DOI:10.28974/idojaras.2022.1.2

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 126, No. 1, January – March, 2022, pp. 27–45

Regionalization of low flow for chosen catchments of the upper Vistula river basin using non-hierarchical cluster analysis

Agnieszka Cupak* and **Grzegorz Kaczor**

University of Agriculture in Cracow
Faculty of Environmental Engineering and Land Surveying
Department of Sanitary Engineering and Water Management
30-059 Cracow, Poland

**Corresponding author E-mail: a.cupak@ur.krakow.pl*

(Manuscript received in final form November 26, 2020)

Abstract— The aim of this work was the regionalization of low flow for chosen catchments located in the upper Vistula river basin using non-hierarchical cluster analysis. Next, with such creative clusters, the regional relationships were determined between the specific low flow discharge q_{95} and the meteorological and physiographic parameters of the catchment. The study evaluated regional regression models for low flow (specific q_{95} discharge) in selected, 30 catchments located in the upper Vistula river basin. The data for calculations were a series of observations of daily discharge from the multiannual period of 1963–2016 and were obtained from the Institute of Meteorology and Water Management – National Research Institute in Warsaw. The study showed, that the k-means method can be used for regional regression determination. The parameters which influenced the catchments grouping in clusters were the specific low flow discharge q_{95}, precipitation, median catchment altitude, mean catchment slope, soil, and land use. The study indicated that k-means method may be an effective tool for evaluating low flow in rivers of the southern parts of Poland.

Key-words: low flow, cluster analysis, morphoclimatic parameters, catchments grouping

1. Introduction

The flow in a river is the sum of natural processes that take place within the catchment, such as: supplying, storing, and outflowing of water. Supply of water depends highly on precipitation, landscape and use of the area, and the roughness coefficient. Storage of water and its flow are dependent on complex physiographic elements of the catchment. The natural factors that affect the low flow in the river are: the type and infiltration capacity of the soil, deposition of aquifers, speed and frequency of water supply, evapotranspiration, management and topography of the area, and the climate. In many cases, ground waters provide supply of water for streams during the time of low flows. Low flows do not unbalance the ecology at such times. Therefore, it is important for aquifer layers to have access to sufficient volume of water, the level of ground waters to be sufficiently low to cross the watercourse, and the size and hydraulic conditions of the aquifer to be sufficient to maintain the flow during dry time. Supply of water for low flows may also come from the nearby surface of the valley bottom, where water is stored in the form of saturated soil, alluvial area, and wetlands. These are places saturated with water during or right after precipitation (*Smakhtin,* 2001; *Ziernicka-Wojtaszek* and *Kaczor*, 2013). The geological structure of the catchment also significantly affects the appearance of low flows. *Armbruster* (1976) and *Smith* (1981) confirmed in their study the direct relationships between the geological structure of the catchment and the speed of outflow at low flow time.

Modern society faces the common phenomenon of shortage of water. This problem is enhanced with the fact that water deficits occur in many parts of the world at the same time (*Bates et al.,* 2008). Shortage of water is related to drought that affects resources of surface and underground waters and can lead to reduction of the supply of water, deterioration of its quality, crop failure, and disturbances in habitats (*Mishra* and *Singh,* 2011).

The water balance is the main and commonly used model to determine low flows in controlled catchments. It requires entering some data, which in most cases are easily available, and the method is relatively simple and easy (*Merz* and *Blöschl*, 2004). In general, hydrometric gauging records are not available at the site of interest. Where these records are available, they may be of short length, leading high uncertainties in the selection of the probability distribution and the estimation of the parameters of selected model. When the observed streamflow records are unavailable or inadequate for a proper local frequency analysis, other approaches must be used (*Ouarda et al.,* 2008). In uncontrolled catchments, the relevant parameters are acquired from other sources of information, such as the neighboring catchments or the data concluded on the basis of literature information (*Merz* and *Blöschl*, 2004; *Wałega* and *Młyński*, 2017). In Poland, in case of controlled catchments, statistical methods based on different type of distribution of extreme value are in use, i.e., Gumbel method for low flow calculation (*Byczkowski,* 1972). For uncontrolled catchments, empirical formulas are used. There are few formulas

to low and average flow calculations, but they were worked out based on hydrological data for years 1950–1980 by the 20th century, for example the Punzet formula (*Punzet,* 1981). At present there is a need of verification or updating of these formulas, especially since actual hydrological data sequence are much longer at present (*Wałęga et al.,* 2014).

The regional frequency analysis is the most commonly used tool for the estimation of extreme hydrological events (floods, droughts) at sites, where little or no reliable data are available (*Vogel* and *Kroll*, 1990; *Tucci et al.* 1995; *Durrans* and *Tomic,* 1996; *Hamza et al.,* 2001; *Ouarda et al.,* 2005; *McLean* and *Watt*, 2005; *Laaha* and *Blöschl*, 2006). In general, a regional frequency analysis procedure is composed of two main steps: the identification of groups of hydrologically homogenous catchments (or regions) and the application of a regional estimation method within each delineated region (*Ouarda et al.,* 2008). Regionalization of catchments is founded on the premise, that catchments of the same climate, geology and topography, vegetation and soils have similar values of low flow parameters. It is possible to define the homogenous regions in a variety of manners: as geographically contiguous region, as geographically non-contiguous regions (*Ouarda et al.,* 2008). Regions grouped in this way are not always strictly homogeneous, but this approach may produce sufficient results when availability of data is limited. A homogeneous region can be also perceived as a group of catchments hydrologically similar, but not necessarily geographically neighboring. Multidimensional statistical analyses are often used to group catchments. cluster analysis is the general name of multidimensional statistical techniques that are used to study, interpret, and classify the data with those of a similar group or groups. The data from one cluster should be as close to each other as possible, whereas parameters from different clusters should differ, if possible.

Non-hierarchical methods of grouping require the initial determination of the number of clusters. They may be classified based on the techniques used to initiate the clusters, the criteria for cluster creation, and the types of the data for which they are appropriate (*Rao* and *Srinivas*, 2008). One of the most frequently used and the best known of the non-hierarchical clustering methods, i.e., for catchments grouping for the sake of flooding, is the k-means method (*Lecce*, 2000; *Burn* and *Goel*, 2000). In India, *Ahuja* (2012) used the k-means method for data regionalization of Godaravi catchments. In Poland, *Cupak* (2017) used it for low flow grouping. This method is effective for grouping large sets of data with numerical attributes. However, there are some limitations to this method in the breaking down of the data into categories. The method is also sensitive to presence of errors (*Rao* and *Srinivas,* 2008).

The objective of this work was the regionalization of low flow in chosen catchments located in in the area of the upper Vistula river basin with use of the non-hierarchical cluster analysis – the k-means method. Next, with such creative clusters, the regional relationships were determined between low flow q_{95} and the meteorological and physiographic parameters of the catchment.

2. Material and methods

The analysis was conducted for 30 selected catchments of the upper Vistula river basin (*Fig. 1*). The source material for the analysis were daily flows from the period 1963–2016 (*Table 1*). It was assumed, as a criterion of catchments' selection, that for analysis, only those catchments will be taken, for which daily streamflows are available with a minimum record length of 20 years. 13 physiographic and meteorological characteristics of catchments were also used (*Table 1*) in the analysis. The data related to daily flows, temperature, and precipitation were obtained from the Institute of Meteorology and Water Management, National Research Institute in Warsaw.

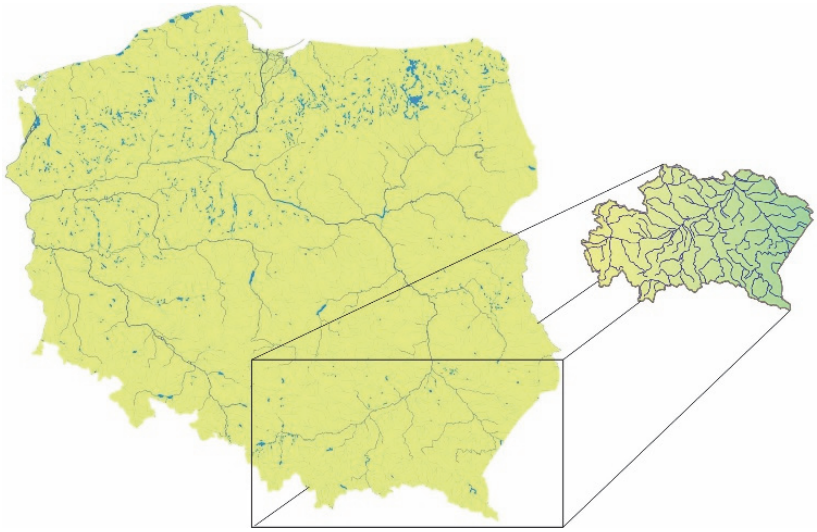

Fig. 1. Location of the upper Vistula river basin (https://pl.wikipedia.org/wiki/Plik:Polska_hydrografia2.jpg)

The first step was to determine $Q_{95\%}$, that is the flow achieved during 95% of days in the studied timeframe. This low flow characteristic is widely used in Europe and was chosen because of its relevance for multiple choices of water management, among other things in case of projection of water supply systems. Then, $Q_{95\%}$ was subsequently standardized by the catchment area resulting in specific low flow discharges q_{95} ($dm^3 \cdot s^{-1} \cdot km^{-2}$). The data were standardised on the basis of Eq.(1) in order to obtain average values expected for the individual variables, which were given in various units.

$$x_{ij} = \frac{w_j}{\sigma_j}\left[f(y_{ij})\right] \qquad for \;\; j = 1, \ldots, n \;\; , \tag{1}$$

where

$f(y_{ij})$ is the function subject to transformation,

y_{ij} is the value of the feature j, in n – dimensional function of the vector y_i,

w_j is the weight assigned to the given feature,

σ_j is standard deviation.

In the k-means method, the first step is to determine the number of clusters. The center is determined for each group, which is defined as the function of the vector between the clusters. After assigning the variables to the clusters, the function of the vector is calculated again to redetermine the location of the center of the cluster. The variables are again assigned to the groups, according to the position of the new cluster (*Dikbas et al.,* 2013). The Euclidean distance was used to calculate the distance of the objects from the centers of the clusters.

The calculation procedure was run four times – for two, three, four, and five clusters. First, two clusters are generated. In the last step, analysis of correlations was conducted and models of correlations were defined. The coefficient of correlation was calculated, which describes the relationship between the unit outflow q_{95} and selected meteorological and physiographic features of the catchment. The coefficient of determination R^2 was also determined for the level of confidence α = 0.05. Regional regression is built as a multiply regression (Eq.(4)), which shows relationships between low flow (as a dependent variable) and morphoclimatic parameters (as independent variables). It is used to identify the parameters that most strongly influence the low flow. To determine the power of regression equation, adjusted coefficient of determination R^2_{adj} for the level of significance 0.05 was calculated. The best results were obtained while using stepwise regression:

$$q_{95} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_{p-1} \cdot x_{p-1}\,, \tag{2}$$

where

x_i are the morphoclimatic parameters of a catchment,

β_i is the regression coefficient.

The statistical calculation were made using STATISTICA 13 software. *Figs. 2* and *4* were made in Inkscape.

2.1. Model performance criteria

The performance measures used in this study were the Nash–Sutcliffe efficiency (*E*), the percent bias (*PBIAS*), and the adjusted coefficient of determination (R^2_{adj}). Additionally to the regression model the goodness of fit was tested in case, when uncontrolled catchment will be included to a region.

The value of percent bias (Eq.(3)) and a root mean sum of squares error (Eq.(4)) were calculated for each clusters obtained with use cluster analysis (*Patel,* 2007).

$$PBIAS = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_{95}^{obs} - q_{95}^{sim}}{q_{95}^{obs}}\right) \cdot 100\% \,, \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(q_{95}^{obs} - q_{95}^{sim}\right)^2} \,, \tag{4}$$

where

q_{95}^{obs} is the observed specific low flow discharge q_{95} for catchment *i,*

$q_{95}^{sim}$is the model prediction.

RMSE and *PBIAS* values of 0 indicate a perfect fit. *PBIAS* measures the average tendency of the simulated data to be larger or smaller than observed ones. The optimal value of PBIAS is 0. Positive values indicate model underestimation, while negative values indicate model overestimation (*Fang et al.,* 2014). For the assessment under *PBIAS* a classification was used suggested by *Van Liew et al.* (2007), described as follows: *PBIAS*<10%: it is a very good model, 10% < *PBIAS* < 15%: the model is good; 15% < *PBIAS* < 25%: the model is satisfactory, and when *PBIAS* ≥ 25%: the model is unsatisfactory model (*Pereira et al.,* 2016).

The *E* value (Eq.(5)) is a normalized statistic, that expresses the relative magnitude of the residual variance compared to the variance of the measured data (*Nash* and *Sutcliffe,* 1970; *Tegegne et al.,* 2017). *E* indicates how well a plot of observed versus simulated data fits a 1:1 line (*Tegegne et al.,* 2017). *E* was recommended for two major reasons: it is recommended for use by *ASCE* (1993) and *Legates* and *McCabe* (1999), and it is very commonly used, which provides extensive information on reported values (*Moriasi et al.,* 2007). It is calculated as:

$$E = 1 - \left(\frac{\sum_{i=1}^{n}(q_{95}^{obs} - q_{95}^{sim})^2}{\sum_{t=1}^{Nn}(q_{95}^{obs} - \overline{q_{95}^{obs}})^2}\right) \,, \tag{5}$$

where

q_{95}^{obs} is the observed specific low flow discharge,

q_{95}^{sim} is the predicted specific low flow discharge,

$\overline{q_{95}^{obs}}$ is the average value.

The range of *E* lies between 1.0 (perfect fit) and −∞. An efficiency of lower than zero indicates that the mean value of the observed time series would have been a better predictor than the model (*Krause et al.,* 2005).

3. The description of the study area

The research included 30 selected catchments located in the upper Vistula river basin (*Fig. 2*). This area is spread within three great Carpathian physiographic units: the Carpathians (40% of the basin area), the Subcarpathian valleys (about 35% of the basin area), and the Małopolska Upland (about 25% of the basin area). The Carpathians and the Upland are the source areas for most of the upper Vistula tributaries, while the Subcarpathian valleys are a transit area for the Vistula and an estuary area for the rivers and streams formed in the Carpathians and Subcarpathian Uplands (*Chełmicki,* 1991).

Fig. 2. Location of analyzed catchments in the upper Vistula river basin, where: 1 – Dłubnia, 2 – Opatówka, 3 – Biała Tarnowska, 4 – Szreniawa, 5 – Wieprzówka, 6 – Łeg, 7 – Tanew, 8 – Biała, 9 – Pszczynka, 10 – Skawa, 11 – Łososinka, 12 – Biała Nida, 13 – Trzebośnica, 14 – Czarna, 15 – Soła, 16 – Wisła, 17 – Dunajec, 18 – Koprzywianka, 19 – Skawica, 20 – Czarna Nida, 21 – Wschodnia, 22 – Ropa, 23 – Jasiołka, 24 – Solinka, 25 – Osława, 26 – Stupnica, 27 – Mleczka, 28 – Łubinka, 29 – Grabinianka, 30 – Wielkopolska

Catchments (*Fig. 2*) chosen for analysis are diverse in respect of analyzed parameters. The average annual precipitation amounted more than 1000 mm for the Carpathian inflows of Vistula river, and in case of other catchments it is about 600 –800 mm (*Table 1*).

Table 1. Statistical summary of catchments' characteristics

Variable	Variable description	Units	Min.	Mean	Max.
A	Catchment area	km^2	66.3	472.8	2034.0
L	Length of the watercourse	km	8.8	33.2	72.0
T	Mean annual air temperature	°C	5.0	7.0	8.0
P	Mean annual precipitation	mm	603.8	796.6	1192.6
I	Mean catchment slope	–	0.002	0.022	0.091
H_{me}	Median catchment altitude	m a.s.l.	202.0	391.5	836.0
LU1	Coniferous forests	%	0.0	16.0	77.6
LU2	Mixed forests	%	0.0	13.0	53.1
LU3	Grassland	%	0.0	8.3	30.0
LU4	Arable land	%	7.4	57.2	87.0
S1	Fluvisols	%	0.0	15.6	34.0
S2	Cambisols	%	0.0	28.0	100.0
S3	Luvisols	%	0.0	17.1	73.4

Catchments with different area were also chosen, from small ones (like Łubinka with an area of 66.3 km^2) to large ones (like Tanew – 2093 km^2 or Biała Tarnowska – 957 km^2). The mean slope is in range from 0.002 for Łęg river to 0.091 for Biała river. In case of some catchments, cambisols and arable land dominates (*Table 1*).

In the analysis, data of precipitation and temperature were also used for the meteorological station located in the area of the upper Vistula river basin (*Table 2*).

Table 2. Meteorological stations used in this study

Station	Altitude	Latitude	Longitude
Annopol	165	50 53	21 50
Frampol	245	50 40	22 40
Gorlice	439	49 40	21 10
Jarocin	187	50 34	22 18
Jasło	228	49 44	21 29
Jawiszowice	265	49 58	19 08
Kalwaria Zebrzydowska	339	49 52	19 42
Kamesznica	821	49 36	19 04
Kańczuga	237	49 59	22 24
Klimontów	258	50 40	21 27
Konieczno	256	50 48	20 03
Kowaniec	672	49 30	20 02
Maków Podhalański	578	49 44	19 41
Osielec	525	49 41	19 45
Pilzno	193	49 59	21 18
Radomyśl Wielki	189	50 12	21 18
Radziemice	249	50 15	20 15
Raków	265	50 41	21 03
Rozdziele	375	49 48	20 27
Rudzica	272	49 51	18 53
Rybotycze	301	49 39	22 39
Sandomierz	217	50 41	21 42
Skoczów	302	49 47	18 47
Szaflary	686	49 25	20 02
Szczawne	463	49 24	22 09
Terka	668	49 18	22 26
Tuchów	223	49 54	21 03
Wadowice	257	49 52	19 30
Zawichost	139	50 48	21 52
Żabnica	824	49 34	19 11

4. Results

To facilitate the identification of clusters, the following symbols were applied for the individual groups:

- for two clusters: 2a, 2b,
- for three clusters: 3a, 3b, 3c,
- for four clusters: 4a, 4b, 4c, 4d,
- for five clusters: 5a, 5b, 5c, 5d, 5e.

As a result of classification, clusters were obtained with the catchments of similar values of the analyzed parameters (*Fig. 3*).

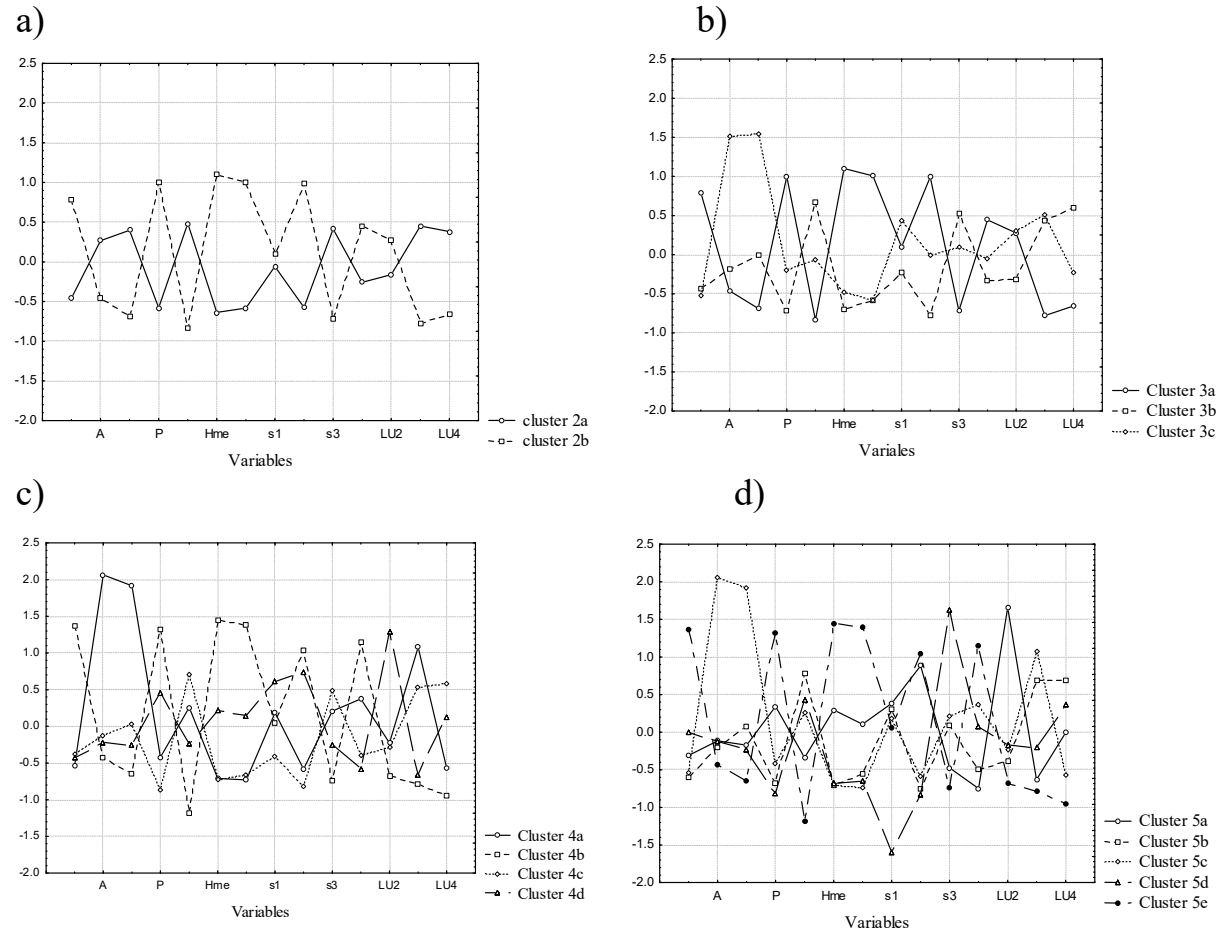

Fig. 3. Chart of the centers of the individual parameters for: a – two clusters, b – three clusters, c – four clusters, d – five clusters.

Fig. 4 presents the distribution of the analyzed catchments assigned to the individual clusters. cluster 2a includes 19 catchments located in the northern and central parts of the upper Vistula river basin. These are catchments varied in terms of area: from middle-size catchments of approximately 160 km^2 to large ones, whose area exceeds 800 km^2. In terms of the length of the streams, this cluster (*Fig. 4a*) includes varied catchments: from 17 km in case of Czarna river to 72 km in case of

Tanew river. In terms of the specific low flow discharge q_{95}, the cathments are in the range from 1 to 4 $dm^3 \cdot s^{-1} \cdot km^{-2}$. The parameters according to which catchments were classified in this cluster are: low median altitude of the catchment, which was up to 380 m a.s.l., the slope of the catchment (<0.03), with luvisols soils dominant in the catchments, as well as the catchment use, where arable lands are dominant, which constitute 64% of the catchment area on average. The parameters, which had the greatest influence on the shaping of the specific outflow in the cluster 2a, were the mean catchment slope and the median catchment altitude, for which the partial correlation coefficient resulted in 73%. A similar partial correlation value of 70% for this cluster was obtained for the mean annual air temperature and luvisols. In the cluster 2b, 11 catchments were included (*Fig. 4a*), similar in terms of median catchment altitude (over 390 m a.s.l.), large catchment slope (over 0.02), as well as soil type – cambisols prevail in the area of the catchments in the cluster 2b. However, similarly to the cluster 2a, catchments varied in terms of area (from 66 to 681 km^2) as well as specific low flow discharge q_{95} (in the range from 1.78 to 7.98 $dm^3 \cdot s^{-1} km^{-2}$). The parameters, which in the cluster 2b and 3b had the greatest influence on the shaping of the q_{95} outflow, had a mean catchment slope for which the partial correlation coefficient was 53%.

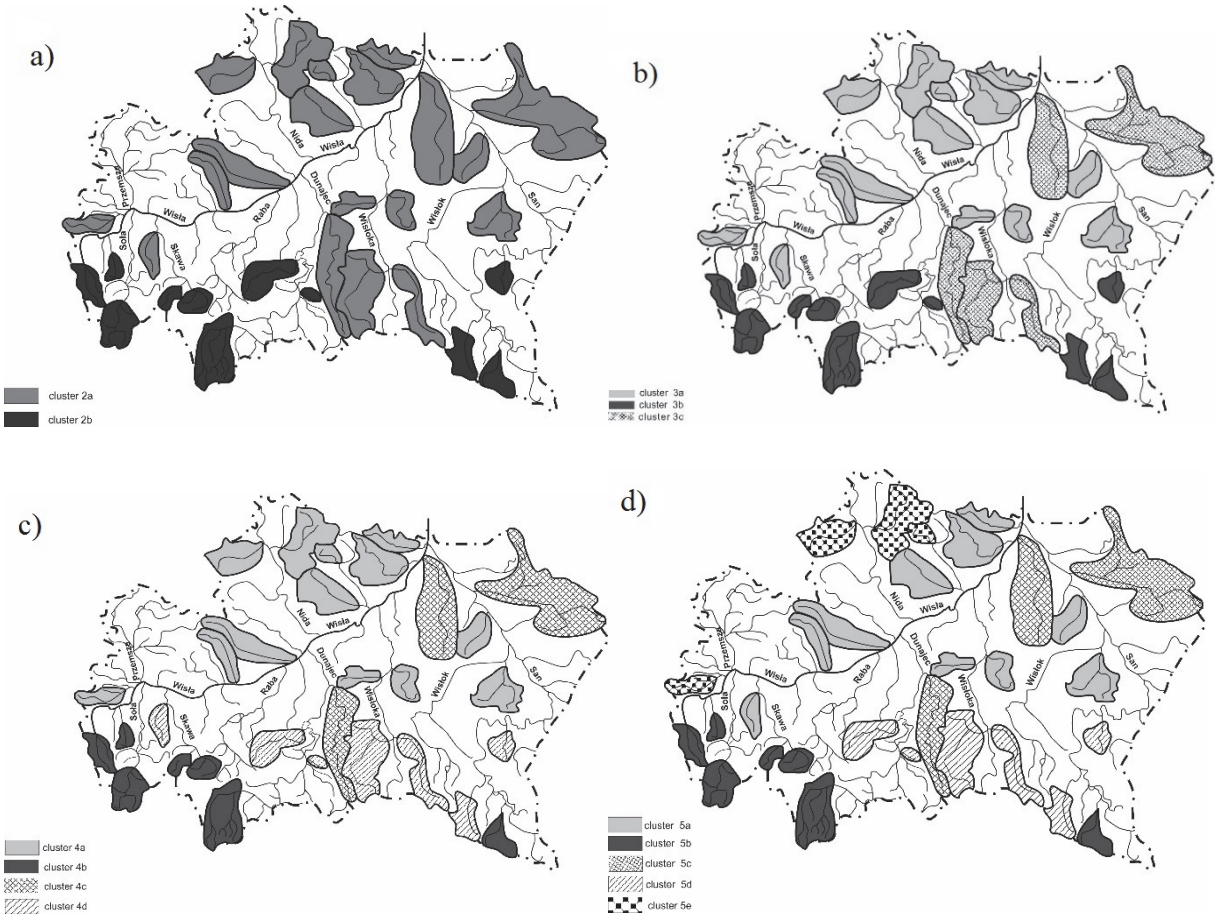

Fig. 4. Localization of the investigated catchments forming the clusters identified with the k-means method for: a – two clusters, b – three clusters, c – four clusters, d – five clusters

In the next step, three clusters were composed. cluster 3b (*Fig. 4b*) included 11 catchments, the same that created the cluster 2b. In the cluster 3a, 14 catchments were included, similar in terms of precipitation (averaged at about 700 mm), median catchment altitude (200 m a.s.l. > H_{me} > 340 m a.s.l.), and with the domination of luvisols and arable land. These catchments were similar also in terms of specific low flow discharge q_{95} (averaged at 2 $dm^3 \cdot s^{-1} km^{-2}$), and catchment area – compared to clusters 3b and 3c, the catchments in this cluster feature average areas from 154 to 755 km^2. cluster 3b including the catchments with the smallest area, and the cluster 3c – the largest ones. Also, the streams were the shortest ones in this cluster (the longest ones were in case of the catchments in the cluster 3c). The parameter, which in the cluster 3a had the greatest influence on the shaping of the specific outflow, was the mean annual air temperature, for which the partial correlation coefficient was 74%.
cluster 3c is composed of 5 catchments similar in terms of specific low flow discharge q_{95} (with the average of 2 $dm^3 \cdot s^{-1} km^{-2}$), precipitation (750 mm on the average), and the largest catchment areas (> 600 km^2), and stream lengths (> 40 km). Further delineation of clusters (four – *Fig. 4c* and five – *Fig. 4d*) was not successful due to the creation of clusters with a small number of catchments. Additionally, the catchments included in the new clusters, whose analyzed parameters, on the basis of which the given catchment was assigning into the group, did not significantly differ from other clusters, e.g., the catchments in the clusters 4b, 4c, and 4d, as well as for the clusters 5a, 5c, and 5d featured very similar values of specific low flow discharge (in both cases in the range from 1 to 4 $dm^3 \cdot s^{-1} km^{-2}$), average annual precipitation (700 mm in case of the clusters 4a, 4c, 5a, and 5c), median catchment altitude (about 260 m a.s.l. in case of clusters 4a, 4c, 5a, 5c, and 5d) (*Fig. 3c* and *3d*). The parameters, which in the cluster 4a had the greatest influence on the shaping of the specific outflow q_{95}, were the coniferous forests (r = 97%). An equally high partial correlation coefficient of 95% was obtained for the length of the watercourse and 93% for luvisols. In turn, for the cluster 5a, for the mean annual air temperature the particle correlation was 76%.

Then regression dependences were determined for selected clusters, for which more than 10 catchments were included (*Table 3*), between the specific discharge q_{95} and the individual parameters.

Table 3. Components of the regional regression model based on the k-means method

cluster	R_2 (%)	R^2_{adj} (%)	Model
2a	87	75	$q_{95} = 6.906^* - 0.0272 \cdot S1^* - 0.003 \cdot P - 75.916 \cdot I^* + 0.018 \cdot S3^* + 0{,}017 \cdot H^*_{me} + 1.245 \cdot T^* - 0.029 \cdot LU4^* - 0.005 \cdot LU2 + 0.026 \cdot S2$
2b, 3b	85	75	$q_{95} = 27.9383 - 50.581 \cdot I - 0.1579 \cdot LU4 + 0.1681 \cdot S1 - 0.0807 \cdot LU2 - 0.013 \cdot P - 0.003 \cdot A - 0.0418 \cdot S2 + 0.0199 \cdot LU1 - 0.0221 \cdot L + 0.0197 \cdot T$
3a	69	59	$q_{95} = -11.1668^* - 0.0385 \cdot S1^* + 1.7797 \cdot T^* + 0.0147 \cdot S3^*$
4a	98	95	$q_{95} = -12.223^* + 0.0022 \cdot P + 0.0912 \cdot LU1^* + 0.0841 \cdot L^* - 0.0016 \cdot A^* + 0.0181 \cdot S3^* + 0.834 \cdot T + 0.0441 \cdot LU3^* + 0.0106 \cdot H^*_{me}$
5a	57	52	$q_{95} = -15.3857^* + 2.2687 \cdot T^*$

R^2_{adj} denotes the goodness of fit coefficient of determination
*Parameter statistically significant at level α=0.05

The best regression model resulted in the case of the cluster 4a (R^2_{adj} = 95%) – *Table 3,* for the level of confidence 0.05. Also, a strong adjusted coefficient of determination, with the value of 75%, was obtained for clusters 2a, 2b, and 3b.

The scatter plots allow a detailed examination of the performance of individual catchments including the existence of outliers and a potential heteroscedasticity of the observations and the predictions (*Laaha* and *Blöschl,* 2006). Overall, the relative scatter of the method (*Fig. 5*) corresponds well with the coefficient of determination in *Table 3*. The model fit was the best for the catchments closest to the diagonal line.

In case of other clusters (with less than 10 catchments each), models of correlation dependence were determined between specific low flow discharge q_{95} and individual independent variables. *Table 4* summarizes only those correlation dependences whose coefficient of correlation (*r*) exceeded the average value (> 0.5). A pretty strong dependence (0.7>*r*>0.9) was obtained in case of the cluster 3c, 5b, and 5d (for such variables as average precipitation, median catchment altitude, mean catchment slope, coniferous forests, and grassland). On the other hand, almost full correlation (for *r*>0.9) was obtained in the case of the clusters 4b and 5c (the variables: catchment area, length of the watercourse, and fluvisols) and the cluster 5d in case of mean annual air temperature.

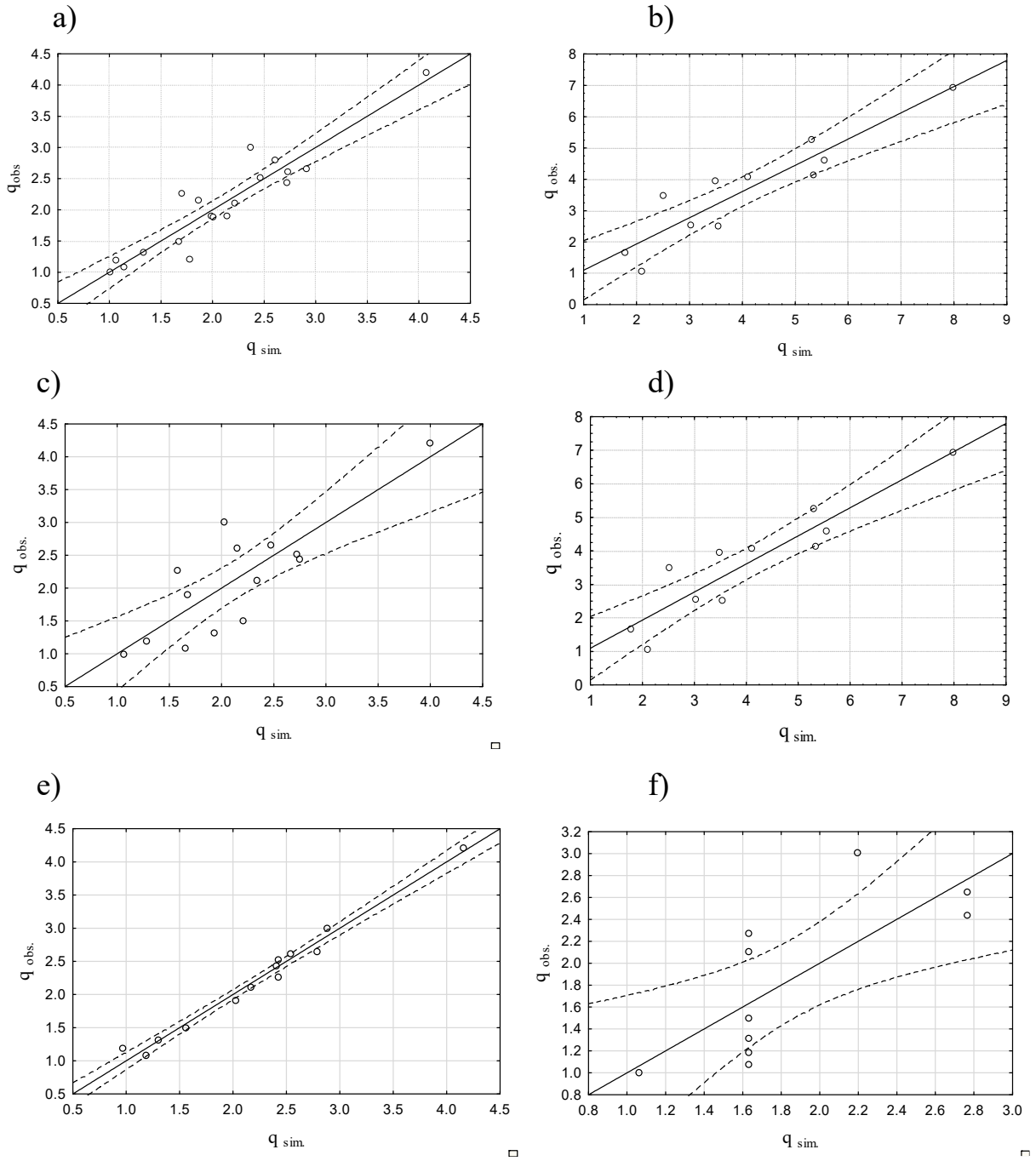

Fig. 5. Differences between observed and calculated q_{95} ($dm^3 \cdot s^{-1} \cdot km^{-2}$) for clusters: a – 2a, b – 2b, c – 3a, d – 3b, e – 4a, and f – 5a.

Table 4. Correlation models between q_{95} and each parameters

cluster	Correlation model	Correlation coefficient	R^2
3c	$q_{95} = 3.31 - 0.067 \cdot S1$	-0.83	0.69
	$q_{95} = 1.174 - 0.0078 \cdot A$	0.78	0.61
4b	$q_{95} = 0.54 + 0.0011 \cdot A$	0.94	0.88
	$q_{95} = -2.57 + 0.73 \cdot L$	0.97	0.94
	$q_{95} = 3.59 - 0.094 \cdot S1$	-0.98	0.96
	$q_{95} = 1.25 + 0.031 \cdot LU1$	0.63	0.40
	$q_{95} = 2.69 - 0.0038 \cdot LU3$	-0.50	0.25
4c	$q_{95} = 2.72 + 42.47 \cdot I$	0.58	0.34
	$q_{95} = 6.88 - 0.048 \cdot LU4$	-0.56	0.31
4d	$q_{95} = 8.01 - 0.007 \cdot P$	-0.58	0.34
	$q_{95} = 1.29 + 0.016 \cdot S2$	0.51	0.26
	$q_{95} = 2,.41 - 0.024 \cdot S3$	-0.55	0.30
	$q_{95} = 2.40 - 0.057 \cdot LU1$	-0.64	0.41
	$q_{95} = 1.19 + 0.029 \cdot LU2$	0.69	0.48
5b	$q_{95} = -5.37 + 1.16 \cdot T$	0.79	0.62
5c	$q_{95} = 0.54 + 0.0011 \cdot A$	0.94	0.88
	$q_{95} = -2.57 + 0.73 \cdot L$	0.97	0.94
	$q_{95} = 3.59 - 0.094 \cdot S1$	-0.98	0.96
	$q_{95} = 1.25 + 0.031 \cdot LU1$	0.63	0.40
	$q_{95} = 2.69 - 0.0038 \cdot LU3$	-0.50	0.25
5d	$q_{95} = 3.90 - 0.0026 \cdot A$	-0.62	0.38
	$q_{95} = -3.88 + 0.01 \cdot P$	0.88	0.77
	$q_{95} = -11.60 + 1.97 \cdot T$	0.95	0.90
	$q_{95} = 17.45 - 0.055 \cdot H_{me}$	-0.79	0.62
	$q_{95} = 3.69 - 113.0 \cdot I$	-0.71	0.50
	$q_{95} = 3.11 - 0.46 \cdot S1$	-0.61	0.37
	$q_{95} = 0.73 + 0.04 \cdot S3$	0.60	0.36
	$q_{95} = 6.13 - 0.19 \cdot LU1$	-0.81	0.66
	$q_{95} = 2.01 + 0.13 \cdot LU3$	0.78	0.61
5e	$q_{95} = 2.72 + 42.47 \cdot I$	0.58	0.34
	$q_{95} = 6.88 - 0.048 \cdot LU4$	-0.56	0.31

A summary of the performance indicator statistics such as percentage bias (*PBIAS*) and the root mean sum of squares error (*RMSE*) for the models presented in this study is given in *Table 5*.

Table 5. Values of *BIAS*, *RMSE* and *E* for clusters obtained with the use of the k-means method

cluster	PBIAS [%]	RMSE [$dm^3 \cdot s^{-1} \cdot km^{-2}$]	E
2a	-22.2	0.68	0.81
2b, 3b	9.9	0.79	0.79
3a	-6.10	0.48	0.69
4a	-1.50	0.11	0.98
5a	-6.62	0.45	0.57

The highest value of *PBIAS* was obtained in the case of cluster 2a (*Table 4*). For clusters 2a, 3a, 4a, and 5a, the predicted values of the specific low flow discharge q_{95} were overestimated, only in case of clusters 2b and 3b, the estimated values were underestimated. According to *Liew et al.* (2007) classification, the models for cluster 2b, 3b, 4a, and 5a are very good, and for cluster 2a they are satisfactory. According to the criterion in (*Pereira et al.,* 2016), the results obtained with the use of cluster analysis for almost every cluster, for which regional regression was made, except cluster 2a, are equal to a very good model. The values of the *E* coefficient were similar to the adjusted determination coefficient R^2. For cluster 4a we got the highest value of *E*, what corresponds with the value of R^2_{adj} (95%) and *PBIAS* (the lowest value equal -1.50), which means that in case of this cluster, we got the best fit of the regression model.

5. Discussion and conclusion

This paper discusses the regionalization of the specific low flow discharges q_{95} in chosen 30 catchments located in the in the area of upper Vistula river basin with use the non-hierarchical cluster analysis - the k-means method. Our research confirms that this method can be used for grouping of watersheds, according to hydrological characteristics. It is also important that this method can be a useful and interesting tool for low flow estimation in uncontrolled catchments. The positive aspect of this method is that we can determine the number of groups. However, when the number of clusters is too large, there is probably no training data in the cluster. Another disadvantage of the method is the lack of an unambiguous criterion on the basis of which the number of clusters can be determined (*Lin* and *Chen*, 2006). In the analysis, 13 physiographic and meteorological characteristics of catchments were also used. We started with two clusters and finished with five. The assumption of two clusters is too small number, because the given group includes catchments varied in terms of some parameters, for example for cluster 2a, in which the catchment areas ranged from about 160 km^2 to more than 800 km^2, while the strean length ranged from 17 km

to 72 km. In turn, defining four and five clusters resulted in the situation when they included catchments, whose analyzed parameters, on the basis of which the given catchment was included in the group, did not significantly differ from the other clusters. Additionally, these clusters featured a low number of catchments, e.g., clusters 4c and 5c had only 3 catchments.

In our research, we got similar relationship to those got by *Cupak* (2017). In case of grassland, coniferous forests, median catchment altitude, mean annual air temperature, and eutric cambisols, the relationship had a positive character, that is the greater the value of each analyzed parameters, the greater the value of low flow is. For parameters, like mean catchment slope and fluvisols, we also got similar – negative relationship, which means, that the value of low flow increases as the value of these parameters decreases.

We also got similar values of R^2, in case of regression models got from cluster analyses, which are in the literature, but for hierarchical cluster analyses. For example in Austria, *Laaha* and *Blöschl* (2006) obtained R^2 varying from 32% to 75% for clusters got with the use of Ward's method between specific low flow discharge q_{95} and catchment characteristics. However, in their study, 325 catchments were taken into research.

In summary, the parameters which influenced the catchments grouping in clusters were the specific low flow discharge q_{95}, precipitation, median catchment altitude, mean catchment slope, soil, and land use.

The best fitting of the model was obtained in the case of cluster 4a, for which the adjusted coefficient of determination and the coefficient *E* rated high, at 95% and 0.98, respectively. The parameters, which had the greatest influence on the shaping of the specific outflow q_{95} in the cluster 4a were the coniferous forests (r = 97%). An equally high partial correlation coefficient of 95% was obtained for the length of the watercourse and 93% for luvisols. However, despite the high value of the obtained coefficients, optimum results (in our opinion: assigning individual catchments to clusters so that the given cluster includes only catchments most similar to each other in terms of hydrological, meteorological, and physiographic parameters, and definitely different from those included in the other catchments), was obtained in case of generation of three clusters, despite lower values of R^2_{adj} (75% and 59%) and the coefficient *E* (0.69 and 0.79). The parameter, which had the greatest influence on the shaping of the specific outflow in the cluster 3a had the mean annual air temperature, for which the partial correlation coefficient was 74%, while in the cluster 3b, the parameter with the greatest influence was the mean catchment slope for which the partial correlation coefficient was 53%.

References

Ahuja, S., 2012: Regionalization of River Basins Using cluster Ensemble. *J. Water Res. Protect. 4,* 560–566.

Armbruster, J.T., 1976: An infiltration index useful in estimating low-flow characteristics of drainage basins. *J. Res., USGS 4,* 533–538.

ASCE, 1993: Criteria for evaluation of watershed models. *J. Irrigation Drainage Eng. 119,* 429–442.

Bates, B.C., Kundzewicz, Z.W., Wu, S., and *Palutikof,, S.,* 2008: Climate change and water. Technical Paper of the Intergovernmental Panel on Climate Change. IPCC Secretariat, Geneva.

Burn, D.H. and *Goel, N.K.,* 2000: The formation of groups for regional flood frequency analysis. *Hydrol. Sci. J. 45,* 97–112.

Byczkowski, A., 1972: Hydrologiczne podstawy projektowania budowli wodno-melioracyjnych. PWRiL, Warszawa (in Polish).

Chełmicki, W., 1991: Położenie, podział I cechy dorzecza. In (eds.: *Dynowska I.* and *Maciejewski M.*) Dorzecze górnej Wisły. PWN Kraków (in Polish).

Cupak, A., 2017: Initial results of nonhierarchical cluster methods use for low flow grouping. *J. Ecol. Engin. 18,* 44–50.

Dikbas, F., Firat, M., Koc, C.A., and *Gungor, M.,* 2013: Defining homogenous regions for streamflow processes In Turkey Rusing k – means clustering method. *Civil Engin. 38,* 1313–1319.

Durrans, S.R. and *Tomic, S.,* 1996: Regionalisation of Low-Flow Frequency Estimates: An Alabama Case Study. *Water Resour. Bull. 32,* 23–37.

Fang, G.H., Yang, J., Chen Y.N., and *Zammit, C.,* 2014: Comparing bias correction methods in downscaling meteorological variables for hydrologic impact study in an arid area in China. *Hydrol. Earth Syst. Sci. Discuss. 11,* 12659–12696.

Hamza, A., Ouarda, T.B.M.J., Durrans, R.S., and *Bobée, B.,* 2001: Développement de modèles de queues et d'invariance d'échelle pour l'estimation régionale des débits d'étiage. *Rev. Canadienne de Génie Civil 28,* 291–304. (In France)

Krause, P., Boyle, D.P., and *Bäse, F.,* 2005: Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci. Europ. Geosci. Union 5*, 89–97.

Laaha, G. and *Blöschl, G.,* 2006: A Comparison of Low Flow Regionalisation Methods—Catchment Grouping. *J. Hydrol. 323,* 193–214.

Lecce, S.A., 2000: Spatial variations in the timing of annual floods in the southeastern United States. *J. Hydrol. 235,* 151–169.

Legates, D.R. and *McCabe, G.J.,* 1999: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resour. Res., 35,* 233–241.

Lin, G.F., Chen, and *L.H.,* 2006: Identification of homogenous regions for regional frequency analysis using the self-organizing map. *J. Hydrol. 324,* 1–9.

McLean, R.K. and *Watt, W.E.,* 2005: Regional Low Flow Frequency Relations for Central Ontario. *Canad. Water Resour. J. 30,* 179–196.

Merz, R. and *Blöschl, G.,* 2004: Regionalisation of catchment model parameters, *J. Hydrol. 298,* 95–123.

Mishra, A.K. and *Singh, V.P.,* 2011: Drought modeling – A review, *J. Hydrol. 403,* 157–175.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., and *Veith, T.L.,* 2007: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE. *Amer. Soc. Agricult. Biol. Engin. 50*, 885–900.

Nash, J.E. and *Sutcliffe, J.,* 1970: River flow forecasting through conceptual models part I – A discussion of principles. *J. Hydrol., 10,* 282–290.

Ouarda, T.B.M.J., Charron, Ch., and St-Hilaire, A., 2008: Statistical models and the estimation of low flows. Canad. Water Resour. J. 33, 195–206.

Ouarda, T.B.M.J., V. Jourdain, N. Gignac, H. Gingras, H. Herrera, and Bobée, B. 2005:

Développement d'un modèle hydrologique visant l'estimation des débits d'étiage pour le Québec habité. INRS-ETE, Rapport de recherche No R-684-f1 (in French).

Patel, J.A., 2007: Evaluation of low flow estimation techniques for ungauged catchments. *Water Environ. J.* 21, 41–46.

Pereira, D. dos R., Martinez, M. A., da Silva, D.D., and *Pruski, F.F.,* 2016: Hydrological simulation in a basin of typical tropical climate and soil using the SWAT Model Part II: Simulation of hydrological variables and soil use scenarios. *J. Hydrol.: Regional Studies 5,*149–163.

Punzet, J., 1981: Empiryczny system ocen charakterystycznych przepływów rzek i potoków w karpackiej części Dorzecza Wisły. *Wiadomości IMGW, zeszyt 1 – 2*, 31 – 39 (in Polish).

Rao, A.R. and *Srinivas, V.V.,* 2008: Regionalization of Watersheds. An approach based on cluster analysis, Springer.

Smakhtin, V.U., 2001: Low flow hydrology: a review. *J. Hydrology 240,* 147–186.

Smith, R.W., 1981: Rock type and minimum 7-day/10-year flow In Virginia streams. Virginia Water Resource Research Center, Virginia Polytechnology Institute and State University, Blacksburg, Bulletin, vol. 116.

Tegegne, G., Park D.K., and *Kim, Y.,* 2017: Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *J. Hydrol.: Regional Studies 14,* 49–66.

Tucci, C., A. Silveira, J. Sanchez, and *Albuquerque, F.,* 1995: Flow Regionalization in the Upper Paraguay Basin, Brazil. *Hydrol. Sci. J. 40,* 485–497.

Van Liew, M.W., Veith, T.L., Bosch, D.D., and *Arnold, J.G.,* 2007. Suitability of SWAT for the conservation effects assessment project: A comparison on USDA-ARS experimental watersheds. *J. Hydrologic Eng., 12,* 173–189.

Vogel, R.M. and *Kroll C.N.,* 1990: Generalised Low-Flow Frequency Relationships for Ungauged Sites in Massachusetts. *Water Resour. Bull. 26,* 241–253.

Wałęga, A., Młyński, D., and *Kokoszka, R.,* 2014: Weryfikacja wybranych metod empirycznych do obliczania przepływów minimalnych i średnich w zlewniach dorzecza Dunajca. *Infrastructure and Ecology of Rural Areas, No II/3*, 825–837 (in Polish).

Wałęga, A. and *Młyński, D.,* 2017. Seasonality of media monthly discharge in selected Carpathian rivers of the upper Vistula basin. *Carpathian J. Earth Environ. Sci. 12,* 61.

Ziernicka-Wojtaszek, A. and *Kaczor, G.,* 2013. Wysokość i natężenie opadów atmosferycznych w Krakowie i okolicach podczas powodzi w okresie maj-czerwiec 2010.. *Acta Sci. Pol., Formatio Circumiectus, 12 (2),* 143 (in Polish).