Rainfall Temporal Distribution in Thrace by Means of an Unsupervised Machine Learning Method

Presentation · July 2018

DOI: 10.13140/RG.2.2.13319.44966

3 authors, including:

Konstantinos Vantas
Aristotle University of Thessaloniki
6 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Marios Vafiadis
Aristotle University of Thessaloniki
19 PUBLICATIONS **376** CITATIONS
SEE PROFILE

Rainfall Temporal Distribution in Thrace by Means of an Unsupervised Machine Learning Method

K. Vantas, E. Sidiropoulos and M. Vafeiadis

Faculty of Engineering Aristotle University of Thessaloniki

Temporal distribution of rainfall

Temporal distribution of rainfall

Knowledge about the temporal distribution of rainfall is essential for integrated and rational water resources management:

- Drainage design
- Erosion control
- Water quality assessment
- Global change impact studies

Typical methodology

A typical methodology includes:

1. The determination of total duration and height of rainfall.
2. The disaggregation of this height using a temporal pattern that represents the expected internal rainfall structure (design hyetograph, DH).

Methods for the production of DH

The four types of methods for the production of DH are:

1. Specification of simple geometrical shapes anchored to a single point of the intensity-duration- frequency (IDF) curve.
2. Use of the entire IDF curve.
3. Standardized profiles obtained directly from rainfall record.
4. Simulation from a stochastic rainfall model.

In practice, the first 3 methods are used.

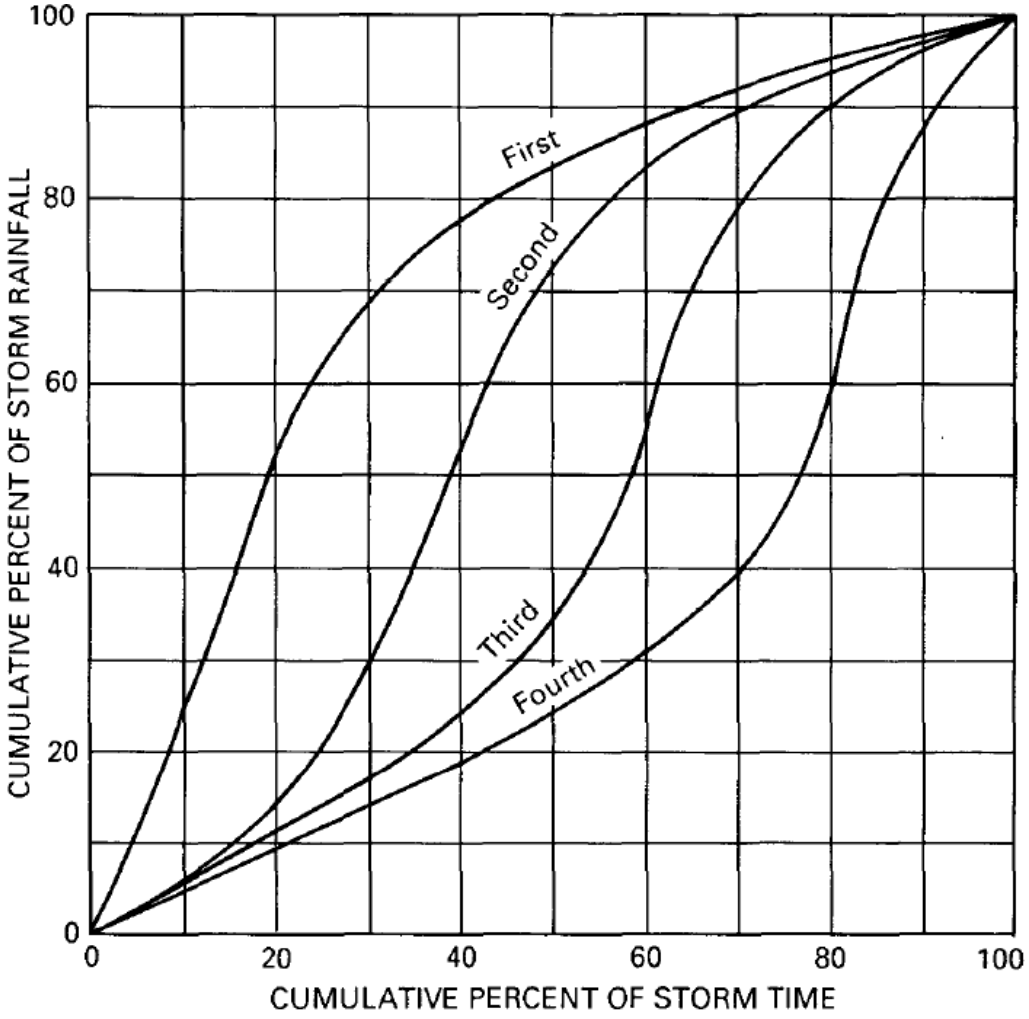
Huff's curves

Huff (1967) presented a method, in which storm data are classified using the quartile where the maximum intensity occurs.

Rainstorm events:

a) Are separated from precipitation time series using a six-hour fixed Critical time Duration (CD) of no precipitation.

b) Are transformed to dimensionless curves by dividing time by the total duration of the event and cumulative rainfall by the total precipitation volume.



Proposed unsupervised method

Proposed method

It is an original, controlled, fully reproducible, unsupervised method that produces automatically and objectively the optimal number of DHs by direct use of precipitation records.

Proposed unsupervised method

1. Raw precipitation data is cleaned from noise and errors.
2. CD is determined on the basis of a Poisson process hypothesis (i.e. the rainstorms events' interarrival times are distributed exponentially).
3. A temporal model of monthly CD values for the area is constructed.
4. Unitless Cumulative Hyetographs (UCH) are compiled and Principal Components Analysis (PCA) is applied to them.
5. Agglomerative Hierarchical Clustering is applied on the principal components (HCPC).
6. The number of clusters is determined by repetitive statistical comparisons between the centers of the clusters already produced at the previous steps.
7. Finally, a limited number of DHs is produced, capable of representing the totality of the rainstorm records.

Storm identification

Algorithm 1: Temporal model of CD

Input: Stations' precipitation time series $\boldsymbol{P_i}$ where $i = 1, \ldots, k$; Critical durations test vector $\boldsymbol{CD} = [120, 180, \ldots, 1800]$ (min); Number of samples that are drawn for parametric bootstrapping $s = 50,000$;

1 **for** *station* $i \leftarrow 1$ *to* k **do**
2 **for** *month* $m \leftarrow 1$ *to* 12 **do**
3 **for** *cd in* $\boldsymbol{CD}$ **do**
4 Compute the vector of interarrival times $\boldsymbol{t_\alpha}$ using inter-month data and n = length($\boldsymbol{t_\alpha}$);
5 **if** $n \geq 50$ **then**
6 Estimate the average storm arrival rate $\widehat{\omega}$ from $\boldsymbol{t_\alpha}$ using Maximum Likelihood Estimation;
7 Obtain the Kolmogorov–Smirnov's p-value for the original sample $\boldsymbol{t_a}$ and the estimated distribution;
8 Generate s samples of size n from the estimated distribution;
9 For each sample compute the one-sample Kolmogorov–Smirnov's p-value using the estimated distribution as theoretical;
10 Use the empirical non-parametric distribution of p-values to obtain the p-value for the original sample $\boldsymbol{t_\alpha}$;
11 Get minimum dry period duration $MDPD_{i,m}$ from $\boldsymbol{CD}[max(p - value)]$;
12 Use $MDPD$ values to fit the smooth sinusoidal model:

$$f(CD) = \theta_1 sin\left(\frac{2\pi}{12}m\right) + \theta_2 sin\left(\frac{4\pi}{12}m\right) + \theta_3 cos\left(\frac{2\pi}{12}m\right) + \theta_4 cos\left(\frac{4\pi}{12}m\right)$$

Result: Monthly values of CD for the area

Optimal number of clusters

Algorithm 2: Optimal number of clusters

Input: tree produced from HCPC algorithm; significance level $\alpha = 0.05$

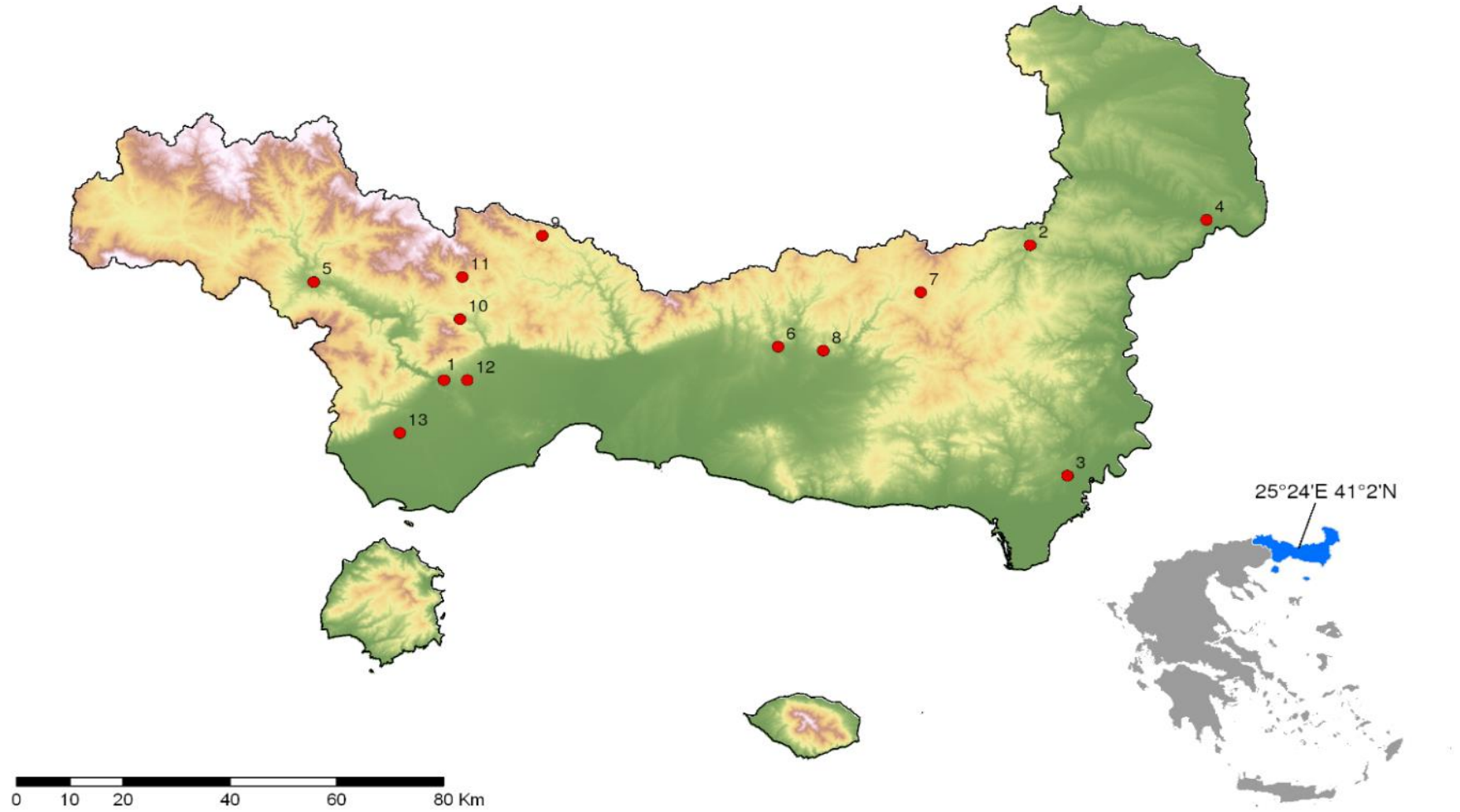
while *all p-values* $< \alpha$ **do**

- moving down the tree cut into q different clusters $q = 1, \ldots, m$;
- calculate the mean values $\bar{x}_q$ of the UCHs that belong to cluster q;
- for all $\bar{x}_q$ obtain the Kolmogorov–Smirnov two sample test, p-values;
- adjust the obtained p-values using Benjamini and Hochberg method;

Result: optimal number of clusters q_{opt} and design hyetographs $\bar{x}_{opt}$

Study Area and Dataset

Study area

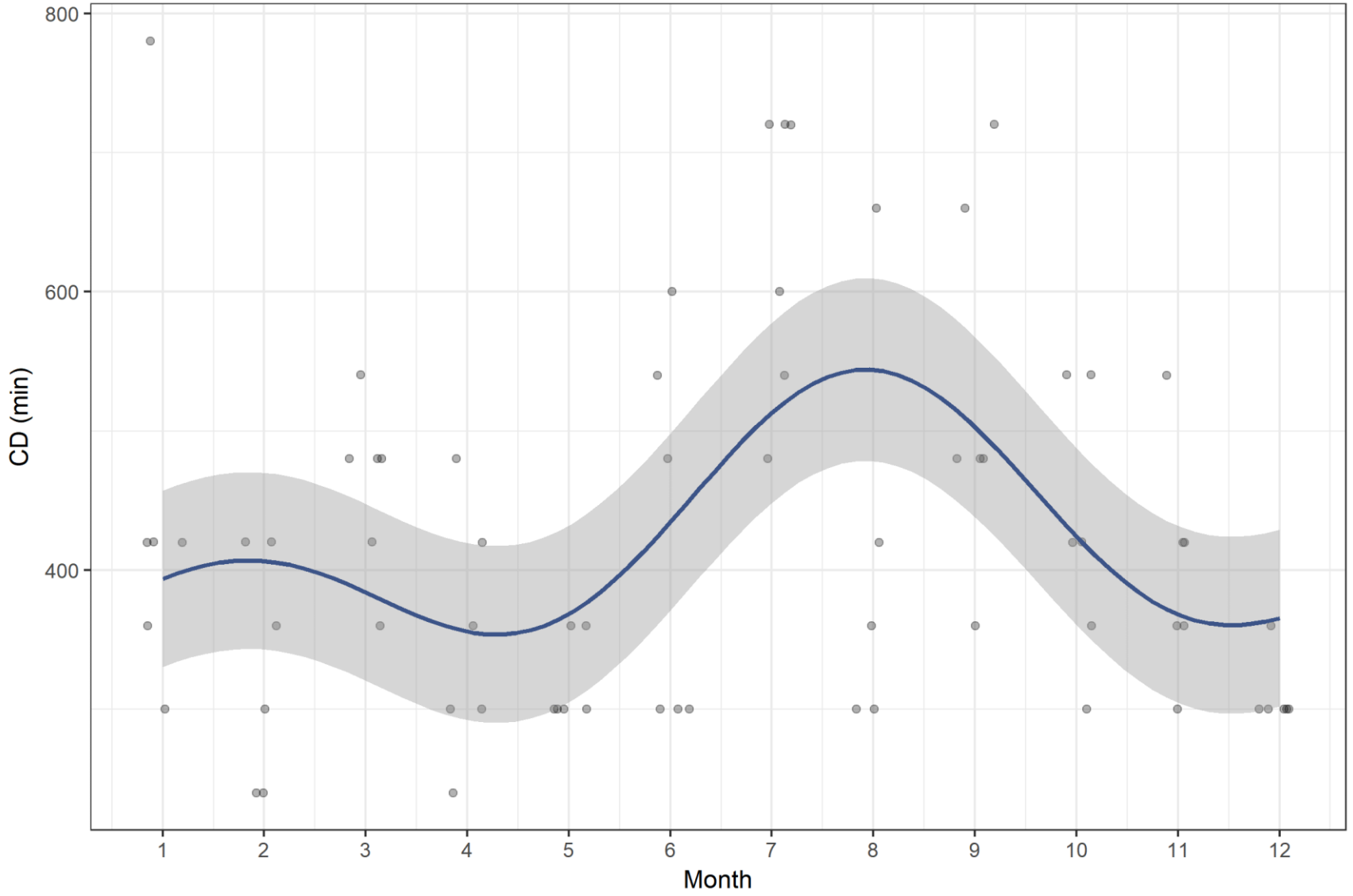


Dataset

	ID	Name	Lat (°)	Long (°)	Elevation (m)	Data Length (yr)	From	To	Data Coverage
1	200249	TOXOTES	41.09	24.79	75	41	1956	1997	62%
2	200259	MIKRO DEREIO	41.32	26.10	116	24	1973	1997	63%
3	200260	FERRES	40.90	26.17	43	35	1962	1997	56%
4	200263	DIDYMOTEIXO	41.35	26.50	25	41	1955	1996	62%
5	200311	PARANESTI	41.27	24.50	122	36	1960	1996	65%
6	500250	GRATINI	41.14	25.53	120	31	1965	1996	21%
7	500251	KECHROS	41.23	25.86	700	31	1965	1996	20%
8	500253	MIKRA KSIDIA	41.13	25.64	70	31	1965	1996	25%
9	500262	THERMES	41.35	25.01	440	31	1965	1996	21%
10	500265	GERAKAS	41.20	24.83	308	31	1965	1996	26%
11	500267	ORAIO	41.27	24.83	656	31	1965	1996	18%
12	500272	SEMELH	41.09	24.84	65	24	1968	1992	21%
13	500273	CHRYSOUPOLI	40.99	24.69	15	26	1966	1992	16%

Analysis

Critical Duration model

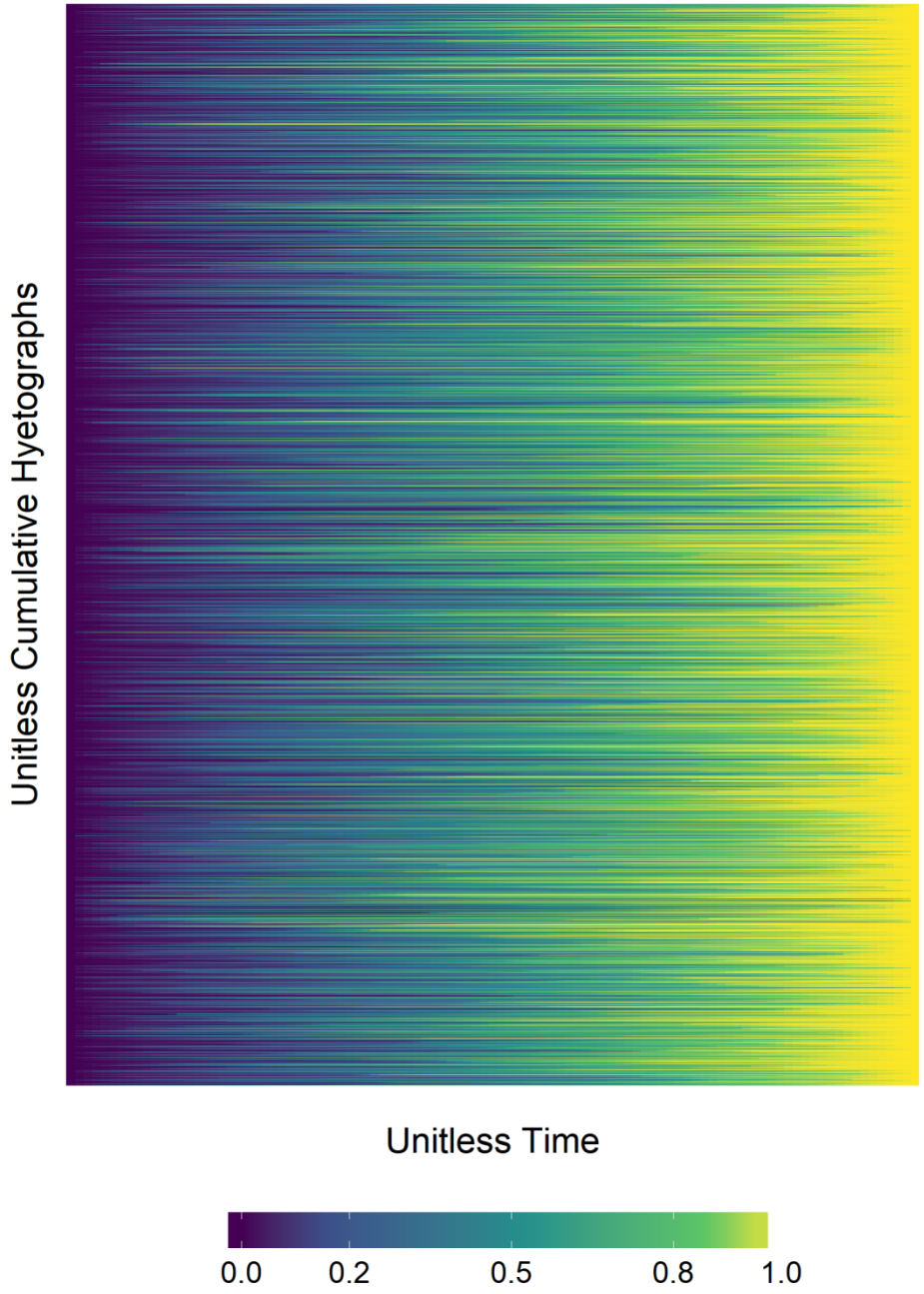


Unitless Cumulative Hyetographs

- The rainstorms are extracted using the CD model.
- Only the events with duration greater than 3 hours and cumulative rainfall greater than 12.7 mm are used in the analysis.
- Cumulative Hyetographs are transformed to unitless form.
- Linear interpolation is applied to compute the unitless cumulative rainfall for every 1% of unitless time values.

Unitless Cumulative Hyetographs

A population of 1,622 out of 25,377 extracted rainstorms met the criteria of minimum duration and cumulative height.

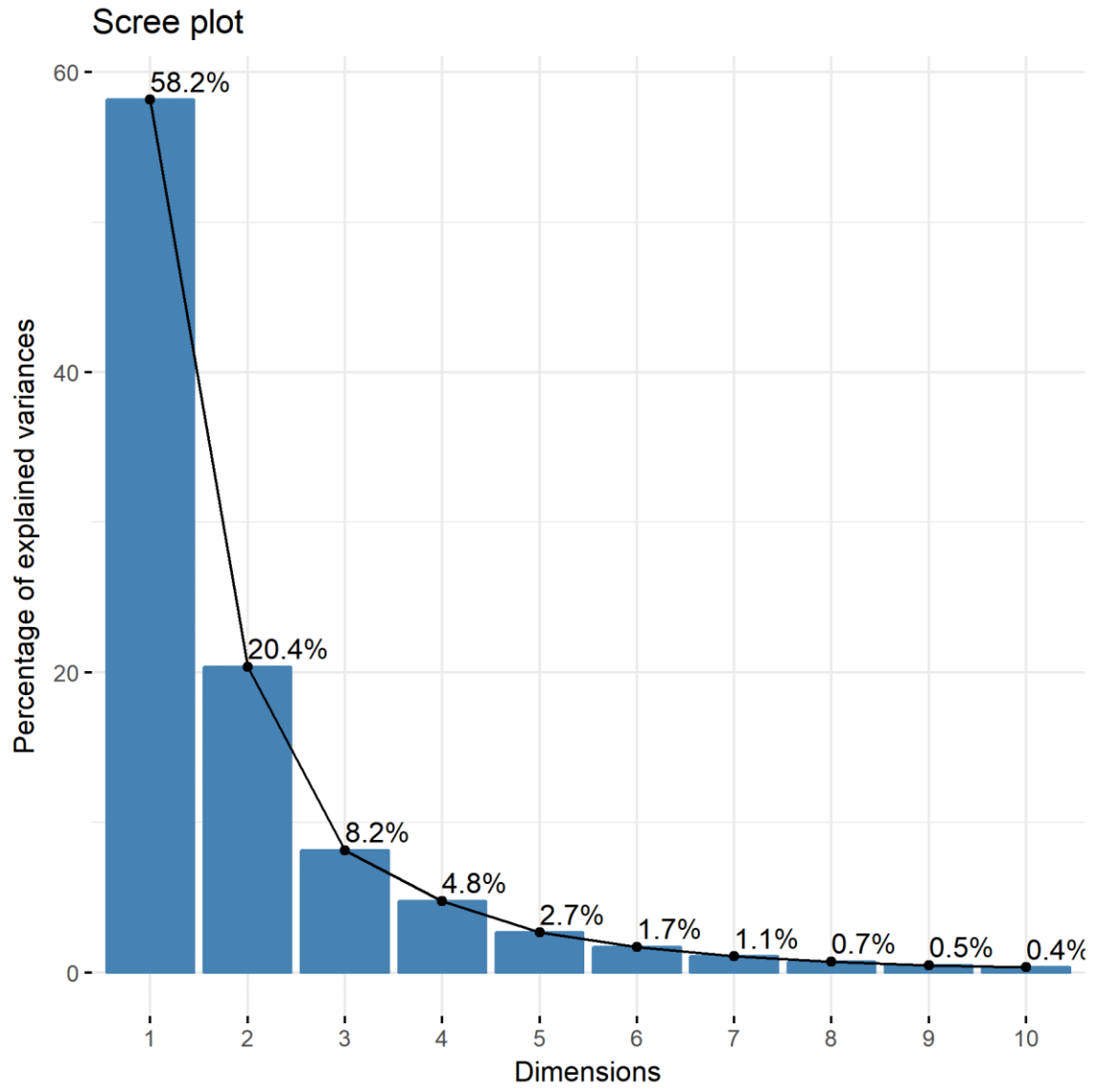


Principal Component Analysis

PCA is applied to reduce the dimensionality of the data to a few dimensions.

From PCA it is concluded that using only the first two dimensions explains 78.5% of total variance and the first 15 explains 99.5%.

The first two dimensions can be used to plot a UCH as a single point.



Results

Clustering Tendency

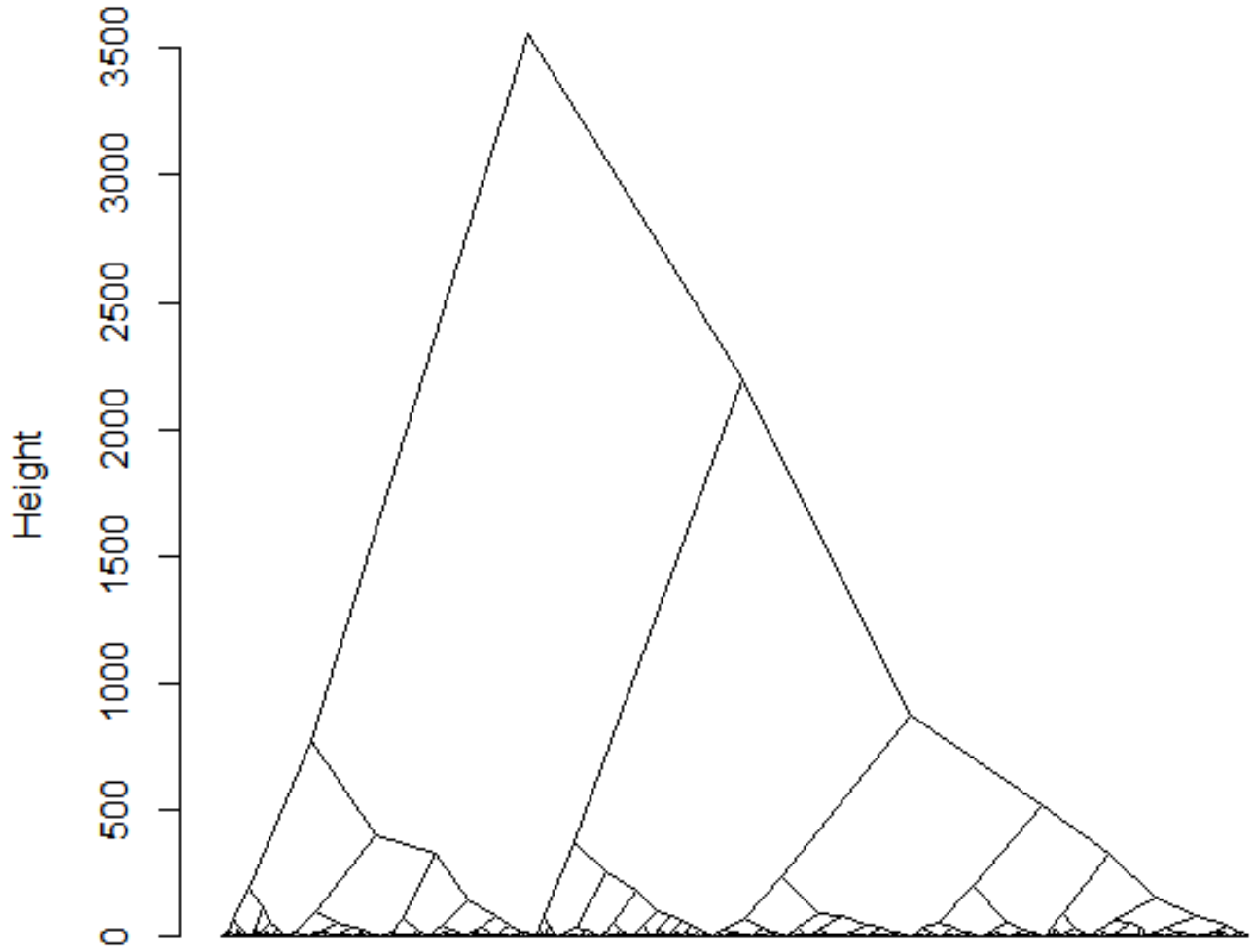
The Hopkins index, H for clustering tendency is applied, because all the clustering algorithms can return clusters even if there was no structure in data.

The computed value of H was 0.88, thus it indicates clustering tendency at the 90% confidence level.

Hierarchical Clustering on Principal Components

The clustering method applied is HCPC, using Ward's minimum variance criterion that minimizes the total within-cluster variance.

The result is a tree-based representation of the UCHs.

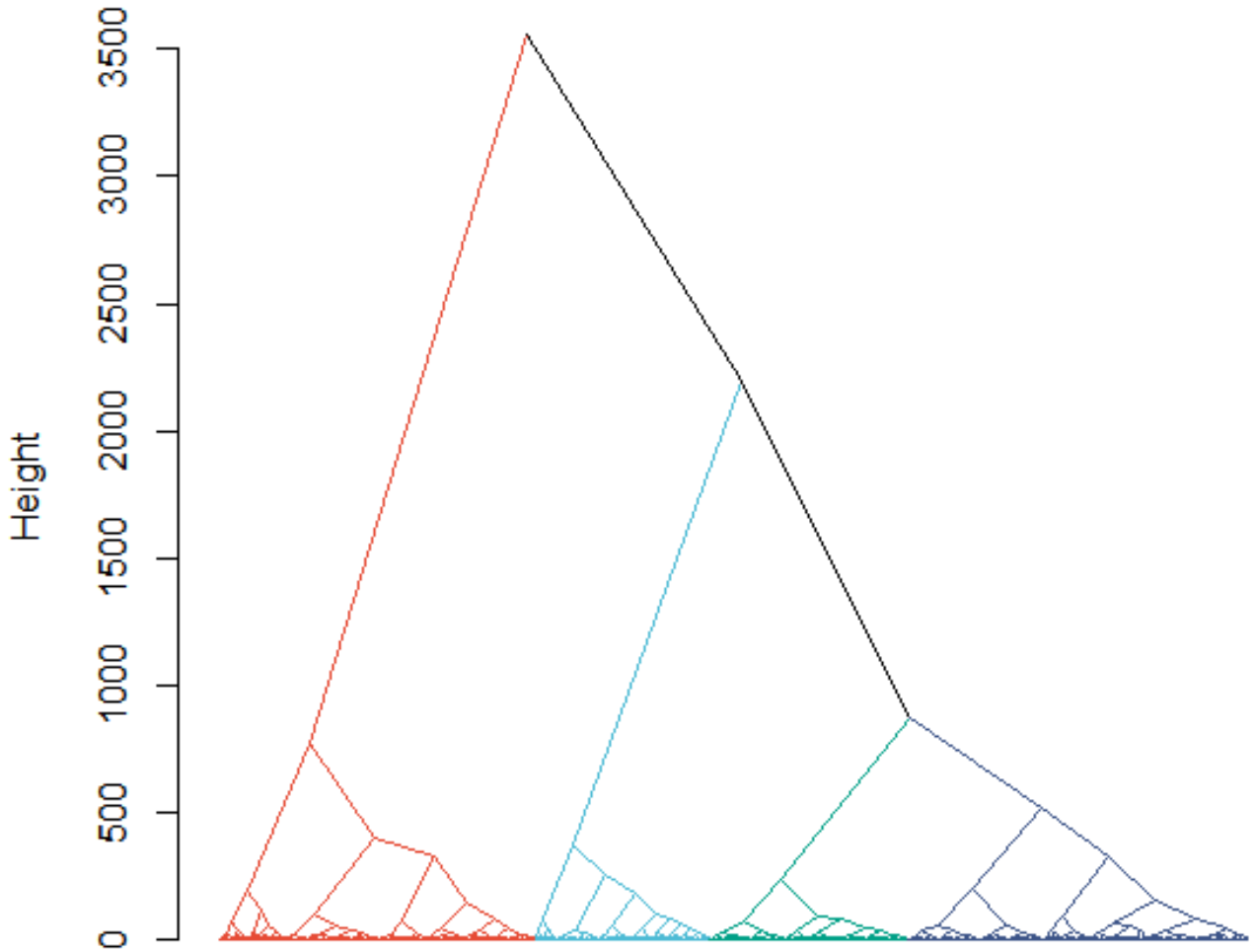


Optimal number of clusters

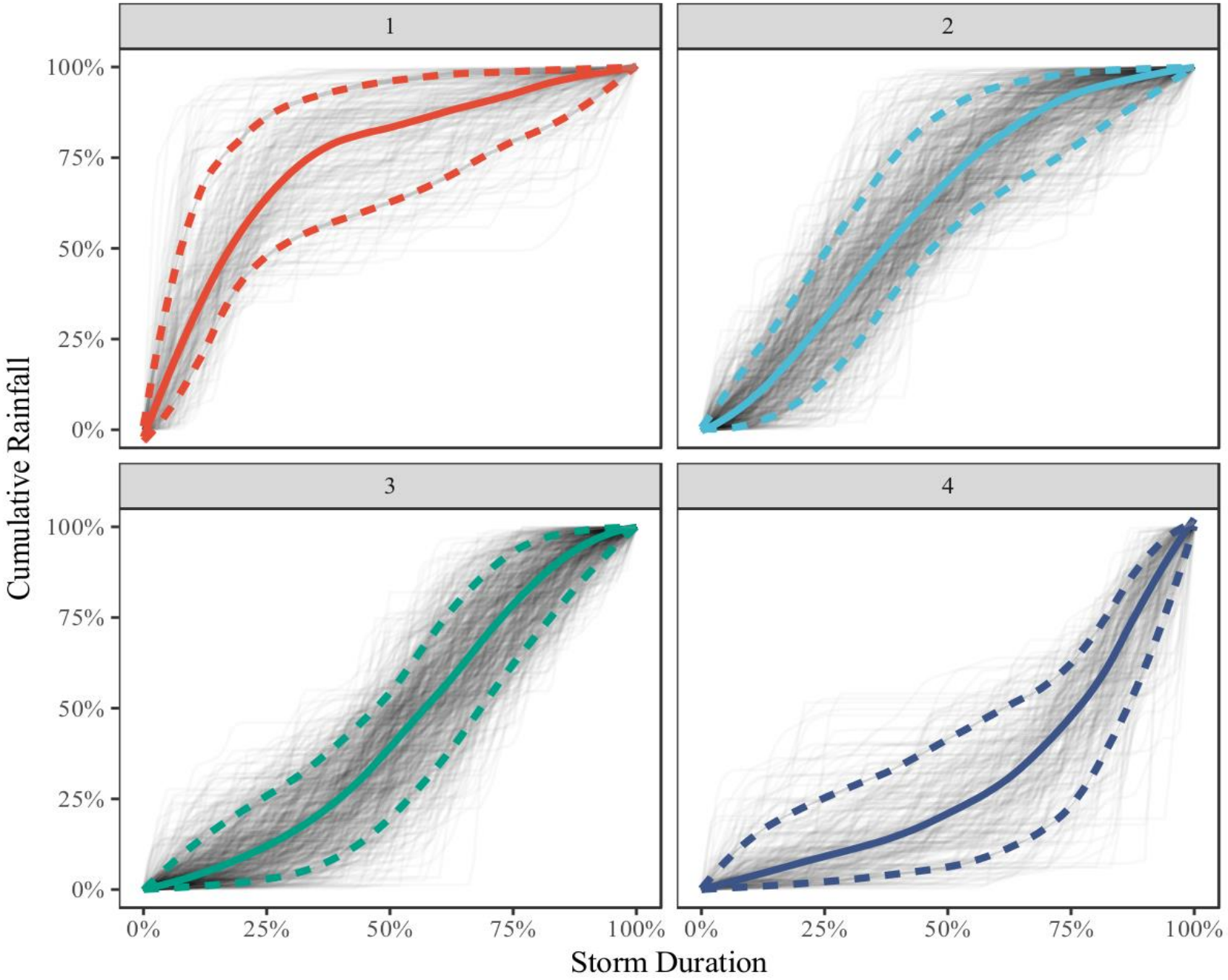
Moving down, the tree is cut into different clusters. The center of each cluster represents a different DH and these DHs are tested if drawn from the same distribution using the two sample Kolmogorov-Smirnov test.

If any of the produced DHs' p-values is not smaller than a predefined significance level the procedure stops.

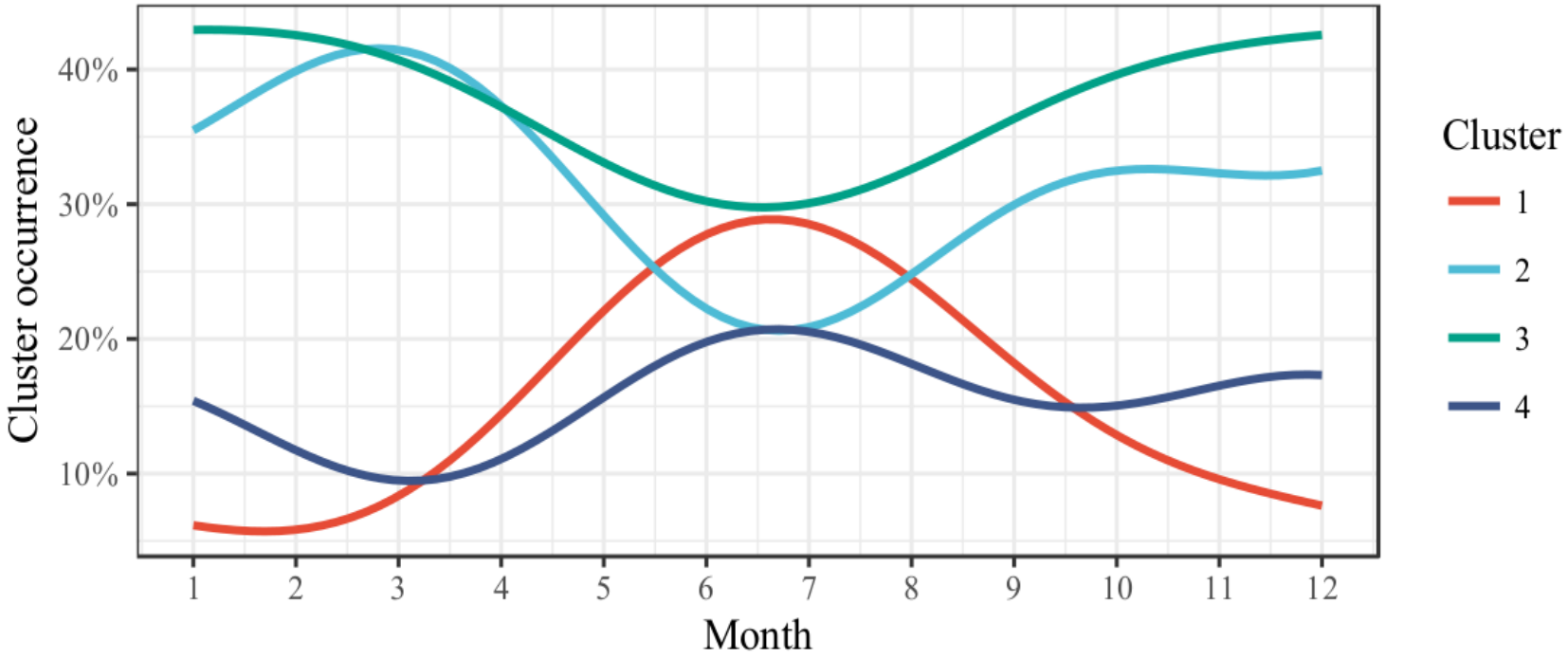
The algorithm identified 4 clusters for $\alpha = 0.05$.



Colored lines are the 10th, 50th and 90th-percentiles dimensionless hyetographs curves derived from the four identified clusters. With grey lines are shown the UCHs of each cluster.



Cluster monthly occurrence



Cluster	Occurrence (%)	Duration (hr)	Prec. (mm)	$I30_{max}$ (mm/hr)
1	12.50	16.25	16.5	20.1
2	32.80	18.75	19.4	13.0
3	39.50	19.5	19.5	12.4
4	15.20	16.5	18.5	16.8

Correlation analysis

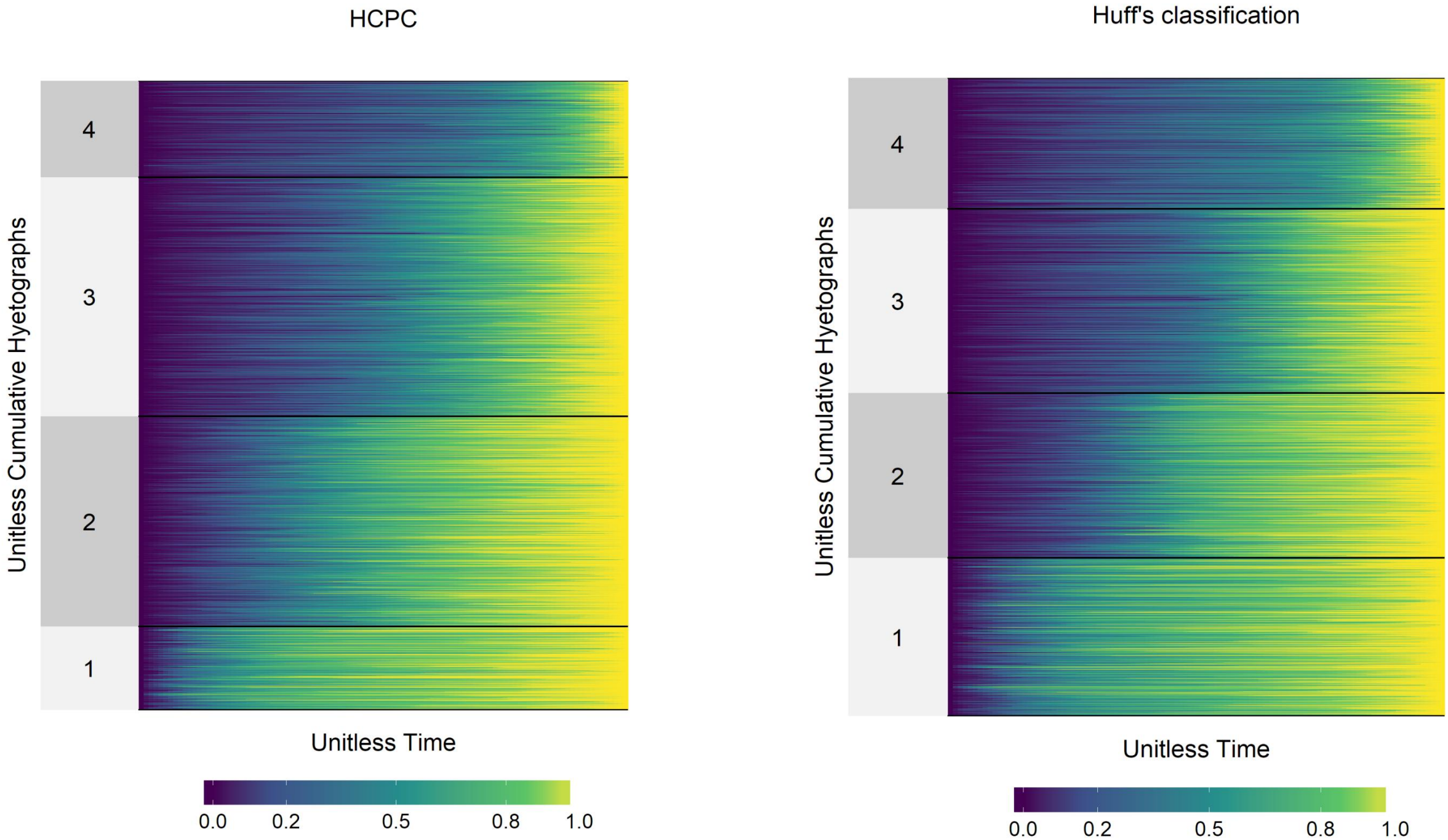
After developing DHs for each station and for every month, correlation matrices were computed, utilizing Pearson's r using the respective UCHs per cluster. These matrices showed very high similarity between:

a) The DHs per station with $r \geq 0.98$ and

b) The DHs per month with $r \geq 0.95$.

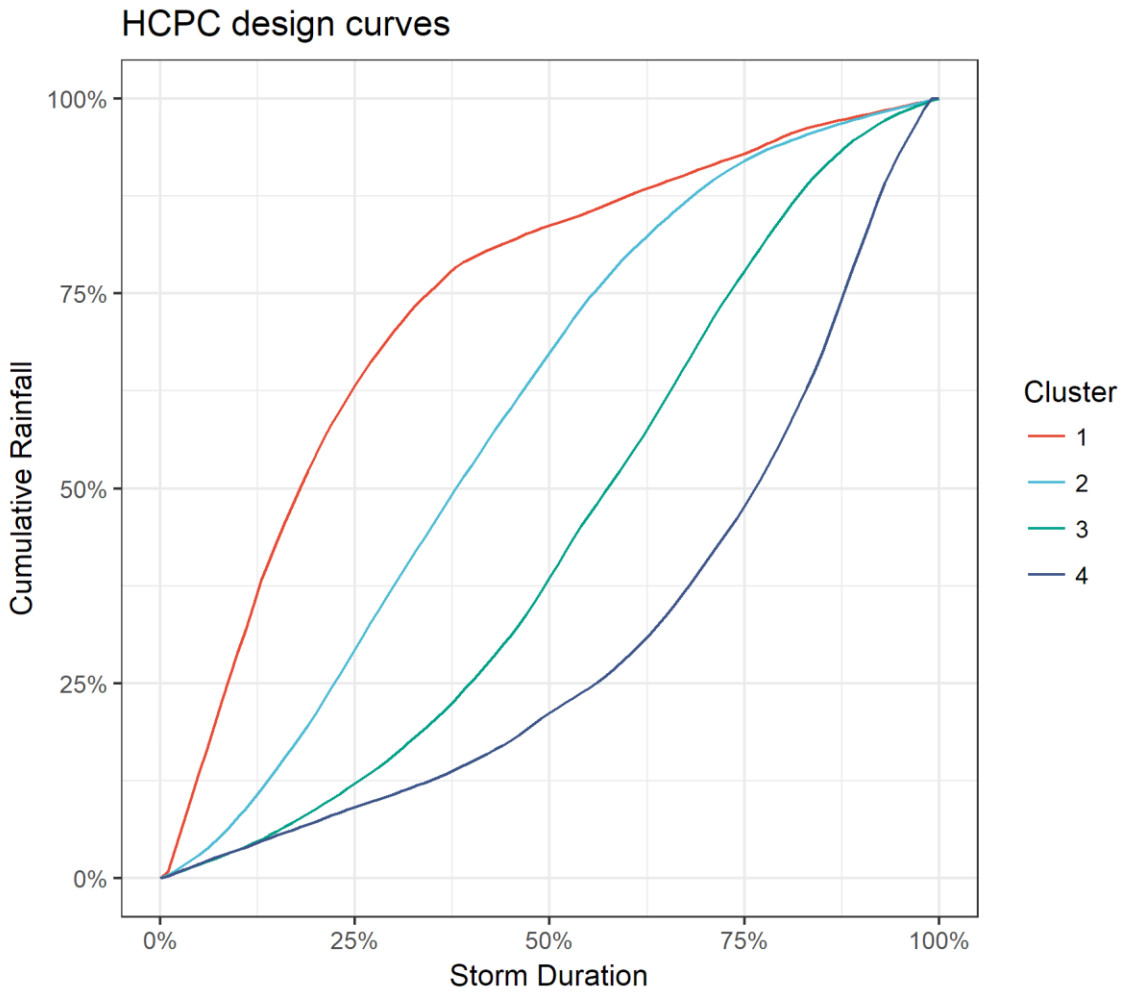
Comparing the proposed method and Huff’s curves

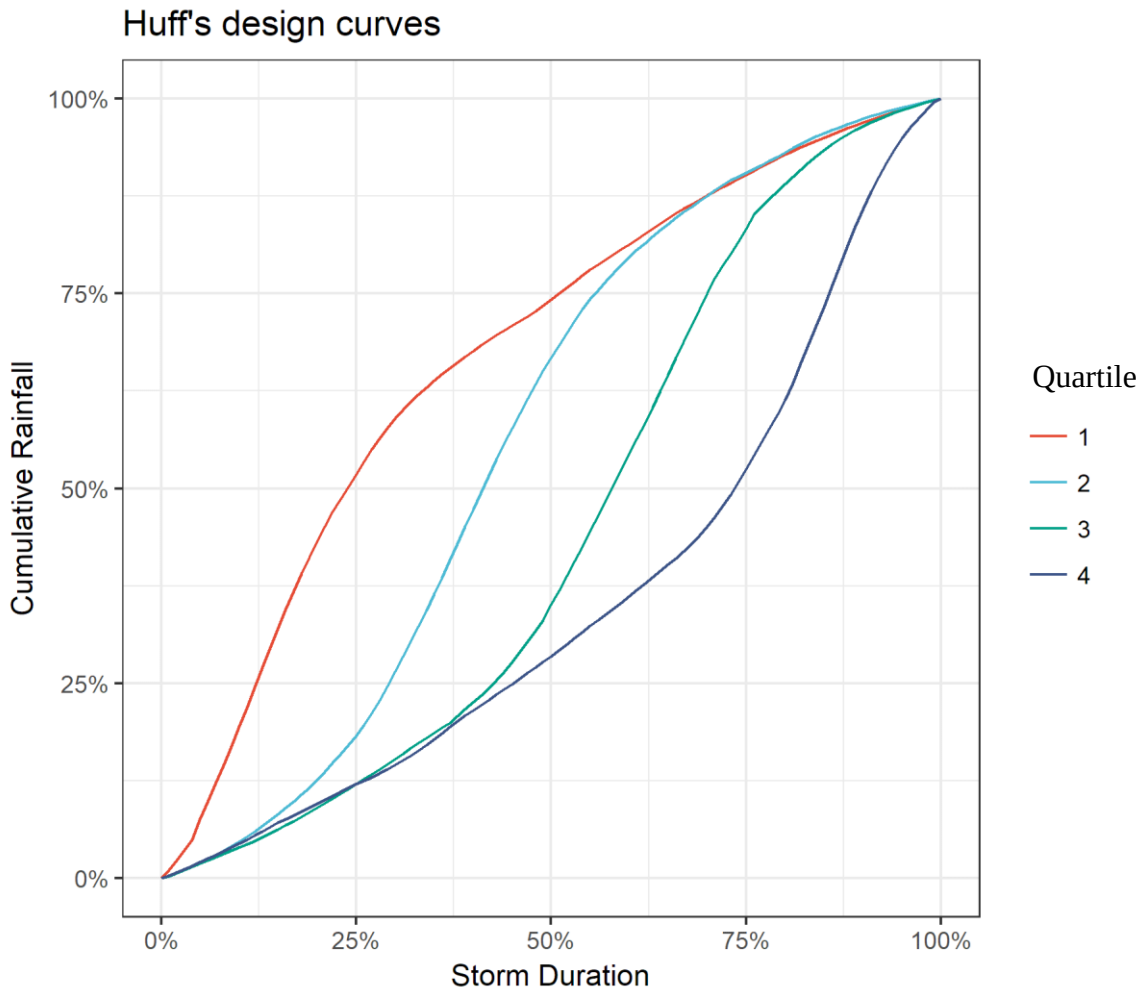
UCH’s classification


HCPC
Unitless Cumulative Hyetographs
4
3
2
1
Unitless Time
0.0
0.2
0.5
0.8
1.0
Huff's classification
Unitless Cumulative Hyetographs
4
3
2
1
Unitless Time
0.0
0.2
0.5
0.8
1.0

Dimensionless Hyetographs Curves

Design Curves





Statistical Analysis

HCPC p-values

	Cluster 2	**Cluster 3**	**Cluster 4**
Cluster 1	0.01	$4 \cdot 10^{-7}$	$2 \cdot 10^{-14}$
Cluster 2	-	$4 \cdot 10^{-2}$	10^{-6}
Cluster 3			$4 \cdot 10^{-2}$

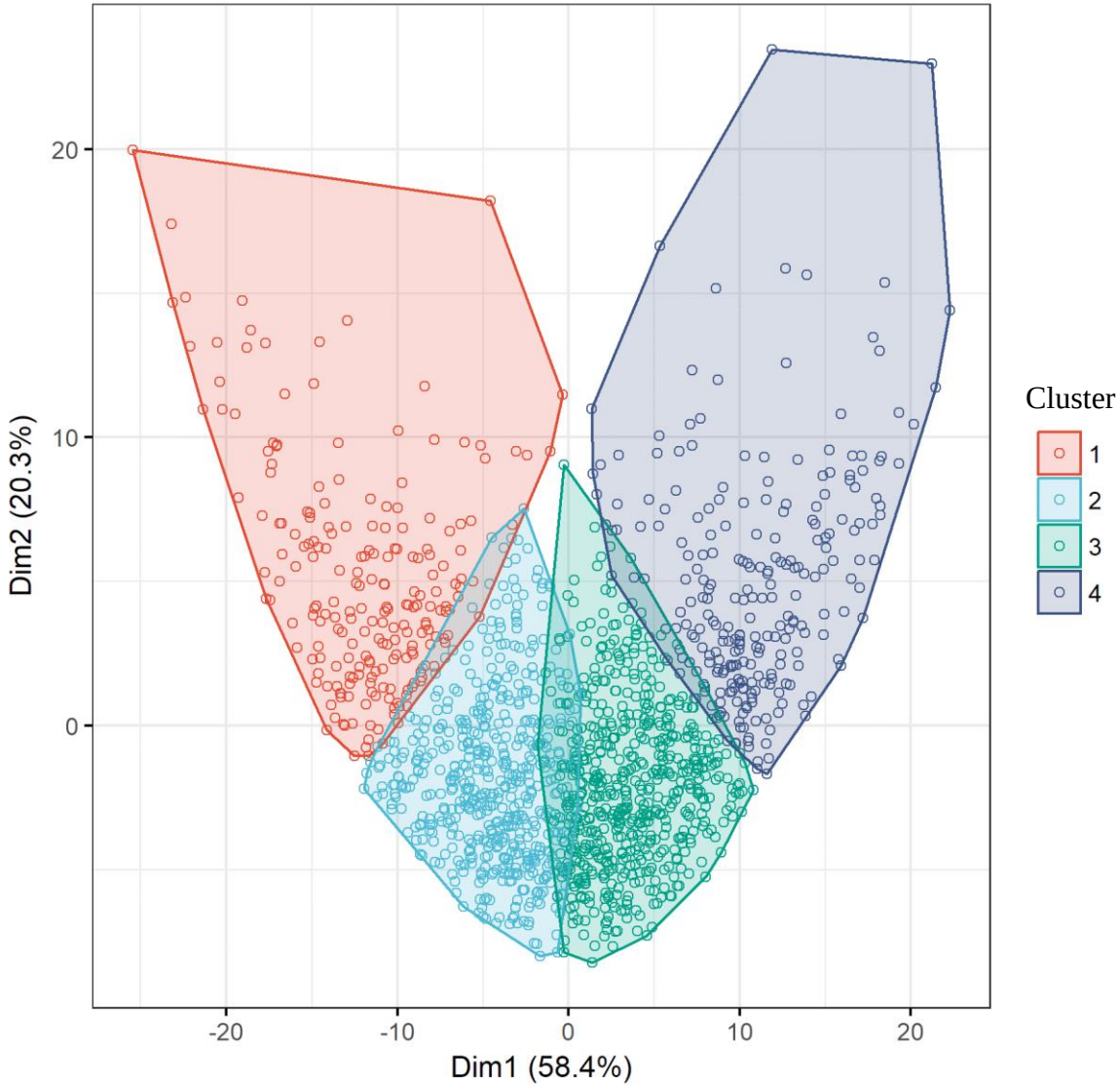
Huff's curves p-values

	Quart 2	**Quart 3**	**Quart 4**
Quart 1	0.10	$4 \cdot 10^{-5}$	$3 \cdot 10^{-10}$
Quart 2	-	0.11	$4 \cdot 10^{-5}$
Quart 3			0.12

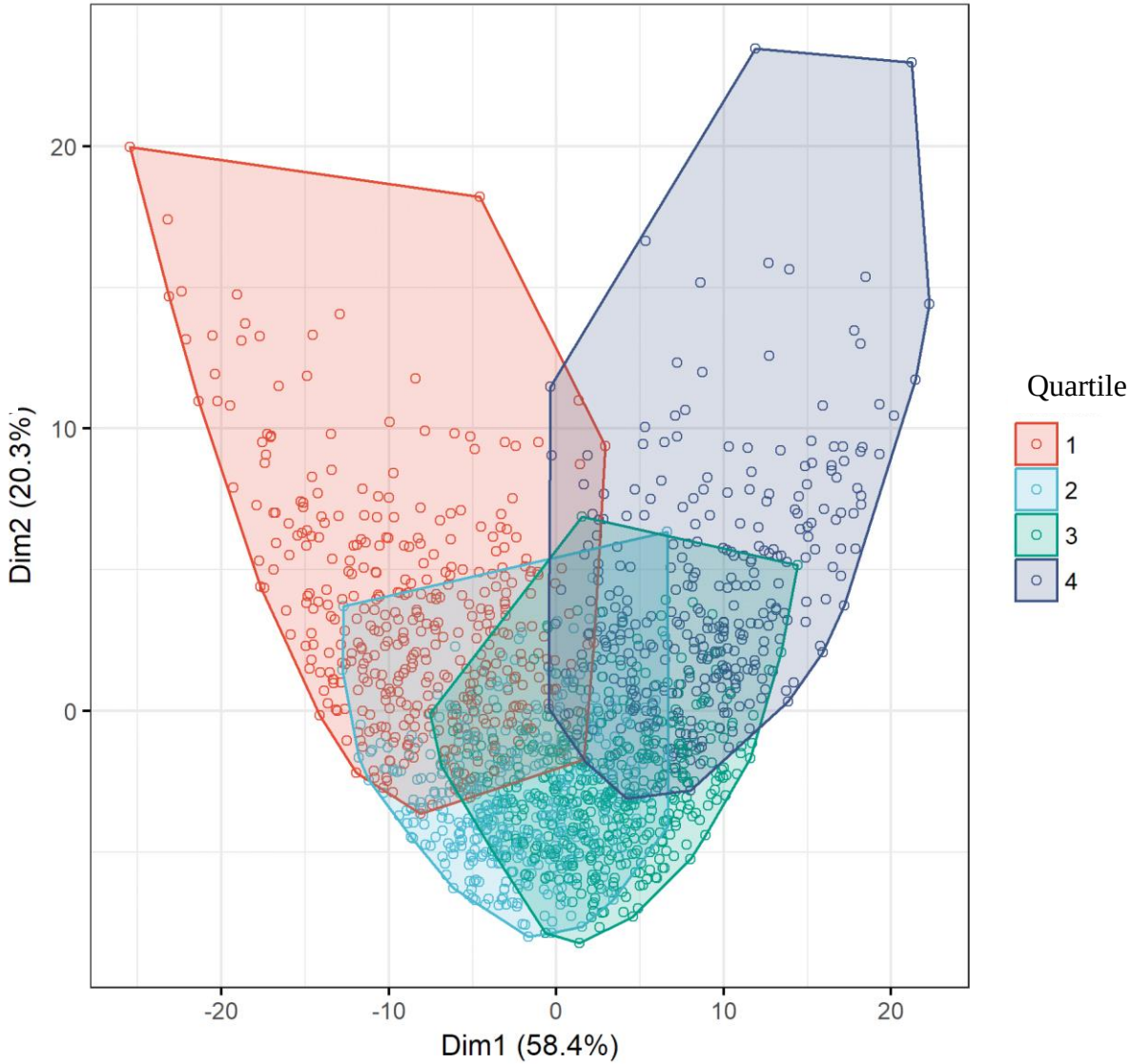
Each possible pairs of curves, are tested if are drawn from the same distribution using the two-sample Kolmogorov-Smirnov test. The p-values are adjusted using the Benjamini and Hochberg method which controls the false discovery rate (because of the multiple statistical tests).
Three pairs of the Huff's curves fail to reject the hypothesis that are drawn from the same distribution for both $\alpha = 0.05$ and $\alpha = 0.10$.

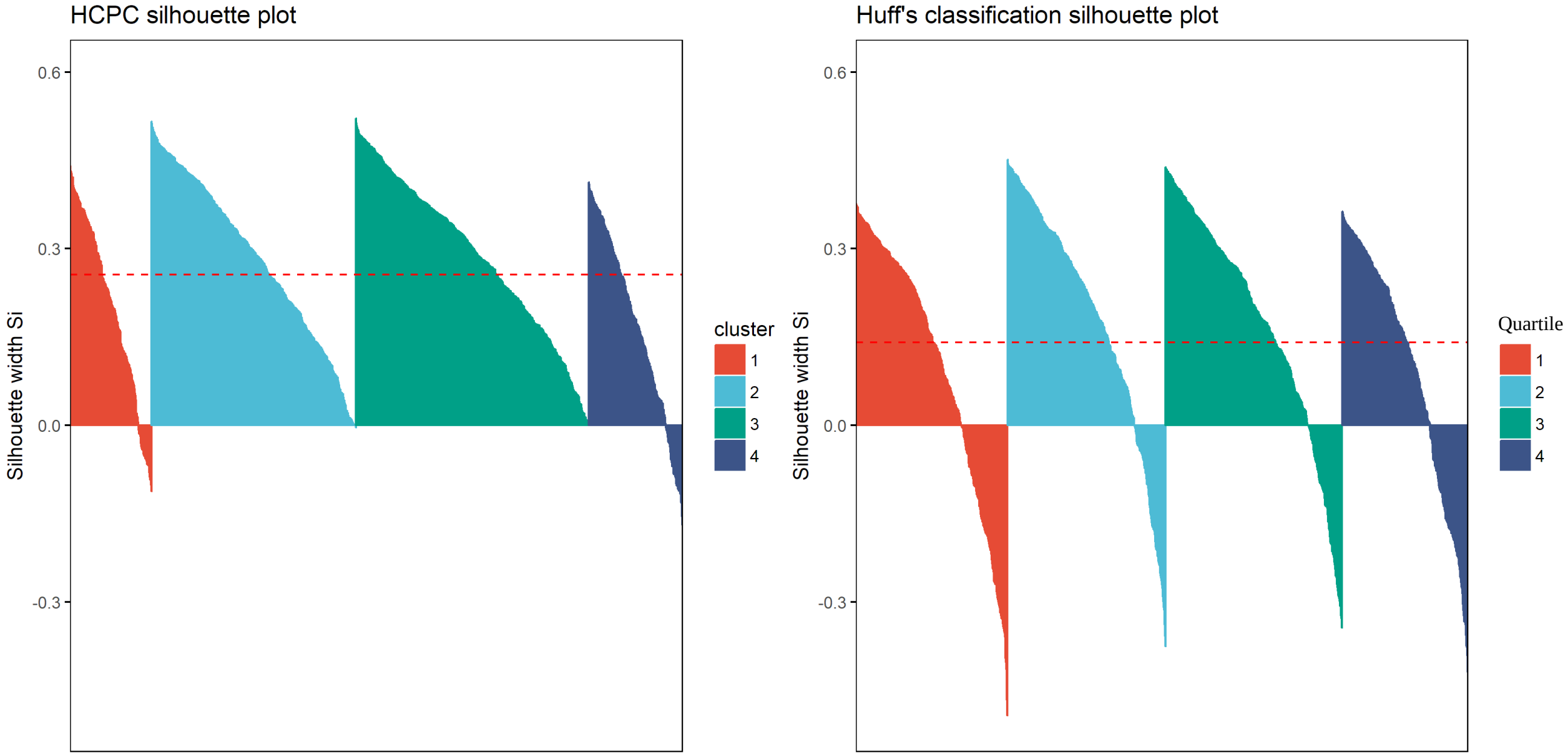
Clustering Validation

HCPC cluster's ellipses



Huff's classification ellipses



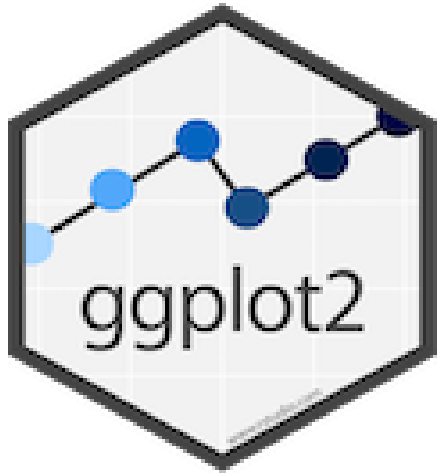

HCPC silhouette plot
Silhouette width Si
0.6
0.3
0.0
-0.3
cluster
1
2
3
4
Huff's classification silhouette plot
Silhouette width Si
0.6
0.3
0.0
-0.3
Quartile
1
2
3
4

Conclusions

Conclusions

1. A temporal model of critical dry duration between rainstorms was introduced in contrast to more simplified approaches of the literature.
2. Only a small number of Principal Components sufficiently explain almost all of the variability of the observed UCHs.
3. Clustering tendency showed that the UCHs data set contains meaningful clusters (i.e. non-random structures).
4. Internal structure validation showed that the proposed method provide better classification of the UCHs.
5. Four representative design hyetographs were produced. Such hyetographs have not been derived in Greece so far, especially in a way that covers the various Water Divisions.
6. The proposed methodology may be utilized for the systematic production of such hyetographs, also based on intensity-duration-frequency curves.
7. This method is fully unsupervised, as no prior empirical knowledge is used.

The analysis and the algorithms were implemented in the R language using the packages: hydroscoper, FactoMineR, factoextra, ggplot2, tidyverse.



Thank you for your attention