www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN POLSKA AKADEMIA NAUK

JOURNAL OF WATER AND LAND DEVELOPMENT

e-ISSN 2083-4535

ITP-PIB

Polish Academy of Sciences (PAN) Institute of Technology and Life Sciences – National Research Institute (ITP – PIB)

JOURNAL OF WATER AND LAND DEVELOPMENT
DOI: 10.24425/jwld.2022.142303
2022, No. 55 (X–XII): 47–55

Regionalisation of watersheds with respect to low flow

Agnieszka Cupak, Bogusław Michalec

University of Agriculture in Krakow, Faculty of Environmental Engineering and Land Surveying,
al. Mickiewicza 21, 31-120 Kraków, Poland

RECEIVED 16.02.2022 ACCEPTED 09.06.2022 AVAILABLE ONLINE 17.11.2022

Abstract: The aim of the study was to compare two grouping methods for regionalisation of watersheds, which are similar in respect of low flow and chosen catchments parameters (physiographic and meteorological). In the study, a residual pattern approach and cluster analysis, i.e. Ward's method, were used. The analysis was conducted for specific low flow discharge q_{95} ($dm^3 \cdot s^{-1} \cdot km^{-2}$). In the analysis, 50 catchments, located in the area of the upper and central Vistula River basin, were taken. Daily flows used in the study were monitored from 1976 to 2016. Based on the residual pattern approach (RPA) method, the analysed catchments were classified into two groups, while using the cluster analysis method (Ward's method) – into five. The predictive performance of the complete regional regression model checked by cross-validation R^2_{cv} was 47% and $RMSE_{cv}$ = 0.69 $dm^3 \cdot s^{-1} \cdot km^{-2}$. The cross validation procedure for the cluster analysis gives a predictive performance equal to 33% and $RMSE_{cv}$ = 0.81 $dm^3 \cdot s^{-1} \cdot km^{-2}$. Comparing both methods, based on the cross-validated coefficient of determination (R^2_{cv}), it was found that the residual pattern approach had a better fit between predicted and observed values. The analysis also showed, that in case of both methods, an overestimation of specific low flow discharge q_{95} was observed. For the cross-validation method and the RPA method, the *PBIAS* was –10%. A slightly higher value was obtained for the cross-validation method and models obtained using cluster analysis for which the *PBIAS* was –13.8%.

Keywords: cluster analysis, low flow, regional regression, residual pattern approach, watershed

INTRODUCTION

Low flows, like maximum flows, are a natural component of the hydrological regime of a river. They may occur in summer as well as in winter [Mandal, Cunnane 2009]. During periods of low flow the watercourse is fed by groundwater. Low flow depends on many factors such as the geology of the catchment, the hydrological regime or climate factors (air temperature and precipitation) [Cupak *et al.* 2017]. Proper estimation of low flow characteristics is an important issue in water management, e.g. for determining water resources, water engineering and management, energy use of watercourses, environmental flow determination issues, but also when a watercourse is used as a receiver of treated wastewater [Cupak *et al.* 2017]. Low flows are also important for economic purposes, for the use of surface water for agricultural irrigation, electricity production and for the protection of the ecosystem and its biodiversity [Jurik 2020; Jurik *et al.* 2016; Ševková *et al.* 2012; Voicu *et al.* 2020; Ziernicka-Wojtaszek, Kaczor 2013].

The most reliable method of estimation of low flow characteristics is direct, statistical methods, based on in-stream flow measurement series. The key problem, however, is uncontrolled sites where observation data is not available. In such cases, hydrological regionalisation techniques, based on information from catchments where streamflow data are available, can be used [Demuth, Young 2004]. These methods are applied in many European countries. In Austria, for example, a procedure for estimating low flows has been developed and depends on the available data (catchment type: controlled, uncontrolled). In Poland, there is no such procedure that would clearly indicate which method should be used, hence there is a need to develop it. For controlled catchments, this problem does not occur. For these, the most reliable statistical methods should be used [Cupak 2020; Wałęga *et al.* 2014].

For uncontrolled catchments, on the other hand, empirical formulas are commonly used. However, it should be taken into account that these methods are characterised by the lowest accuracy of obtained results. Another aspect of empirical formulas is the year of their development, e.g. Punzet's formula was developed by the author in the 1980s [Cupak 2020; Wałęga *et al.* 2014]. They, therefore, need to be verified using data now coming from much longer observation sequences [Wałęga *et al.* 2014].

Verification of existing formulas seems to be important also in view of the progressing climate change, its warming, which translates into changes in the water cycle and thus into changes in water resources [Gutry-Korycka, Jokiel 2017]. It is predicted that in the area of Poland the temperature will increase as well as a change in the amount of precipitation in particular seasons of the year will occur. An increase in precipitation is predicted in the winter, while a decrease is predicted in the summer.

The intensity and frequency of phenomena such as floods and droughts are also expected to increase [Suchożebrski 2018]. These forecasts are confirmed by the results of measurements of the average air temperature in Poland. In 2020 it amounted to 9.9°C and was 1.6°C higher than the mean annual multi-year temperature value for the climatological normal period 1981–2010. Since 1951, it has been observed that on the territory of Poland the air temperature increased by about 2.0°C. Over the last 70 years, the air temperature has increased by 2.1°C in the lowlands, Sub-Carpathians and Carpathians. An increase in air temperature is also observed for the winter and summer periods, and since the beginning of the second half of the 20th century, the temperature has increased by 2.5 and 1.9°C, respectively [IMGW 2021]. Also Krajewski *et al.* [2021], in their study on the impact of land use change and climate change on runoff changes in a small agricultural catchment observed an increasing trend in mean annual air temperature. At the same time, they observed both a decrease in runoff depth and annual precipitation. Changes in runoff volume were caused by shifts in climatic variables. Studies of the relationships between flow intermittence and climate, carried out for 452 catchments located across Europe, indicate a strong spatial variability of the seasonal patterns of intermittence and the annual and seasonal number of zero-flow days. Most of the detected trends indicate an increasing number of zero-flow days, especially in southern Europe [Tramblay *et al.* 2020].

The regionalisation method is one of the most common techniques used to extrapolate hydrological information at uncontrolled sites using information from controlled sites [Lin, Wang 2006; Riggs 1973]. Regionalisation includes two tasks: delineation of hydrologically similar regions and identification of regional models for these regions [Lin, Wang 2006]. This method is based on the assumption that catchments with similar climatic and physiographic parameters will be characterised by similar flow, for example, unit outflow or distribution of mean monthly flow. However, in terms of geographic location, the catchments may not necessarily be next to each other. Regression relationship models are developed for the resulting regions. It is important to choose the right clustering method for the variables.

Many studies have analysed different methods of determining the set of water gauging stations which may be regarded as forming a region of sufficient homogeneity of extreme flow characteristics. In their study, Laaha and Blöschl [2006] and Vezza *et al.* [2010] used, among others, the residual pattern approach and statistical method, a hierarchical cluster analysis. In clustering, variables are divided into groups so that, within one cluster there are the most similar variables, and within the other, as dissimilar as possible. Another technique, is the residual pattern approach, which is based on the residuals extracted from a regression model, which is developed for all analysed catchments and their characteristics, without grouping [Laaha, Blöschl 2006]. The regionalisation technique to predict streamflow in ungauged catchments in Mexico has been used by Arsenault *et al.* [2019]. They tested three methods: multiple linear regression (MLR), spatial proximity (SP), and physical similarity (PS).

The aim of the study was to compare two grouping methods for regionalisation of watersheds, which are similar in respect of low flow and chosen catchment parameters (physiographic and meteorological). And to determine the optimum grouping method to be applied in uncontrolled catchments to estimate low flow. For the determination of groups of catchments that are similar in terms of specific low flow discharge q_{95} and meteorological and physiographic parameters, two methods were used: residual pattern approach and cluster analysis, i.e. Ward's method.

MATERIALS AND METHODS

STUDY AREA

The analysis takes into account 50 catchments that are located in the upper and central Vistula basin (Fig. 1), which is physiographically quite diverse. The area of the upper and central Vistula basin covers about 154,579 km^2, which is ca. 50% of the total Poland's area.

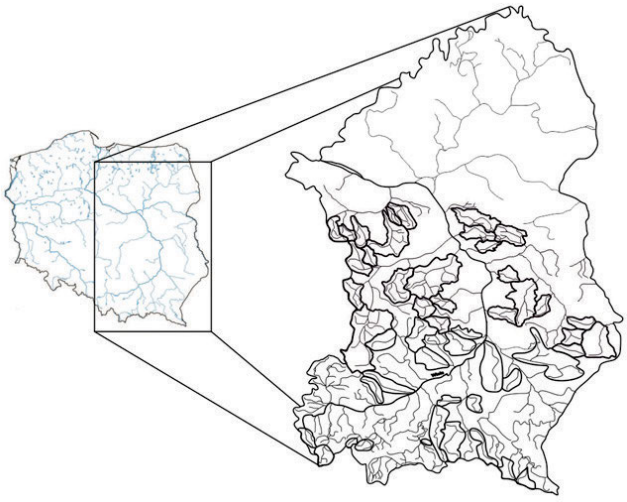

Fig. 1. Location of analysed catchments in the area of upper and central Vistula basin; source: own elaboration

According to Kondracki [2000], the analysed area spreads within three physiographic units: Carpathian Mountains, Non-Alpine Central Europe and East European Lowland. The median catchment altitude of analysed catchments varies from about 119 m a.s.l. for the Sucha catchment (in cross-section Nowa Sucha), which is located in the centre of the central Vistula River

basin. The highest median catchment altitude, about 836 m a.s.l., is observed in the Dunajec catchment in the Nowy Targ cross-section. The catchment is located in the southwestern part of the analysed area of the upper and middle Vistula River basin. Climate, in particular temperature and precipitation, depends on the altitude. In general, as altitude increases, the temperature decreases, and the climate becomes more humid. The average annual temperature varies between 6.0 and 8.5°C. The warmest part of the country is the Silesia Lowlands, where the average annual temperature is about 8.5°C. The coldest season of the year is winter, with an average annual temperature on Kasprowy Wierch of 0.8°C. The average annual precipitation in the analysed area is about 600 mm, with the highest value, about 1400 mm, recorded in the catchments located in the south, and the lowest, about 500 mm, in the centre (lowland catchments) [Cedro, Walczakiewicz 2017]. In the analysis, catchments of different areas were used, from small (like Łubinka of 66.3 km^2) to large (Pilica – 2548.67 km^2 or Tanew – 2034.00 km^2). The average catchment slope ranges from 0.009 for the Wolbórka River to 0.085 for the Wieprz River (Tab. 1). Two soil groups (luvisols and cambic arenosols) were included in the analysis, as other soil groups were not present in most catchments. As there are lakes in the northeastern part of the analysed basin, this area was not included in the analysis.

DATA

For the analysis, daily flows monitored from 1976 to 2016 were taken and collected for 50 catchments located in the upper and central Vistula River basin (Fig. 1). Also for these catchments, 12 chosen physiographic and meteorological parameters were specified (Tab. 1). As a criterion for the selection of catchments, it was assumed that only those catchments would be included in the analysis for which daily flows for at least 20 years are available. Data on daily flows, temperature and precipitation were obtained from the Institute of Meteorology and Water Management National Research Institute – National Research Institute (Pol. Instytut Meteorologii i Gospodarki Wodnej – Państwowy Instytut Badawczy, IMGW-PIB) in Warsaw. Data from the IMGW-PIB was processed. Parameters such as soils and land cover were determined on the basis of the soil map of Poland [Dobrzański *et al.* 1972] and Corine Land Cover 2012 base [CLC undated], morphometric parameters – on the basis of Kondracki [2000], and physiographic parameters were determined in QGIS program, using WMS service. Low flows were quantified based on $Q_{95\%}$, i.e. flow that occurs for 95% of the analysed time. This characteristic is commonly used, among other things, for water management choices, including water supply design. $Q_{95\%}$ was standardised by the catchment area and the specific low flow discharge q_{95} (dm^3·s^{-1}·km^{-2}) was calculated.

CLASSIFICATION OF CATCHMENTS

Regional regression model

The regional regression is constructed as a multiple regression (Eq. 1), which shows the relationship between a specific low flow discharge q_{95} (as dependent variable) and morphoclimatic parameters (independent variables):

$$q_{95} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} \quad (1)$$

where: x_i = analysed catchment characteristics, β_i = regression coefficient.

Stepwise regression was used to plot the regression model because it is most commonly used and gives the most accurate results (Eq. 1). The Mallow's C_p was used for the stepwise regression procedure to get the best model. It is a metric used when there are several variables that can be used in a regression model. It can be used to determine the optimal model in terms of prediction error [Laaha, Blöschl 2006]. It is calculated as [Statology 2021]:

Table 1. Statistical summary of catchments' characteristics

Variable	Symbol	Unit	Value		
			min.	mean	max.
Catchment area	A	km^2	66.63	615.86	2548.67
Length of the watercourse	L	km	11.20	45.57	121.98
Mean annual air temperature	T	°C	5.00	7.39	8.23
Mean annual precipitation	P	mm	536.00	670.05	1192.57
Mean catchment slope	I	–	0.001	0.022	0.085
Median catchment altitude	H_{me}	m a.s.l.	119.50	286.36	836.00
Forests	LU1	%	3.00	27.34	68.17
Grassland	LU2	%	0.00	8.33	34.00
Arable land	LU3	%	6.30	50.92	86.00
Built – up area	LU4	%	0.00	5.95	27.80
Luvisols	S1	%	0.00	51.95	100.00
Cambic arenosols	S2	%	0.00	18.37	73.40

Source: own elaboration.

$$Cp = RSS_p/S^2 - N + 2(P_v + 1) \quad (2)$$

where: RSS_p = the residual sum of squares for a model with p predictor variables, S^2 = the residual mean square for the model (estimated by mean square error – *MSE*), N = the sample size, P_v = the number of predictor variables.

Model performance criteria

The regression model attempted to combine all morphoclimatic variables under the following assumptions: no multicollinearity, the significance of the independent variables, homoscedasticity and normality of residuals. The last two parameters were checked by plotting the normality of the residuals. For this purpose, the Anderson–Darling test and the Shapiro–Wilk test were also used, as well as the White's test to check homoscedasticity. The quality of the model fit was also checked, for the regions obtained, and therefore to what degree q_{95} is explained by the independent variables. The efficiency measures used in the study were the coefficient of determination R^2 and R^2_{adj} (adjusted coefficient of determination), Nash–Sutcliffe efficiency (E) and percentage bias (*PBIAS*).

Additionally, for the regression model, the goodness of fit, in cases when uncontrolled catchment will be included in a region, was tested [Vezza *et al.* 2010]. Determination of the coefficient of determination (R^2, R^2_{adj}) is valued by finding the best model from among the others, but cannot be applied to compare models of various nature. For that reason, a cross-validation method was carried out. On the basis of cross-validation, the coefficient of determination (R^2_{cv}) can be calculated (Eq. 3):

$$R^2_{cv} = \frac{\mathrm{var}(q_{95}) - V_{cv}}{\mathrm{var}(q_{95})} \quad (3)$$

where: V_{cv} = the root mean square residual error, var(q_{95}) = the spatial variance of the flow characteristics.

The value of *PBIAS* (Eq. 4) and a root mean sum of squares error (*RMSE*) – Equation (5) was calculated [Patel 2007]:

$$PBIAS = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{q^i_{95} - \hat{q}^i_{95}}{q^i_{95}}\right) \cdot 100\% \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(q^i_{95} - \hat{q}^i_{95})} \quad (5)$$

where: q^i_{95} = the observed specific low flow discharge q_{95} for catchment i, $\hat{q}^i_{95}$ = the forecast of the model.

RMSE and *BIAS* values equal to 0 indicate a perfect fit. The *PBIAS* indicates the average bias of the simulated data whether it is higher or lower compared to the observed values. A *BIAS* percentage of 0.0 indicates adequate model fit. A positive value shows that the model is underestimated, whereas a negative value shows that the model is overestimated [Fang *et al.* 2014]. The classification proposed by van Liew *et al.* [2007] was used to assess *PBIAS*, described as follows: model is classified as very good when *PBIAS* < 10%; model is good when *PBIAS* is in the range 10–15%; model is satisfactory when *PBIAS* is in the range 15–25%, and model is classified as unsatisfactory when *PBIAS* ≥ 25% [dos Pereira *et al.* 2016].

The E value (Eq. 6) is a normalised statistic which gives the relative magnitude of the residual variance versus the variance in the measured data [Nash, Sutcliffe 1970; Tegegne *et al.* 2017]. The E indicates how well both the observation and computation data fit the 1:1 line [Tegegne *et al.* 2017]. The E was used because it is recommended for use by ASCE [Moriasi *et al.* 2007]. It is calculated as:

$$E = 1 - \frac{\sum_{i=1}^{N}(q_{95} - \hat{q}_{95})^2}{\sum_{i=1}^{N}(q_{95} - \overline{\hat{q}_{95}})^2} \quad (6)$$

where: q_{95} = the observed low flow, $\hat{q}_{95}$ = the calculated low flow, $\overline{\hat{q}_{95}}$ = the average value.

E takes values ranging from 1.0 (indicating perfect fit) to −∞. Values of coefficient less than 0 indicate that the model is useless [Krause *et al.* 2005].

Classification methods

The first applied method was the residual pattern approach (RPA), in which the residual, between the low flow values, calculated from the global regression model, and the observed values, was estimated. Geographically contiguous regions are then plotted manually on a map [Cupak 2020]. In the RPA, in the first step, a global regression model should be determined using stepwise regression, and then the residuals obtained from the global model should be plotted on a map in geographical space. In the last step, if patterns of residuals are visible, then regions with similar signs and magnitude of residual values should be identified [Vezza *et al.* 2010].

Another method was cluster analysis, in which a set of feature vectors is divided into clusters or groups in such a way that for one cluster the feature vectors are as similar as possible and as different as possible from other adjacent clusters [Cupak 2020].

For the calculation, the Euclidean distance, as a measure of similarity, was used. It is a measure of the distance between objects defined by relevant features. Next, the delineation of homogeneous regions (cluster agglomeration) was carried out on the basis of the relationship between the catchment characteristics and the specific low flow q_{95}. For this purpose, the hierarchical cluster analysis method, i.e. Ward's method, was used. The purpose of Ward's algorithm [Ward 1963] is to minimise the sum of squares of the deviations from the centroids of their clusters [Rao, Srinivas 2006]. Among the hierarchical cluster analysis methods, the Ward's algorithm is the most commonly used. It is characterised by a trend to form equal-value spherical clusters and it performs well in recovering the cluster structure. This makes Ward's algorithm a useful tool for identifying homogeneous regions for regionalisation [Rao, Srinivas 2006]. However, as with other hierarchical clustering techniques, Ward's algorithm does not provide for the reassignment of feature vectors that may not have been correctly classified at the start of the analysis [Rao, Srinivas 2006].

RESULTS AND DISCUSSION

The analysis began by defining a global regression model using the stepwise regression method. The initial regression model consisted of five variables. Then, to avoid overestimation, the

variables were manually verified and these variables (three variables) that had the least influence on the performance of the model were all rejected (R^2 decreased from 65% to 59%) and the resulting model is defined by Equation (7). Watercourse length (L) and median catchment elevation (H_{me}) were found to be the most relevant variables for low flow regionalisation. The model parameters (L and H_{me}) are statistically significant at the 0.05 level. The coefficient of determination for this model was 59%, while R^2_{adj} was 57%, and R^2_{cv} was 45%.

$$q_{95} = 0.352 + 0.011L + 0.0048H_{me} \quad (7)$$

The residuals were tested against the general assumptions of multiple regression, non-linearity and homoscedasticity. The model assumption of normality of residuals and heteroscedasticity was checked using the Shapiro–Wilk test (p-value was 0.263), the Anderson–Darling test (p-values were 0.41) and diagnostic plots (Fig. 2), and from these, it can be concluded that the residuals have a normal distribution. The White's test showed homoscedasticity of the residuals (p-value was 0.375).

The residual map is shown in Figure 3. It was observed that the specific low flow discharge q_{95} for the catchments analysed in this study had low values, and therefore the differences between them were not large, in contrast to e.g. Laaha and Blöschl [2006] or Vezza *et al.* [2010]. Therefore, for further analysis, the low residuals were assumed to be 20% of the average specific low flow discharge q_{95}.

The approach using residuals suggests that the analysed catchments can be divided into two major sub-units. The first one included lowland catchments and upland catchments. In this group, the residuals were mostly small (<0.5 $dm^3 \cdot s^{-1} \cdot km^{-2}$), and their distribution seemed to be random. The second group consisted of catchments located in mountainous and upland areas of the Upper Vistula Water Region. This group was characterised by higher values of residuals (>0.5 $dm^3 \cdot s^{-1} \cdot km^{-2}$) and the distribution of residuals was random. Regression models were developed for the groups obtained using the RPA method (Tab. 2). The parameter that influenced low flow in both regression models was the median catchment altitude. The value of the coefficient of determination calculated for the second group was high, at 82% (R^2_{adj} = 81%). In contrast, a much lower value of R^2 = 47% was found for group 1. On the other hand, the cross-validation coefficient of determination calculated for the RPA method and all catchments was R^2_{cv} of 47% and $RMSE_{cv}$ = 0.69 $dm^3 \cdot s^{-} \cdot km^{-2}$. Figure 4 shows the classification of the groups determined by the RPA method. An overestimation of low flows (*PBIAS* has a negative sign) was noted in the models obtained. For group 1, the *PBIAS* value was ten times higher

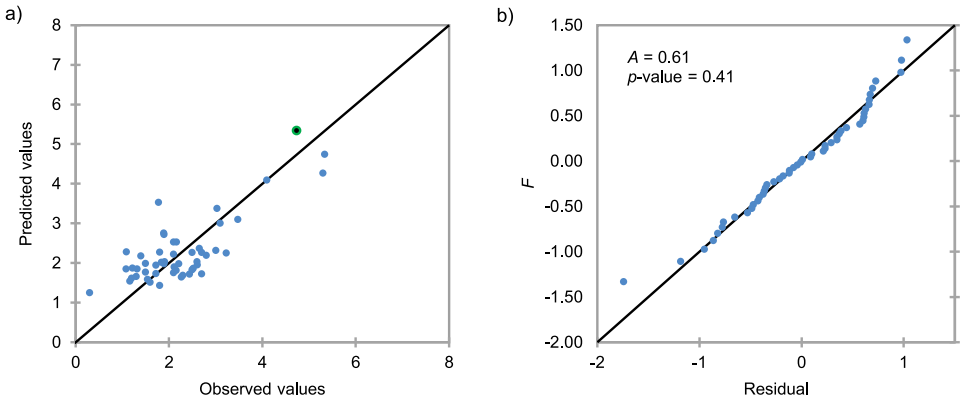

Fig. 2. Global regression model: a) scatter plot of regression model, b) residuals normal plot; A = the value of the Anderson–Darling test, F = cumulative distribution function; source: own study

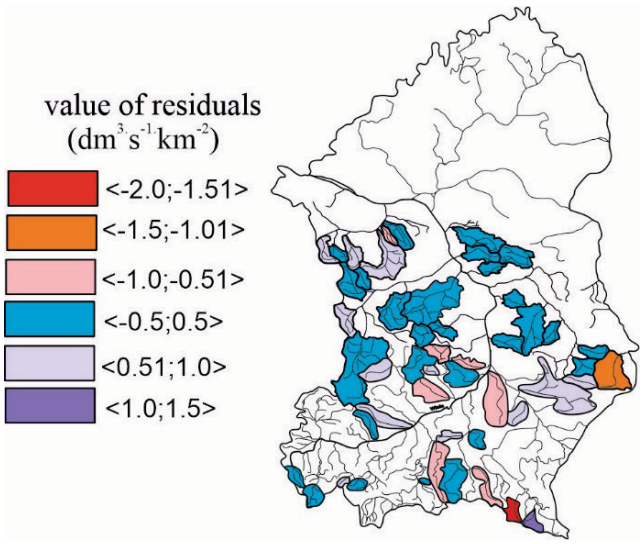

Fig. 3. Residual pattern of the global regression model: source: own study

compared to the second group. According to the criterion proposed by van Liew *et al.* [2007], the model for the first group was unsatisfactory, and for the second one, it was classified as very good. The Nash–Sutcliff E-coefficient for group 1 gave a good model fit (E = 0.47), while for group 2 the model fit was excellent (E = 0.82) – Table 2. Taking into account that the grouping of catchments in the RPA method is based on the calculated residuals (their size and sign), while the catchment characteristics are omitted, it is suggested that uncontrolled catchments should be assigned to a given group based on their geographical location. For the catchments located in the central Vistula River basin, it is proposed to allocate the catchment to group 2. Whereas the catchment, which is located within the Upper Vistula Water Region, should be allocated to group 1.

The second clustering method used in this study was cluster analysis. According to Laaha and Blöschl [2006] and Vezza *et al.* [2010], the best is Ward's method. Euclidean distance was adopted as the measure of distance between clusters. It gives the most preferable classification of clusters. Ward's method resulted

Table 2. Models based on residual pattern approach (RPA) method

Group	Model	R^2 (%)	R^2_{adj} (%)	*RMSE* ($dm^3 \cdot s^{-1} \cdot km^{-2}$)	*PBIAS* (%)	*E*
1	$q_{95} = 0.982 + 0.004H_{me}$	47	44	0.893	-28.2	0.47
2	$q_{95} = 0.344 + 0.01L + 0.004H_{me}$	82	81	0.305	-2.23	0.82

Explanations: R^2 = coefficient of determination, R^2_{adj} = adjusted coefficient of determination, *RMSE* = root mean sum of squares error, *PBIAS* = percentage bias, *E* = Nash–Sutcliffe efficiency, q_{95} = specific low flow discharge that occurs for 95% of the analysed time.
Source: own study.

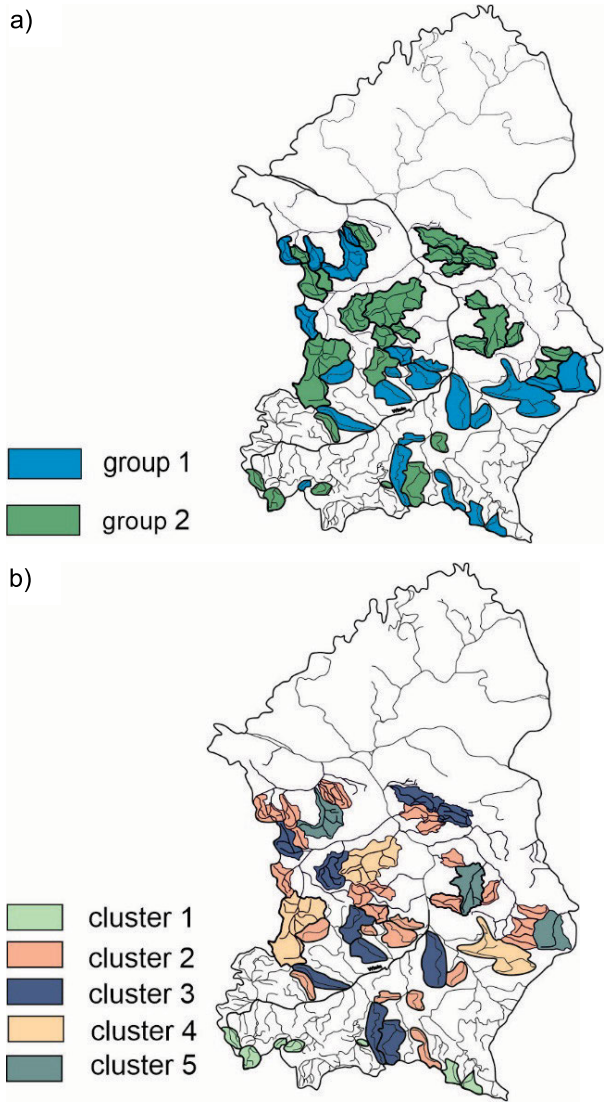

Fig. 4. Groups of catchments based on: a) residual pattern approach (RPA) method, b) cluster analysis; source: own study

in the agglomeration of the analysed catchments into five clusters. Cluster 1 included 7 catchments, the 2 included 27 catchments, and the 3 included 10 catchments. For clusters 4 and 5, the same number of catchments was recorded: 3 catchments. Smakhtin [2001] states that the catchment characteristics and meteorological parameters that most frequently influence low flows include catchment area, precipitation, and the presence of areas covered by forest or underlain by surface standing water. Equally important are lithological parameters, existing soils or the density of the river network. Gustard and Irving [1994], in their study, linked soil data with hydrological data, dividing the soils presented in the United Kingdom into classes. They distinguished 12 classes, for which they defined regression models for low flow [Smakhtin 2001]. The parameters which influenced catchment grouping in this study were catchments' area, length of the watercourse, mean catchment slope and mean annual precipitation. Also, in terms of soil, differences between clusters were observed. Clusters 1 and 2 were dominated by luvisols and clusters 4 and 5 by cambic arenosols.

Then, after identifying clusters, a map of their localisation in the analysed area was made (Fig. 4). It can be observed that the catchments forming a given cluster are adjacent to each other, and only in the case of mountain catchments included in the cluster 1 and lowland catchments, their location is dispersed. This provides information that, in the case of cluster analysis, there is a link between the continuity of the region and the catchment characteristics in spatial terms. For the clusters obtained, the models were influenced by different parameters (Tab. 3). For cluster 1, the parameter that had the greatest influence on the variability of the low flows was the median catchment altitude, and was statistically significant at the 0.05 significance level. The model parameters for cluster 2, selected by the regression analysis, were median catchment altitude and luvisols, and they were significant in statistical terms at the 0.05 significance level. For the models in groups 1 and 2, it can be seen that the specific low flow discharge q_{95} was affected by the median altitude of the catchment. Also, the soils for groups 2 and 3 proved to be significant parameters included in the regression equation

Table 3. Models based on cluster analysis

Group	Model	R^2 (%)	R^2_{adj} (%)	*RMSE* ($dm^3 \cdot s^{-1} \cdot km^{-2}$)	*PBIAS* (%)	*E*
1	$q_{95} = -1.74 + 0.0084H_{me}$	71	65	0.84	–6.34	0.71
2	$q_{95} = 1.514 + 0.0042H_{me} - 0.0098S1$	36	31	0.52	–14.6	0.36
3	$q_{95} = 2.29 - 0.018S2$	38	30	0.49	–5.09	0.46

Explanations: H_{me}, S1, S2 as in Tab. 1, the other as in the Tab. 2.
Source: own study.

(luvisols and cambisols, respectively). However, for clusters 4 and 5, due to the small number of catchments (3 catchments per group), it was not possible to determine the regression relationships. Clusters 2 and 3 are rather poorly explained by the respective multiple regression models (groups 2 and 3 with R^2_{adj} = 31% and R^2_{adj} = 30%). This suggests that the models do not completely reflect the forecasting characteristics for ungauged catchments. In the case of cluster 1, however, model performance was good (R^2_{adj} = 65%), suggesting that there may be heterogeneity in processes associated with low flow in this group. The catchment characteristics used in the cluster analysis did not fully reflect the regional anomalies in the low flow pattern. Also, the *E*-coefficient for cluster 1 was 0.71, which corresponds to a very good model, but for clusters 2 and 3, the models were classified as good (*E* was 0.36 and 0.38 respectively). Under the classification suggested by VAN LIEW *et al.* [2007], the regression models for clusters 1 and 2 are classified as very good and for cluster 2 as satisfactory. The cross-validation procedure for cluster analysis gave a predictive efficiency (R^2_{cv}) of 33% and $RMSE_{cv}$ = 0.81 $dm^3 \cdot s^{-1} \cdot km^{-2}$. The result obtained was significantly worse than the method based on the residuals. For models obtained by cluster analysis, an overestimation of the predicted values of low flows was observed (*PBIAS* had a negative sign). According to VAN LIEW *et al.* [2007], the model for clusters 1 and 3 was classified as very good, while for cluster 2, it was classified as good.

Determination of flow characteristics in uncontrolled catchments based on regional correlation models is one of the most frequently used methods. It uses the relationships between a given low-flow characteristic, e.g. specific low flow discharge q_{95}, and catchment parameters [VEZZA *et al.* 2010]. The regression model determined for all catchments, without catchments grouping, gave R^2_{cv} = 45% for cross-validation and was lower compared to the calculated coefficient of determination, which was 59%. Considering the two clustering methods, it should be noted that it was clear that the residuals method gives better results compared to the cluster analysis method. The R^2_{cv} value for the RPA method was slightly higher than the global regression model across the study area. The RPA method gave a slight improvement in performance (R^2_{cv} = 47%) compared to the global model. We also obtained a lower R^2_{cv} value for the RPA method than VEZZA *et al.* [2010], which was 53% for their study, and much lower compared to LAAHA and BLÖSCHL [2006], who obtained R^2_{cv} = 63%. For cluster analysis, we obtained a cross-validation coefficient of determination (30%) and this was twice as low compared to other studies. VEZZA *et al.* [2010], in their study, obtained 68% while LAAHA and BLÖSCHL [2006] obtained 59%.

The final step in the evaluation of catchment pooling methods was to examine scatter diagrams of predicted and observed specific low flow discharge q_{95} (Fig. 5). The scatter diagrams provide detailed information about the results for each analysed catchment, such as the existence of outliers and possible heteroscedasticity in the observations and forecasts [LAAHA, BLÖSCHL 2006]. It should be stated that the scatter plots developed for the catchment grouping methods used in this study correspond to the coefficient of determination calculated on the basis of the cross-validation method for these methods. Clearly, of the two methods, the residual pattern approach performed better than the cluster analysis. Also, for the cross-validation method, it

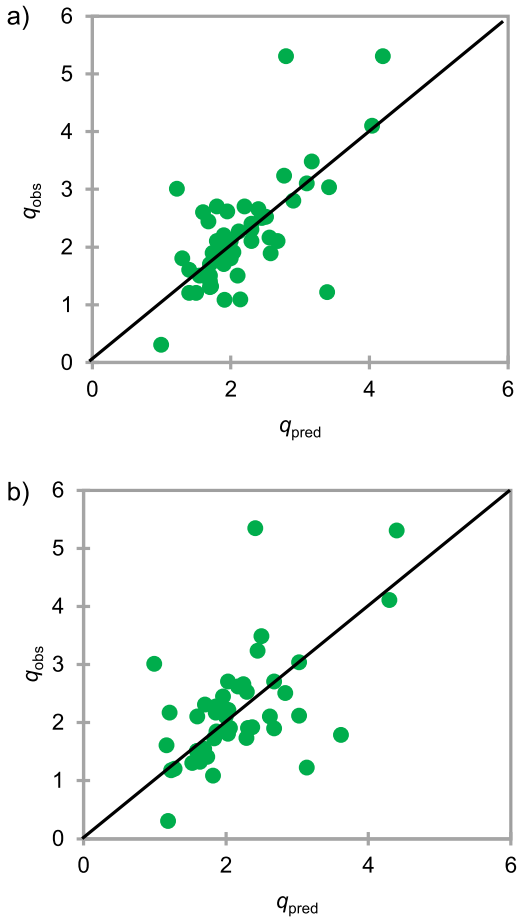

Fig. 5. Cross-validation scatter plots observed and predicted specific low flow according two methods: a) residual pattern approach (RPA), b) cluster analysis; source: own study

was checked whether the models produced underestimated or overestimated results. For the RPA method, the *PBIAS* was –10%, which indicated an overestimation of specific low flow discharge q_{95} and according to van LIEW *et al.* [2007], it gave a good model. For the cross-validation method and models obtained using cluster analysis, the *PBIAS* was –13.8% and also indicated overestimation of specific low flow discharge q_{95} so the model was classified as good.

CONCLUSIONS

The paper compares two catchment pooling methods in terms of their performance in predicting specific low flow discharges q_{95}. The analysis was made using the RPA method, and cluster analysis, i.e. Ward's method. The analysis carried out showed that the residual pattern approach had a better fit between calculated and observed values. Based on this method, the catchments were divided into two groups. The first group included lowland catchments and some hilly catchments, and the second group consisted of upland and hilly catchments. The predictive efficiency of the regional regression model for the RPA, tested by the R^2_{cv} cross-validation method, was 47% and $RMSE_{cv}$ = 0.69 $dm^3 \cdot s^{-1} \cdot km^{-2}$. In contrast, five clusters were obtained using Ward's method. However, in the case of clusters 2 and 3, the obtained coefficient of determination had a low value

(about 40%). For region 1 the model performance was good. The cross-validation procedure gave a predictive efficiency of 33% and $RMSE_{cv}$ = 0.81 $dm^3 \cdot s^{-1} \cdot km^{-2}$. This was significantly worse than the RPA. Comparing the results obtained from the two methods, it can be concluded that the RPA method is more optimal for use in Polish conditions.

The conducted analysis allows us to conclude that methods of low-flow regionalisation may give promising results of low flow calculation in ungauged catchments in Polish conditions. However, in our opinion, it will be very interesting to use other clustering methods to support the critical analysis of the results and to select the best model, as well as analyse the accuracy of the results obtained. Another thing is that in our study we studied one low flow index (q_{95}) to see if the effectiveness of the grouping methods will be the same if other features are analysed. Furthermore, considering climate warming that affects the water cycle and, consequently, water resources, it will be interesting to examine in future studies whether it affects low flows.

REFERENCES

Arsenault R., Breton-Dufour M., Poulin A., Dallaire G., Romero-Lopez R. 2019. Streamflow prediction in ungauged basins: analysis of regionalization methods in a hydrologically heterogeneous region of Mexico. Hydrological Sciences Journal. Vol. 64 (11) p. 1297–1311. DOI 10.1080/02626667.2019.1639716.

Cedro A., Walczakiewicz S. 2017. Podstawy meteorologii i klimatologii. W: Odnawialne źródła energii w Polsce ze szczególnym uwzględnieniem województwa zachodniopomorskiego [Basics of meteorology and climatology. In: Renewable energy sources in Poland with particular reference to the West Pomeranian Voivodeship]. Eds. M. Świątek, A. Cedro. Szczecin. Wydaw. ZAPOL Sobczyk p. 29–44.

CLC undated. CORINE Land Cover – CLC 2012 [online]. [Access 10.01.2022]. Available at: https://clc.gios.gov.pl/index.php/clc-2012/metadane

Cupak A. 2020. Regionalization methods for low flow estimation in ungauged catchments – A review. Acta Scientiarum Polonorum. Formatio Circumiectus. Vol. 19(1) p. 21–35. DOI 10.15576/ASP.FC/2020.19.1.21.

Cupak A., Wałęga A., Michalec B. 2017. Cluster analysis in determination of hydrologically homogeneous regions with low flow. Acta Scientiarum Polonorum. Formatio Circumiectus. Vol. 16(1) p. 53–63.

Demuth S., Young A.E. 2004. Regionalisation procedures. In: Hydrological drought: Processes and estimation methods for streamflow and groundwater. Eds. L.M. Tallaksen, H.A.J. van Lanen. Developments in Water Science. Vol. 48. Amsterdam. Elsevier p. 307–344.

Dobrzański B., Witek T., Kowaliński S., Królikowski L., Kuźnicki F., Siuta J., ..., Zawadzki S. 1972. Polska mapa gleb. Warszawa. Wydaw. Geologiczne.

Dos Pereira D.R., Martinez M.A., da Silva D.D., 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. Journal of Hydrology: Regional Studies. Vol. 5 p. 149–163. DOI 10.1016/j.ejrh.2015.11.008.

Fang G.H., Yang J., Chen Y.N., Zammit C. 2014. Comparing bias correction methods in downscaling meteorological variables for hydrologic impact study in an arid area in China. Hydrology and Earth System Sciences Discussions. Vol. 11 p. 12659–12696. DOI 10.5194/hessd-11-12659-2014.

Gustard A., Irving K.M. 1994. Classification of the low flow response of European soils. FRIEND: Flow Regimes from International Experimental and Network Data. IAHS Publication. No. 221 p. 113–117.

Gutry-Korycka M., Jokiel P. 2017. Projekcje ewolucji zasobów wodnych Polski w wyniku zmian klimatu i wzrastającej antropopresji. W: Hydrologia Polski [Projections of the evolution of Poland's water resources as a result of climate change and increasing anthropopression. In: Hydrology of Poland]. Eds. P. Jokiel, W. Marszelewski, J. Pociask-Karteczka. Warszawa. Wydaw. Nauk. PWN p. 301–305.

IMGW 2021. Klimat Polski 2020. Raport [Climate of Poland 2020. Report]. Warszawa. Instytut Meteorologii i Gospodarki Wodnej Państwowy Instytut Badawczy pp. 24.

Jurik L. 2020. Sucho v krajine a vodné stavby [Drought in the landscape and water structures]. Vodohospodàrsky spravodajca. No. 7–8 p. 5–7.

Jurik L., Kaletová T., Halaj P. 2016. Water management for sustainable growth strategies. Visegrad Journal on Bioeconomy and Sustainable Development. Vol. 1 p. 31–35. DOI 10.1515/vjbsd-2016-0006.

Kondracki J. 2000. Geografia regionalna Polski [Regional geography of Poland]. Warszawa. Wydaw. Nauk. PWN. ISBN 83-01-13050-4 pp. 440.

Krajewski A., Sikorska-Senoner A.E., Hejduk L., Banasik K. 2021. An attempt to decompose the impact of land use and climate change on annual runoff in a small agricultural catchment. Water Resources Management. Vol. 35 p. 881–896. DOI 10.1007/s11269-020-02752-9.

Krause P., Boyle D.P., Bäse F. 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences. Vol. 5 p. 89–97. DOI 10.5194/adgeo-5-89-2005.

Laaha G., Blöschl G. 2006. A comparison of low flow regionalization methods-catchment grouping. Journal of Hydrology. Vol. 323 p. 193–214. DOI 10.1016/j.jhydrol.2005.09.001.

Lin G.F., Wang C.M. 2006. Performing cluster analysis and discrimination analysis of hydrological factors in one step. Advances in Water Resources. Vol. 29 p. 1573–1585. DOI 10.1016/j.advwatres.2005.11.008.

Mandal U., Cunnane C. 2009. Low-flow prediction for ungauged river catchments in Ireland [online]. Irish National Hydrology Seminar. [Access 10.12.2021]. Available at: https://hydrologyireland.ie/wp-content/uploads/2016/12/4-Low-flow-prediction-for-ungauged-river-catchments-in-Ireland.pdf

Moriasi D.N., Arnold J.G., van Liew M.W., Bingner R.L., Harmel R.D., Veith T.L. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE. Vol. 50(3) p. 885–900. DOI 10.13031/2013.23153.

Nash J.E., Sutcliffe J. 1970. River flow forecasting through conceptual models part I – A discussion of principles. Journal of. Hydrology. Vol. 10 p. 282–290. DOI 10.1016/0022-1694(70)90255-6.

Patel J.A. 2007. Evaluation of low flow estimation techniques for ungauged catchments. Water and Environment Journal. Vol. 21 p. 41–46. DOI 10.1111/j.1747-6593.2006.00044.x.

Rao A.R., Srinivas V.V. 2006. Regionalization of watersheds by hybrid cluster analysis. Journal of Hydrology. Vol. 31(1–4) p. 57–79. DOI 10.1016/j.jhydrol.2005.06.003 .

Riggs H.C. 1973. Regional analysis of streamflow characteristics. Techniques of Water Resources Investigations of the United States Geological Survey. Book 4. Chapt. B3. Washington DC. USGS pp. 15.

Smakhtin V.U. 2001. Low flow hydrology: A review. Journal of Hydrology. Vol. 240 p. 147–186. DOI 10.1016/S0022-1694(00)00340-1.

Statology 2021. What is Mallows' Cp? (defintion & example) [online]. Statology. Statistics. Simpilified. [Access 10.01.2022]. Available at: https://www.statology.org/mallows-cp/

Ševková A., Sabo M., Kohnová S. 2012. Pooling of low flow regimes using cluster and principal component analysis. Slovak Journal of Civil Engineering. Vol. 20(2) p. 19–27.

Suchożebrski J. 2018. Zasoby wodne Polski. W: Zarządzanie zasobami wodnymi w Polsce [Water resources of Poland. In: Water resources management in Poland]. Discussion Paper. Global Compact Network Poland p. 92–96.

Tegegne G., Park D.K., Kim Y. 2017. Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. Journal of Hydrology: Regional Studies. Vol. 14 p. 49–66. DOI 10.1016/j.ejrh.2017.10.002.

Tramblay Y., Rutkowska A., Sauquet E., Sefton C., Laaha G., Osuch M., ..., Datry T. 2020. Trends in flow intermittence for European rivers. Hydrological Sciences Journal. Vol. 66(1) p. 37–49. DOI 10.1080/02626667.2020.1849708.

Van Liew M.W., Veith T.L., Bosch D.D., Arnold J.G. 2007. Suitability of SWAT for the conservation effects assessment project: A comparison on USDA-ARS experimental watersheds. Journal of Hydrologic Engineering. Vol. 12(2) p. 173–189. DOI 10.1061/(ASCE)1084-0699(2007)12:2(173).

Vezza P., Compoglio C., Rosso M., Viglione A. 2010. Low flows regionalization in North-Western Italy. Water Resources Management. Vol. 24 p. 4049–4074. DOI 10.1007/s11269-010-9647-3.

Voicu R., Radecki-Pawlik A., Tymiński T., Mokwa M., Sotir R., Voicu L. 2020. A potential engineering solution to facilitate upstream movement of fish in mountain rivers with weirs: Southern Carpathians, the Azuga River. Journal of Mountain Science. Vol. 17 p. 501–515. DOI 10.1007/s11629-019-5572-y.

Wałęga A., Młyński D., Kokoszka R. 2014. Weryfikacja wybranych metod empirycznych do obliczania przepływów minimalnych i średnich w zlewniach dorzecza Dunajca [Verification of selected empirical methods for the calculation of minimum and mean flows in catchments of the Dunajec basin]. Infrastruktura i Ekologia Terenów Wiejskich. Vol. II/3 p. 825–837.

Ward Jr J.H. 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association. Vol. 58 p. 236–244.

Ziernicka-Wojtaszek A., Kaczor G. 2013. The intensity and amount of precipitation in both the city of Krakow and the neighbouring areas during the May–June 2010 flood. Acta Scientiarum Polonorum. Formatio Circumiectus. Vol. 12(2) p. 143–151.