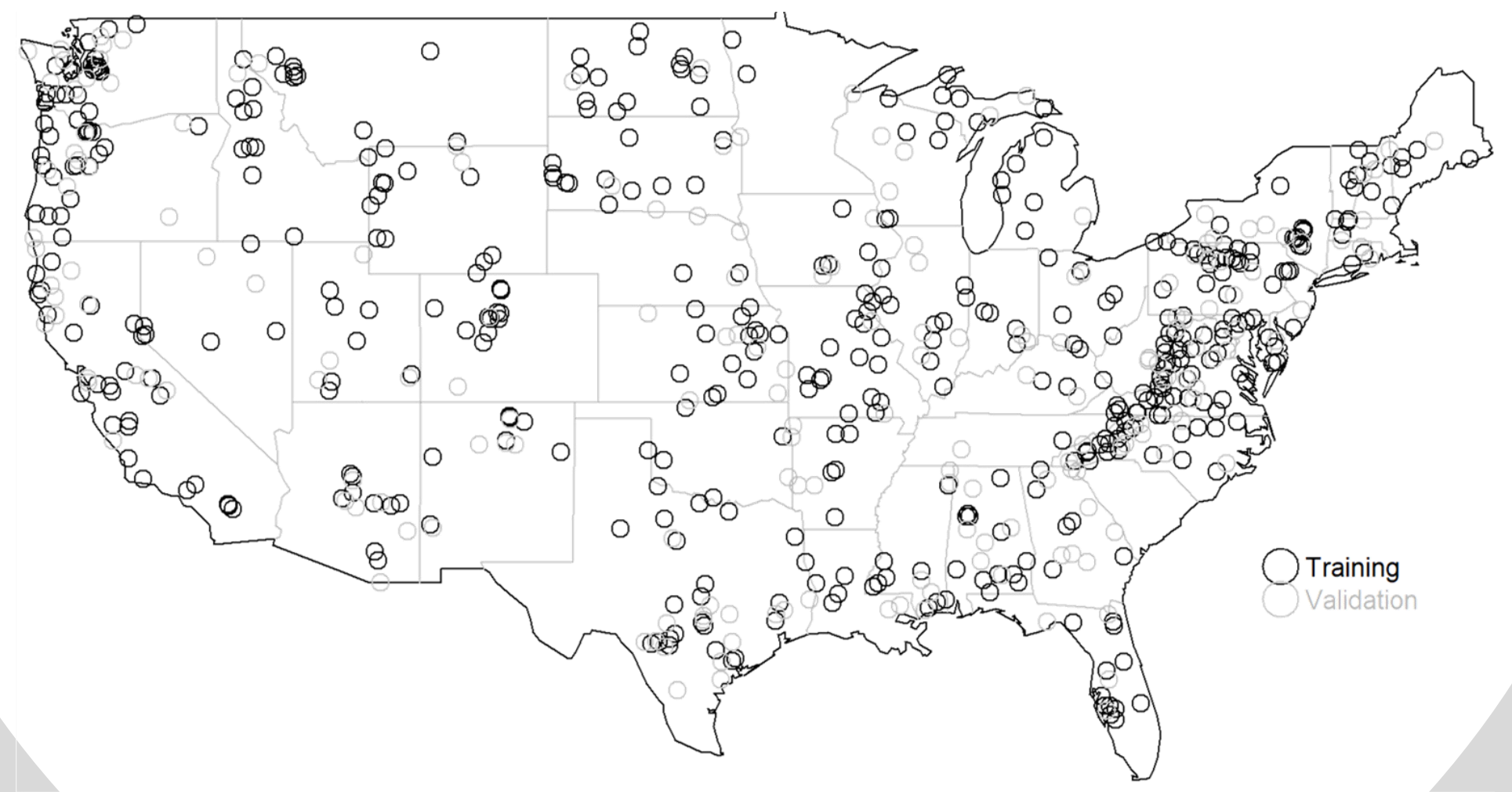
To determine input variables that control streamflow predictability within each group of catchments and over different runoff attributes representing particular hydrological conditions

2 Where?

CAMELS

Catchment **A**ttributes and **ME**teorology for **L**arge-sample **S**tudies
Addor et al., 2017, HESS

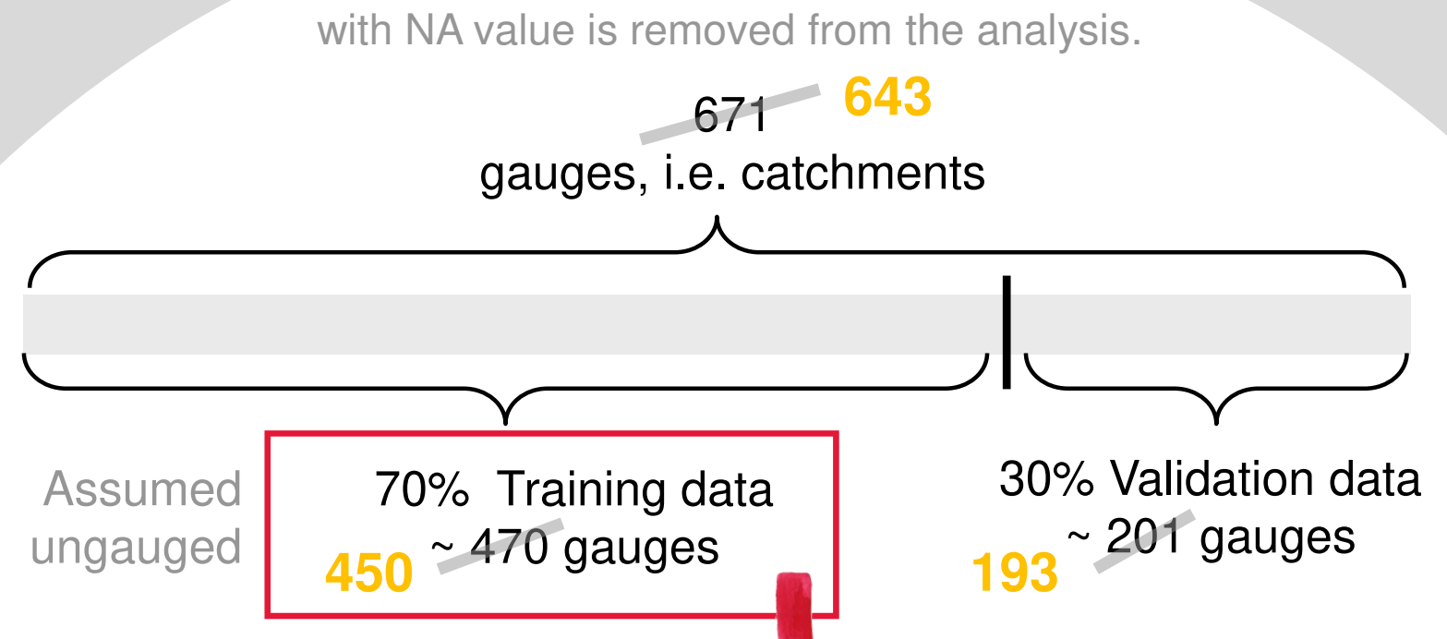
671 watersheds across continental USA
(unimpacted / less impacted by anthropogenic changes)



Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017) **The CAMELS data set: catchment attributes and meteorology for large-sample studies**. Hydrology and Earth System Sciences, 21(10), 5293-5313. https://www.hydrol-earth-syst-sci.net/21/5293/2017/

3 Data

Catchments with at least 1 attribute with NA value is removed from the analysis.

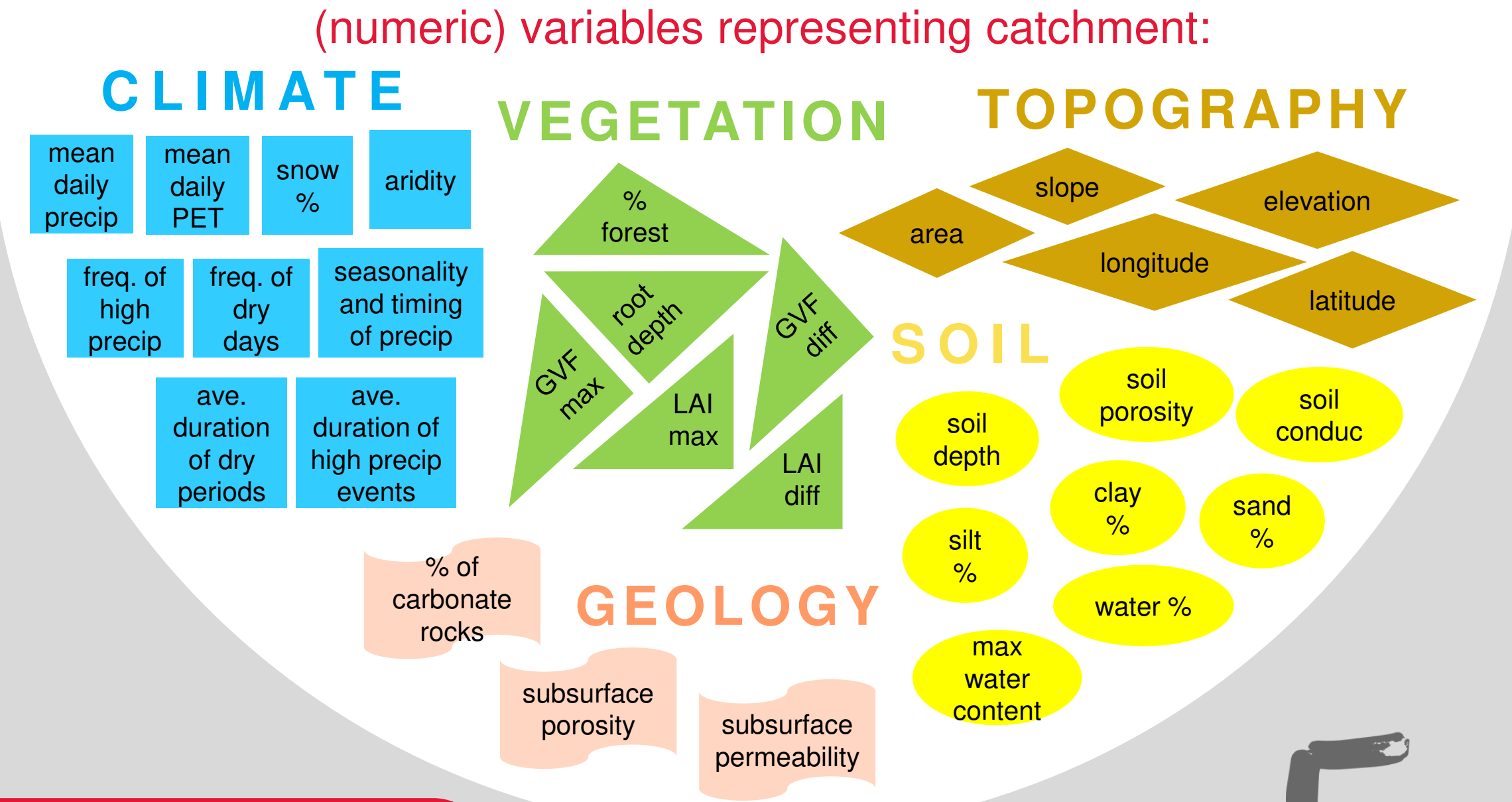


~~671~~ 643 gauges, i.e. catchments

Assumed ungauged: 70% Training data ~ ~~470~~ 450 gauges

30% Validation data ~ ~~201~~ 193 gauges

Clustering is performed on input space consisting of **31** (numeric) variables representing catchment:



CLIMATE: mean daily precip; mean daily PET; snow %; aridity; freq. of high precip; freq. of dry days; seasonality and timing of precip; ave. duration of dry periods; ave. duration of high precip events

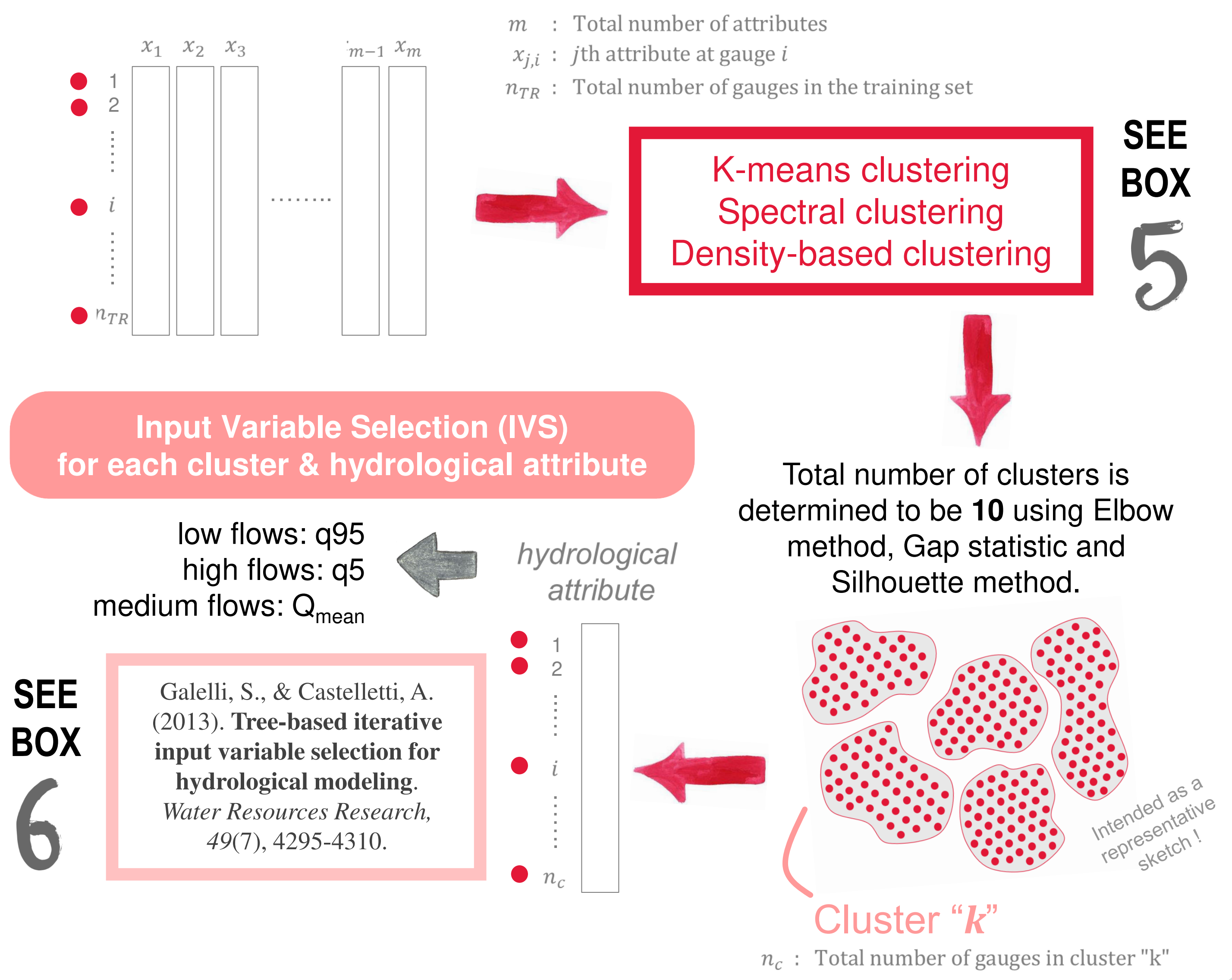
VEGETATION: % forest; root depth; GVF max; LAI max; GVF diff; LAI diff

TOPOGRAPHY: slope; elevation; area; longitude; latitude

SOIL: soil porosity; soil conduc; soil depth; clay %; sand %; silt %; water %; max water content

GEOLOGY: % of carbonate rocks; subsurface porosity; subsurface permeability

4 Methodology



Clustering of catchments using available topography, soil, geology, vegetation and climate attributes

m : Total number of attributes
$x_{j,i}$: jth attribute at gauge i
n_{TR} : Total number of gauges in the training set

K-means clustering
Spectral clustering
Density-based clustering

SEE BOX 5

Total number of clusters is determined to be **10** using Elbow method, Gap statistic and Silhouette method.

n_c : Total number of gauges in cluster "k"

Input Variable Selection (IVS) for each cluster & hydrological attribute

low flows: q95
high flows: q5
medium flows: Q_{mean}

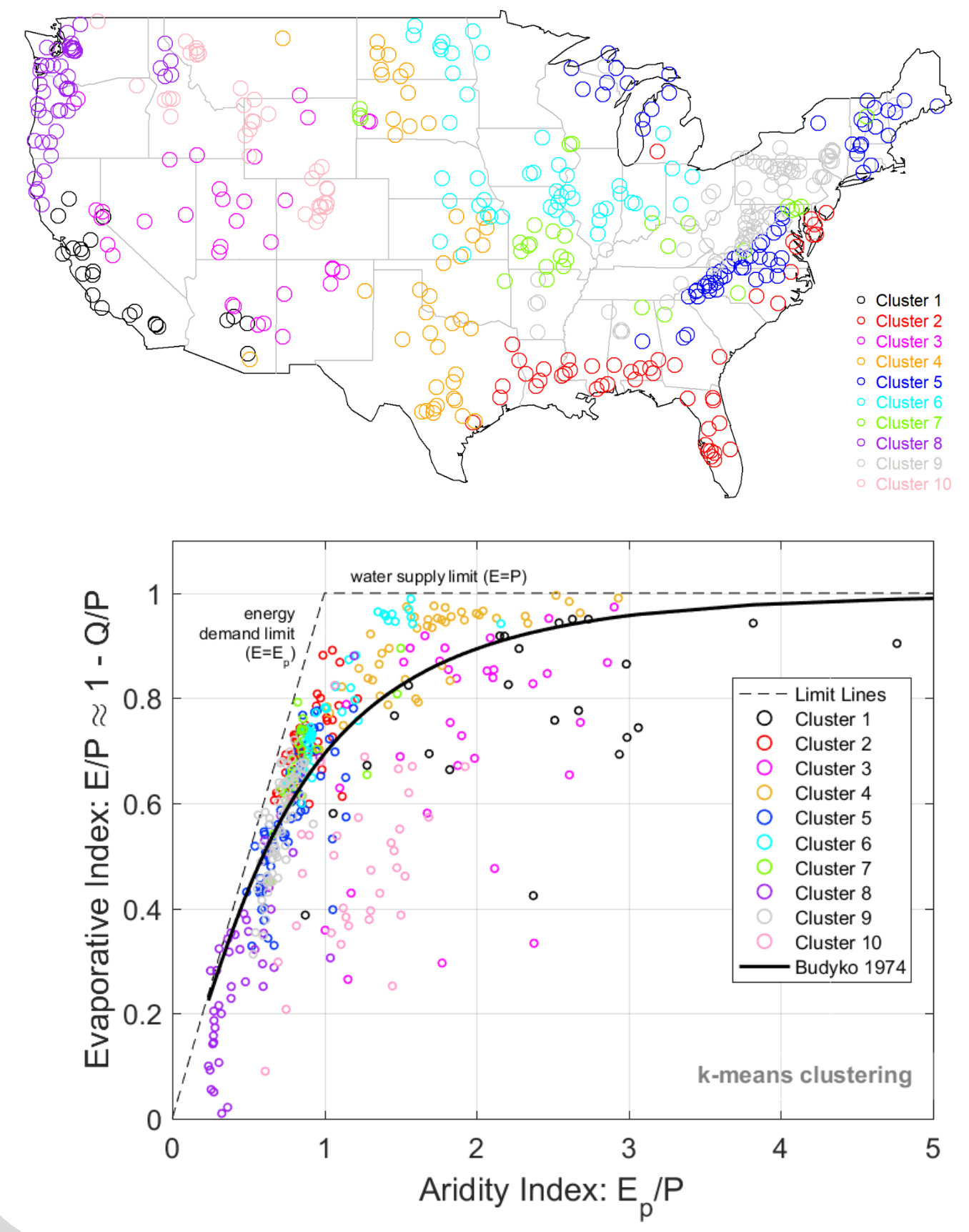
SEE BOX 6

Galelli, S., & Castelletti, A. (2013). **Tree-based iterative input variable selection for hydrological modeling**. *Water Resources Research*, *49*(7), 4295-4310.

5 Catchment clustering

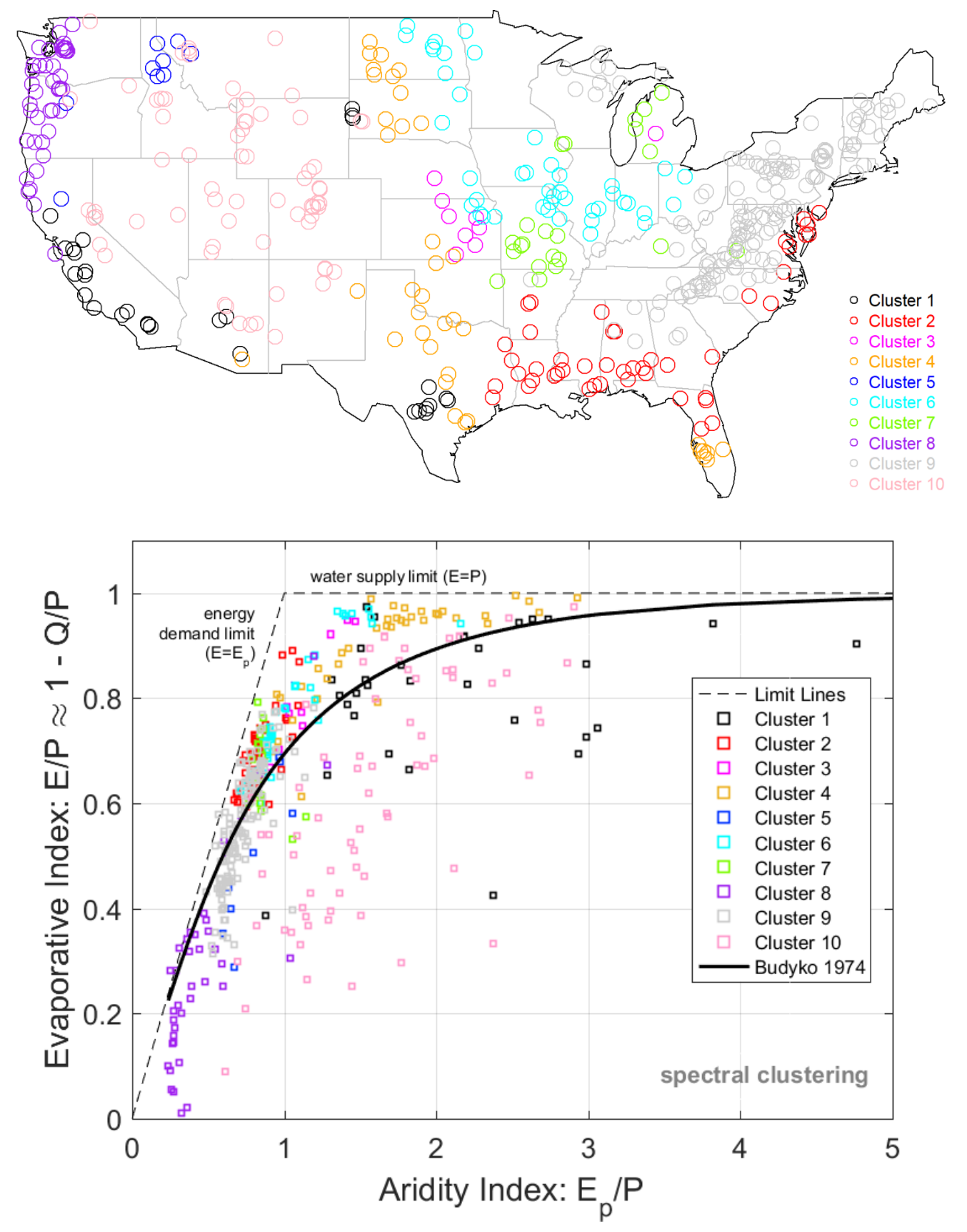
K-means clustering

- Most well-known traditional clustering method based on a **center-based partitional algorithm**. Number of clusters, k, is a user-specified parameter.
- Tends to produce **clusters of roughly equal size**. Clustering depends greatly on the initial choice of cluster centers.
- Emphasizes **homogeneity rather than separation**; it is usually more successful regarding small within-cluster dissimilarities than regarding finding gaps between clusters.
- **Not capable of forming clusters with non-convex shapes**.
- Efficient for **large data sets**, only works on numerical data.
- **Unable to handle noisy data and outliers**.
- The algorithm of Hartigan and Wong (1979) with **k= 10** is used.



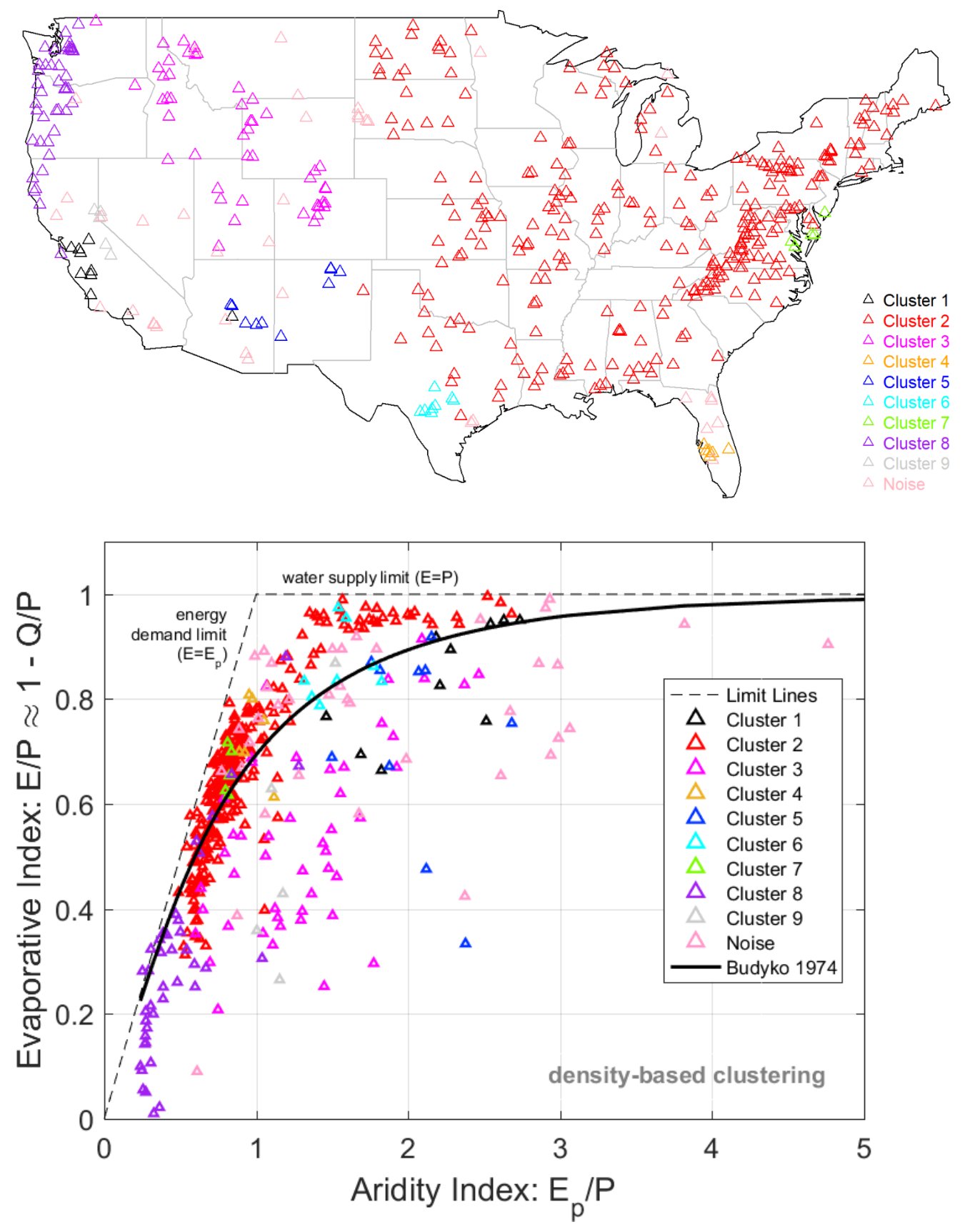
Spectral clustering

- A hybrid clustering method based on singular value decomposition and **k-means.**
- Clustering is performed by embedding the data into **the subspace of the eigenvectors of an affinity matrix**.
- Concerned with the **similarity between data points in different clusters**, rather than dispersion within a cluster.
- Aims to cluster data that is connected but not necessarily compact or clustered within convex boundaries. **Allows clusters to have arbitrary shapes.**
- Extensive validation on real world applications remains a big challenge due to its **high computational cost**.
- The algorithm of Ng, Jordan and Weiss (2002) with **k =10** is used.



Density-based clustering

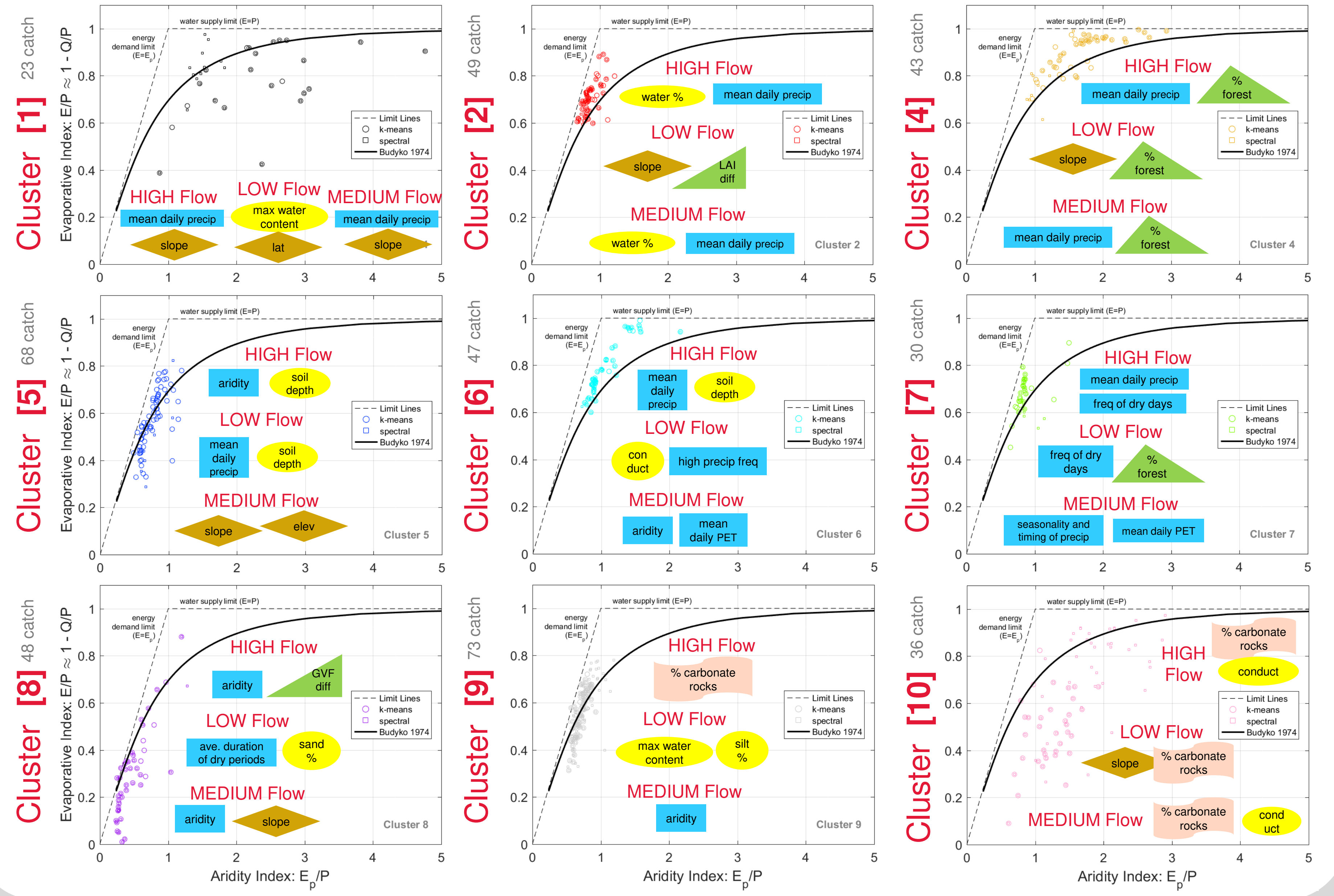
- Able to find **arbitrarily shaped clusters**, where clusters are defined as dense regions separated by low density regions.
- Based on the concepts of **density reachability and connectivity**.
- Threshold of neighbourhood of a data point must be specified by the modeller. Density in a neighbourhood for a data should be high enough if it belongs to a cluster.
- Can **handle noisy data.**
- The algorithm used is of Ester et al. (1996) – DBSCAN (Density-Based Spatial Clustering of Applications with Noise):
  Eps = 0.7 (based on the k-nearest Neighbour distance plot) – max radius of the neighbourhood)
  minPts = 5 (default) – min number of points within the Eps neighbourhood.



6 Input Variable Selection (IVS)

Iterative Input Selection method (IIS) method proposed by Galelli and Castelletti (2013) is used. In this method, one input variable is selected at each iteration on the basis of the partial dependence between each input variable, and the output relies on a tree-based ranking method to estimate the information gained from the data. As a regression method, Extra-Trees (Geurts et al., 2006) is employed for both ranking and modelling. MATLAB toolbox is available through the IVS4EM project (Galelli et al., 2014).

The results are presented for the clusters identified by the k-means method for three hydrological attributes: 5% flow percentile for **HIGH** flows, 95% flow percentile for **LOW** flows, mean daily discharge for **MEDIUM** flows.



Summary

- IVS is important for hydrological predictions in ungauged catchments! And clustering of input data space helps!
- Input variables that control streamflow predictability at ungauged locations can vary significantly:
  - over different runoff attributes representing particular hydrological conditions.
  - among different groups of catchments as identified by catchment clustering.

Come also to see my poster "Input variable selection for hydrological predictions in ungauged catchments: with or without clustering?" @A.7 on Wed, 11 Apr, 17:30–19:00 Hall A (HS 1.10 Large sample hydrology).

- The effect of clustering method choice needs to be carefully explored for analysing catchment similarity.
- Next step **(1)** — Try hierarchical clustering as another benchmark clustering method and compare results.
- Next step **(2)** — Train data-driven models for each cluster for predicting hydrological attribute of interest on validation dataset.

Contact

Elcin Kentel
ekentel@metu.edu.tr
Supervisor

Nilay Dogulu
ndogulu@metu.edu.tr
blog.metu.edu.tr/e149313
@DoguluNilay
PhD Candidate

Inci Batmaz
ibatmaz@metu.edu.tr
Co-Supervisor

Geophysical Research Abstracts
Vol. 20, EGU2018-13146, 2018
EGU General Assembly 2018

Clustering approaches for analysing similarity in ungauged catchments: input variable selection for hydrological predictions

Nilay Dogulu (1), Inci Batmaz (2), and Elcin Kentel (3)

(1) Middle East Technical University (METU), Civil Engineering Dept., Water Resources Laboratory, Ankara, Turkey (ndogulu@metu.edu.tr), (2) Middle East Technical University (METU), Statistics Dept., Ankara, Turkey (ibatmaz@metu.edu.tr), (3) Middle East Technical University (METU), Civil Engineering Dept., Water Resources Laboratory, Ankara, Turkey (ekentel@metu.edu.tr)

Catchments are hydrological units that exhibit unique but distinct features that greatly contribute to heterogeneity and complexity of rainfall-runoff processes. While the lure of understanding such diversity has underpinned the focus of many research efforts in hydrology, including predictions in ungauged basins, there is still room for improving our ability to benefit from this diversity in the context of data-driven hydrologic regionalization. An outstanding issue in this line of research concerns enhanced utilization of knowledge on dominant factors affecting catchments' hydrologic response behaviour under different types of streamflow. Our study addresses this issue by grouping similar catchments across continental USA using the CAMELS dataset (Addor et al., 2017) for the purpose of determining input variables that control streamflow predictability within each group of catchments. To this aim, we explore the performance of different clustering methods in identifying similar catchments based on available topography, soil, geology, vegetation and climate attributes, and then evaluate the set of variables which characterize hydrological attribute of interest (95% flow percentile for low flows, mean daily discharge for medium flows, and 5% flow percentile for high flows) using iterative input variable selection method (Galelli and Castelletti, 2013). We compare three clustering approaches that belong to different family of methods: partitional clustering algorithm (k-means clustering), density-based clustering algorithm, and spectral clustering algorithm. We discuss the results from the perspective of underlying assumptions and capabilities of these methods, and provide insights into effects of clustering method choice in analysing variability of catchment similarity with respect to high, medium and low flows.

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017) The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences, 21(10), 5293-5313.

Galelli, S., & Castelletti, A. (2013). Tree-based iterative input variable selection for hydrological modeling. Water Resources Research, 49(7), 4295-4310.