CLUSTERING OF A RUNOFF REGIME IN SLOVAKIA USING THE PCA METHOD

Zuzana Sabová*,[1], Anna Liová[1], Silvia Kohnová[1]

*zuzana.sabova@stuba.sk
[1]Department of Land and Water Resources Management, Faculty of Civil Engineering, Slovak University of Technology in Bratislava, Radlinského 11, 810 05 Bratislava, Slovak Republic

Abstract
A new scheme has been created for pooling data on the long-term average monthly runoff in Slovakia using information from 57 gauging stations from 1991 to 2020. A statistical analysis has determined the optimal number of clusters from the normalised data, and the Principal Component Analysis (PCA) method and K-means clustering were used to group basins into five pooling groups. Finally, the critical characteristics of these pooling groups were determined, and a typical regime was determined.

Keywords
pooling groups, PCA method, K-means clustering, average monthly discharges, runoff regime

1 INTRODUCTION

Climate change is currently the most topical environmental issue addressed worldwide. The impact of climate change on hydrological characteristics is significant. Various analyses have been performed to assess the impact of climate change on the hydrological regimes of watercourses; these include modelling future changes in a hydrological regime or an analysis of a hydrological regime by determining indicators of changes in runoff formation [1]. Regular changes in the water level, flow rate, and flow rate at a specific time are characteristic changes of river runoff regimes. Elements of climate change (such as the air temperature and air humidity, and the wind speed and its frequency) play essential roles in supplying watercourses with water and altering their conditions, so they can also reduce the amount of water flowing into watercourses [2]. In order to better evaluate the effects of climate change on hydrological regimes, the pooling method is used. Runoff pooling is a method of determining a basin's hydrological characteristics without any direct observation. It is a spatial variant of the classification system, which divides the territory into pooling groups with similar or uniform groups of hydrological characteristics. The pooling aims to define hydrologically similar territories where the same calculation methods can be used to determine hydrological quantities [3], [4].

There is currently a growing interest in hydrological assessments to understand alterations in climate conditions. In addition to uncertainties resulting from climate models and emission scenarios, a primary source of uncertainty can also be the variability of local climates. A study that addressed this issue in central Norway found significant changes in runoff regimes [5]. The authors used a set of daily precipitation and daily mean temperatures from stochastic weather generators trained on historical data, along with climate change information obtained from a regional climate model. Anthropogenic activities also affect changes in a runoff regime. On the middle and lower Yellow River (China), there have been changes since 1970 in the runoff regime due to human activity. The results were obtained through wavelet analysis, which is used to analyse the effects of human activity on a runoff regime [6].

In the territory of Slovakia, the first pooling of a runoff regime was carried out in 1957. Dub singled out three primary areas according to the percentage share of the first half of the year: high-mountain, mid-mountain and highland-lowland [7]. In 1980, according to the distribution of runoff during the year and the dominant source of water, five areas were distinguished: high-mountain, two mid-mountain areas and two highland-lowlands [8]. Grešková [4] processed the pooling of streamflows in hydrology and hydrogeography. Hanušin [9] used the months with the maximum and minimum average monthly discharges, along with the coefficient of the variation of the average monthly discharges during the year, to derive five types of a runoff regime.

The 1980s are considered a breakthrough period in the development of hydroclimatic variables, during which a decrease in runoff for the upper Hron basin was detected in the winter. The causes of the changes in a runoff regime are a significant increase in air temperature, a decrease in the snow cover depth, and changes in the seasonal distribution of precipitation [1].

DOI 10.13164/juniorstav.2024.24100

An important study that deals with the regionalisation of the runoff regime in the territory of Slovakia is a study by the Slovak Hydrometeorological Institute and Institute of Hydrology of the Slovak Republic [10]. In this work, the cluster analysis method of the long-term average monthly discharges in the reference period of 1961–2000 from 209 gauging stations was used. For selecting the basins, the condition was determined that the area would be up to 300 km^2 and that the basins would have at least 20 years of discharge measurements. The results from determining individual types of runoff regimes were based on the percentages of the long-term monthly discharges within the hydrological year. The authors also used spatial extrapolation of the regional types established in selected basins in Slovakia, which increased the number of basins studied to 1,441.

This paper applies the Principal Component Analysis (PCA method) and K-means clustering for grouping the average monthly discharges for selected gauging stations in Slovakia. A new reference period (1991–2020) was used. The results of the analysis are presented in graphic and tabular form and can be used for the arbitrary classification of new basins into regional types.

2 METHODOLOGY

In this study, the data of the long-term average monthly discharges from 57 gauging stations in the territory of Slovakia, which were provided by the Slovak Hydrometeorological Institute (Fig. 1), were used. The selected basin areas ranged from 7.25 km^2 (5130 – the Spariská gauging station on the Vydrica stream) to 11,474.30 km^2 (9670 – the Streda nad Bodrogom gauging station on the Bodrog stream).

Before conducting cluster analysis, a crucial consideration is whether the data should be standardized. It is essential to acknowledge that many distance measures are susceptible to scale choices, resulting in varying numerical magnitudes. Standardization encompasses both character and object standardization:

- Character standardization involves the most commonly applied method, normalizing each character to its Z-score by subtracting the mean and dividing by the standard deviation. This form of standardization is referred to as Z-function normalization.
- Object standardization, at times, proves advantageous by transforming data to mitigate variance inconsistency, such as addressing skewness in data distribution [11].

The data from the gauging stations were normalized using standard normalization and then processed using the PCA method and K-means clustering. QGIS (version 3.26.3) and R Studio (version 2022.12.0) created the outputs.

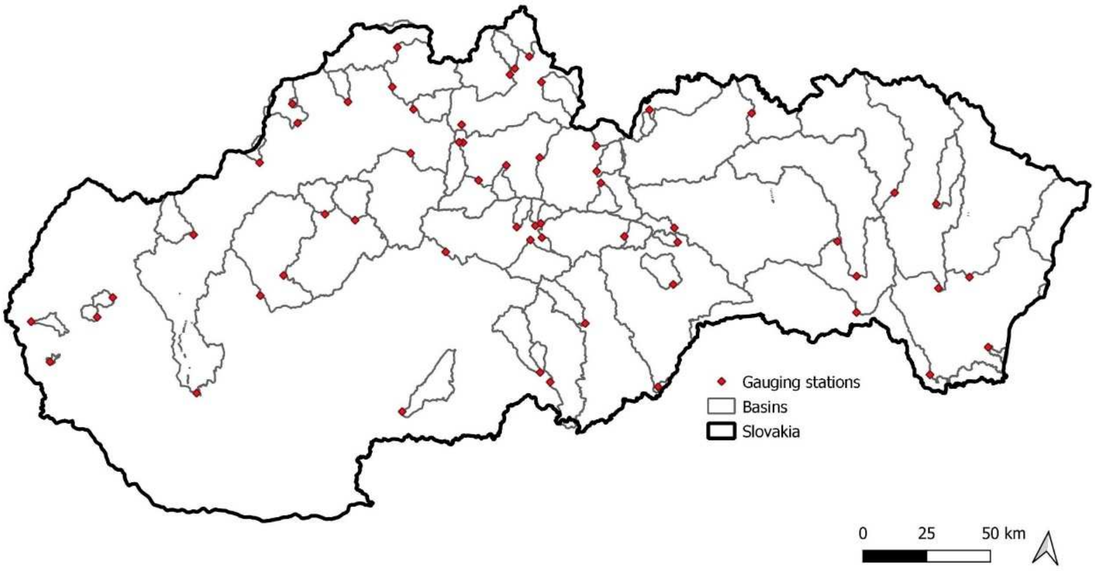

Fig. 1 Location of the selected gauging stations in Slovakia.

Cluster analysis is one of the techniques employed to explore the likeness among multiparametric objects, which involves utilizing numerous variables. The process involves categorizing these objects into classes, each constituting a cluster. A cluster is a collection of objects where the measured distance (or dissimilarity) between them is smaller than the distance between objects that do not belong to the same cluster. This method is beneficial in situations where objects naturally tend to group [11].

Principal Component Analysis (PCA) method is a multivariate statistical technique that is among the most widely used methods of data exploration and analysis in multiple fields of inquiry. PCA analysis is a dimensionality reduction method and is most useful when large amounts of data are available, i.e., there are multiple observations per variable. When examining data using this method, it is important to identify a reduced set of characters representing the original data in a lower-dimension subspace with minimal information loss [12].

The primary objective of the PCA is to streamline the representation of a set of features that exhibit mutual linear dependence or correlation. This involves breaking down the original data matrix into structural and noise matrices. PCA can be defined as a technique for linearly transforming the initial features (variables) into fresh, uncorrelated variables known as principal components. Each principal component signifies a linear amalgamation of the original characteristics. The fundamental trait of each principal component lies in its variability or dispersion level [11].

The K-means method, introduced by MacQueen in 1967, stands as one of the earliest clustering algorithms and continues to be widely adopted, primarily owing to its straightforward nature. Unlike hierarchical clustering approaches, this method requires the pre-specification of the number of clusters to be identified. It operates iteratively and may only sometimes converge to a singular solution. Following each iteration, every object is allocated to the cluster whose centre is nearest, and new cluster centres are computed based on the following relationship:

$$C_i = \frac{1}{|S_i|} \sum_{x_k \epsilon S_i} x_k \tag{1}$$

where S_i denotes the i-th cluster, C_i is the centre of the i-th cluster, and x_k are all objects that belong to the S_i cluster. The symbol $|S_i|$ indicates the number of cluster elements S_i. The procedure is repeated as long as the division of objects into clusters is still changing [13].

When deciding the number of clusters in the analyses, the following statistical methods were used, e.g., the Average Silhouette Width and the Total within the Sum of Squares. When evaluating these results, using the so-called "elbow" method, we set a suitable number of clusters to 5. The elbow method is where the most significant disruption in the metric is sought; see Fig. 2.

Based on the PCA method (Cumulative Proportion), the main components which performed as input to the K-means clustering analysis were selected. The standard use of the PCA is to reduce the dimension of the task, that is, to reduce the number of characters without much loss of information by using only the first few main components. In this study, the Cumulative Proportion was determined to be 0.98 (or 98%), and seven main components based on it were used (Tab. 1). By extracting the principal components, the correlations and multicollinearity from the data were removed.

Tab. 1 Importance of components in the PCA method to preserve information during an analysis.

Importance of component												
	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**	**PC11**	**PC12**
Standard deviation	2.70	1.40	1.08	0.83	0.58	0.43	0.41	0.32	0.22	0.17	0.10	0.00
Proportion of Variance	0.61	0.16	0.10	0.06	0.03	0.02	0.01	0.01	0.00	0.00	0.00	0.00
Cumulative Proportion	0.61	0.77	0.87	0.93	0.95	0.97	0.98	0.99	1.00	1.00	1.00	1.00

Finally, K-means clustering was used to create homogeneous pooling groups. The basis of K-means clustering is the division of the data set on the intra-annual runoff distribution into clusters of the runoff regime. This division aims to minimize the total sum of the square of the difference between the runoff distribution in an individual basin year and the average runoff distribution within the given cluster. The main condition for the application of the mentioned method is the pre-definition of the resulting number of clusters [14].

In order to identify the border separating individual clusters, the Support Vector Machine (SVM) model with Radial Basis Function (RBF) kernel was used. The RBF kernel was selected to account for spherical nature of clusters created by K-means clustering algorithm. For every cluster (5 clusters in total), one SVM model was trained separating observations in this cluster from all other observations (one-vs-rest setup). Normalized data of the normalized data of the long-term average monthly discharges was used for training, and trained model was used to calculate the importance of the characteristic features of individual clusters. The statistical method, the permutation feature importance, was used for this step and two most important features were selected as cluster descriptors. The described methodology was implemented in R language using R Studio environment.

DOI 10.13164/juniorstav.2024.24100

3 RESULTS

The results of the analysis are processed in tabular, map, and graphic formats. Based on the statistical methods used, the analysis was based on the division into 5 clusters (Fig. 2). The number of clusters was chosen based on the Average Silhouette Width and the Total Within Sum of Square tests.

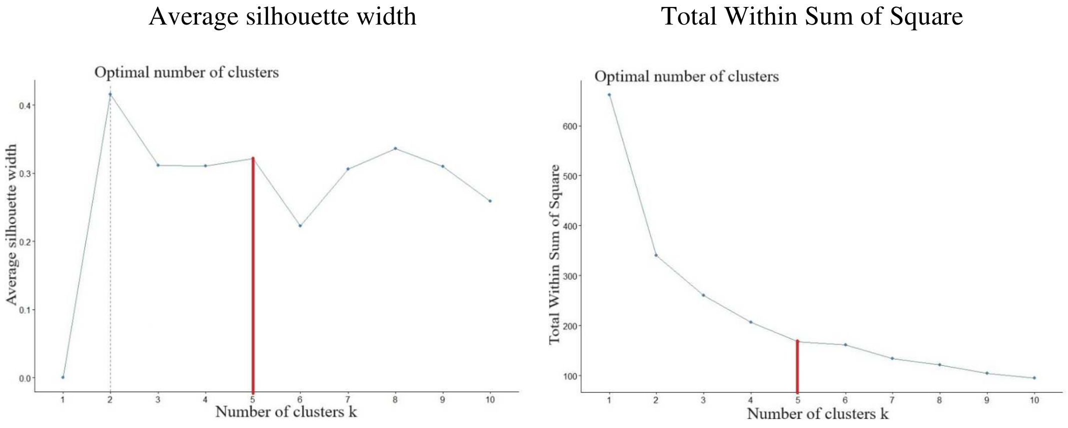

Fig. 2 Results of the statistical methods when calculating the number of clusters for PCA method and K-means clustering.

The results of the analysis with the boundaries of the clusters according to the most significant months are shown in Tab. 2, together with the indicated limit values of the normalised data of the long-term average monthly discharges.

Tab. 2 Division of gauging stations into clusters according to the analysis and their threshold values of the normalized data of the long-term average monthly discharges.

Cluster	ID		Month	Min.	Max.
1	*5740*	*7015*	*January*		
	5790	*7045*	*February*		
	5800	*7160*	*March*		
	5810	*7730*	*April*	*1.75*	*2.51*
	5820	*7820*	*September*		
	6130	*7860*	*October*	*-0.7*	*-0.84*
	6150		*November*		
2	*5100*	*9650*	*January*		
	5880		*February*	*-0.66*	*0.71*
	6480		*March*	*1.01*	*2.25*
	8870		*April*		
	8930		*September*		
	9290		*October*		
	9500		*November*		
3	*5330*	*8320*	*January*		
	5400	*8690*	*February*		
	5550		*March*		
	5730		*April*	*0.24*	*1.8*
	5780		*September*		
	5840		*October*		
	7930		*November*	*-0.83*	*-0.19*
4	*5310*		*January*		
	6950		*February*		

DOI 10.13164/juniorstav.2024.24100

Cluster	ID			Month	Min.	Max.
4	7060			March		
	7065			April		
	7070			September		
	7660			October	-0.62	-0.31
	8530			November	-0.06	0.09
5	5130	6400	7440	January	-0.24	0.45
	5250	6450	7480	February		
	5260	6470	7600	March		
	6180	6540	9410	April	0.62	1.61
	6200	6620	9620	September		
	6360	6640	9670	October		
	6390	6730		November		

To better evaluate the results of the analysis, a spatial visualisation of the basins was also created, see Fig. 3. The final runoff regime typical of each cluster is visualised in Fig. 4a and Fig. 4b with a graph of the dependence between the two most important months that were decided when splitting the clusters.

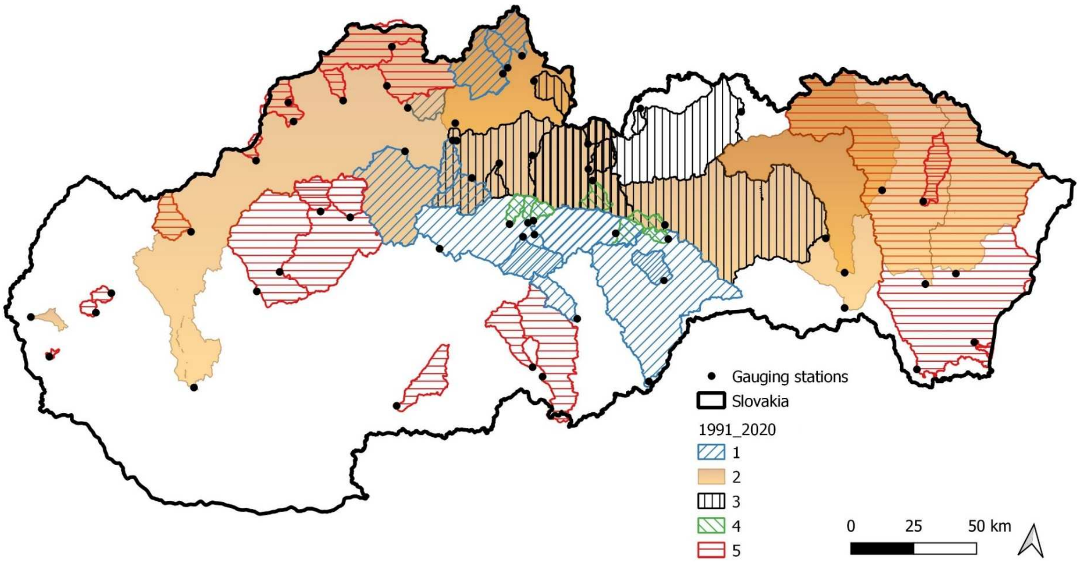

Fig. 3 Spatial representation of the results of the analysis for Slovakia's basins in 1991–2020.

The first group of gauging stations (Cluster 1) are stations characterised by the maximum values of the normalised data of the long-term average monthly discharges in March and April. The low values of the normalised data of the long-term average monthly discharges occur in September. Cluster 1 includes basins located in the south of Slovakia and in central Slovakia. The most significant months for Cluster 1 are April and November.

The second group of gauging stations (Cluster 2) is represented by the normalised data values of the long-term average monthly discharges, which reach a maximum in March and April and a minimum in the autumn. The boundary of this cluster was determined by February and March. Basins in the northwest and northeast of Slovakia represent Cluster 2.

The next group of gauging stations is Cluster 3, which is mainly located in the northern part of central Slovakia. The maximum values of the normalised data of the long-term average monthly discharges are reached in May, and the lowest values in December, January and February. April and November are the most important months for Cluster 3. These basins represent part of the High Tatras, and snowmelt affects the highest flow rates.

Gauging stations located in central Slovakia are located in Cluster 4. They reach the highest normalised data values of the long-term average monthly discharges in April and May. They show the lowest long-term average monthly discharges in January and February. The decisive months for this group of gauging stations are October and November.

The last group created is Cluster 5, in which the highest values of the normalised data of the long-term average monthly discharges are reached in March, and the lowest values of the long-term average monthly discharges are in August and September. The important months, in this case, are January and April. Cluster 5 has the largest group of gauging stations.

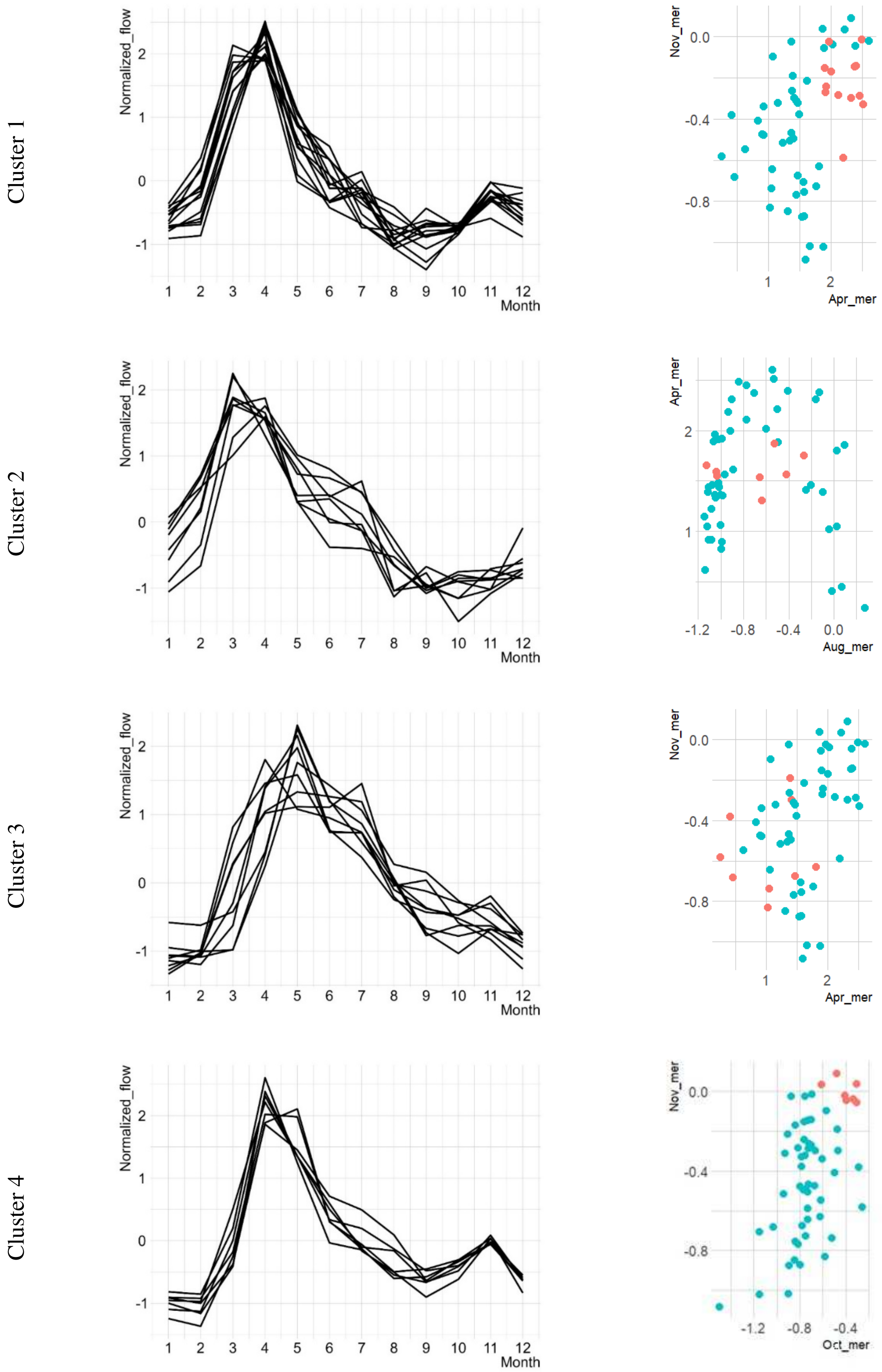

Fig. 4a Visualization of the runoff regime in each cluster and the dependence graph of the most important months for the clusters in 1991–2020.

DOI 10.13164/juniorstav.2024.24100

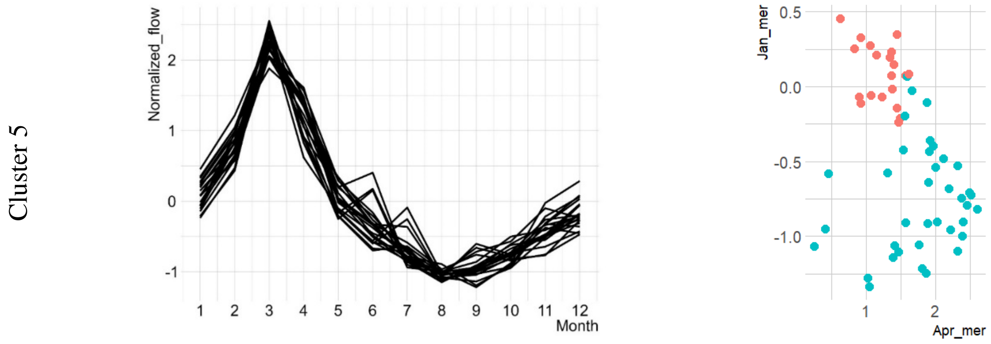

Fig. 4b Visualization of the runoff regime in each cluster and the dependence graph of the most important months for the clusters in 1991–2020.

4 DISCUSSION

From the runoff regime typical of each cluster, we can see that the most significant changes in the flow regime occur where snowfall becomes lighter due to higher temperatures. Thus, winter runoff increases, and spring discharges decrease. These changes are noticeable in many parts of Eastern Europe. In western maritime Europe, low flows will decrease, but further easterly, minimum flows will increase as discharges rise during the current low flow season, i.e., winter [15]. For the territory of Norway, the temperature is the dominant source of variability in the colder months, as it directly affects precipitation and snowmelt. In May and June, the temperature and precipitation contribute to the high variability of the runoff regime approximately equally. Summer and autumn are affected by the overland runoff due to the precipitation [5]. The predictions for the future of the territory of Slovakia indicate that the most significant changes will occur in northcentral Slovakia, where the high mountainous character of the country is concerned. As the air temperature increases and precipitation changes, the highest average monthly discharges will shift from May to April [16].

The study aims to create new pooling groups of the long-term average monthly discharges in the territory of Slovakia in the new reference period 1991-2020. The analytical method created can be used in further studies, in which climate scenario data can also be used to predict future flow regime changes. The continuation of the study for clustering by the PCA method and K-means clustering will be extended to the basins of Austria in the future.

5 CONCLUSION

The presented study aims to create new pooling groups of the long-term average monthly discharge regime in Slovakia. The analyses used data from 57 selected gauging stations from the period 1991-2020. Based on the PCA method, the main principal components, which performed as an input to the K-means clustering analysis, were selected. The data, which was modified by standard normalisation, were divided into 5 clusters. Using the R Studio program, the most important characteristic features of the individual clusters of the gauging stations created were also analysed, which could help to classify other gauging stations into derived pooling groups.

The results of the work consist of the division of the basins selected into five groups, each having a typical runoff regime in the territory of Slovakia. Cluster 1 is characteristic for the south of central Slovakia and central Slovakia. The lowest long-term normalised average monthly discharges are in September and the highest in April. Cluster 2 is for the northwest and northeast Slovakia. The long-term normalised average monthly discharges reach their maximum in March and their minimum in autumn. Cluster 3 is for northcentral Slovakia, and Cluster 4 is for central Slovakia. Both clusters have the highest long-term normalised average monthly discharges in May and the lowest in the winter. Cluster 5 is for the east, south and west of Slovakia. This cluster is characterised by an increase in the long-term normalised average monthly discharges in March and a decrease in the long-term normalised average monthly discharges in August and September.

Acknowledgement

This study was supported by the Slovak Research and Development Agency under Contract No. APVV-20-0374 and VEGA Grant Agency No 1/0782/21. The authors thank the agencies for their research support.

References

[1] BLAHUŠIAKOVÁ, Andrea and Milada, MATOUŠKOVÁ. Rainfall and runoff regime trends in mountain catchments (Case study area: the upper Hron River basin, Slovakia). *Journal of Hydrology and Hydromechanics* [online]. September 2015, 63(3), pp. 186–192. [accessed 25 October 2023]. DOI 10.1515/johh-2015-0030

[2] CRNOGORAC, Cedomir and Vesna, RAJCEVIC. Climate Change and Protection Against Floods. In: FILHO, Walter Leal, Goran, TRBIĆ and Dejan, FILIPOVIC. *Climate Change Adaptation in Eastern Europe. Managing Risk and Building Resilience to Climate Change* [online]. Springer, 2019, pp. 127–136. [accessed 25.10.2023]. ISBN 978-3-030-03383-5. DOI https://link.springer.com/chapter/10.1007/978-3-030-03383-5_9

[3] SLOVAK HYDROMETEOROLOGICAL INSTITUTE. Processing of the hydrological characteristics – average annual discharges, precipitation totals per basin. Final report of a partial research and development task 3030-01. Bratislava. 2005, p. 82

[4] GREŠKOVÁ, Anna. Regionalisation of characteristics of low flow in hydrology and hydro-geography. *Geographical Journal* [online]. 1998, 50(2), pp. 157–174. [accessed 26 October 2023]. Available at: https://www.sav.sk/journals/uploads/05031153GC_1998_2_5_Greskova.pdf

[5] YUAN, Qifen, Thordis L., THORARINSDOTTIR, Stein, BELDRING, Wai Kwok, WONG, and Chong-Yu, XU. Assessing uncertainty in hydrological projections arising from local-scale internal variability of climate. *Journal of Hydrology* [online]. March 2023, 620(129415), p. 13. [accessed 25 October 2023]. DOI 10.1016/j.jhydrol.2023.129415

[6] SANG, Yan-fang, Dong, WANG, Ji-chun, WU, Qing-ping, ZHU and Ling, WANG. Human impacts on runoff regime of middle and lower Yellow River. *Water Science and Engineering* [online]. March 2011, 4(1), pp. 36–45. [accessed 26 October 2023]. DOI 10.3882/j.issn.1674-2370.2011.01.004

[7] DUB, Oto. Hydrológia, hydrografia, hydrometria (Hydrology, hydrography, hydrometry). Slovak Publishing House of Technical Literature, Bratislava: SVTL. 1957, p. 488

[8] ŠIMO, E. and M., ZAŤKO. Types of runoff regime. In: Mazúr (Ed.): *Atlas of the Slovak Republic. Ministry of the Environment*, Bratislava, Environmental Agency, Banská Bystrica. 2002, ISBN 80- 88833-27-2

[9] HANUŠIN, Ján. Typification of runoff regime on example of a set of chosen catchments of Slovakia. *Geographical Journal* [online]. 1999, 51(1), pp. 97–108. [accessed 25 October 2023]. Available at: https://www.sav.sk/journals/uploads/05031221GC_1999_1_6_Hanusin.pdf

[10] SLOVAK HYDROMETEOROLOGICAL INSTITUTE. Determination of regional types of outflow regime for the territory of Slovakia. Final report. Bratislava. 2005, p. 44.

[11] MELOUN, Milan, MILITKÝ, Jiří, and Martin, HILL. Computer analysis of multidimensional data in examples in the fields of natural, technical and social sciences. Academia Praha, 2005, p. 450. ISBN 80-200-1335-0

[12] KHEFIR, Ferath and Adeliya, LATYPOVA. Principal component analysis. Chapter 12. In: MECHELLI, Andrea and Sandra, VIEIRA. *Machine Learning. Methods and Applications to Brain Disorders* [online]. *Academic Press*, 2020, pp. 209–225. [accessed 26.10.2023]. ISBN: 978-0-12-815739-8. DOI 10.1016/B978-0-12-815739-8.00012-2

[13] MACQUEEN, J. B. Some Methods for classification and Analysis of Multivariate Observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* [online]. University of California Press. 1967, pp. 281–297. [accessed 16 January 2024]. Available at: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings- of- the- Fifth- Berkeley- Symposium- on- Mathematical- Statistics- and/chapter/Some-methods-for-classification-and-analysis-of- multivariate-observations/bsmsp/1200512992

[14] HARTIGAN, John and Anthony M., WONG. Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics* [online]. 1979, 28(1), pp. 100–108. [accessed 25 October 2023]. DOI 10.2307/2346830

[15] NIGEL, Arnell. The effect of climate change on hydrological regimes in Europe: a continental perspective. *Global Environmental Change* [online]. 1999, 9, pp. 5–23. [accessed 26 October 2023]. DOI 10.1016/S0959-3780(98)00015-6

[16] SABOVÁ, Zuzana and Silvia, KOHNOVÁ. Seasonal and spatial changes in mean monthly discharges in selected gauging stations of Slovakia. In: KALICZ, Péter, HLAVČOVÁ, Kamila, KOHNOVÁ, Silvia, SZÉLES, Borbála, RATTAYOVÁ Viera, et al. *HydroCarpath 2022. Hydrology of the Carpathian Basin: synthesis of data, driving factors and processes across scales.* Sopron: University of Sopron Press, 2022, p. 104. ISBN 978-963-334-452-1

DOI 10.13164/juniorstav.2024.24100