Input variable selection for hydrological predictions in ungauged catchments: **with or without clustering?**

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

Background

1

A key step in data-driven environmental modelling is **input variable selection** to ensure that the **least number of variables with minimum redundancy** are used to characterize the inherent relationship between inputs and outputs.

Hydrological predictions in ungauged catchments is one such area where the information on influential predictors of runoff signatures guides in understanding **dominant controls of meaningful information transfer** from gauged to ungauged locations (i.e. regionalization). This understanding is valuable especially for the analysis of **hydrological similarity among ungauged catchments**, e.g. to identify reference (donor) catchment(s).

Large-sample hydrology can help to gain useful insights on how these significant predictors change over different runoff signatures representing particular hydrological conditions as well as within different groups of similar catchments.

Research Objective

To explore the added value of clustering for input variable selection

Where?

2

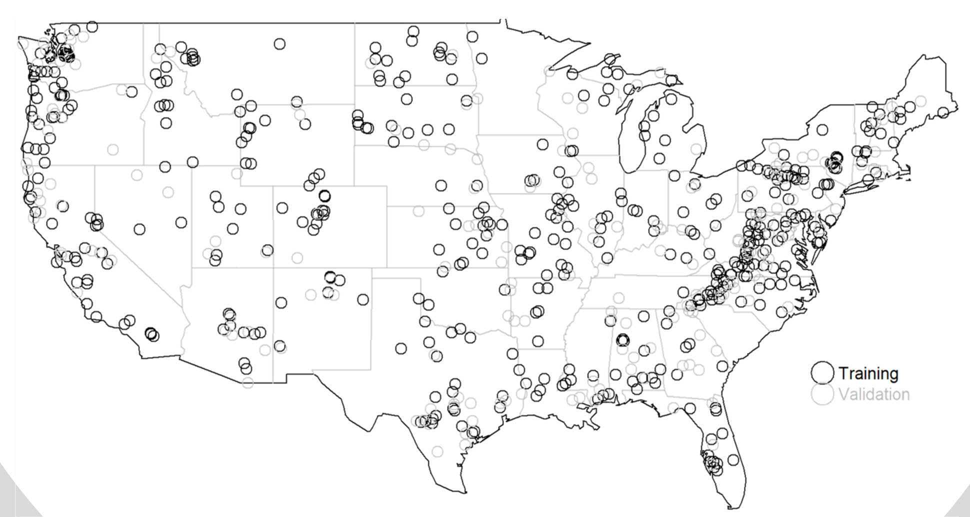
CAMELS

Catchment **A**ttributes and **ME**teorology for **L**arge-sample **S**tudies

Addor et al., 2017, HESS

671 watersheds across continental USA

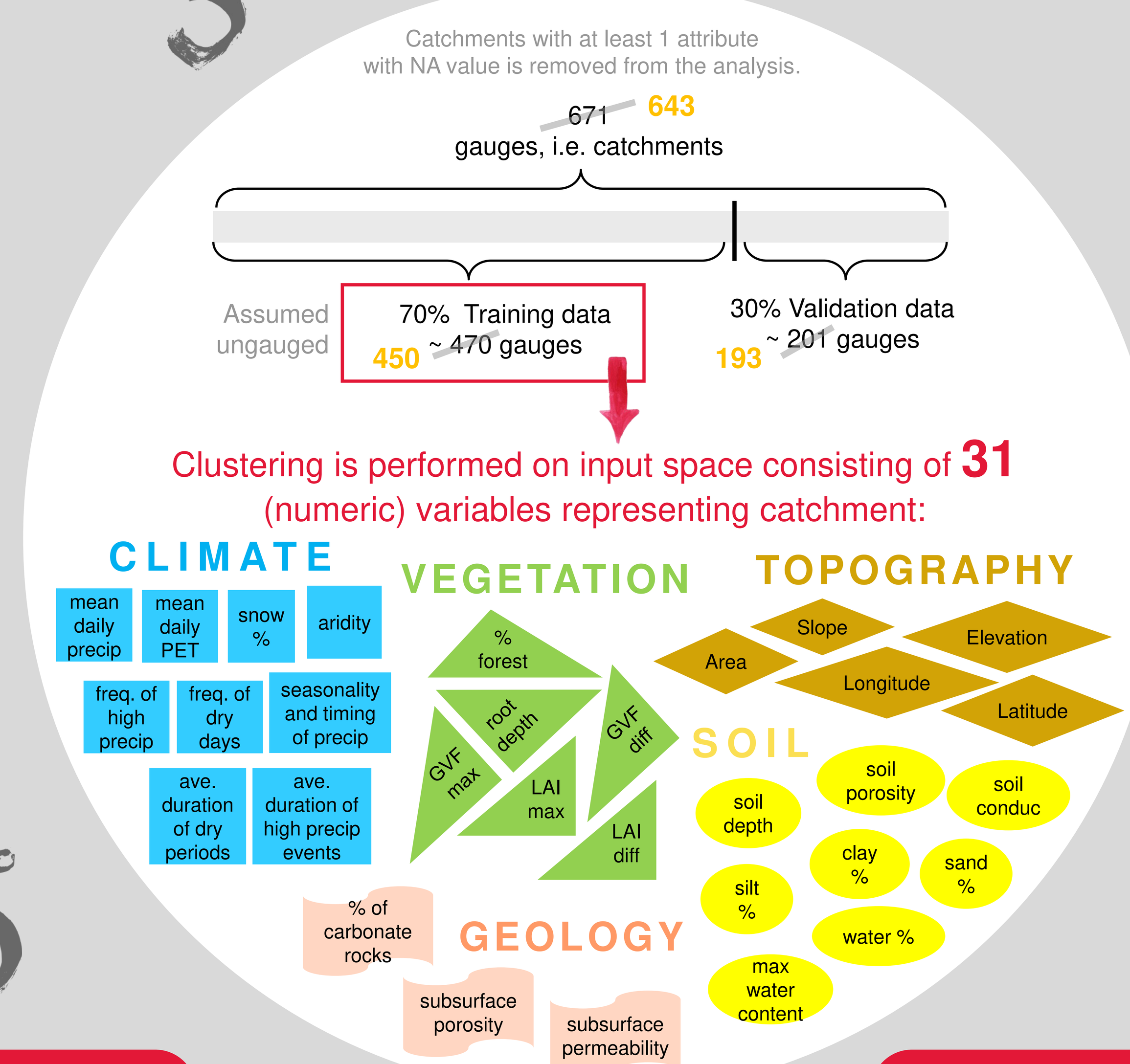
(unimpacted / less impacted by anthropogenic changes)



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/

Data

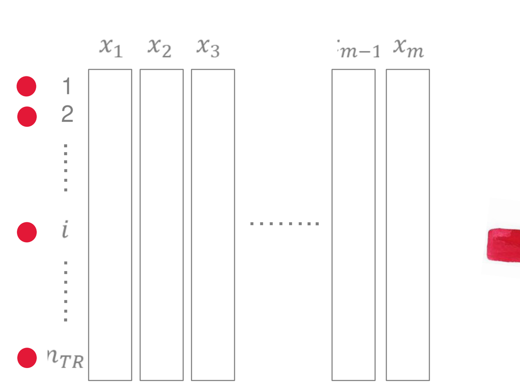
3



Methodology

4

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

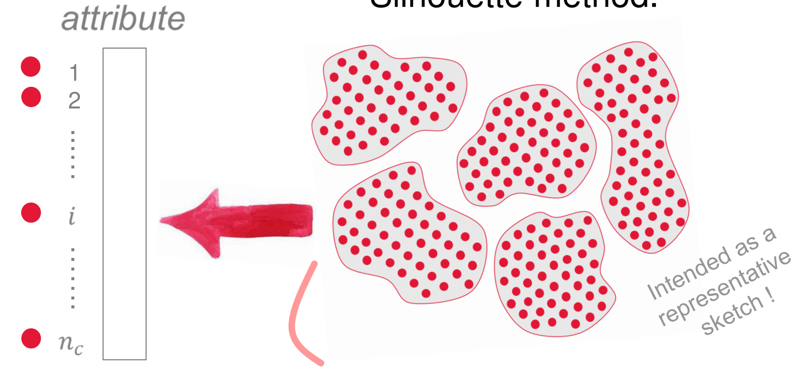
SEE BOX 6

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

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

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

hydrological attribute



Cluster "k"

n_c : Total number of gauges in cluster "k"

Partial Mutual Information (PMI)
Partial correlation Input Selection (PCIS)
Iterative Input Selection (IIS)

SEE BOX 5 and 7

IVS methods

5

Partial Mutual Information (PMI)

- Information theory based filter algorithm developed by Sharma (2000), later modified by Bowden et al. (2005) and May et al. (2008).
- It is a model-free approach capable of accounting for input redundancy.
- Inputs are selected in a stepwise procedure based on the estimation of the PMI, which measures the partial dependence between each input variable and the output, conditional on the inputs that have already been selected. The term "partial" implies the partial or additional dependence the new predictor can add to the existing prediction model.
- Nonparametric kernel methods are used to characterize joint *pdf* of variables.

Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - a strategy for system predictor identification. *Journal of Hydrology 239*, 232-239.
Bowden, G. J., Dandy, G. C., & Maier, H. R. (2005). Input determination for neural network models in water resources applications. Part 1—background and methodology. *Journal of Hydrology, 301*(1-4), 75-92.

Partial Correlation Input Selection (PCIS)

- Linear correlation-based filter algorithm introduced by May et al. (2008).
- PCIS algorithm is aimed at finding the partial linear correlation between two variables after removing the effects of other variables. Pearson correlation coefficient is used to estimate the strength of the relationship between inputs and output and a multiple linear regression based on least squares. Algorithm is terminated when the selection of additional inputs no longer results in an improvement (increase). No tuning is required.
- R code is available through IVS4EM project (Galelli et al., 2014).

May, R. J., Maier, H. R., Dandy, G. C., & Fernando, T. G. (2008). Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling & Software, 23*(10), 1312-1326.
Galelli, S., Humphrey, G. B., Maier, H. R., Castelletti, A., Dandy, G. C., & Gibbs, M. S. (2014). An evaluation framework for input variable selection algorithms for environmental data-driven models. *Environmental Modelling & Software, 62*, 33-51.

Iterative Input Selection (IIS)

- A hybrid filter-wrapper IVS method proposed by Galelli and Castelletti (2013).
- 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. Extra-Trees (Geurts et al., 2006) is used for both ranking and modelling as a regression method. MATLAB toolbox is available through the IVS4EM project (Galelli et al., 2014).
- No assumption on dependence between input and output variables is needed.

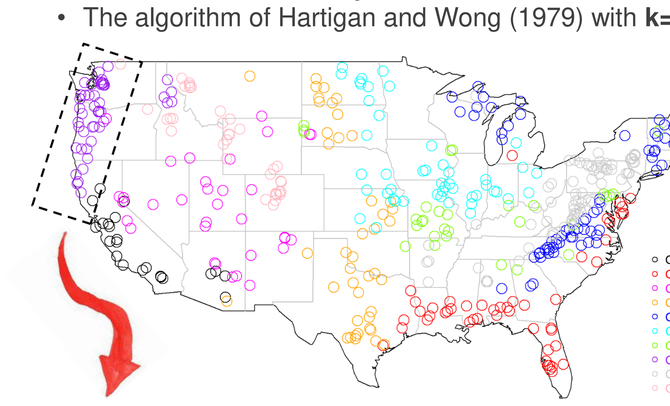
Galelli, S., & Castelletti, A. (2013). Tree-based iterative input variable selection for hydrological modeling. *Water Resources Research, 49*(7), 4295-4310.
Geurts, P., D. Ernst, and L. Wehenkel (2006), Extremely randomized trees, *Mach. Learn., 63*(1), 3-42.

Clustering

6

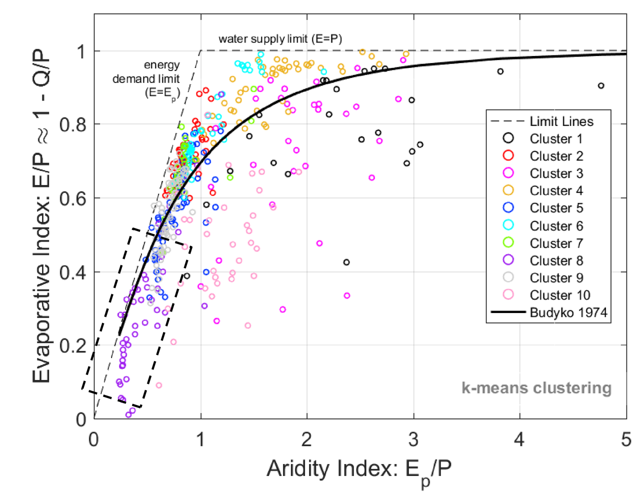
K-means clustering

- 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.



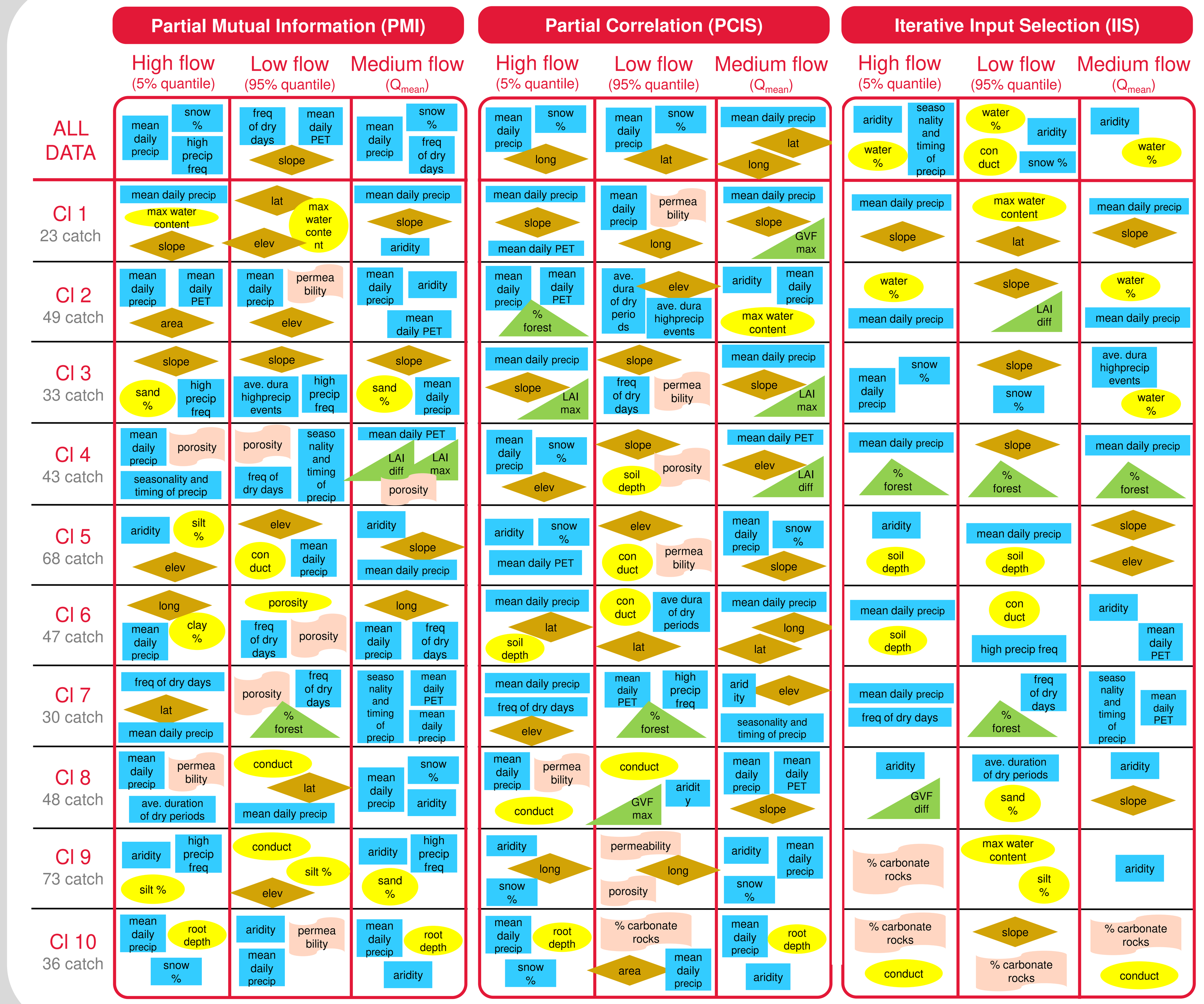
By looking at the Budyko Curve, one can say that the differences in annual water balance behaviour of catchments in Clusters 2, 5, 7 and 9 are not very significant.

Catchments in north and mid western coastal regions are identified by Clusters 1 and 8, respectively, the latter representing energy-limited catchments with WET climate (i.e. aridity index <1) as can be seen on the Budyko Curve.



Input Variable Selection (IVS)

7



Summary

- IVS is important for hydrological predictions in ungauged catchments! And clustering of input data space adds value!
- Climate and topography related variables are more frequently selected compared to soil, geology and vegetation attributes. However, there are cases where geology becomes more important as a predictor of the system.
- We found that selected variables for predicting hydrological attributes are not the same for different clusters.
- Moreover, selected variables for a given cluster vary depending on the hydrological attribute of interest.
- More in-depth analysis of IVS results is in progress.
- The effect of clustering method choice could be further explored.
- Next step: training data-driven models for each cluster to make hydrological predictions on validation set.

Contact

Elcin Kentel
ekentel@metu.edu.tr
Nilay Dogulu
ndogulu@metu.edu.tr
blog.metu.edu.tr/e149313
@DoguluNilay
Inci Batmaz
ibatmaz@metu.edu.tr
Supervisor
PhD Candidate
Co-Supervisor

Geophysical Research Abstracts
Vol. 20, EGU2018-14438, 2018
EGU General Assembly 2018

Input variable selection for hydrological predictions in ungauged catchments: with or without clustering?

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)

A key step in data-driven environmental modelling, including for hydrological purposes, is input variable selection (IVS) to ensure that the least number of variables with minimum redundancy are used to characterize the inherent relationship between inputs and outputs. Hydrological predictions in ungauged catchments is one such area where the information on influential predictors of runoff signatures guides in understanding dominant controls of meaningful information transfer from gauged to ungauged locations (i.e. regionalization). This understanding is valuable especially for the analysis of hydrological similarity among ungauged catchments, e.g. to identify reference (donor) catchment(s). Large-sample hydrology can help to gain useful insights on how these significant predictors change over different runoff signatures representing particular hydrological conditions as well as within different groups of similar catchments. This study explores the added value of clustering for input variable selection in the case of catchments across continental USA using the CAMELS dataset (Addor et al., 2017). We employ the method of k-means clustering on the input space consisting of topography, soil, geology, vegetation and climate attributes as a way to deal with heterogeneity and complexity in hydrological processes, and thus, aiming to identify similar groups of catchments. The input variables (for predicting selected hydrological attributes) are determined by three IVS filter algorithms (partial mutual information, partial correlation input selection, and iterative input selection), then evaluated and compared for among different clusters and when no clustering method is applied. We present and discuss the results for three hydrological attributes – 95% flow percentile (low flows), mean daily discharge (medium flows), and 5% flow percentile (high flows) – in order to account for variability in hydrological conditions, and to investigate the dominant catchment and/or climate characteristics controlling low/medium/high flow predictability at ungauged locations. The findings of our study have implications for reference (donor) catchment selection and hydrological model independent regionalization (aka hydrostatistical or data-driven methods), and are particularly relevant to understanding hydrological similarity and catchment classification in the absence of local runoff observations.

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.