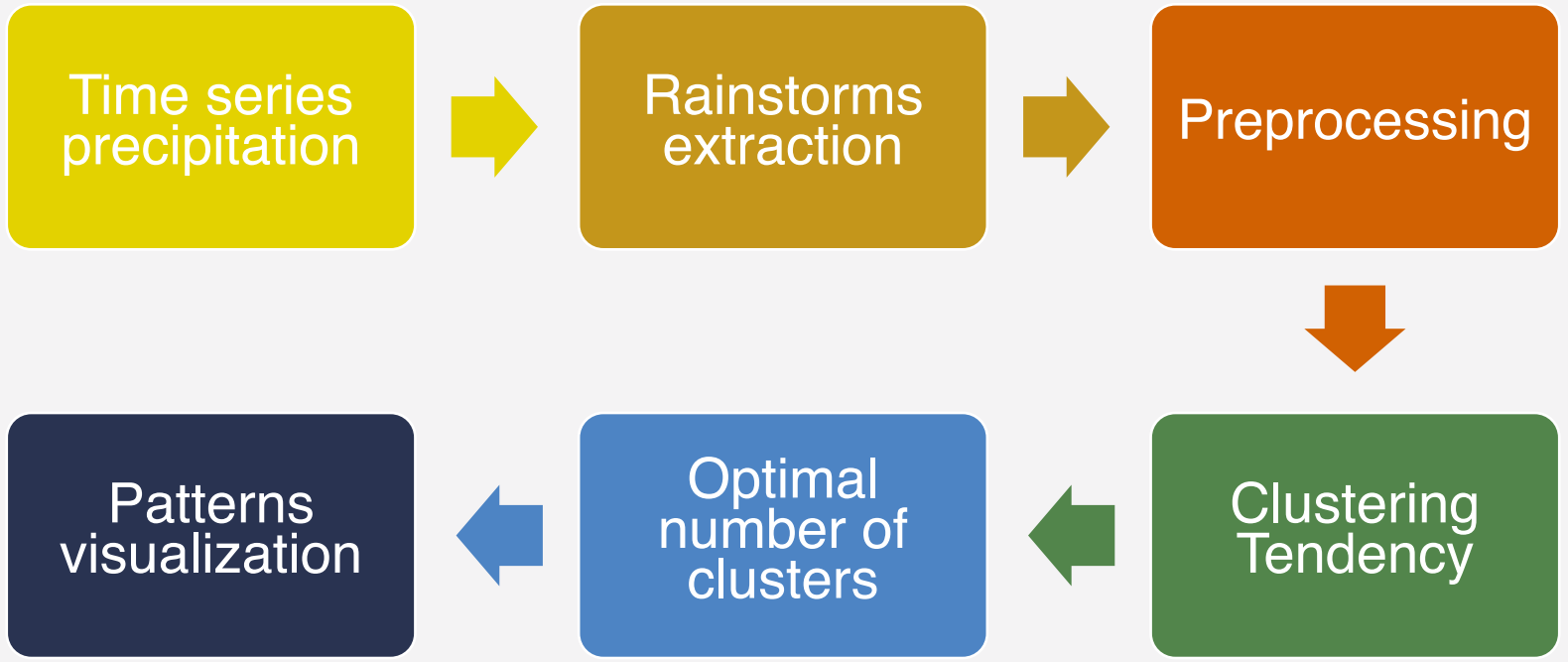
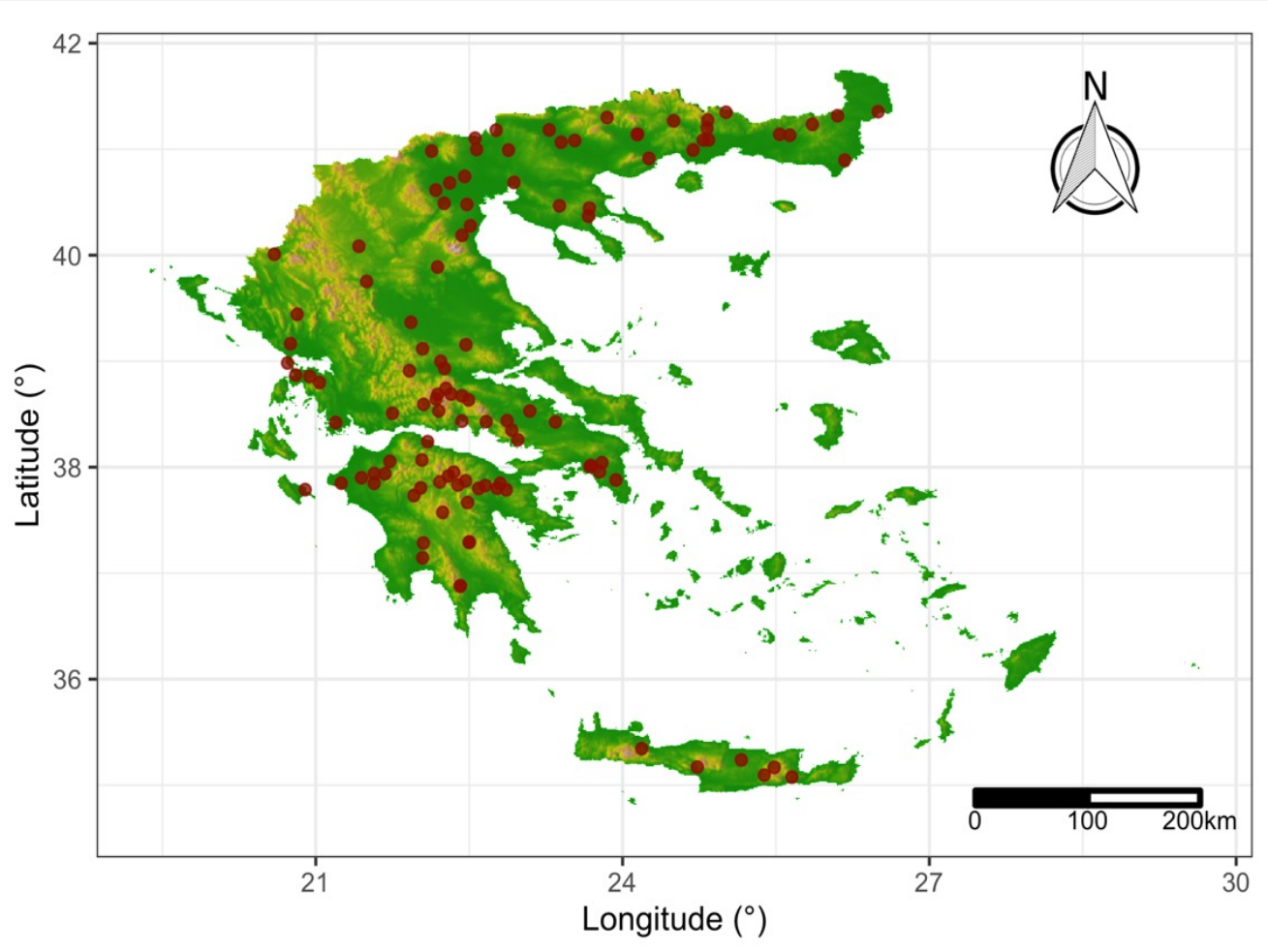
Knowledge discovery using clustering analysis of rainfall timeseries

Konstantinos Vantas and
Epaminondas Sidiropoulos

EGU2021: Clustering in Hydrology

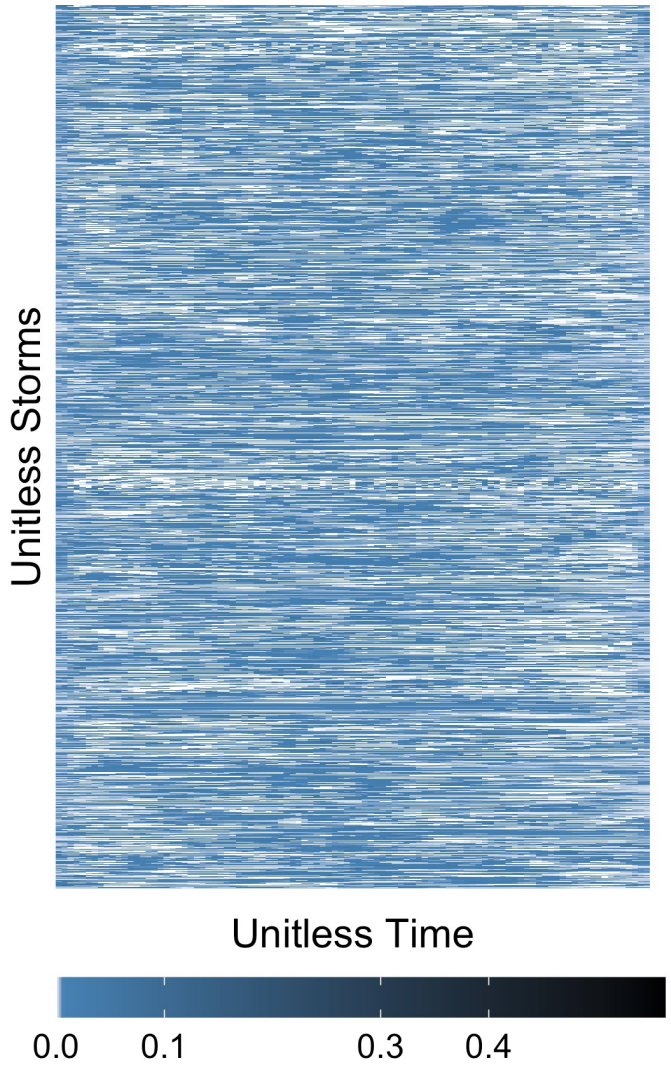
Investigating the presence of intra-storm temporal patterns using rainfall timeseries





Dataset

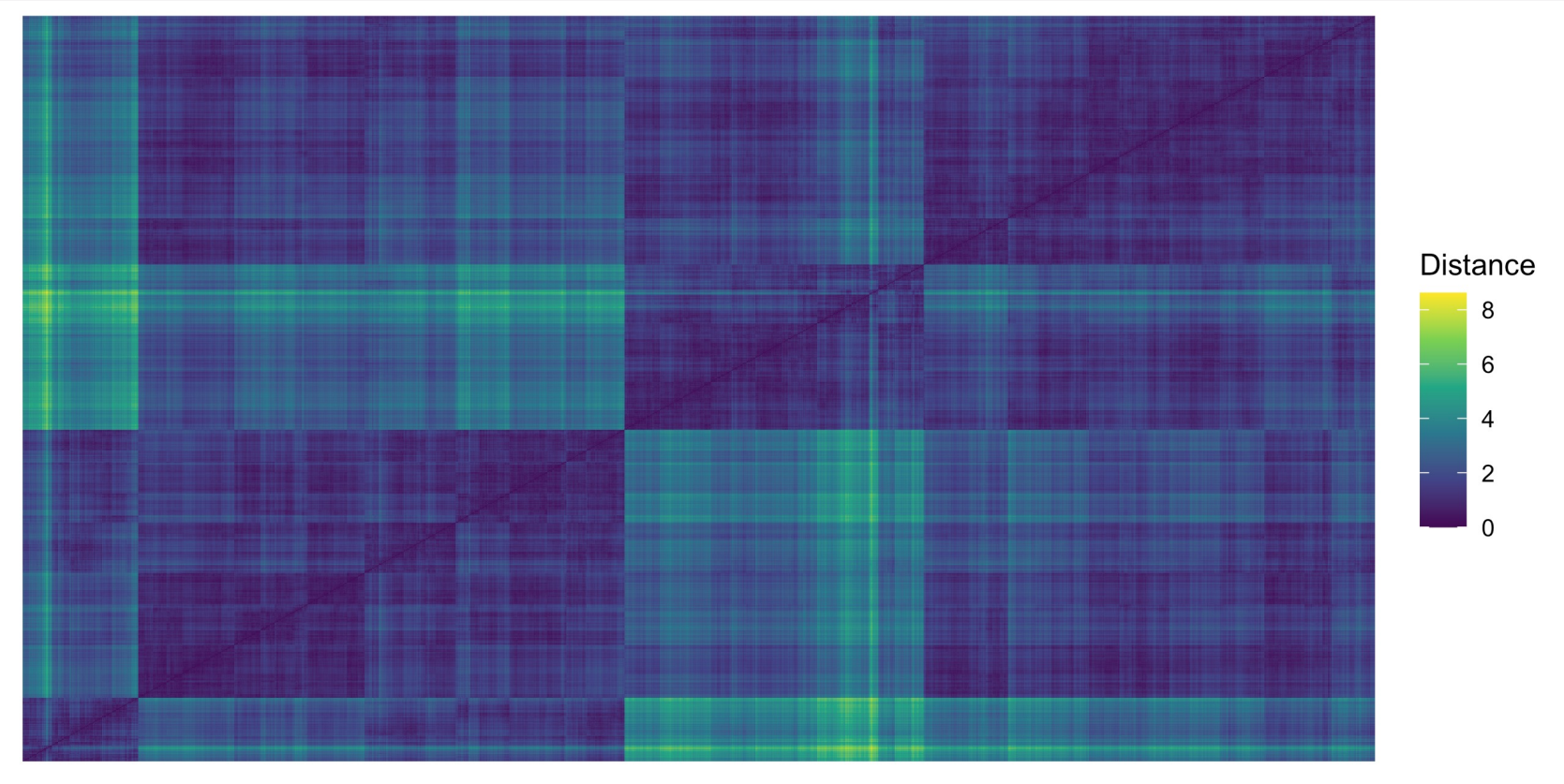
- 108 meteorological stations across Greece.
- The timeseries comprise a total of 2926 years of 30 min records with a mean length of 26.6 years per station.



Rainstorms extraction and preprocessing

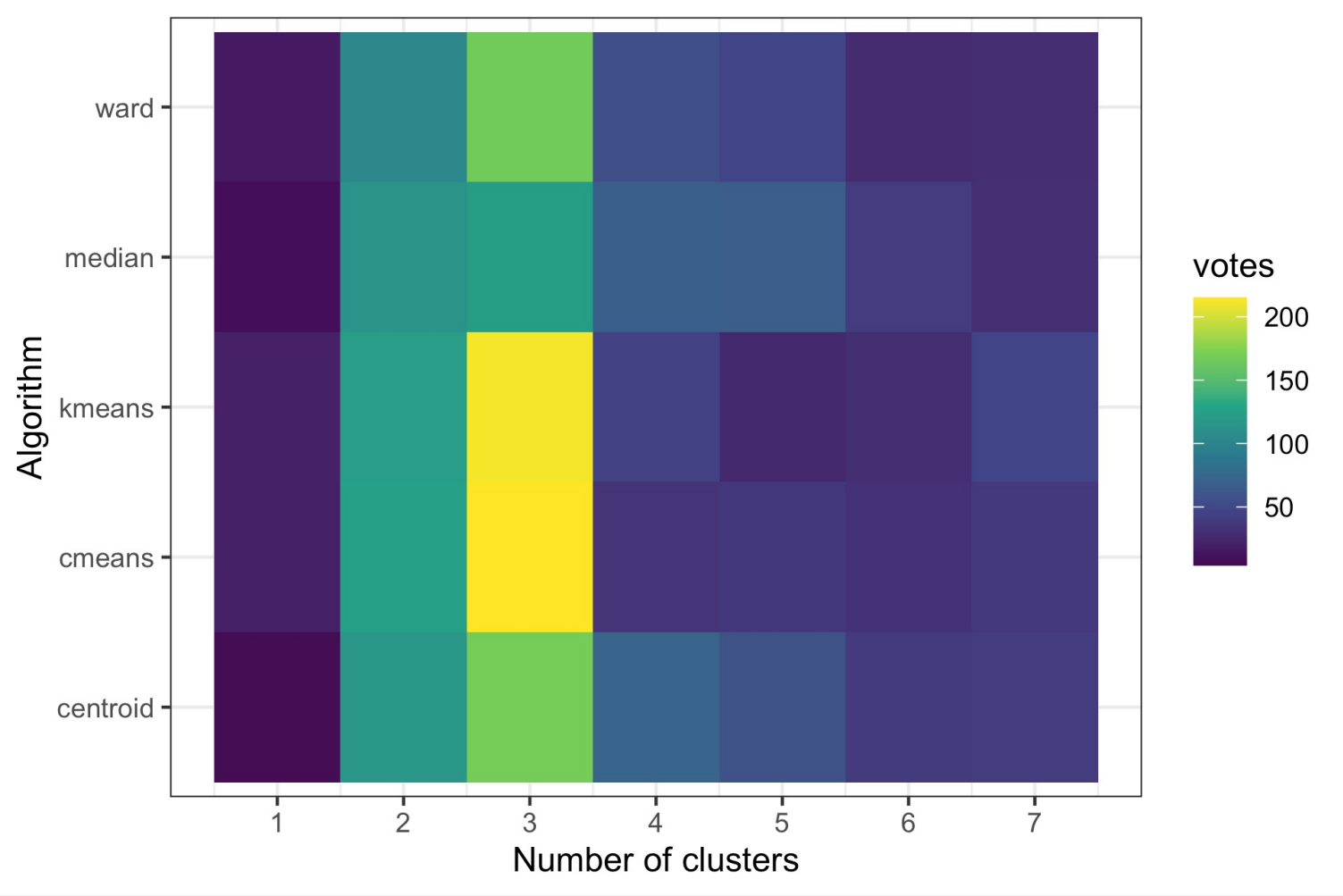
1. A Poisson process hypothesis is assumed for the separation of the precipitation timeseries to statistically rainstorm events.
2. The independent events are scaled to unitless form.
3. Linear interpolation is applied to compute the unitless rainfall for every 1% of unitless time values.
4. A set of 174,883 rainstorms were extracted and 26,678 are used in the analysis with cumulative rainfall greater than 12.7 mm.

Clustering Tendency



1. All clustering algorithms can return clusters, even if there is no structure in the used data.
2. It is advisable to have a preliminary look into the dataset in order to detect any existing clustering tendency.
3. The visual assessment of cluster tendency (VAT) is used.
4. The VAT created an image matrix, where the clusters are indicated as at least four darker blocks along the diagonal.

Optimal number of clusters



- The relevant number of clusters is a-priori unknown.
- A voting scheme with random subsets of the data used the recommendations based on 26 indicators.
- Different algorithms, produce different results. Most recommendations are between two and four.

Optimal number of clusters

A custom developed, domain-specific iterative algorithm via fuzzy clustering is utilized. The optimal number for α = 0.05 is **four**.

Algorithm 2: Optimal number of clusters using FCM

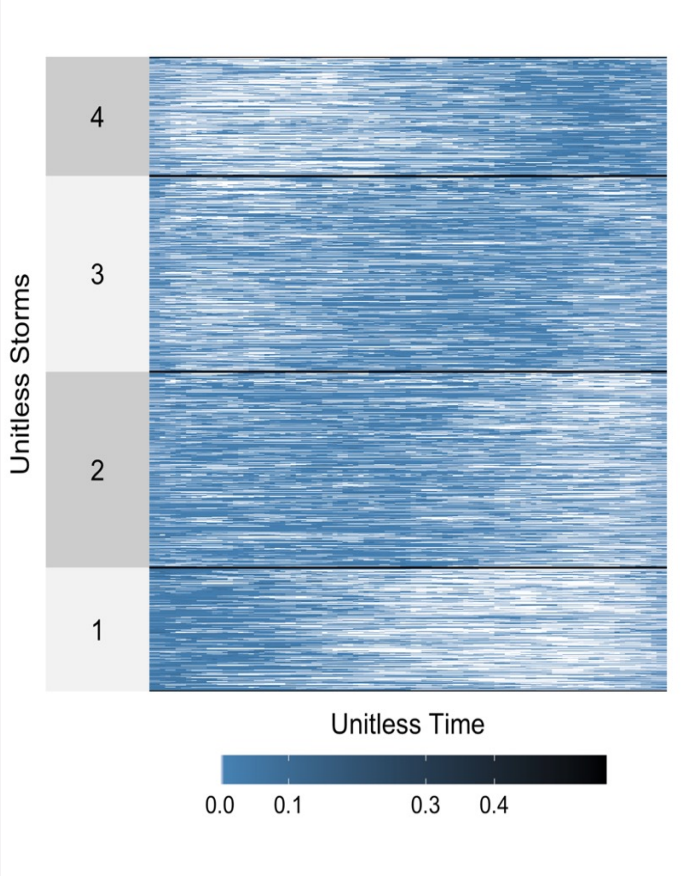
Input: unitless rainstorms $\boldsymbol{U}$; maximum number of clusters to test n_{max}; significance level $\alpha = 0.05$

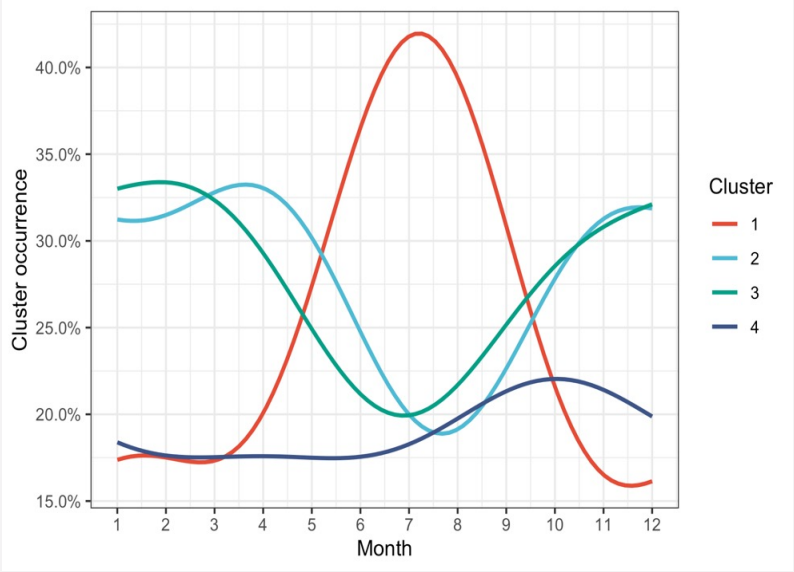
1 compute the row-wise cumulative values $\boldsymbol{U'}$ of $\boldsymbol{U}$
2 **for** $i \leftarrow 2$ *to* n_{max} *and all p-values* $< \alpha$ **do**
3 apply FCM on $\boldsymbol{U'}$ for $c = i$ and compute cluster centers C;
4 for all pairs in C obtain the Kolmogorov–Smirnov two sample test, p-values;
5 adjust the obtained p-values using Benjamini and Hochberg method;

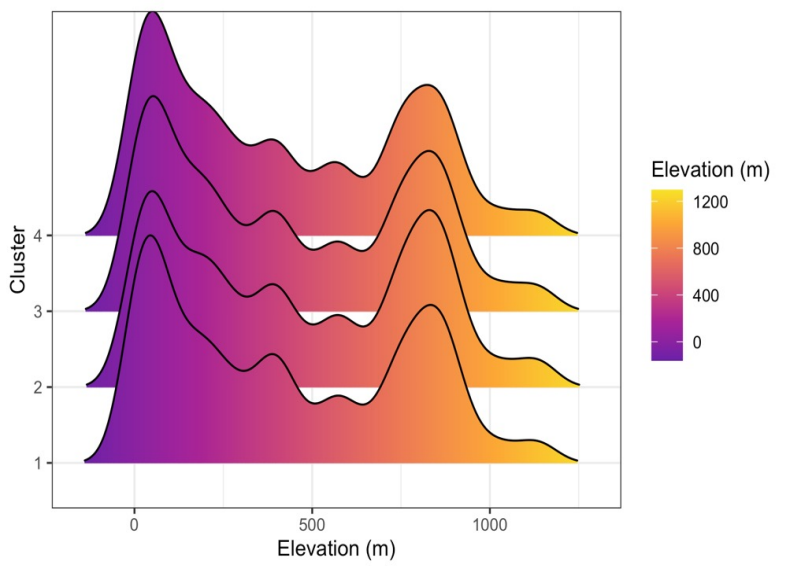
Result: optimal number of clusters c_{opt} and clustering results

Clustering Results

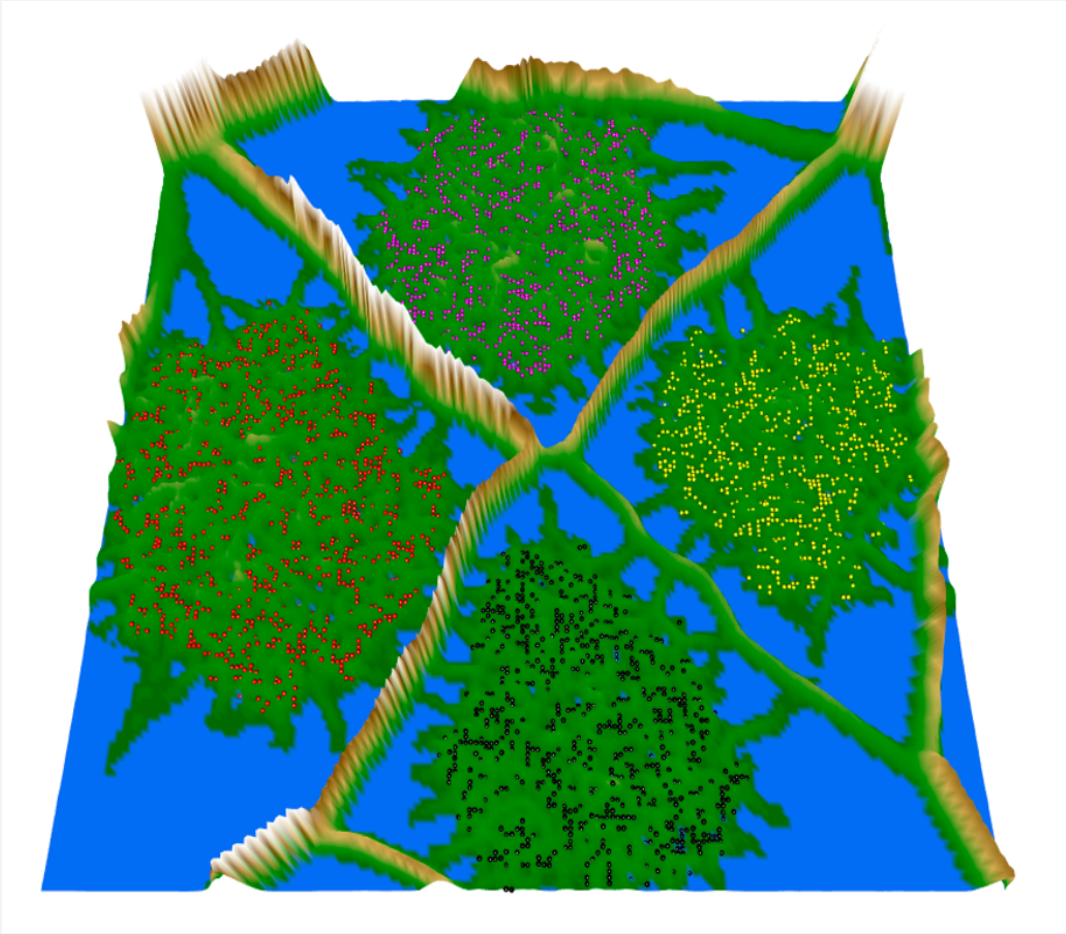
- The temporal patterns per station and given cluster showed very high similarity with $r \geq 0.974$.
- Cluster one has the same seasonality with convective activity over the country.
- Cluster four, with its relatively higher intensity values and higher ratio of occurrence during the wetter winter months in Greece, can be utilized in hydrologic design.
- Despite the rich topography and variability of micro-climates in Greece, the clusters have regional stability in long distances.



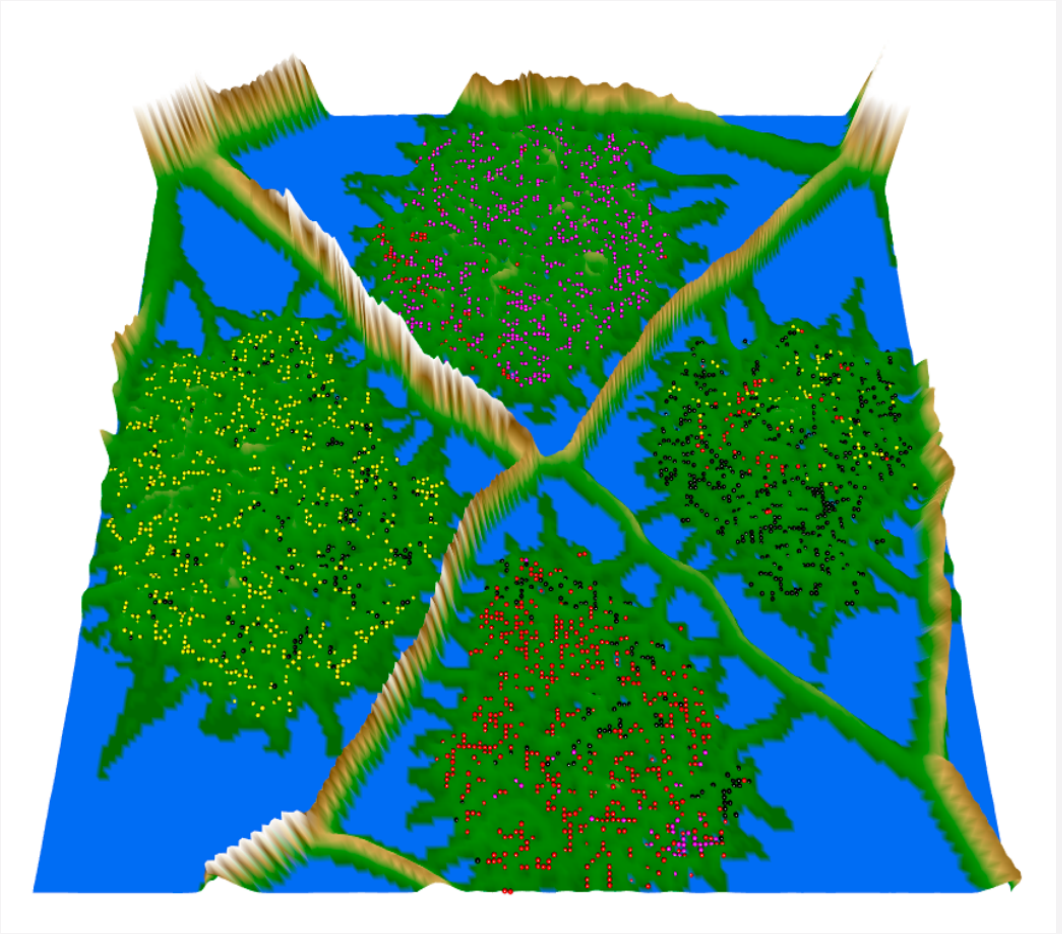




Projecting Data Using Non-Linear Mapping



a. FCM results with membership > 0.75 in order to remove equivocation from the data



b. Huff's classification in which rainstorm data are classified by the quartile where the maximum intensity occurs

Conclusions

1. A domain specific algorithm proposed the presence of four clusters.
2. The analysis revealed intra-storm patterns that have seasonality, are independent of the elevation and, also, have regional stability.
3. Non linear mapping indicates the presence of four clusters and that the commonly used Huff's method misclassifies the data.

Vantas, K.; Sidiropoulos, E. Intra-Storm Pattern Recognition through Fuzzy Clustering. *Hydrology* **2021**, *8*, 57. https://doi.org/10.3390/hydrology8020057

The data importing, analysis and presentation were done using the open source R language for statistical computing and graphics using the packages: hydroscoper, hyetor, e1071, FCPS, GeneralizedUmatrix, factoextra and ggplot2

knvantas@law.auth.gr @kvantas