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
GR- 54124 Thessaloniki, Macedonia, Greece

*Corresponding author e-mail: nontas@topo.auth.gr, tel : +302310996143

Abstract

An unsupervised method that utilizes a combination of statistical and machine learning techniques is presented in order to classify statistically independent rainstorm events and create a limited number of design hyetographs for the Water Division of Thrace in Greece. The whole process includes the necessary steps from importing raw precipitation time series data to producing the initially unknown optimal number of representative design hyetographs. These hyetographs can be used for stochastic simulation, water resources planning, water quality assessment and global change studying. The present type of analysis is applied for the first time on data from a Greek region and, in addition, it presents certain characteristics of a more general applicability. Namely, the method employed is fully unsupervised, as no empirical knowledge of local rainfalls is implicated or any arbitrary introduction of quartiles for grouping. Also, the critical time duration of no precipitation between rainstorm events is not defined in advance, as is the case in the pertinent literature.

Keywords

rainfall temporal distribution; design hyetographs; unsupervised machine learning; hierarchical clustering; Principal Components Analysis

1. INTRODUCTION

Knowledge about the temporal distribution of rainfall is essential in current methods of water resources management such as drainage design, erosion control, water quality assessment and global change studies. A typical methodology includes the determination of total duration and height of rainfall and disaggregation of this height using a temporal pattern that represents the expected internal rainfall structure, the design hyetograph (DH). Veneciano and Villani (1999) provided categorization of methods for the production of design hyetographs, distinguishing four types. The first two methods are based on intensity-duration-frequency curves, the third method is based on standardized profiles derived from rainfall records and the last method relies on stochastic rainfall models via simulation. The first three methods are used more frequently.

Huff (1967) presented a probabilistic method, in which storm data are classified using the quartile where the maximum intensity occurs. More details about the development and utility of Huff's curves in disaggregation and stochastic simulation can be found in the literature (Bonta and Rao, 1987; Bonta and Shahalam, 2003; Bonta, 2004a, 2004b; Vandenberghe et al., 2010). A necessary step prior to the construction of Huff's curves is the extraction of individual rainstorm events from precipitation time series. Huff used a six-hour fixed Critical time Duration (CD) of no precipitation to separate these events, and many researchers followed the same approach (Loukas and Quick, 1996; Williams-Sether et al., 2004; Azli and Rao, 2010; Dolšak et al., 2016), although Bonta (2001)

showed that CD has seasonal variability. The determination of rainfall temporal distribution is dealt with in this paper by means of machine learning methods.

Applications of machine learning using hydro-meteorological data, in general, has been dealt within the literature, in terms of supervised methods trained on big datasets, such as infilling erosivity values (Vantas and Sidiropoulos, 2017) or to create more accurate models than widely used formulae, such as the flow velocity prediction (Kitsikoudis et al., 2015). The use of unsupervised methods in relation to the special issue of temporal distribution of rainfall is scarce. Self-organized maps have been applied to a small data-set to estimate design storms (Lin and Wu, 2007) and k-means clustering has been used to create a predefined number of rainfall patterns (Nojumuddin and Yusop, 2015).

This paper presents an original, controlled, fully reproducible, unsupervised method that produces automatically and objectively the optimal number of DHs using precipitation records. This method comprises of the following steps: a) Raw precipitation data is cleaned from noise and errors. b) CD is determined on the basis of a Poisson process hypothesis. c) A temporal model of CD is constructed with the above rainfall data. d) Unitless Cumulative Hyetographs (UCH) are compiled and Principal Components Analysis (PCA) is applied to the UCH's. e) Agglomerative hierarchical clustering is applied on the principal components (HCPC). f) The number of clusters is determined by repetitive statistical comparisons between the centers of the clusters already produced at the previous steps. g) Finally, a limited number of DHs is produced that represents the rainstorm records.

2. MATERIAL AND METHODS

2.1 Study area and Dataset

The study region, located to the north-east Greece (Fig 1.), extends to an area of 11,243 km^2 that covers the Water Division of Thrace. It is delimited by the boundaries of Greece, Bulgaria and Turkey on the north and east, by the Thracian Sea on the south and by the watershed of Nestos River on the west.

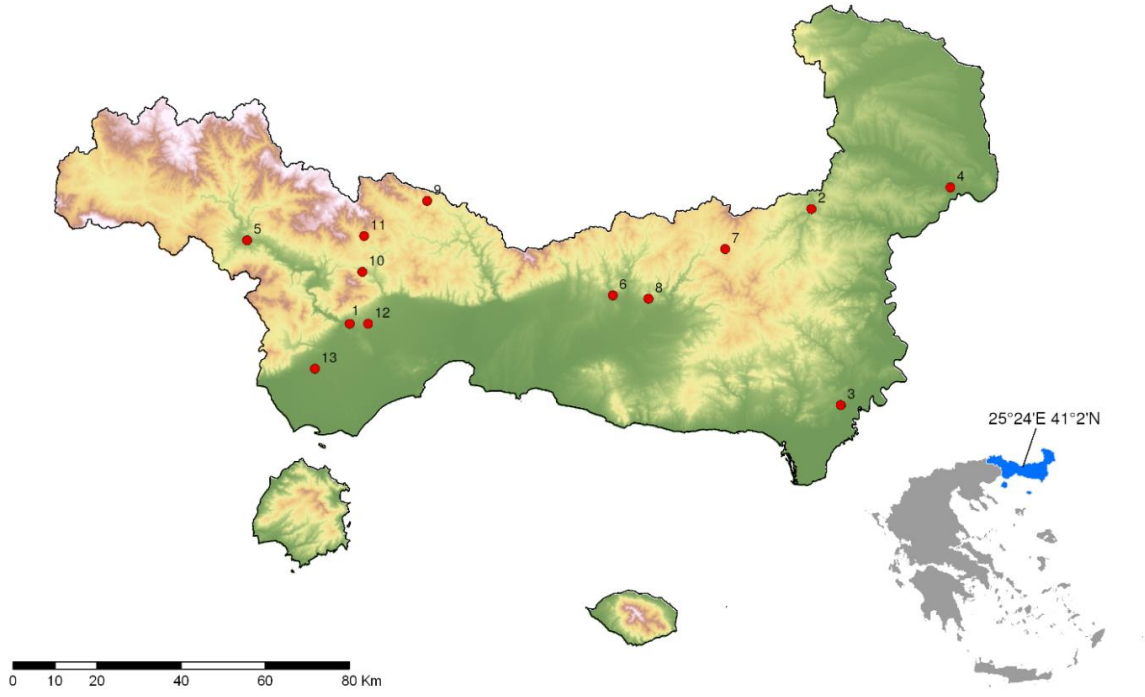

Figure 1. Location of the study area and the 13 meteorological stations from the Greek National Databank for Hydro-meteorological Information.

The climate is predominantly Mediterranean and annual rainfall ranges from 500 mm in coastal and insular areas to 1000 mm in the northern mountainous areas (Ministry of Environment and Energy, 2013). The data utilized in the analysis was taken from the Greek National Databank for Hydro-meteorological Information (Vafeiadis et al., 1994) and came from 13 meteorological stations. The data coverage was 37%, on average (Table 1). The time series comprised a total of 413 years of pluviograph records with a time step of 30 minutes for the time period from 1956 to 1997. The time series rainfall records were checked for consistency and errors which were: a) There were repetitive values, where the same rainfall was recorded over a long-time period, and these were set to zero, b) there were records of aggregated values, where the time step was larger than 30 min, and these were removed, c) there were records where the time step was 5 min and these where aggregated to 30 min, d) probably due to the initial digitization of the pluviometers' bands, there were values near zero (i.e. $\ll 0.01$ mm) which were set to zero.

TABLE 1. Meteorological stations location, pluviograph records data coverage and duration.

	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%

2.2 Storm identification

A Poisson process hypothesis is assumed for the division of the precipitation time series to statistically-independent rainstorm events, in which: a) the events' interarrival times t_α that come from the same month are distributed exponentially, b) the events are separated by a monthly, constant, minimum Critical time Duration of no precipitation, CD, and c) there is a seasonal pattern for CD in the area of interest. The probability density function of t_α is (Restrepo-Posada and Eagleson, 1982):

$$f(t_a) = \omega \cdot e^{-\omega \cdot t_\alpha}, \quad t_a \geq 0 \tag{1}$$

where ω is the average storm arrival rate and:

$$t_a = t_r + t_b \tag{2}$$

where t_r is the storm duration and t_b is the dry time between rainstorms. The estimation of CD is based on an iterative procedure of statistical tests where inter-month data per station are used to ensure homogeneity (Koutsoyiannis and Xanthopoulos, 1990). In Algorithm 1, Appendix, a vector of test CD values is used to compute t_α values and $\widehat{\omega}$ is estimated from this sample of values. Then a non-parametric bootstrap method (Babu and Rao, 2004) that utilizes the one-sample Kolmogorov-Smirnov test (William, 1971) is applied to test the goodness-of-fit, only if the sample size is moderate to large (i.e. ≥ 50), because the data suffer from significant proportions of missing values.

Finally, a temporal, sinusoidal model for the Water Division's CD values per month is fitted (Equation 4, Algorithm 1).

2.3 Development of Unitless Cumulative Hyetographs and Principal Components Analysis

The rainstorms are extracted from the dataset using the monthly CD values obtained from the fitted model of Algorithm 1. The general approach given by Bonta (2004) is followed and only the events with duration greater than 3 hours and cumulative rainfall greater than 12.7 mm are used in the analysis. The hyetographs of the rainstorms that meet these criteria are transformed to unitless form in which a) the time expresses the percentage of the rainstorm duration and b) the cumulative rainfall expresses the percentage of total rainstorm height. Because the UCHs' vectors in this form have variable length, linear interpolation is applied to compute the unitless cumulative rainfall for every 1% of unitless time values. Finally, a matrix of UCHs, $\boldsymbol{U}$ is produced with the values of unitless cumulative rainfall, with every row representing the rainstorm and every column the unitless time values.

Because the time variables (i.e. the m columns of the $\boldsymbol{U}$ matrix) are highly correlated, Principal Component Analysis (PCA, Pearson, 1901) is applied to reduce the dimensionality of the data to a few dimensions. The number of dimensions to retain is determined using the proportion of total variance of the data explained (Jolliffe, 1986). In this analysis this level is set to 99.5%, to ensure that almost all the information from UCHs will be preserved.

2.4 Clustering Analysis

The Hopkins index, H (Lawson and Jurs, 1990), for clustering tendency is applied, because all the clustering algorithms can return clusters even if there was no structure in the $\boldsymbol{U}$ matrix. The computed value of H was 0.88, thus it indicates clustering tendency at the 90% confidence level (Han et al., 2011). The clustering method applied is Hierarchical Clustering on Principal Components (HCPC), using Ward's minimum variance criterion that minimizes the total within-cluster variance (Ward, 1963). This criterion was utilized because it is based on the minimum variance as is PCA (Husson et. al, 2010). The result is a tree-based representation of the UCHs.

The number of clusters is selected from the produced hierarchical tree using the top-down iterative Algorithm 2, Appendix. At each step of the iteration the dendrogram is cut into different groups of UCHs. The center of each group represents a different design hyetograph and these hyetographs, for all possible pairs, are tested if are drawn from the same distribution using the two-sample Kolmogorov-Smirnov test (William, 1971). Because of the multiple pairwise tests, the p-values that resulted are adjusted using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995), which controls the false discovery rate. If any of the produced design hyetographs' p-values is not smaller than a predefined significance level α, the procedure stops and the optimal number of clusters is found. Silhouette analysis (Rousseeuw, 1987) was applied to validate the internal structure of clustering.

3. RESULTS

The relation between the p-values and CD-values was found to have a global maximum for every station and month, which is a desirable feature of Algorithm 1. The fitted monthly sinusoidal model of CD shows a temporal variation during summer months, with an average value of 9 hours, while for the rest of the year the same quantity averages 6.5 hours. Using the calculated CD-values a population of 1,622 out of 25,377 extracted rainstorms met the criteria of minimum duration and cumulative height. 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 application of Algorithm 2 identifies four clusters and some of their statistics are presented in Table 2. The percentiles' values of the DHs are given in Table 3. The first cluster has the highest variance in monthly occurrence, and the highest average value of maximum 30 min duration's intensity. In Figure 2 the clusters' 10th, 50th

and 90th percentiles are shown with the UCHs that belong to them and in Figure 3 the clusters' monthly occurrence.

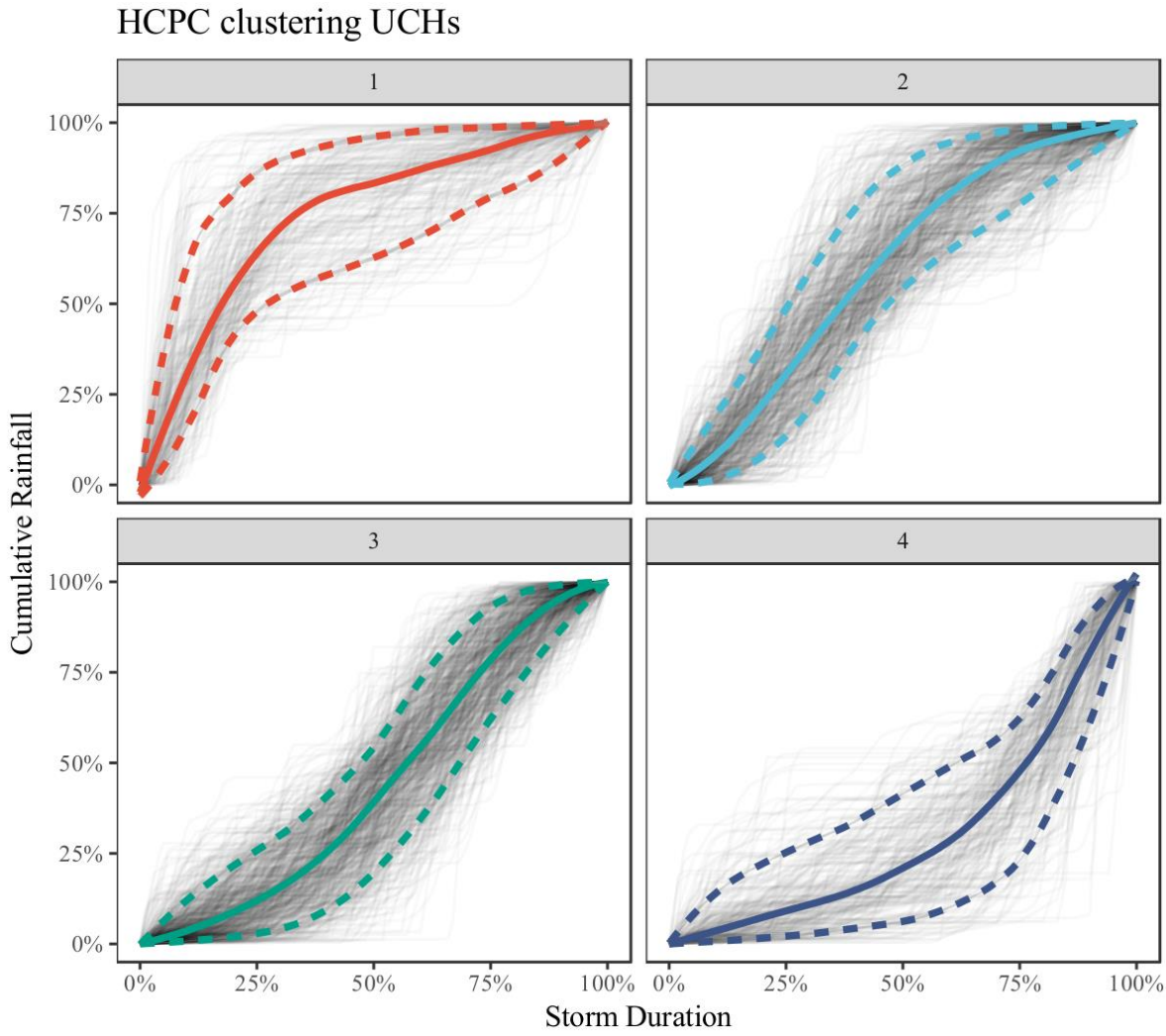

Figure 2: Results from Algorithm 2. At the top the 10th, 50th and 90th-percentiles dimensionless hyetographs curves derived from the four identified clusters. With grey lines are shown the UCHs of each cluster.

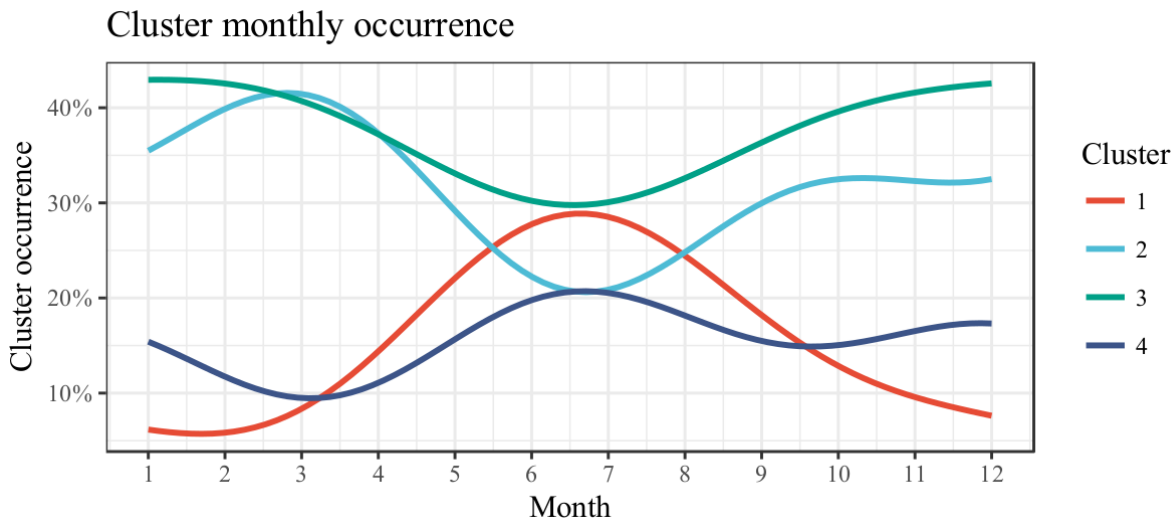

Figure 3. The plot presents the variability of clusters' monthly occurrence.

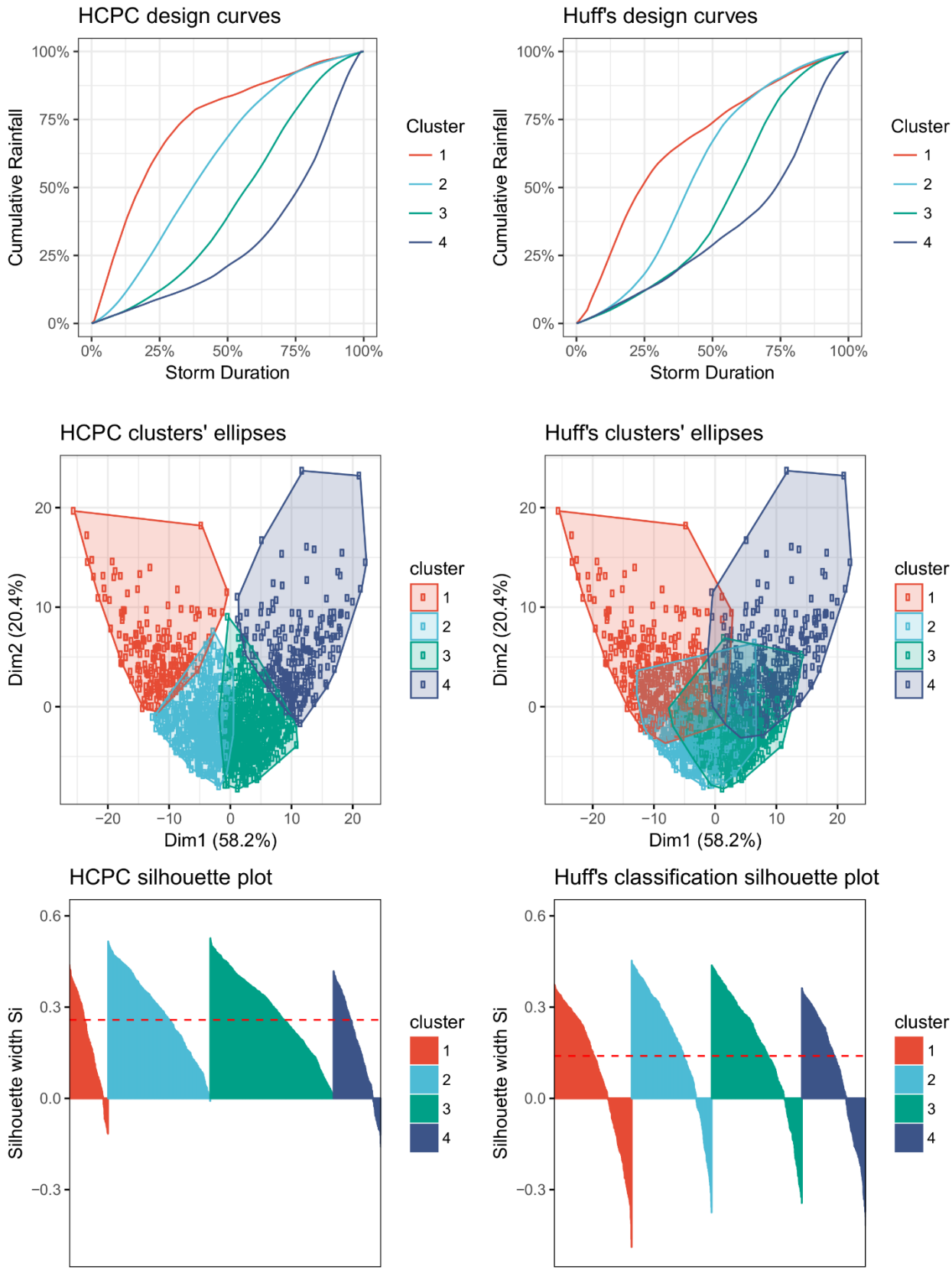

Figure 4: A comparison between the results from HCPC and Huff's quartiles clustering. At the top the derived dimensionless hyetographs curves are shown. In the middle the UCH's plots are shown using the first two principal components and ellipses around the clusters. At the bottom the

silhouette plots are shown and with red, dashed line the average silhouette width of the clustering methods

TABLE 2. Average values of occurrence of clusters, duration, precipitation height and maximum 30 min duration's intensity of clusters' rainstorms.

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

After developing DHs for each station and for every month, correlation matrices were computed, utilizing Pearson's r coefficient (Helsel and Hirsch, 1992), 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 $r \geq 0.95$. A comparison among HCPC and Huff's curves is shown in Figure 4. 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 $\alpha = 0.10$. HCPC results in the clear separation of UHCs, as its clusters ellipses are not overlapping and it creates clusters with better internal structure, as average sill width is almost two times better.

4. CONCLUSIONS

A temporal model of critical dry duration between rainstorms was introduced and implemented and a seasonal variability of rainfall patterns is simulated by the proposed method, in contrast to more simplified approaches of the literature. The unitless cumulative hyetographs produced were subjected to Principal Components Analysis in order to investigate if they can be compressed to a few dimensions, due to high correlation values, and it turned out that only a small number of them sufficiently explain almost all of the variability. Hierarchical Clustering on Principal Components was subsequently applied that yielded a small number of clusters. Clustering tendency and internal structure validation was appropriately investigated and documented. Finally, based on the clustering analysis four representative design hyetographs were produced. These hyetographs do not exist in Greece, especially in a way that covers the various Water Divisions. The proposed methodology may be utilized for the systematic production of such hyetographs, also based on intensity-duration-frequency curves. This method is fully unsupervised, as no prior empirical knowledge is used.

TABLE 3. Design Hyetographs

Storm duration (%)	Cluster 1			Cluster 2			Cluster 3			Cluster 4		
	10th	50th	90th	10th	50th	90th	10th	50th	90th	10th	50th	90th
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5.0	3.9	15.7	38.4	0.4	3.1	10.0	0.4	1.7	6.6	0.4	1.9	8.1
10.0	15.0	31.4	63.5	1.6	8.1	19.2	0.8	3.6	12.4	0.8	3.7	14.3
15.0	29.8	44.9	73.0	3.9	14.9	28.5	1.3	6.0	17.3	1.1	5.5	18.7
20.0	41.8	55.7	80.4	8.0	22.4	38.2	2.0	8.9	21.8	1.6	7.4	22.2
25.0	47.8	64.5	86.8	13.5	30.5	48.6	2.9	12.0	26.0	2.2	9.2	25.3
30.0	52.0	71.4	89.6	20.7	38.9	58.6	4.2	15.7	30.2	2.7	10.9	28.2
35.0	55.8	76.7	92.2	29.9	46.7	67.6	6.2	20.1	35.1	3.5	12.7	31.0
40.0	57.5	80.0	93.7	39.5	54.5	76.9	9.6	25.5	41.1	4.3	14.9	33.8
45.0	60.3	81.8	95.2	47.5	61.8	83.5	13.8	31.6	47.3	5.2	17.6	37.7
50.0	62.7	83.5	96.2	54.0	68.7	88.2	19.6	39.2	54.2	6.1	21.1	41.7
55.0	65.2	85.3	96.9	60.0	75.1	92.2	26.9	47.1	63.4	7.7	24.4	45.0
60.0	68.7	87.4	97.8	64.7	80.4	94.5	34.4	54.2	73.3	10.1	28.3	49.4
65.0	71.8	89.1	98.3	69.0	84.9	96.0	43.4	62.4	80.9	13.0	33.6	52.6
70.0	76.0	90.8	98.5	72.8	89.1	97.4	53.1	71.0	87.8	16.9	40.3	56.4

Storm duration (%)	Cluster 1			Cluster 2			Cluster 3			Cluster 4		
	10th	50th	90th	10th	50th	90th	10th	50th	90th	10th	50th	90th
75.0	79.5	92.5	98.7	77.6	92.3	98.3	62.2	78.5	93.3	22.8	47.7	62.3
80.0	82.0	94.8	99.1	82.1	94.3	98.9	70.7	85.4	96.3	32.4	56.1	70.8
85.0	85.5	96.6	99.3	86.7	96.0	99.3	79.0	91.3	98.1	46.2	67.0	81.3
90.0	89.5	97.7	99.5	91.2	97.5	99.5	86.9	95.5	99.1	60.2	81.1	91.7
95.0	94.6	98.8	99.7	95.6	98.8	99.8	94.0	98.3	99.6	78.0	92.7	98.3
100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

APPENDIX

The analysis and the algorithms were implemented in the R language (R Core Team, 2018) using the packages: hydroscoper (Vantas, 2018), FactoMineR, (Lê et al., 2008) and factoextra (Kassambara and Mundt, 2017).

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 $s = 50,000$;

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

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

Result: Monthly values of CD for the area

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 $q = 1, ..., m$;
  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 $\bar{x}_{opt}$

REFERENCES

1. Azli, M. and Rao, A. R. (2010), 'Development of Huff curves for peninsular Malaysia', **Journal of Hydrology** 388(1-2), pp. 77-84.
2. Babu, G. J. and Rao, C. (2004), 'Goodness-of-fit tests when parameters are estimated', **Sankhya** 66(1), pp. 63-74.
3. Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', **Journal of the Royal Statistical Society**. Series B (Methodological) pp. 289-300.
4. Bonta, J. (2001), 'Characterizing and estimating spatial and temporal variability of times between storms', **Transactions of the ASAE** 44(6), pp. 1593-1601.
5. Bonta, J. (2004a), 'Development and utility of Huff curves for disaggregating precipitation amounts', **Applied Engineering in Agriculture** 20(5), pp. 641-643.
6. Bonta, J. (2004b), 'Stochastic simulation of storm occurrence, depth, duration, and within-storm intensities', **Transactions of the ASAE** 47(5), pp. 1573-1584.
7. Bonta, J. and Rao, A. (1987), 'Factors affecting development of Huff curves', **Transactions of the ASAE** 30(6), 1689-1693.
8. Bonta, J. and Shahalam, A. (2003), 'Cumulative storm rainfall distributions: Comparison of Huff curves', **Journal of Hydrology (New Zealand)** pp. 65-74.
9. Dolšak, D., Bezak, N. and Šraj, M. (2016), 'Temporal characteristics of rainfall events under three climate types in Slovenia', **Journal of Hydrology** 541, pp. 1395-1405.
10. Han, J., Pei, J. and Kamber, M. (2011), Data mining: Concepts and techniques, Elsevier.
11. Helsel, D. R. and Hirsch, R. M. (1992), Statistical methods in water resources, Vol. 49, Elsevier.
12. Huff, F. A. (1967), 'Time distribution of rainfall in heavy storms', **Water Resources Research** 3(4), pp. 1007-1019
13. Husson, F., Josse, J. and Pages, J. (2010), 'Principal component methods-hierarchical clustering partitional clustering: Why would we need to choose for visualizing data', Technical Report-Agrocampus.
14. Jolliffe, I. T. (1986), Principal component analysis and factor analysis, in 'Principal Component Analysis', Springer, pp. 115128.
15. Kassambara, A. and Mundt, F. (2017), factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5.
16. Kitsikoudis, V., Sidiropoulos, E., Iliadis, L. and Hrissanthou, V. (2015), 'A machine learning approach for the mean flow velocity prediction in alluvial channels', **Water Resources Management** 29(12), pp. 4379-4395.
17. Koutsoyiannis, D. and Xanthopoulos, T. (1990), 'A dynamic model for short-scale rainfall disaggregation', **Hydrological Sciences Journal** 35(3), 303-322.
18. Lawson, R. G. and Jurs, P. C. (1990), 'New index for clustering tendency and its application to chemical problems', **Journal of Chemical Information and Computer Sciences** 30(1), pp. 36-41.
19. Lê, S., Josse, J. and Husson, F. (2008), 'FactoMineR: A package for multivariate analysis', Journal of Statistical Software 25(1), pp. 1-18.
20. Lin, G.-F. and Wu, M.-C. (2007), 'A SOM-based approach to estimating design hyetographs of ungauged sites', **Journal of Hydrology** 339(3-4), pp. 216-226.
21. Loukas, A. and Quick, M. C. (1996), 'Spatial and temporal distribution of storm precipitation in southwestern British Columbia', **Journal of Hydrology** 174(1-2), pp. 37-56.
22. Ministry of Environment and Energy (2013), Management plan of Thracian Water Division, Technical report.
23. Nojumuddin, N. S. and Yusop, Z. (2015), 'Identification of rainfall patterns in Johor', **Applied Mathematical Sciences** 9(38), pp. 1869-1888.

24. R Core Team (2018), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.
25. Restrepo-Posada, P. J. and Eagleson, P. S. (1982), 'Identification of independent rainstorms', **Journal of Hydrology** 55(1), pp. 303-319.
26. Rousseeuw, P. J. (1987), 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', **Journal of Computational and Applied Mathematics,** 20, pp. 53-65.
27. Vafeiadis, M., Tolikas, D. and Koutsoyiannis, D. (1994), HYDROSCOPE: The new Greek national database system for meteorological, hydrological and hydrogeological information, in 'WIT Transactions on Ecology and the Environment', pp. 1-8.
28. Vandenberghe, S., Verhoest, N., Buyse, E. and De Baets, B. (2010), 'A stochastic design rainfall generator based on copulas and mass curves', **Hydrology and Earth System Sciences** 14(12), pp. 2429-2442.
29. Vantas, K. (2018), 'hydroscoper: R interface to the Greek National Data Bank for Hydrological and Meteorological Information'. **Journal of Open Source Software**, 3(23), 625.
30. Vantas, K. and Sidiropoulos, E. (2017), 'Imputation of erosivity values under incomplete rainfall data by machine learning methods', **European Water** 57, pp. 193-197.
31. Veneziano, D. and Villani, P. (1999), 'Best linear unbiased design hyetograph', **Water Resources Research** 35(9), pp. 2725-2738.
32. Ward, J. (1963), 'Hierarchical grouping to optimize an objective function', **Journal of the American Statistical Association,** 58(301), pp. 236-244.
33. William, J. C. (1971), Practical Nonparametric Statistics, John Wiley & Sons, New York.
34. Williams-Sether, T., Asquith, W. H., Thompson, D. B., Cleveland, T. G. and Fang, X. (2004), 'Empirical, dimensionless, cumulative-rainfall hyetographs developed from 1959-86 storm data for selected small watersheds in Texas', Technical report, US Geological Survey.