Unsupervised learning applications in hydroclimatology

Arpita Mondal*

Data driven modeling is not a new concept in the field of hydroclimatology that concerns processes in the atmosphere, land surface and hydrology. The behavior and predictability of these physical processes at different spatio-temporal scales are not yet fully understood; therefore, data driven modeling offers a significant complementary approach to understand them.

Supervised learning involves finding a relationship between the dependent and independent variables, and applications of supervised learning have been plenty in this field. Estimation of statistically downscaled rainfall from large-scale atmospheric variables under climate change, or prediction of river discharge from rainfall are examples of supervised learning.

Unsupervised learning involves reducing dimensions or finding patterns in the independent variables without any target variable. In other words, there are attributes or features, but no predictands. This study discusses some recent applications of unsupervised learning in hydroclimatology.

The first application involves understanding patterns of extreme temperature using Empirical Orthogonal Function (EOF) analysis (alternatively called Principal Component Analysis) over India. Extreme temperatures are responsible for heat waves in India and with climate change their occurrence is believed to be more intense and more frequent. The primary EOF mode presents regions of temperature variability and can inform identification of important heat wave regions in the country [1]. The first EOF and the corresponding PC explains the largest amount of variability in the data. In this study, open-source pre-monsoon (MAMJ) gridded daily maximum temperature data over the Indian region is analysed to obtain extreme temperature hot spots.

In the second application, another popular unsupervised learning technique, the k-means clustering algorithm, is used to define regimes of extreme rainfall in the Western Himalayas in India. The Western Himalayas represent a significant ecological and geographical region in the country and extreme rainfall in this region has important implications on lives, societies and environment. Using large-scale climate features, distinct clusters are identified to segregate mild, moderate and wet regimes [2].

In the third and final application, fuzzy c-means clustering and its advanced versions such as Random Forests-based Fuzzy c-means are used for catchment classification. In hydrology, catchment classification is used for regionalization that implies identification of 'homogeneous' hydrological regions in terms of their flood-generating mechanisms. Regionalization forms the first step for flood estimation and prediction in ungauged basins. Within each homogeneous region, flood information can be pooled to 'trade space for time', in order to estimate design flood for the ungauged or poorly gauged target site. Here, for ease of data availability, fuzzy c-means clustering is applied for catchment classification on 243 watersheds in the state of Indiana, USA. Further, this study explores sensitivity of these homogeneous regions to the choice of distance metrics used in the clustering algorithm, such as the Euclidean, Mahalanobis, Cosine, Minknowski or Cityblock distances.

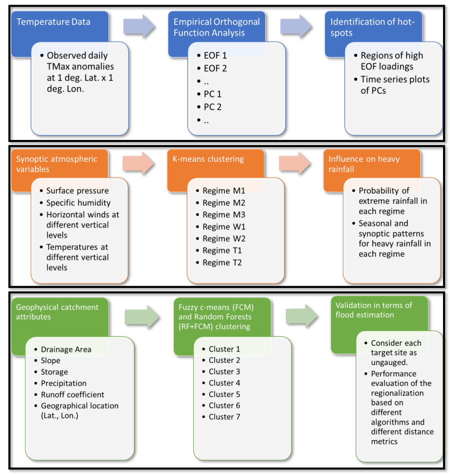

Figure 1: Schematic of the applications of unsupervised learning

References

[1] Ratnam, J. V., Behera, Swadhin K., Ratna, Satyaban B., Rajeevan, M. and Yamagata, Toshio, Anatomy of Indian heatwaves, *Scientific Reports*, 6(1), 2016.

[2] Suma Bhanu Battula, Steven Siems and Arpita Mondal, Dynamical and thermodynamical interactions in daily precipitation regimes in the Western Himalayas, *International Journal of Climatology*, 42(9), 2021.

[3] A. Ramachandra Rao and V.V. Srinivas, *Regionalization of Watersheds: An Approach Based on Cluster Analysis*, Springer New York, 58, 2008.

[4] Arpita Mondal, Reema Roy and Chingka Kalai, Regionalization for Flood Frequency Analysis: Sensitivity to Choice of Clustering Algorithm and Distance Metric, In *World Environmental and Water Resources Congress*, pp. 353-366, 2023.

*A. Mondal is with the Indian Institute of Technology Bombay email: marpita@iitb.ac.in.