



# Mixed statistical and data mining analysis of river flow and catchment properties at regional scale

Marek Giełczewski<sup>1</sup> · Mikołaj Piniewski<sup>1</sup> · Paweł D. Domański<sup>2</sup>

Accepted: 3 January 2022 / Published online: 22 January 2022  
© The Author(s) 2022

## Abstract

This paper presents an application of the L-moments and L-moment ratio diagrams (LMRD) to the analysis of hydrological data at regional (country) scale. Existing research focuses on two main areas of the analysis: statistical analysis using LMRD and regression analysis. Further research mixes both approaches applying regression analysis to L-moments. Another direction of the research is clustering of the climatic and physiographic catchment properties and its validation using LMRD. However, LMRD plots can be separately used as the clustering domain. It is proposed to decompose the features into some classes, and then present these results on the LMRD. Such plots constitute the source for the clustering. Obtained clusters are then validated against k-means clustering performed in the LMRD diagram domain. Results show that statistical L-moments analysis can be improved with data mining clustering algorithms. Such combination delivers a new perspective for the interpretation of the results. It is shown that clustering in the LMRD domain is consistent with the K-means clustering. It is another argument showing that L-moments diagrams can be considered as a very powerful and informative tool for hydrologists enabling the comparison on the regional basis with respect to various catchment properties. The method is validated on data consisting of daily river flow data from 290 gauges covering entire Poland.

**Keywords** Hydrology · Extremes · L-moments · Kappa distribution · Clustering · River flow · Catchment properties

## 1 Introduction

Investigating the effects of catchment properties related to climate and physiography on river flow characteristics is a frequent area of research in hydrology. Such an analysis is mostly based on statistical approaches (Helliwell et al. 2007; Calver et al. 2009; Merz and Blöschl 2009; Salinas

et al. 2014; Singh 2017). Actually, the analysis can be conducted on the single gauge basis (Popat et al. 2020; Negi et al. 2021) or on the higher regional level (Smith et al. 2015; Kar et al. 2017). Regional analysis uses different approaches. Generally, there are two leading methodologies. Application of the method of L-moments in the regional frequency analysis of extreme events for river flows, apart from other natural sciences, is considered to be one possible approach (Hosking and Wallis 1993; Kjeldsen et al. 2017; Simková 2017). The second methodology uses regression analysis (Merz and Blöschl 2009; Mašiček and Toman, Palàt 2011; Lun et al. 2021). There is also research mixing both approaches (Di Baldassarre et al. 2006; Salinas et al. 2014), i.e. application of the regression analysis to correlate the L-moment diagram data.

Both approaches have their advantages and shortcuts. L-moments ratio diagrams (LMRD) origin from the traditional moment ratio diagrams (MRD) introduced by Karl Pearson in the beginning of the XIX century. The idea is to use and graphically represent the relationship between the first four moments of the distribution. There are various versions of MRD, but the diagram showing the relationship

---

Marek Giełczewski and Mikołaj Piniewski contributed equally to this work.

✉ Paweł D. Domański  
pawel.domanski@pw.edu.pl

Marek Giełczewski  
marek\_gielczewski@sggw.edu.pl

Mikołaj Piniewski  
mikolaj\_piniewski@sggw.edu.pl

<sup>1</sup> Department of Hydrology, Meteorology and Water Management, Warsaw University of Life Sciences, Nowoursynowska 159, Warsaw 02-776, Poland

<sup>2</sup> Institute of Control and Computation Engineering, Warsaw University of Technology, Nowowiejska 15/19, Warsaw 00-665, Poland

between the fourth and the third moment, i.e. kurtosis versus skewness is the most popular one. MRD of the variation factor versus skewness is the second popular representation.

MRDs can show empirical data or analytical theoretical curves. Theoretical relations can be evaluated for the majority of univariate distributions (Vargo et al. 2010). MRD can be used for several purposes, such as:

1. to quantify the distance between various univariate distributions,
2. to visualize distribution versatility within the range of its moments,
3. to select the best fit distribution to the empirical data,
4. to visualize and test the relationships between distribution families,
5. and to identify homogeneous process expressed by the empirical moments relations.

Historical formulation uses classical moments definitions as for normal Gaussian distribution. Introduction of the L-moments by Hosking (1990) enabled to increase the approach applicability. L-moments offer better moments estimation properties, especially for small sample number and fat-tailed distributions (outliers). It has been shown that they might even outperform other estimation approaches, like the standard algorithm of moments, TL- or LQ-moments (Šimková 2017). L-moments and respective L-moments diagrams are widely used in the extreme events analysis in climatology, hydrology (Katz et al. 2002; Maeda et al. 2013), astronomy (Podladchikova et al. 2003) or medicine (Louzada et al. 2016). LMRD is the most common tool supporting the identification of a suitable frequency distribution of empirical samples (Kjeldsen and Prosdocimi 2015). Typical diagram represents the L-kurtosis ( $\tau_4$ ) versus L-skewness ( $\tau_3$ ). Empirical data may be easily confronted with the theoretical plots for the selected candidate probabilistic density functions (PDFs). The proximity of the empirical data to the theoretical ones might play the role of a selection criterion to choose the best fit distribution (Peel et al. 2001).

Apart from distribution fitting, L-moments diagrams are often utilized to compare various samples originating from different sources, like for instance river discharge, in search for their homogeneity. There are defined various discordance measures used to recognize sources, whose sample's L-moments are marked contrarily from the others (Hosking and Wallis 1993; Khan et al. 2017). The task associated with the homogeneity testing is realized in the 2-dimensional (2D) space spanned by L-skewness and L-kurtosis. Verification of the homogeneity hypothesis may be performed by the visual inspection or may be supported by the dedicated distance measures.

L-Moment Ratio Diagrams deliver new perspective to the statistical analysis of various time series properties. These properties are described by the aggregated position of the time series on the diagram. Thus, they allow to perform regional analysis comparing different features associated with the data. This feature is the strong advantage of the approach. However, the method does not give any insight into the specific statistical properties. Regional analysis is limited to the general clustering of the observations without connection to the internal (mean, L-Cv, L-skewness and L-kurtosis) moments.

Regression analysis (Garmdareh et al. 2018; Desai and Ouarda 2021) is applied to address this issue. It aims at finding a quantitative relationship between certain catchment properties and respective flow time series L-moments. This approach is quite natural and promising. However, one has to take into account that standard regression using least squares (LS) has its limitations. It assumes that the process behind data is stationary, affected by Gaussian noise and without outliers (Hawkins 1980; Domański 2020). If these conditions are not met, the LS regression estimation will be biased (Rousseeuw and Leroy 1987; Huber and Ronchetti 2009). Moreover, LS regression assumes that the observations are relatively uniformly distributed over the domain number of observations and their number is high enough to satisfy convergence towards the limiting probabilistic density function (PDF). Unfortunately, hydrological data does not meet these assumptions. The process underlying statistical process are neither stationary, nor Gaussian. Outliers are frequent, what is visible in the tails fatness and the resulting need for the extreme analysis. The length of the time series is not long enough to assure the convergence. Therefore, obtained estimations can be susceptible to the bias.

Researchers use different approaches to minimize the effect, as for instance weighted moving average (WMA) (Salinas et al. 2014). The present study does not focus on that aspect, though sample regression analysis in different context is included. It is proposed to use robust regression performance index in form of the least median square (LMS) (Rousseeuw 1984).

Following above arguments considering LMRDs and regression analysis, the first approach (LMRD clustering) is selected as it opens a new research opportunity. This opportunity arises from the fact that LMRD plots enable various approaches to the data comparison in the 2D space of the moment ratio diagram. Such a statistical procedure resembles the 2D grouping tasks well known in the data mining research. One may find dozens of clustering algorithms that aim at data grouping, finding cluster centers or identifying the delimitation borders (Aggarwal and Reddy 2014; King 2015). Thus, the use of clustering methods in the grouping task in the L-moments diagrams data seems to

be natural and obvious. As the grouping itself is not an ultimate focus of this research, classical k-means algorithm is used (Lloyd 1982), as it is one of the simplest and popular unsupervised machine learning methods.

Recent literature review shows some reports on clustering approaches used in the hydrological analysis. The Ward's data agglomeration method is a hierarchical approach and uses analysis of variance as the measure of the distance between clusters. It has been used for catchment hydrological analysis (Cupak 2017; Cupak et al. 2017). A principal component analysis and a subsequent clustering of the principal components has been used to cluster catchment data (Jehn et al. 2020), while authors of Sharghi et al. (2018) compared four approaches of clustering, K-means, Ward, SOM and Wavelet-Entropy-based method in a similar task. K-means clustering has been applied in the analogous research in Dikbas et al. (2013), Aytac (2020). Step-wise cluster analysis hydrological model has been proposed in Wang et al. (2021). Actually, current research is frequently organized as follows: researchers select a clustering algorithm, do catchments' clustering by their features and finally observe the results in the LMRD. Proposed approach is opposite. First, LMRD are plotted and the K-means clustering is performed in the diagram 2D domain. Clustering results are compared with expert and statistical features classification.

Concluding, this study aims at the analysis of climatic and physiographic catchment properties in the domain of the L-moments ratio diagram based on the river daily flow data from 290 flow gauges from Polish rivers. Catchment properties are compared with the L-moments of the river flow data and the best fitted PDFs. The analysis is supported by the k-means clustering to identify leading properties and to allow the generalization of observations. The study starts with Sect. 2 presenting the considered geographical area and respective data. Regional clustering and analysis of properties is included in Sections 4, preceded by Sect. 3 describing applied methods and algorithms and followed by Sect. 5 with concluding remarks and open research items.

## 2 Study area and data

Regional hydrological analysis is performed using daily river flow data [ $m^3/s$ ] provided by the Institute of Meteorology and Water Management - National Research Institute (IMGW-PIB) from 290 gauges located in Poland. The analysis uses available historical data (daily discharges) starting from November 1950 till the end of October 2019. The map showing the area of interest is presented in Fig. 6. Moreover, the analysis uses a set of catchment climatic and

physiographic properties, i.e. the catchment area [ $km^2$ ], mean annual precipitation [mm], mean annual of the minimum and maximum daily temperatures [deg. C], catchment slope [-], elevation [m asl], dominant land use classes (fractions), dominant soil hydrological groups (fractions), soil permeability classes (fractions) and fraction of lakes [-]. All the properties were calculated in ArcGIS software using readily available climatic and physiographic data for the respective catchment areas. Climatic data were derived from the G2DC-PL+ data set (Piniewski et al. 2021), and physiographic data were derived from the input maps used in the recently developed hydrological model SWAT covering the Polish territory (Marcinkowski et al. 2021).

At first, catchment properties are briefly analyzed, i.e. any possible effect of these features on the catchment flow is checked. Such an analysis is not straightforward, as there are at least two challenges. First, the appropriate aggregate measure of the river flow variable has to be chosen. Three candidates are taken into account: mean value as the classical statistical shift estimator, median representing robust shift estimator and the maximum value. In order to account for the direct effect of catchment size on river flows, selected flow indices (mean, median and max) were normalized, i.e. divided by the respective catchment area. They are referred to as "specific" flows, e.g. a mean specific flow, etc., and their unit is  $m^3 \cdot s^{-1} \cdot km^{-2}$ .

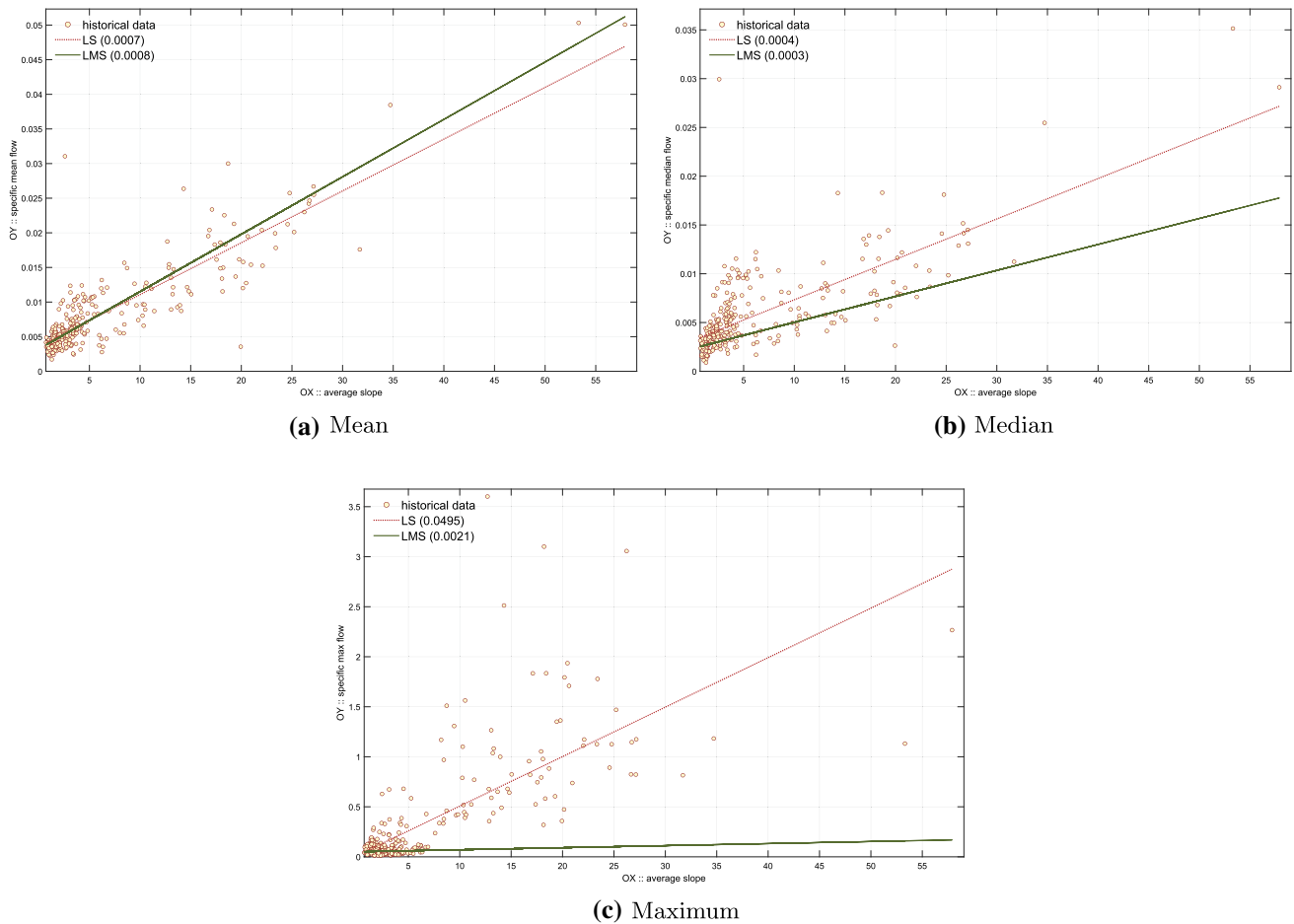
Second issue is associated with the regression analysis itself. First of all, the number of observations is very low, which may bias any obtained estimation. Moreover, the underlying statistical process behind data cannot be assumed neither Gaussian, nor stationary. Thus, apart from classical mean shift estimator, the robust version is used (Rousseeuw 1984). Classical regression uses least squares (1) as the performance index, while the robust one – the LMS (2). The example of this effect is shown in sample Fig. 1.

$$LS = \text{mean}_k (\epsilon_k)^2, \quad (1)$$

$$LMS = \text{median}_k (\epsilon_k)^2, \epsilon_k \text{ -- sample regression error.} \quad (2)$$

The estimators differ. LS index takes into account even outlying observations, while the robust one concentrates of data inliers (majority). The decision, which one should be taken into consideration as the resulting index is not straightforward.

As expected, the effect of catchment area on selected specific flow indices evaluated by robust regression is close to 0 due to standardization process (Table 1). Two categories of catchment properties stand out from others:



**Fig. 1** Example regression plots illustrating the effect of average slope on the mean, median and maximum specific flows

topography and climate. Catchment slope has the highest correlation with mean specific slope across all catchment properties, followed by mean annual precipitation and average elevation. The higher the slope, elevation and precipitation, the higher the mean, median and maximum specific flows. In contrast, average (min and max) temperature have negative correlation with mean and median specific flow, which can be explained by two facts: that higher elevation is associated not only with higher precipitation but also with lower temperature, and that lower temperature results in lower evapotranspiration. Other categories of catchment properties such as land cover, soils and hydrography exhibit less pronounced relationship with specific flow, particularly when assessed with robust regression. Linear regression shows that fraction of arable land has negative correlation with all specific flows, whereas fraction of forests, urban land and grassland has positive correlation. Catchments having a high fraction of impermeable soils have positive correlation with specific flow due to higher surface runoff. In general, the robust regression has significantly different results for maximum

specific flows, perhaps because the outliers are most abundant there.

The values of each property are grouped into clusters. The number of clusters is set to 4 (as the results of the selection of quantiles:  $Q_1$ ,  $Q_2$  and  $Q_3$ ). The division is realized in two ways: using hydrological expert knowledge and using statistical quantiles ( $Q_1 = 25\%$ ,  $Q_2 = 50\%$  and  $Q_3 = 75\%$ ) for respective feature. Data are presented in Table 2. These properties are compared with the best fitted PDFs and L-moments properties.

### 3 Methods and algorithms

All the applied methods are specified and introduced in this section. Statistical approach covers used distributions, Q-Q plots, L-moments and moment ratio diagrams. Data mining approach is represented by the k-means clustering algorithm.

**Table 1** Catchment features impact on specific mean, median and maximum flows (normalized, i.e. divided by the respective catchment area): analysis using linear and robust regression

Feature	Linear regression			Robust regression		
	Mean	Median	Max	Mean	Median	Max
Catchment area	– 10.4	– 8.5	– 11.2	0.6	0.3	– 0.5
Mean annual precipitation	28.4	23.1	26.7	24.5	14.8	1.1
Average max temperature	– 33.8	– 30.1	– 19.5	– 7.9	– 9.6	1.3
Average min temperature	– 32.1	– 24.8	– 26.0	– 9.1	– 16.8	– 0.6
Average elevation	33.1	25.5	31.1	14.3	21.7	6.3
Average slope	37.6	29.3	35.3	41.6	18.8	1.5
Arable land	– 21.7	– 19.1	– 18.2	– 3.0	– 9.2	0.5
Urban	11.7	12.8	14.6	– 1.5	– 6.5	2.4
Grassland	17.8	13.5	13.2	1.3	2.3	1.6
Forests	15.1	13.0	13.0	2.9	8.3	– 0.7
Soil group A	– 8.9	– 5.6	– 11.1	0.4	– 0.6	– 0.6
Soil group B	– 10.9	– 7.5	– 11.7	– 1.2	– 1.9	0.3
Soil group C	13.5	9.1	15.0	– 0.9	1.4	0.1
Soil group D	9.9	6.0	16.2	3.2	3.9	– 4.2
Fraction of lakes	– 3.7	1.9	– 12.2	1.2	2.1	– 1.3
Permeability: very low	9.4	5.3	16.9	– 0.8	3.0	2.9
Permeability: low	– 10.0	– 10.3	– 5.2	– 0.5	– 2.6	1.0
Permeability: average	– 9.4	– 5.6	– 14.2	– 0.3	– 3.5	– 0.1
Permeability: variable	– 10.2	– 6.6	– 16.6	– 1.8	– 1.4	– 1.0
Permeability: high + very high	– 11.9	– 8.4	– 15.1	– 0.4	– 0.9	1.0

Relation is presented in form of the linear regression slope angle – higher angle absolute value means stronger impact

### 3.1 Distributions used for the extreme flow analysis

Analysis of the regional river flow data starts with the review of properties for each flow gauge. Hydrographs are plotted together with the histograms. Considered discharge time series are characterized by the periodicity and frequent extreme high value incidents.

Histograms for such time series are highly asymmetric. One may find various probabilistic distribution functions appropriate in such a case. The idea of this research was to minimize unnecessary assumptions. Therefore the analysis takes into account seven different distributions being frequently used in the considered research context: Gamma (GAM), lognormal (LGN), exponential (EXP), Weibull (WEI), Generalized Extreme Value (GEV), generalized Pareto (GPD) and four-parameter Kappa (K4D). Respective probabilistic density functions and their parameters are shown in Table 3.

Each PDF should be fitted to the empirical data. The fitting is achieved using maximum likelihood estimation except the four parameter Kappa distribution, for which the L-moments approach is applied. Sample plot showing fitting, i.e. data histogram and fitted PDFs is sketched in Fig. 2.

### 3.2 Q-Q plots and the best fit PDF estimation

In general quantile-quantile (so called Q-Q) plots enable to compare different distributions by plotting their quantiles against each other. A point on the diagram corresponds to a certain quantile coming from an empirical and theoretical distribution. They may be used as the heuristics for testing "goodness-of-fit" between some empirical data and a theoretical PDF. It allows to check the assumption the empirical data are derived by the chosen distribution. The fitting is relatively simple. We need to find such a theoretical curve that appears to be the closest to the empirical points. The distance between both curves determines the "goodness-of-fit" index.

Formal PDF fitting is done using the maximum likelihood (MLE) approach. It gives reliable results and is simple. It is frequently used for large datasets (as in our case) (Coles and Dixon 1999). The Q-Q plot is used to select the best fit function. The fitting measure is robust median absolute distance between empirical and theoretical data. It plays the role of the "goodness of fit" index supporting a decision about the best fitting distribution. Figure 3 shows a sample Q-Q plot for sample empirical data versus seven theoretical distributions.

**Table 2** Clustering regions for selected catchment features. Quantiles Q1, Q2 and Q3 refer to each catchment feature in the respective row

Feature	Unit	Expert clustering			Quantiles		
					Q1	Q2	Q3
Catchment area	[km <sup>2</sup> ]	100	1000	10000	260.5	517.7	1203.8
Mean annual precipitation	[mm]	650	800	1000	660.7	713.4	843.2
Average max temperature	[°C]	11	12	13	11.3	12.0	12.6
Average min temperature	[°C]	2	3	4	3.1	3.5	3.9
Average elevation	[m]	200	400	600	132.1	185.6	346.7
Average slope	[%]	3	10	20	2.0	3.3	7.6
Arable land	[%]	20	40	60	25.0	39.1	54.0
Urban	[%]	3	8	15	2.8	4.8	7.4
Grassland	[%]	5	10	15	5.8	8.9	13.4
Forests	[%]	20	40	60	26.2	37.5	52.2
Soil group A	[%]	5	20	40	1.1	21.6	42.2
Soil group B	[%]	20	40	60	22.4	39.3	54.0
Soil group C	[%]	10	25	50	9.8	21.5	47.7
Soil group D	[%]	5	10	15	0.3	2.1	6.4
Fraction of lakes	[%]	1	2	4	0.03	0.3	1.9
Permeability: very low	[%]	5	15	30	0.2	2.0	15.3
Permeability: low	[%]	5	15	30	5.6	13.0	25.3
Permeability: average	[%]	5	15	30	0.8	10.1	19.0
Permeability: variable	[%]	5	15	30	0.3	5.2	8.2
permeability: high + very high	[%]	5	15	30	0.2	13.3	21.1

### 3.3 L-moments

L-moments have been introduced as linear combinations of order statistics (Hosking 1990). This approach significantly improves conventional methodology and definitions. It gives new characterization of the shape of a probability density function and allows to estimate the distribution parameters. Unlike product moments, L-moments deliver almost unbiased L-moments statistics, even for very small samples. Additionally they are less sensitive to the distribution's tails (Peel et al. 2001). Above properties suite them almost perfectly and allow to describe environmental data commonly characterized by a moderate or high skewness.

Evaluation of L-moments is realized through the following procedure. At first, the data  $\{x(1), \dots, x(N)\}$ ,  $N$  - number of samples, ranked in ascending order from 1 to  $N$ . The sample L-moments  $(l_1, \dots, l_4)$ , the sample L-skewness  $\tau_3$  and L-kurtosis  $\tau_4$  are evaluated as

$$l_1 = \beta_0, \quad (3)$$

$$l_2 = 2\beta_1 - \beta_0, \quad (4)$$

$$l_3 = 6\beta_2 - 6\beta_1 + \beta_0, \quad (5)$$

$$l_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0, \quad (6)$$

$$\tau_2 = \frac{l_2}{l_1}, \quad (7)$$

$$\tau_3 = \frac{l_3}{l_2}, \quad (8)$$

$$\tau_4 = \frac{l_4}{l_2}, \quad (9)$$

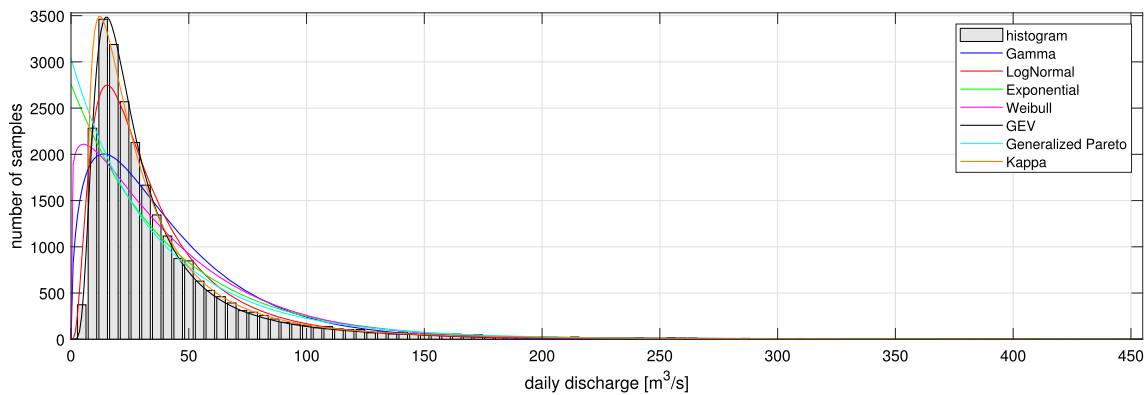
where

$$\beta_j = \frac{1}{N} \sum_{i=j+1}^N x_i \frac{(i-1)(i-2) \cdots (i-j)}{(N-1)(N-2) \cdots (N-j)} \quad (10)$$

Therefore, the data are summarized and described by the sample L-location (L-shift)  $l_1$ , L-scale  $l_2$ , L-Cv (dimensionless measure of variability)  $\tau_2$ , L-skewness  $\tau_3 \in (-1, 1)$  and L-kurtosis  $\tau_4 \in (-\frac{1}{4}, 1)$ . L-moments might be used to fit a distribution to a dataset. It is done through fitting of the sample empirical L-moments to the exact theoretical L-moments of the distribution. L-moments  $\tau_3$  and  $\tau_4$  may play the role of the goodness-of-fit measure. L-moments can be analytically calculated for different distributions. Theoretical relationships for univariate PDFs can be found in Hosking (1990), Hosking (1992), Kjeldsen et al. (2017).

**Table 3** PDF functions used in the analysis

Name	Probabilistic density function	Parameters
GAM	$\mathcal{F}_{\beta,\xi}(x) = \frac{x^{\xi-1} \exp(-\frac{x}{\beta})}{\beta^{\xi} \Gamma(\xi)}$	$\beta > 0, \xi > 0, \Gamma(\cdot)$ Gamma function
LGN	$\mathcal{F}_{\mu,\sigma}(x) = \frac{\exp(-\frac{(\log x - \mu)^2}{2\sigma^2})}{x\sigma\sqrt{2\pi}}, x > 0$	$\mu \in \mathbb{R}, \sigma > 0$
EXP	$\mathcal{F}_{\mu}(x) = \frac{\exp(-\frac{x}{\mu})}{\mu}$	$\mu \in \mathbb{R}$
WEI	$\mathcal{F}_{\beta,\xi}(x) = \frac{\xi}{\beta} \left(\frac{x}{\beta}\right)^{\xi-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\xi}\right)$	$\beta > 0, \xi > 0$
GEV	$\mathcal{F}_{\mu,\beta,\xi} = \frac{1}{\beta} t(x)^{\xi+1} \exp(-t(x))$ $t(x) = \begin{cases} \left(1 + \left(\frac{x-\mu}{\beta}\right)^{\xi}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ \exp\left(-\frac{x-\mu}{\beta}\right) & \text{if } \xi = 0 \end{cases}$	$\mu \in \mathbb{R}, \beta > 0, \xi \in \mathbb{R}$
GPD	$\mathcal{F}_{\mu,\beta,\xi} = \frac{1}{\beta} \exp(-(1-\xi)t(x))$ $t(x) = \begin{cases} -\xi^{-1} \log\left\{1 - \frac{\xi(x-\mu)}{\beta}\right\} & \text{if } \xi \neq 0 \\ \frac{x-\mu}{\beta} & \text{if } \xi = 0 \end{cases}$	$\mu \in \mathbb{R}, \beta > 0, \xi \in \mathbb{R}$
K4D	$\mathcal{F}_{\mu,\beta,k,h}(x) = \frac{1}{\beta} \left[1 - \frac{k}{\beta}(x-\mu)\right]^{1/k-1} \cdot \left\{1 - h\left[1 - \frac{k}{\beta}(x-\mu)\right]^{1/k}\right\}^{1/h-1}$	$\mu \in \mathbb{R}, \beta > 0, h > 0, k > 0$



**Fig. 2** Sample PDF fitting performance for seven theoretical distributions versus empirical data histogram for Chalupki measurement site data

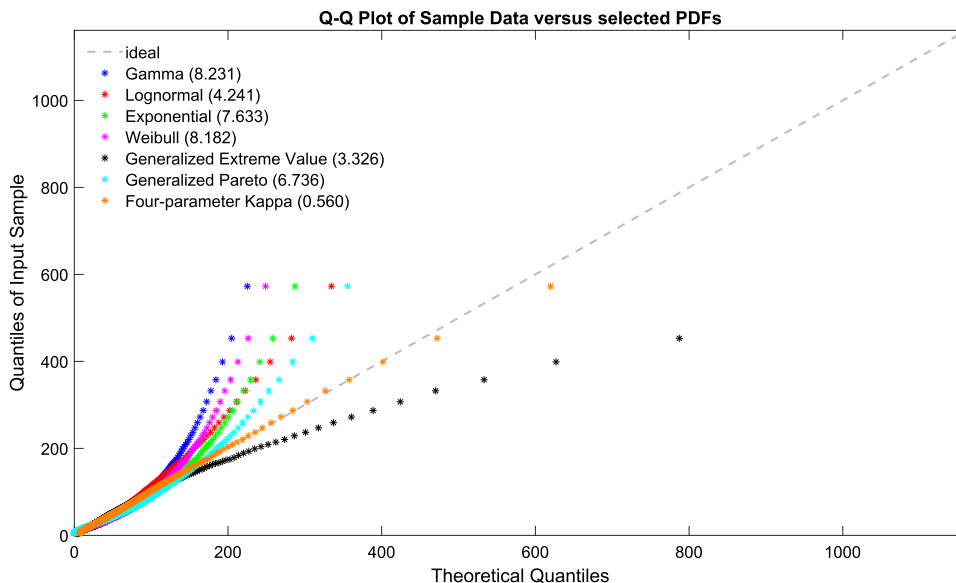
**3.4 Moment ratio diagrams**

MRDs have been used in practical applications to answer questions about sample’s statistical properties, such as theoretical distribution fitting, comparison of the distribution’s shapes or PDFs classification according to certain categories (Bobeje et al. 1993). The MRD is a graphical representation in a Cartesian coordinates of a pair of standardized moments. In general, there are two types of moment ratio diagrams: a graph of skewness versus kurtosis and a graph of a scale factor versus skewness for

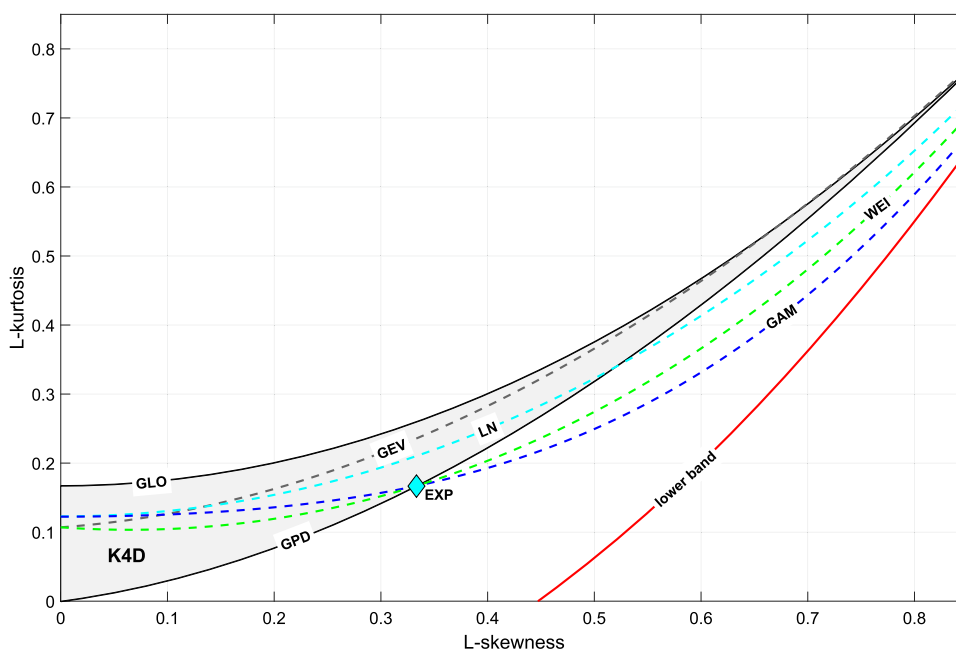
common univariate probability distributions (Vargo et al. 2010). The literature shows that the first one has gained significant popularity and is adopted in current research.

Introduction of L-moments has been naturally used by the MRDs and the L-moment diagrams have been extensively analyzed. They form the graphical representation of a relationship between L-kurtosis and L-skewness. Figure 4 shows theoretical relationship for selected distributions. One may see that exponential distribution is represented as a single point, four parameters Kappa as a region, while the rest of them by polynomial curves. As

**Fig. 3** Sample Q-Q plot showing empirical data versus seven theoretical distributions for Chałupki measurement site data: four-parameter Kappa (K4D) exhibits the best fitting



**Fig. 4** L-moment diagram for theoretical distributions, GLO is a Generalized Logistic. Red line depicts the lower band – a limit of all distributions. Grey color covers K4D area



GPD, GEV and Generalized Logistic (GLO) distributions are the special cases of the K4D, they belong to its region; GLO as the upper band and GPD the lower. Polynomials used to approximate theoretical relationships can be found in Tallaksen and Van Lanen (2004), Kjeldsen et al. (2017).

Position of any point on the L-moment diagram informs about statistical properties of respective time series data. Left-bottom part of the L-moment diagram is occupied by datasets with low L-skewness  $\tau_3$  and L-kurtosis  $\tau_4$ . L-skewness closer to zero means that the PDF is more symmetrical. Simultaneously, lower L-kurtosis reflects lighter tails of the respective distribution. In contrary, top-right diagram area depicts data, which are highly skewed

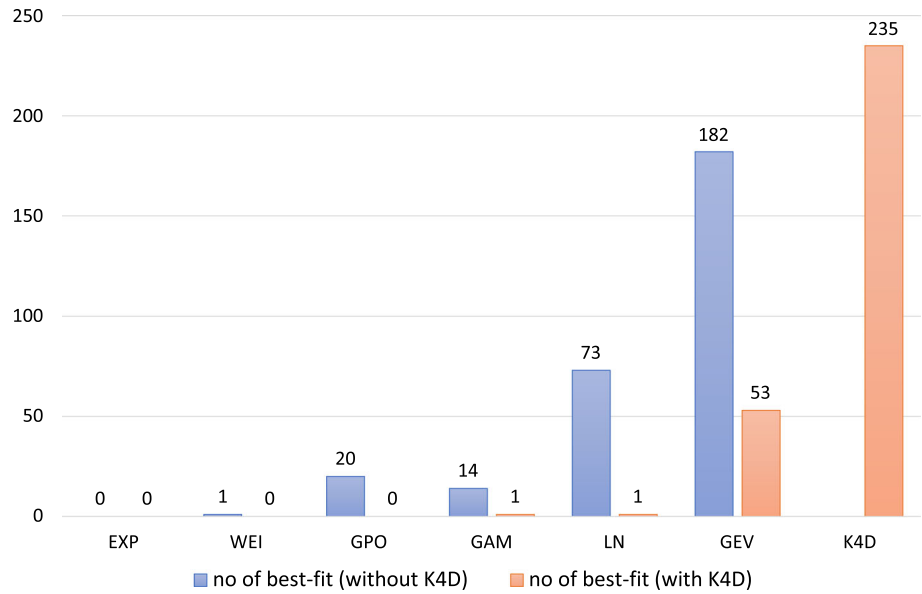
with significantly fat tails. Therefore, positioning on the L-moment diagram allows fast data interpretation and classification.

### 3.5 Clustering

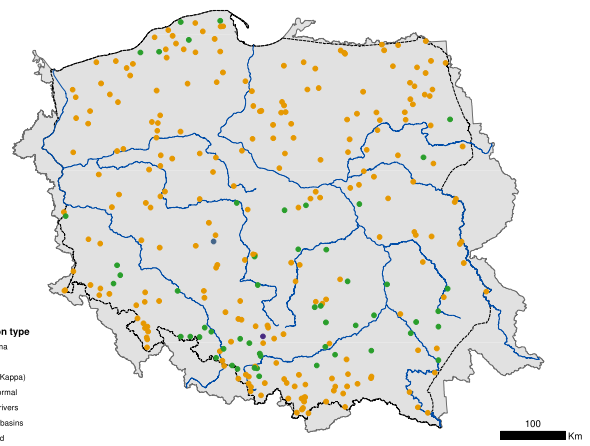
Data clustering algorithms belong to the machine learning area of science and are used to group data that lie close to each other and may be classified as more similar to each other in some sense, in contrast to other ones. The literature shows dozens of clustering articles. The one used herein is a k-means algorithm. It is one of the simplest unsupervised approaches as it makes inferences from data using only



**Fig. 5** PDF fitting analysis using Q-Q plots



**(a)** Analysis without K4D



**(b)** Analysis with K4D

**Fig. 6** Maps showing flow gauges with the best fit PDFs

input vectors without referring labeled results (Jin and Han 2010). The algorithm works according to the following steps:

1. Specify the number of clusters  $M$ .
2. Initialize centroids randomly selecting  $M$  points for the centroids without replacement.
3. Iterate until the assignment of points to clusters does not change.
4. Compute the performance index being the sum of the squared distance between points and centroids.
5. Assign each point to the closest data cluster.
6. Evaluate new centroids taking the average of all data belonging to each cluster.

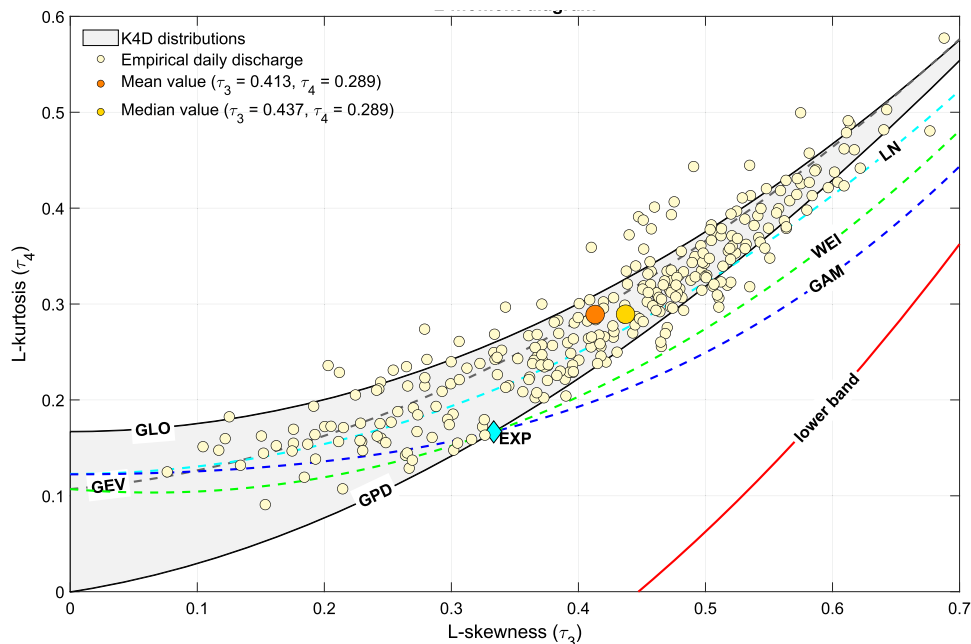
## 4 Regional clustering results

The results of the regional clustering are presented in the below paragraphs, starting from the PDF goodness-of-fit analysis, followed by the clustering in the domain of L-moments diagrams.

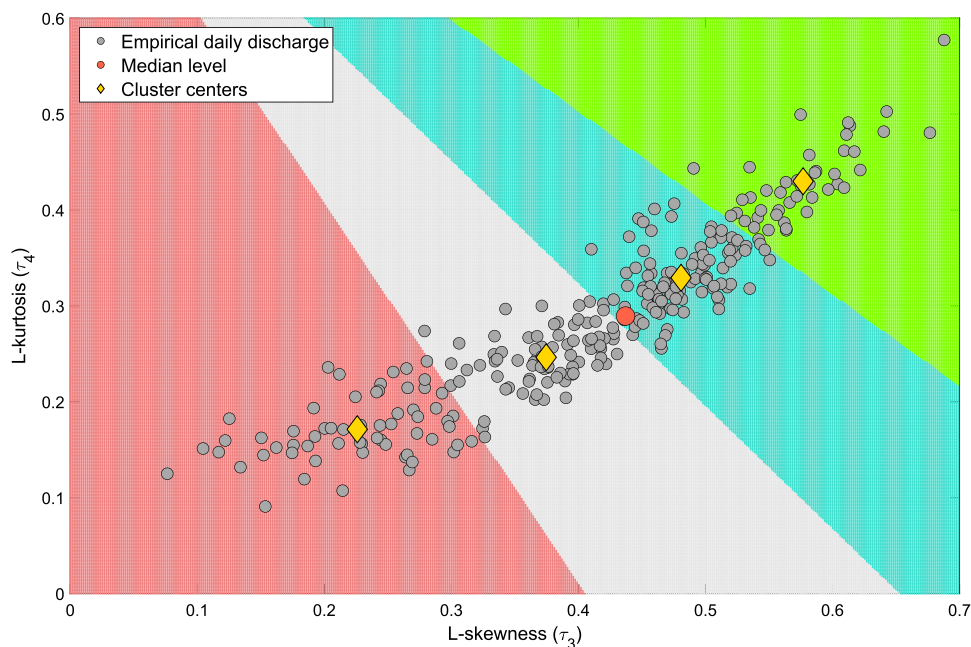
### 4.1 PDF goodness-of-fit analysis

The analysis is grouped into formal steps. At first the goodness-of-fit assessment is done to assign the most appropriate distributions to each flow gauge. In contrast to the majority of reports, the Q-Q diagram approach is used. It allows better rationalization of the choice between four-parameter Kappa distribution and other PDFs, of which

**Fig. 7** L-moment diagram showing empirical data and theoretical curves



**Fig. 8** L-moment diagram showing ML-clustering to four regions: classes denoted as (red)—ML1, (grey)—ML2, (blue)—ML3, (green)—ML4



some belong to the same family (GEV, GPD, EXP). The results are presented in a graphical form. Figure 5 shows the bar plot with the PDF fitting results. Two versions are presented. In the first one all the distributions are used. As it is seen the overwhelming majority of gauges is well described by four-parameter Kappa (81%) and GEV (18%). Once K4D is excluded from the analysis, GEV distribution covers the majority of sites, however LGN, GAM and GPD PDFs still match quite a significant number of points.

Nevertheless, the concluding remark is clear: four-parameter Kappa distribution should be considered as the best

fitting probability density function in case of Polish daily river flow data.

### 4.2 L-moment diagram analysis

At first, L-moments are calculated for each flow time series. Next, they are plotted on the L-diagram (see Fig. 7). Additionally, empirical mean and median L-moments are evaluated and respective points added to the diagram. The plot confirms the fact that K4D distributions covers the

majority of Polish flow gauges, as the points mostly lie within the area belonging to four-parameter Kappa region.

### 4.3 k-means clustering

Empirical points of the L-moments diagram can be treated as any two-dimensional data on a plane and as such can be subject to clustering. Figure 8 shows how the k-means algorithm allocates data to the four regions. The number of regions is equal to the number of classes defined for each catchment property as in Table 2. Expert classification is done into four classes. Similarly straightforward approach using basic quantiles Q1, Q2 and Q3 also gives classification into 4 groups.

The clustering plot depicts evaluated regions' clusters. As one can notice, the median of empirical points lies very close to the border between two middle clusters, which is as expected. For clarity, these clusters are called ML-clusters (machine learning clusters). Machine learning clustering is further compared with expert based classification depending on catchment area categorization.

### 4.4 Effects of catchment properties

At that point all important elements are prepared and catchment properties analysis can be performed. Each property listed in Table 2 will be analyzed separately according to the same proposed methodology.

- (1) Two L-moment plots are prepared for both types of feature cluster definitions, i.e. expert- and quantile-based. The points are classified according to the respective cluster membership. Expert clustering is called EX-clusters, while quantile-based QU-clusters.
- (2) Median point for each cluster is evaluated.

- (3) Binary membership matching between each feature class point belonging either to EX-class or the QU-class, and selected ML-cluster is performed. According to that the number of each EX- or QU-cluster point and ML-clusters is evaluated showing how the feature data fit into clusters.

#### 4.4.1 Catchment area

The analysis starts with the catchment area. Respective L-moment diagrams are shown in Fig. 9. Medians for each cluster are fitting some pattern, i.e. flow gauges with the largest catchment area tend to group themselves in the left-bottom corner of the diagram, while those with small catchment area are grouped in the right-top part of the diagram. The meaning of these regions is as follows. Left-bottom part of the L-moment diagram depicts observations with low L-skewness  $\tau_3$  and low L-kurtosis  $\tau_4$ . L-skewness is closer to zero, which reflects the PDF to be near symmetrical. Lower L-kurtosis values depict lighter tails of the distribution.

Since L-kurtosis for normal distribution (supposed to have no fat-tails) equals to  $\tau_4 = 0.1226$ , observed C4 category median  $\tau_4 = 0.2357$  means that catchment data belonging to C4 are less skewed and do not exhibit too fat tails.

Therefore, the plot brings forward the following observation. Large catchment area means more symmetrical and less tailed distribution. This fact can be simply explained by the large water accumulation and catchment area inertia. In contrast, sites with low catchment area can respond to the rainfall events more abruptly, which results in asymmetric PDF, fatter and longer tails, and finally in more persistent behavior.

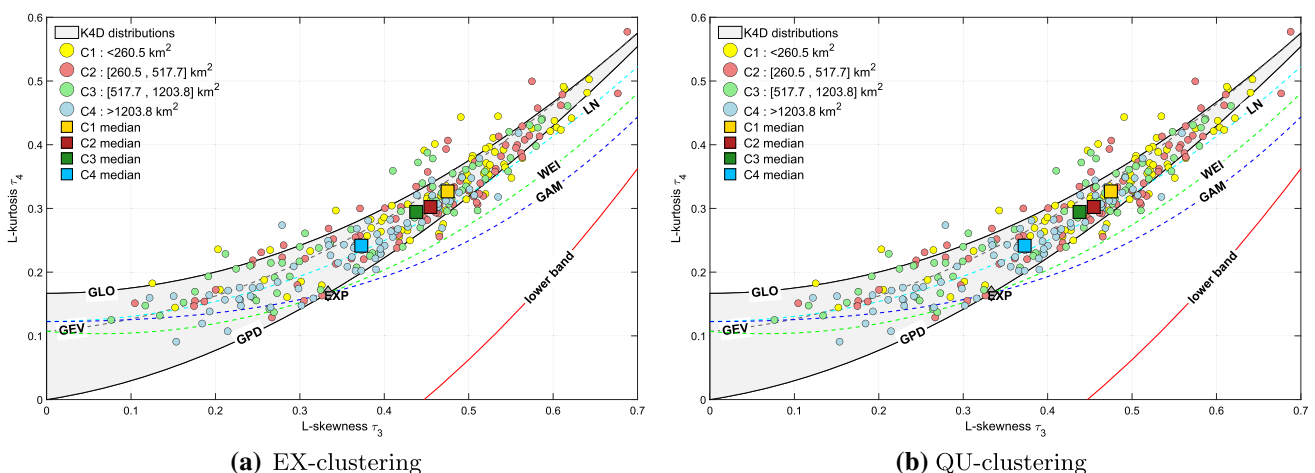
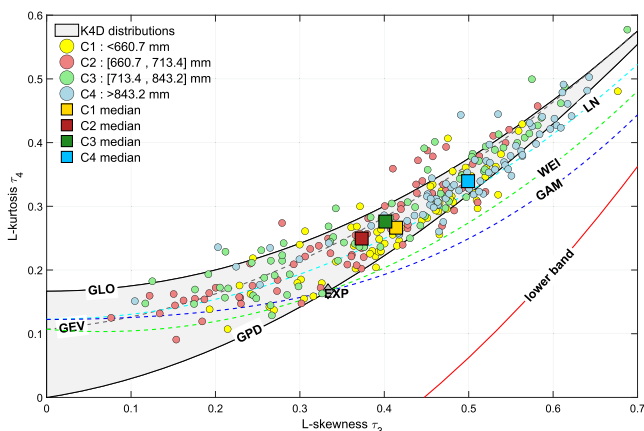


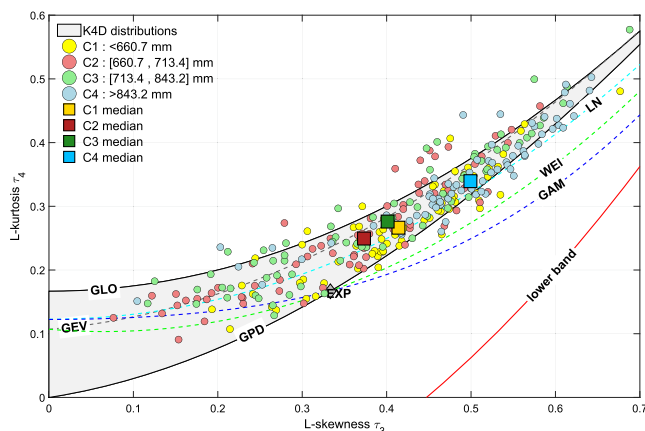
Fig. 9 L-moments diagrams with clustering by catchment area

**Table 4** Cluster matching for catchment area: italic—two largest values, bold—the largest one

	ML-cluster					ML-cluster					
	ML1	ML2	ML3	ML4		ML1	ML2	ML3	ML4		
<b>EX-class</b>	C1	2	4	<b>13</b>	6	<b>QU-class</b>	C1	8	14	<b>31</b>	19
	C2	34	39	<b>69</b>	38		C2	13	17	<b>26</b>	17
	C3	23	<b>32</b>	22	2		C3	18	16	<b>30</b>	8
	C4	2	<b>4</b>	0	0		C4	22	<b>32</b>	18	1



(a) EX-clustering



(b) QU-clustering

**Fig. 10** L-moments diagrams with clustering by mean annual precipitation

**Table 5** Cluster matching for mean annual precipitation: italic—two largest values, bold—the largest one

	ML-cluster					ML-cluster					
	ML1	ML2	ML3	ML4		ML1	ML2	ML3	ML4		
<b>EX-class</b>	C1	10	<b>22</b>	21	3	<b>QU-class</b>	C1	11	<b>29</b>	28	4
	C2	35	<b>48</b>	45	14		C2	21	<b>27</b>	21	4
	C3	16	7	<b>17</b>	<b>17</b>		C3	<b>24</b>	16	19	13
	C4	0	2	<b>21</b>	12		C4	5	8	<b>36</b>	24

Finally, the ML-clusters matching analysis is done. Table 4 presents, how each catchment area category fits into clusters obtained using machine learning (ML-clusters). It is visible that the largest values, for both classifications (EX and QU) are positioned on the reverse diagonal, i.e. C1 matches ML4 and C4 ML1. Though this matching is not exact it confirms observations in a qualitative way.

An additional observation is that expert matching is not even, as for instance EX-class C1 has only 6 observations. That fact may also affect perception of obtained results.

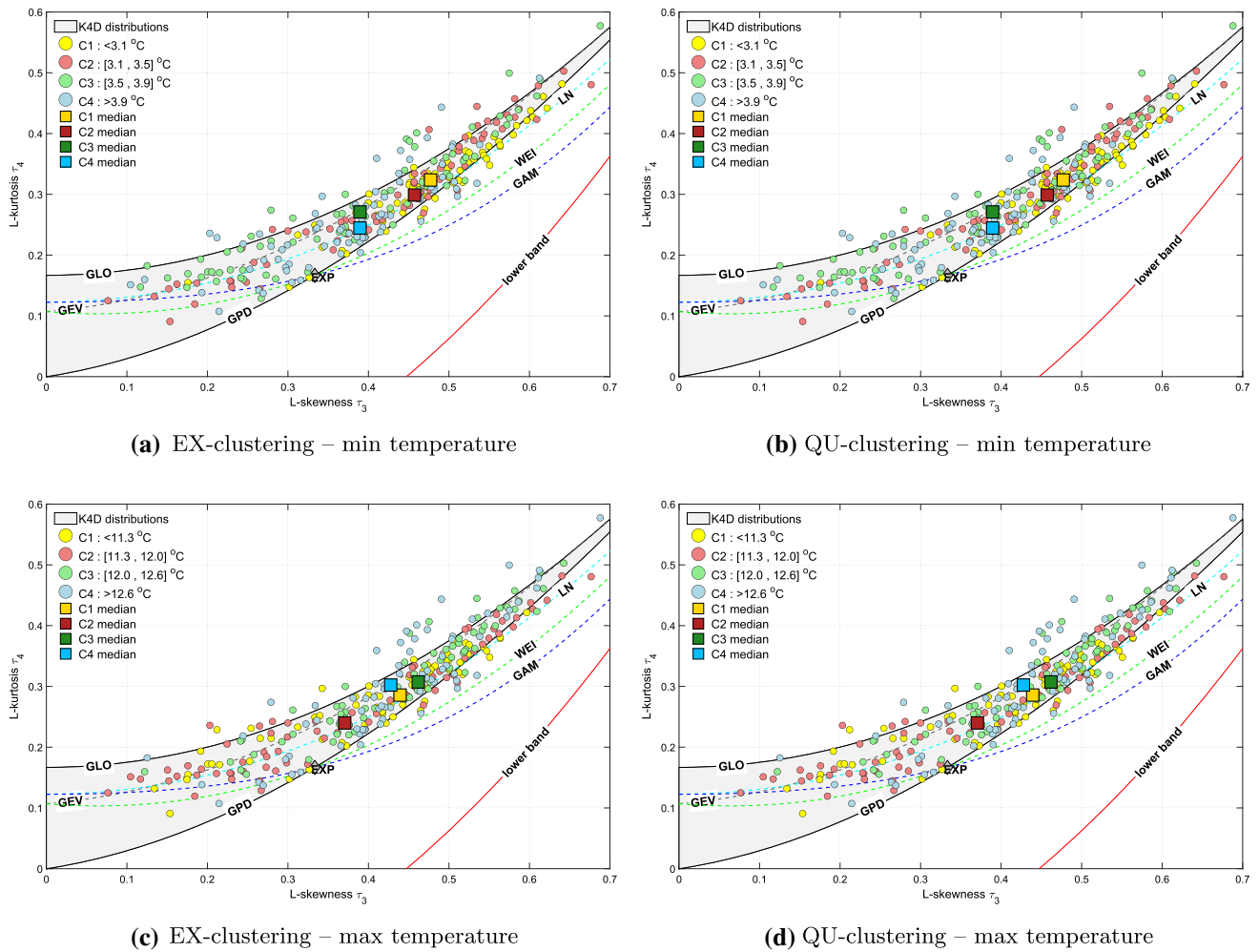
**4.4.2 Mean annual precipitation**

The analysis continues with mean annual precipitation. Respective LMRDs are shown in Fig.10. Median values for each cluster are fitting opposite pattern to the catchment area. Flow gauges with the highest mean annual

precipitation tend to group themselves in the top-right corner of the diagram, while smaller ones in the opposite side. Obtained relationship is not as explicit as in the previous case.

Large values of mean annual precipitation may cause larger flood events, skewing the distribution and generating the tail. Low precipitation has much lower probability for such extreme events.

Finally, the ML-clusters matching analysis is done. Table 5 presents the results. It is visible that the largest values, for both classifications (EX and QU) are on the diagonal, however not as clearly as in previous case. Probably getting better matching requires more exact realization of classification and/or clustering.



**Fig. 11** L-moments diagrams with clustering by maximum and minimum temperatures

**Table 6** Cluster matching for minimum and maximum temperatures: italic—two largest values, bold—the largest one

	ML-cluster					ML-cluster					
	ML1	ML2	ML3	ML4		ML1	ML2	ML3	ML4		
Average minimum temperature											
<b>EX-class</b>	C1	5	<i>15</i>	<b>28</b>	3	<b>QU-class</b>	C1	15	<i>21</i>	<b>33</b>	3
	C2	<b>38</b>	20	20	16		C2	<b>31</b>	12	<i>16</i>	14
	C3	15	<i>35</i>	<b>46</b>	24		C3	8	<i>20</i>	<b>29</b>	15
	C4	3	9	<b>10</b>	3		C4	10	25	<b>26</b>	12
Average maximum temperature											
<b>EX-class</b>	C1	0	1	<b>16</b>	5	<b>QU-class</b>	C1	2	<i>16</i>	<b>38</b>	<i>16</i>
	C2	2	<i>14</i>	<b>21</b>	10		C2	16	15	<b>25</b>	<i>17</i>
	C3	43	36	<b>49</b>	28		C3	<b>23</b>	17	22	10
	C4	16	<b>28</b>	<i>18</i>	3		C4	20	<b>31</b>	19	3

**4.4.3 Average maximum and minimum temperatures**

In the next step clustering by catchment-averaged annual means of daily minimum and maximum temperatures is investigated. Figure 11 shows respective diagrams for

minimum and maximum temperatures and both classifications. The relationships differs between both temperatures. Average minimum values follow visible pattern, as the lower minimum temperatures are more persistent. They exhibit fat tails and have more skewed distributions.

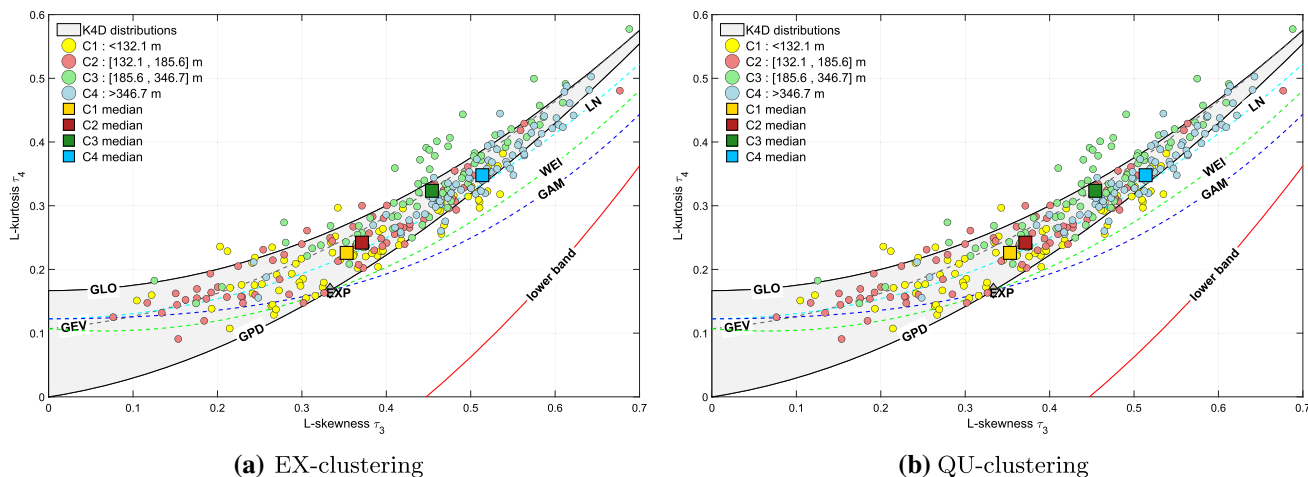


Fig. 12 L-moments diagrams with clustering by elevation

Table 7 Cluster matching for catchment elevation: italic—two largest values, bold—the largest one

	ML-cluster					ML-cluster					
	ML1	ML2	ML3	ML4		ML1	ML2	ML3	ML4		
<b>EX-class</b>	C1	56	<b>60</b>	37	4	<b>QU-class</b>	C1	<b>28</b>	25	18	1
	C2	5	14	<b>32</b>	22		C2	26	<b>29</b>	15	3
	C3	0	1	<b>15</b>	<b>15</b>		C3	4	19	<b>35</b>	14
	C4	0	4	<b>20</b>	5		C4	3	6	<b>36</b>	28

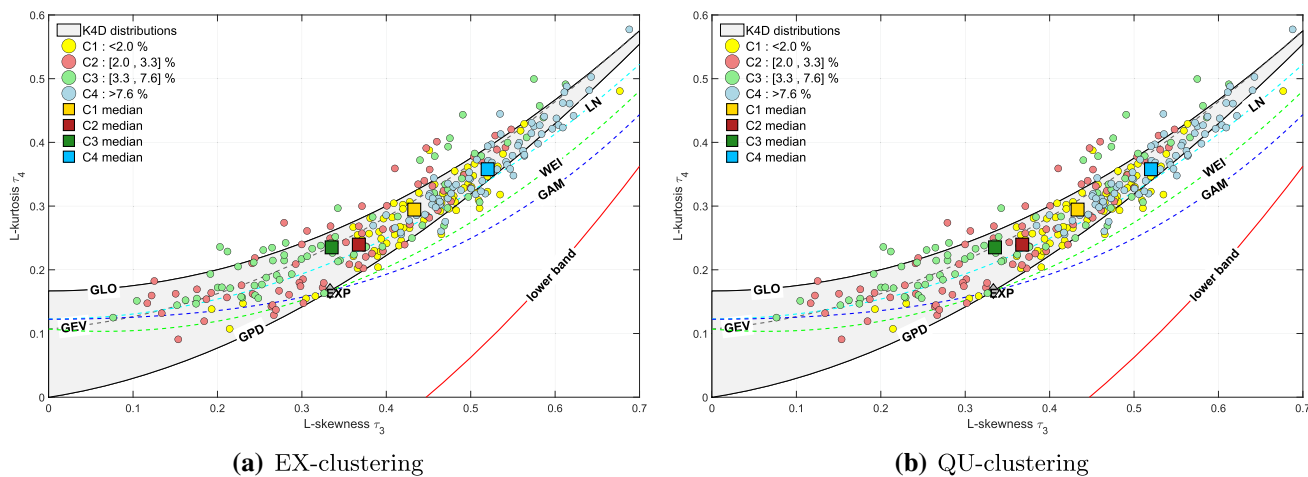


Fig. 13 L-moments diagrams with clustering by average slope

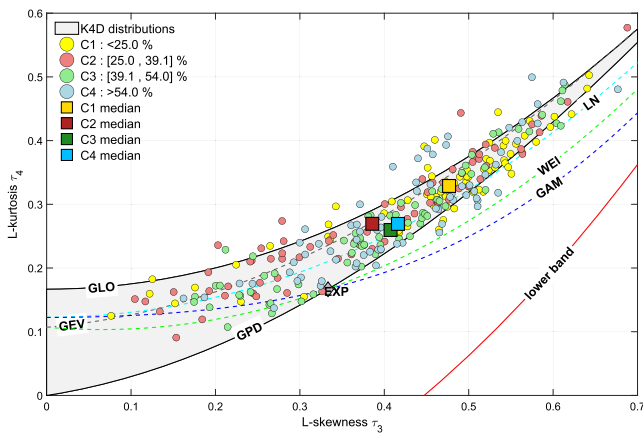
Average maximum temperatures are not so clear in interpretation. However, an interesting observation can be made. All classes are grouped close to each other, except (11, 12)°C and in that region flow gauges exhibit properties closer to Gaussian, than for other temperature ranges. There is no obvious explanation to this fact, since relationship between temperature and flow regime is not straightforward but indirect. It might be that certain

temperature ranges, such as (11, 12)°C, are observed in some regions and other features of those regions directly affect flow levels. Definitely, this observation requires further attention.

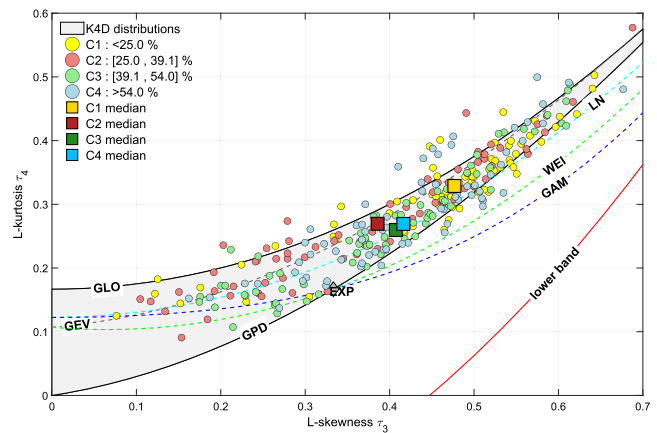
Table 6 presents the results for cluster matching. As it was already shown no pattern is visible as well for the maximum temperature, while minimum temperature slightly depicts diagonal pattern.

**Table 8** Cluster matching for average slope: italic—two largest values, bold—the largest one

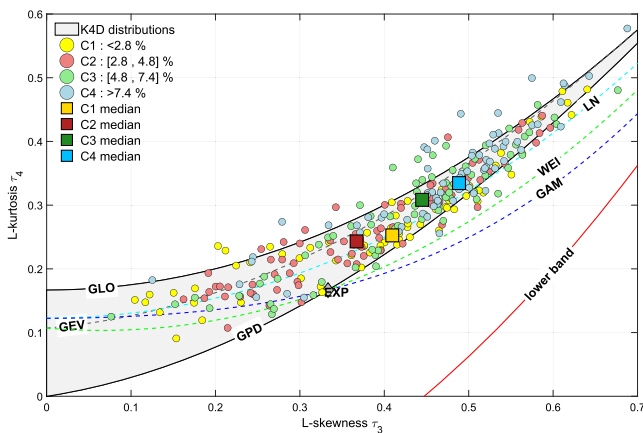
	ML-cluster					ML-cluster					
	ML1	ML2	ML3	ML4		ML1	ML2	ML3	ML4		
<b>EX-class</b>	C1	24	<b>49</b>	<b>49</b>	6	<b>QU-class</b>	C1	5	29	<b>34</b>	4
	C2	<b>37</b>	25	22	16		C2	<b>26</b>	25	18	4
	C3	0	3	<b>19</b>	<b>19</b>		C3	<b>30</b>	<i>19</i>	16	7
	C4	0	2	<b>14</b>	5		C4	0	6	<b>36</b>	<i>31</i>



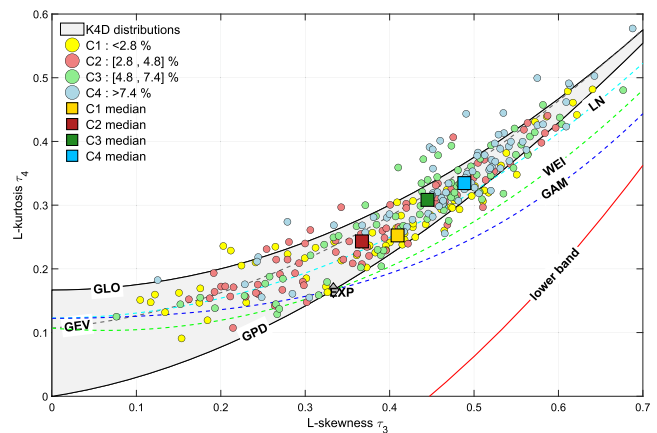
(a) EX-clustering – arable land



(b) QU-clustering – arable land



(c) EX-clustering – urban



(d) QU-clustering – urban

**Fig. 14** L-moments diagrams for land use fractions - part I

**4.4.4 Mean elevation**

Figure 12 shows obtained LMRDs for the mean catchment elevation, while Table 7 respective cluster matching properties. Obtained results are in line with expectations. The higher the elevation, the more persistent, skewed and tailed the distribution is. It is natural that catchments with higher mean catchment elevation exhibit more frequent and higher rainfall events which translates into more flashy flow regime.

Observations from the L-moment diagrams follow in a qualitative way the Table 7 clustering matching. Highlighted the most matching points lie within the diagonal of the table.

**4.4.5 Average slope**

Figure 13 shows obtained L-moment ratio diagram for the average slope. Observed results are less evident than for mean elevation. Probably catchment with low and medium slopes cannot be easily recognized on the diagrams,

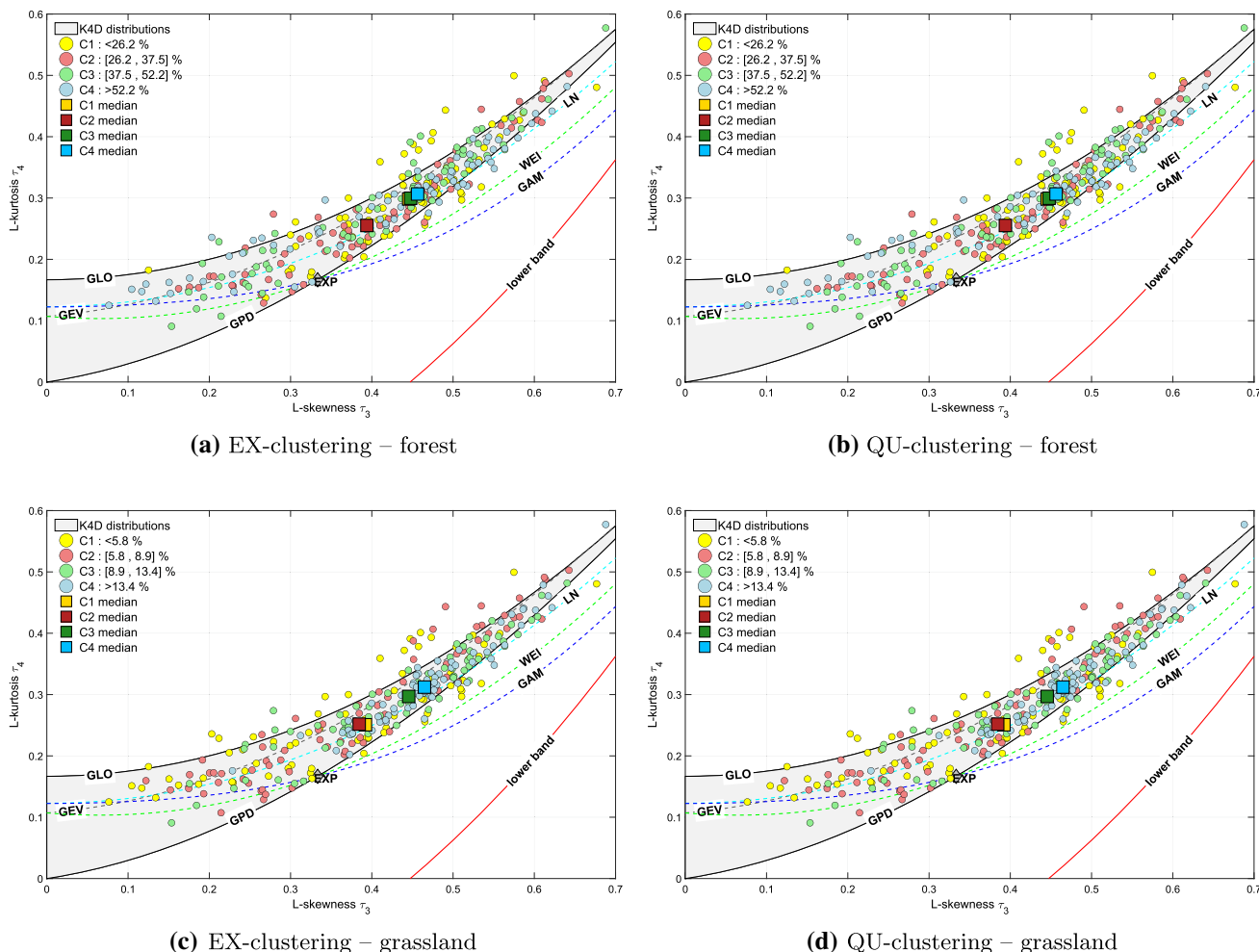


Fig. 15 L-moments diagrams for land use classes - part II

however high-slope catchments (typically in mountainous region) are clearly associated with distribution asymmetry and fat tails.

Similarly to the LMRDs, the matching table is unclear. The matching seems to be fine, as the pattern is similar to the one observed with L-moments diagrams.

4.4.6 Land use

The analysis of the dominant land use is more complex, as it involves categorical rather than quantitative data. The following classes are used: arable land, urban, grassland and forests. Therefore, the analysis has to be performed in a different way. First, the results are evaluated similarly to previous categories. Finally, the L-moments diagram is plotted showing different types of land use, not the categories within the class (Table 8).

Figure 14 shows L-moment diagrams for arable and urban land uses. Figure 15 presents grassland and forests.

Catchments are divided into four groups of different land use types in each case.

Observations confirm expectations and intuition. Low fraction of arable land use means more persistent behaviour. Possible explanation for that observation might be the result of the extensive drainage of agricultural land in Poland. An opposite situation occurs for urban land use fraction. Low urbanization allows more water retention and decreases data persistence, while higher fraction of urban area means more impervious surfaces and faster catchment response to rainfall. It is reflected by skewed distribution with fatter and longer tails.

Forests have clear impact. Firstly, all the categories are quite close to each other, so it is hard to distinguish statistical properties. L-moment diagram shows that the most natural (with the least probability of extreme incidents) is category C2 depicting the forest fraction of (20, 40)%.

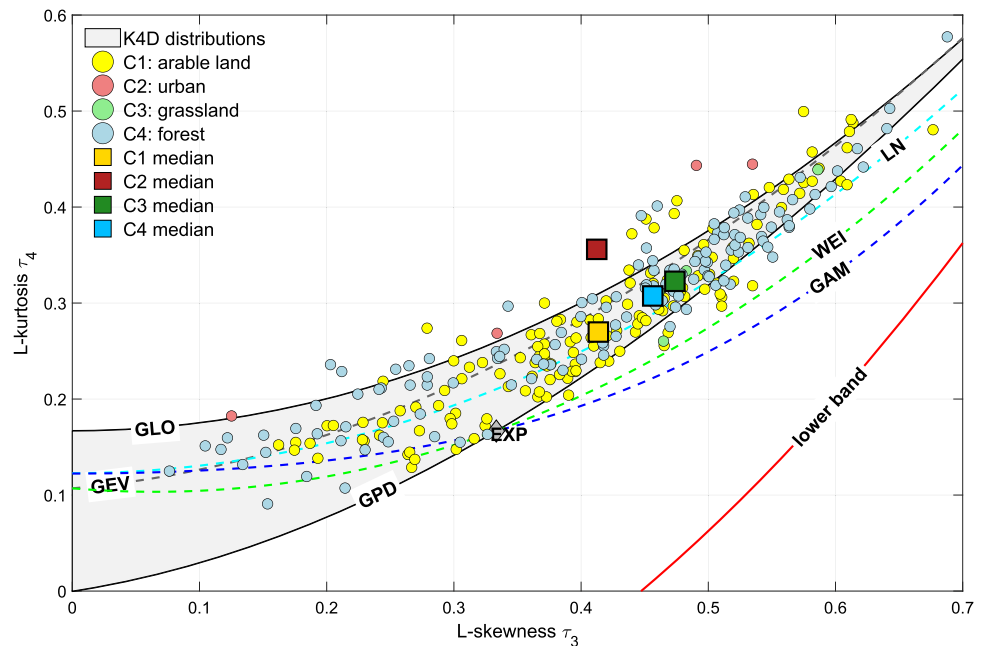
Selection of the categories for grassland shows that categories C1 and C2 exhibit very close to each other, therefore meadows fraction of (0, 10)% are characterized



**Table 9** Cluster matching for land use classes: italic—two largest values, bold—the largest one

		ML-cluster				ML-cluster					
		ML1	ML2	ML3	ML4	ML1	ML2	ML3	ML4		
Arable land											
<b>EX-class</b>	C1	4	7	<b>27</b>	9	<b>QU-class</b>	C1	9	10	<b>36</b>	17
	C2	32	20	<b>35</b>	17		C2	<b>23</b>	18	22	10
	C3	23	<b>32</b>	17	14		C3	20	<b>23</b>	19	10
	C4	5	19	<b>25</b>	4		C4	9	<b>28</b>	27	9
Urban											
<b>EX-class</b>	C1	20	<b>31</b>	22	8	<b>QU-class</b>	C1	20	<b>25</b>	19	8
	C2	34	36	<b>52</b>	20		C2	<b>25</b>	24	18	6
	C3	5	9	<b>21</b>	13		C3	12	17	<b>32</b>	11
	C4	2	3	<b>9</b>	5		C4	7	12	<b>35</b>	19
Grassland											
<b>EX-class</b>	C1	17	16	<b>20</b>	4	<b>QU-class</b>	C1	22	20	<b>25</b>	5
	C2	<b>37</b>	26	21	18		C2	<b>26</b>	19	13	15
	C3	9	20	<b>29</b>	13		C3	12	20	<b>27</b>	13
	C4	1	16	<b>34</b>	9		C4	1	20	<b>39</b>	13
Forests											
<b>EX-class</b>	C1	2	13	<b>19</b>	3	<b>QU-class</b>	C1	5	27	<b>30</b>	10
	C2	26	<b>41</b>	39	21		C2	18	<b>26</b>	18	11
	C3	22	18	<b>27</b>	13		C3	21	12	<b>28</b>	11
	C4	11	7	<b>19</b>	9		C4	20	13	<b>28</b>	12

**Fig. 16** L-moment diagram with clustering using dominant land use



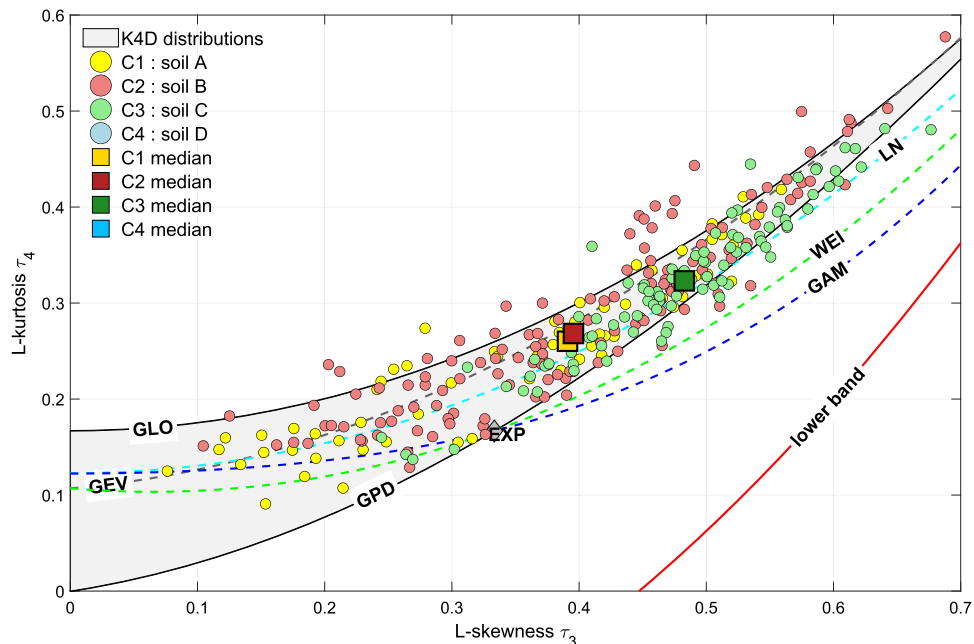
by low skewness and kurtosis. In opposite, higher fraction of pastures significantly encourages more extreme behavior. One explanation for that might be connected with the extensive drainage of grassland areas observed in Poland.

Cluster matching is presented in a single Table 9. Observation of these results and their comparison confirms qualitatively previous results. Figure 16 presents regional analysis of land use classes by means of L-moments diagram. Interestingly, urban land use class lies outside of the

**Table 10** Cluster matching for dominant soil hydrological groups: italic—two largest values, bold—the largest one

		ML-cluster				ML-cluster					
		ML1	ML2	ML3	ML4	ML1	ML2	ML3	ML4		
Soil group A											
<b>EX-class</b>	C1	1	9	<b>44</b>	<i>34</i>	<b>QU-class</b>	C1	0	5	<b>34</b>	<i>33</i>
	C2	5	<b>21</b>	<b>21</b>	3		C2	10	27	<b>32</b>	4
	C3	<b>28</b>	24	18	6		C3	<b>24</b>	<b>24</b>	18	6
	C4	<b>27</b>	25	21	3		C4	<b>27</b>	23	20	3
Soil group B											
<b>EX-class</b>	C1	5	6	<b>33</b>	<i>19</i>	<b>QU-class</b>	C1	11	8	<b>32</b>	<i>21</i>
	C2	25	<b>32</b>	22	6		C2	19	<b>28</b>	<i>21</i>	5
	C3	<b>28</b>	<b>28</b>	26	12		C3	21	19	<b>23</b>	9
	C4	6	<i>12</i>	<b>23</b>	7		C4	13	23	<b>27</b>	10
Soil group C											
<b>EX-class</b>	C1	<b>24</b>	23	20	7	<b>QU-class</b>	C1	<b>24</b>	22	19	7
	C2	25	25	<b>31</b>	6		C2	20	24	<b>25</b>	4
	C3	12	<b>20</b>	<i>19</i>	12		C3	17	<b>21</b>	<b>21</b>	13
	C4	0	11	<b>34</b>	<i>21</i>		C4	0	12	<b>39</b>	22
Soil group D											
<b>EX-class</b>	C1	44	67	<b>79</b>	15	<b>QU-class</b>	C1	3	23	<b>38</b>	8
	C2	<b>13</b>	9	7	5		C2	16	<b>27</b>	26	4
	C3	3	2	6	<b>10</b>		C3	<b>29</b>	22	17	4
	C4	1	1	<i>12</i>	<b>16</b>		C4	13	7	23	<b>30</b>

**Fig. 17** L-moment diagram for different classes of soil hydrological groups



region limited by the four parameter Kappa distribution, as it is the least natural and mostly affected by human action. Concluding, arable land exhibits less extreme events while areas with dominant grassland are characterized by the highest skewness and kurtosis, i.e. have more persistent behavior.

**4.4.7 Dominant soil hydrological group**

Dominant soil hydrological groups analysis is performed analogously to the dominant land use analysis. Soils are

**Table 11** Cluster matching for soil permeability: italic—two largest values, bold—the largest one

		ML-cluster				ML-cluster					
		ML1	ML2	ML3	ML4			ML1	ML2	ML3	ML4
Very low permeability											
<b>EX-class</b>	C1	53	<b>62</b>	53	7	<b>QU-class</b>	C1	<b>28</b>	23	19	2
	C2	7	<i>12</i>	<b>18</b>	5		C2	<b>24</b>	<b>24</b>	21	4
	C3	0	3	<b>16</b>	<i>12</i>		C3	11	26	<b>29</b>	6
	C4	1	2	<i>17</i>	<b>22</b>		C4	1	5	<b>35</b>	32
Low permeability											
<b>EX-class</b>	C1	22	9	<b>25</b>	10	<b>QU-class</b>	C1	22	11	<b>29</b>	10
	C2	21	<b>31</b>	<b>31</b>	12		C2	18	<b>24</b>	<i>21</i>	10
	C3	18	<i>24</i>	<b>26</b>	10		C3	16	<b>24</b>	23	9
	C4	3	<i>14</i>	<b>22</b>	12		C4	8	<i>19</i>	<b>30</b>	16
Average permeability											
<b>EX-class</b>	C1	1	15	<b>48</b>	<i>34</i>	<b>QU-class</b>	C1	0	6	<b>38</b>	28
	C2	25	29	<b>32</b>	6		C2	14	<i>20</i>	<b>27</b>	12
	C3	<b>28</b>	23	20	4		C3	21	<b>26</b>	23	2
	C4	<i>10</i>	<b>11</b>	4	0		C4	26	<b>27</b>	16	4
Variable permeability											
<b>EX-class</b>	C1	12	24	<b>63</b>	<i>42</i>	<b>QU-class</b>	C1	0	7	<b>33</b>	32
	C2	<b>35</b>	33	27	4		C2	12	<i>19</i>	<b>32</b>	10
	C3	<i>10</i>	<b>16</b>	7	0		C3	<b>28</b>	22	19	3
	C4	4	6	7	0		C4	<i>21</i>	<b>31</b>	20	1
High permeability											
<b>EX-class</b>	C1	1	14	<b>44</b>	<i>34</i>	<b>QU-class</b>	C1	0	6	<b>38</b>	28
	C2	<b>23</b>	<b>23</b>	21	5		C2	14	<i>24</i>	<b>25</b>	10
	C3	<b>32</b>	<i>30</i>	25	6		C3	<b>31</b>	25	14	2
	C4	5	<i>12</i>	<b>14</b>	1		C4	16	<i>24</i>	<b>28</b>	5

**Table 12** Cluster matching for the fraction of lakes: italic—two largest values, bold—the largest one

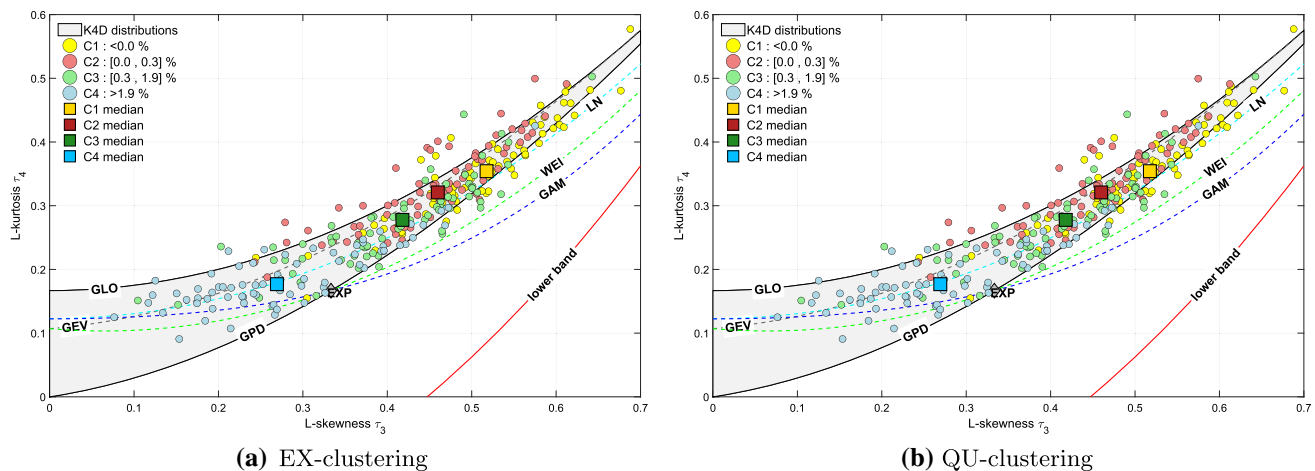
		ML-cluster				ML-cluster					
		ML1	ML2	ML3	ML4			ML1	ML2	ML3	ML4
<b>EX-class</b>	C1	9	<i>46</i>	<b>93</b>	44	<b>QU-class</b>	C1	2	9	<b>35</b>	26
	C2	8	<b>11</b>	8	1		C2	3	<i>20</i>	<b>36</b>	14
	C3	<b>21</b>	<i>16</i>	3	0		C3	10	28	<b>29</b>	5
	C4	<b>23</b>	6	0	1		C4	<b>46</b>	22	4	1

grouped into four categories, denoted as A, B, C and D. Similar analysis has been performed also in that case. In view of similar L-moments plots and the wish to avoid overloading the document with graphs, only a table showing matching properties (Table 10) is presented together with a combined L-moments diagram (see Fig. 17) showing relationships between all considered categories.

Taking into account former experience with L-moment diagrams and matching properties the conclusions might be derived using the matching tables only. Soil hydrological groups A and B (typically sands and loamy sands), characterized by highest permeability, behave in a diagonal

way, i.e. lower fraction of these soils (and thus a higher fraction of impermeable soils) results in higher site time series persistence, i.e. resulting in more extreme behavior.

L-moment diagram showing regional comparison between dominant soil hydrological group shows that group D does not dominate in any catchment area. It also confirms previous observations that soils with lower permeability class contribute to the higher probability of extreme events. Distributions for such catchment areas are characterized by higher skewness and kurtosis.



**Fig. 18** L-moments diagrams with clustering by fraction of lakes

#### 4.4.8 Soil permeability

Soil permeability is classified into five categories, i.e. very low, low, average, variable, high and very high. Similarly to the soil hydrological grouping, only Table 11 with clusters matching is presented. Observed results are in line with intuition. The lower soil permeability is, the higher persistence in distribution is noticed (higher skewness and kurtosis, i.e. fatter and longer tails are observed). In contrast high and very high soil permeability stimulate infiltration and thus protect against extreme events. Average permeability category allows to revert classification matching, which means that only high, very low and low permeable soils encourage persistence.

#### 4.4.9 Fraction of lakes

Fraction of lakes as a single category is analyzed in a more simple way. Table 12 shows the comparison. Table 12 is characterized by counter diagonal relationship. Lower fraction of lakes very significantly moves the category towards ML3 and ML4, which are responsible for extreme properties (high kurtosis and skewness). As opposed, high fraction of lakes introduces accumulation and inertia into the process. Presence of lakes in a catchment results in a buffering effect on river flows, putting the PDF properties closer to the normal Gaussian distribution. This relation is also clearly confirmed by the L-moment diagrams (Fig. 18). This relationship is the strongest among all studied catchment properties.

This classification is the most distinct considering the distance between the categories with respect to the difference in L-skewness and L-kurtosis. The range of variability in L-skewness  $\tau_3$  is from 0.27 to 0.51, while the kurtosis is in the range of  $\tau_4 \in (0.18, 0.35)$ . It is noteworthy that the category C4 denoting the highest fraction of lakes results in

the median being the closest to the normal distribution, out of all the considered features. Thus, presence of lakes offers the biggest positive impact protecting against river flow extremes.

## 5 Conclusions and further research

This work presents statistical regional daily river flow analysis. L-moment diagrams are used to investigate an impact of various catchment properties (catchment area, average precipitation, temperatures, slope, elevation, land use, dominant soil group, soil permeability and fraction of lakes) on the probability of flow extreme events. Decomposition of the catchment features into classes and their presentation using LMRD allows to perform clustering. Moreover, such classification is compared with direct use of the k-mean clustering.

Results show that statistical L-moments analysis can be improved with the use of data mining clustering algorithm. First of all, applied methodology works well and gives an insight into the nature of daily river flow time series for a given set of catchments. Obtained results are clear and confirm common hydrological knowledge. Such confirmation allows to use machine learning based clustering in the LMRD domain, apart from the existing approaches, as an alternative tool for the hydrological data analysis.

This work is interdisciplinary. Although the simplest clustering algorithm is applied, obtained results are in line with the hydrology-based categorization. Such results are promising. More research is required on better customization of the ML-based clustering algorithm. However, once it is fit and tuned, the hydrology analyst will get entirely new investigation methodology. Deeper coordination with the hydrological properties will allow not only to determine catchments that are most vulnerable to

extreme events, but also enable root cause analysis giving recommendations for protective actions. Sample regression analysis presented in Sect. 2 is still preliminary and it also requires further analysis, as initial investigation is promising.

Presented machine learning clustering approach is independent on data space dimension. In current research it is applied to the standard 2D L-moment ratio diagrams. However, there are no formal obstacles to conduct similar analysis in more dimension. Therefore it is planned to prepare 3D L-moment diagrams (L-Cv / L-skewness / L-kurtosis) and do clustering in such a space.

Presented results are even more valuable, in times of rapid climate change and increasingly frequent extremely dangerous flood events in Europe (Piniewski et al. 2017).

**Acknowledgements** The authors acknowledge the Institute of Meteorology and Water Management-National Research Institute (IMGW-PIB) for providing river flow data.

## Declaration

**Conflict of interest** The authors did not receive support from any organization for the submitted work. The authors have no financial or proprietary interests in any material discussed in this article.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Aggarwal CC, Reddy CK (2014) Data clustering. Algorithms and applications. CRC Press, Taylor & Francis Group, Boca Raton
- Aytaç E (2020) Unsupervised learning approach in defining the similarity of catchments: Hydrological response unit based k-means clustering, a demonstration on Western Black Sea region of Turkey. *Int Soil Water Conserv Res* 8(3):321–331
- Bobee B, Perreault L, Ashkar F (1993) Two kinds of moment ratio diagrams and their applications in hydrology. *Stoch Hydrol Hydraul* 7:41–65
- Calver A, Stewart E, Goodsell G (2009) Comparative analysis of statistical and catchment modelling approaches to river flood frequency estimation. *J Flood Risk Manage* 2(1):24–31
- Coles S, Dixon MJ (1999) Likelihood-based inference for extreme value models. *Extremes* 2:5–23
- Cupak A (2017) Initial results of nonhierarchical cluster methods use for low flow grouping. *J Ecol Eng* 18(2):44–50
- Cupak A, Walega A, Michalec B (2017) Cluster analysis in determination of hydrologically homogeneous regions with low flow. *Acta Scientiarum Polonorum Formatio Circumiectus* 1:53–63
- Desai S, Ouarda TBMJ (2021) Regional hydrological frequency analysis at ungauged sites with random forest regression. *J Hydrol* 594:125861
- Di Baldassarre G, Castellarin A, Brath A (2006) Relationships between statistics of rainfall extremes and mean annual precipitation: an application for design-storm estimation in northern central Italy. *Hydrol Earth Syst Sci* 10(4):589–601
- Dikbas F, Firat M, Cem Koc A, Gungor M (2013) Defining homogeneous regions for streamflow processes in Turkey using a k-means clustering method. *Arab J Sci Eng* 38:1313–1319 (**International Conference on Technological Advancements in Materials Science and Manufacturing**)
- Domański PD (2020) Study on statistical outlier detection and labelling. *Int J Autom Comput* 17(6):788–811
- Garmdareh ES, Vafakhalh M, Eslamian SS (2018) Regional flood frequency analysis using support vector regression in arid and semi-arid regions of Iran. *Hydrol Sci J* 63(3):426–440
- Hawkins DM (1980) Identification of outliers. Chapman and Hall, London
- Helliwell RC, Coull MC, Davies JLL, Evans CD, Norris D, Ferrier RC, Jenkins A, Reynolds B (2007) The role of catchment characteristics in determining surface water nitrogen in four upland regions in the UK. *Hydrol Earth Syst Sci Discuss* 1(1):356–371
- Hosking JRM (1990) L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J Roy Stat Soc: Ser B (Methodol)* 52(1):105–124
- Hosking JRM (1992) Moments or L-Moments? an example comparing two measures of distributional shape. *Am Stat* 46(3):186–189
- Hosking JRM, Wallis JR (1993) Some statistics useful in regional frequency analysis. *Water Resour Res* 29(2):271–281
- Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley
- Jehn FU, Bestian K, Breuer L, Kraft P, Houska T (2020) Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrol Earth Syst Sci* 24(3):1081–1100
- Jin X, Han J (2010) K-means clustering. In: Sammut C, Webb GI (eds) *Encyclopedia of Machine Learning*. Springer, Boston, MA, pp 563–564
- Kar KK, Yang S-K, Lee J, Khadim FK (2017) Regional frequency analysis for consecutive hour rainfall using l-moments approach in Jeju Island, Korea. *Geoenviron Disasters* 4:18
- Katz RW, Parlange MB, Naveau P (2002) Statistics of extremes in hydrology. *Adv Water Resour* 25(8):1287–1304
- Khan SA, Hussain I, Faisal M, Muhammad Y, Shoukry A, Hussain T (2017) Regional frequency analysis of extremes precipitation using L-Moments and Partial L-Moments. *Adv Meteorol*. Article ID 8727951
- King RS (2015) Cluster analysis and data mining: an introduction. Mercury Learning and Information LLC, Dulles; Boston; New Delhi
- Kjeldsen TR, Prosdocimi I (2015) A bivariate extension of the hosking and wallis goodness-of-fit measure for regional distributions. *Water Resour Res* 51(2):896–907
- Kjeldsen TR, Ahn H, Prosdocimi I (2017) On the use of a four-parameter kappa distribution in regional frequency analysis. *Hydrol Sci J* 62(9):1354–1363
- Lloyd S (1982) Least squares quantization in PCM. *IEEE Trans Inf Theory* 28(2):129–137
- Louzada F, Ramos P, Perdoná G (2016) Different estimation procedures for the parameters of the extended exponential geometric distribution for medical data. *Comput Math Methods Med*. Article ID 8727951

- Lun D, Viglione A, Bertola M, Komma J, Parajka J, Valent P, Blöschl G (2021) Characteristics and process controls of statistical flood moments in Europe - a data-based analysis. *Hydrol Earth Syst Sci* 25(10):5535–5560
- Maeda EE, Arevalo Torres J, Carmona-Moreno C (2013) Characterisation of global precipitation frequency through the L-moments approach. *Area* 45(1):98–108
- Marcinkowski P, Kardel I, Placzkowska E, Osuch P, Okruszko T, Venegas-Cordero N, Ignar S, Piniewski M (2021) A high-resolution simulated water balance and streamflow data set for 1951–2020 for the territory of Poland. *Geosci Data J* (2021)
- Mašiček T, Toman F, Palàt M (2011) Using the step linear regression at the analysis of hydrological conditions of the Frysavka drainage basin. *Infrast Ecol Rural Areas* 11:71–86
- Merz R, Blöschl G (2009) Process controls on the statistical flood moments: a data based analysis. *Hydrol Process* 23(5):675–696
- Merz R, Blöschl G (2009) A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria. *Water Resour Res* 45(1)
- Negi A, Rawat KS, Nainwal A, Shah MC, Kumar V (2021) Quality analysis of statistical and data-driven rainfall-runoff models for a mountainous catchment. *Mater Today Proc* 46:10376–10383 (**International Conference on Technological Advancements in Materials Science and Manufacturing**)
- Peel M, Wang Q, McMahon T (2001) The utility L-moment ratio diagrams for selecting a regional probability distribution. *Hydrol Sci J* 46:147–155
- Piniewski M, Szcześniak M, Kundzewicz ZW, Mezghani A, Hov Ø (2017) Changes in low and high flows in the Vistula and the Odra basins: Model projections in the European-scale context. *Hydrol Process* 31(12):2210–2225
- Piniewski M, Szcześniak M, Kardel I, Chattopadhyay S, Berezowski T (2021) G2DC-PL+: a gridded 2 km daily climate dataset for the union of the Polish territory and the Vistula and Odra basins. *Earth Syst Sci Data* 13(3):1273–1288. <https://doi.org/10.5194/essd-13-1273-2021>
- Podladchikova O, Lefebvre B, Krasnoselskikh V, Podladchikov V (2003) Classification of probability densities on the basis of pearson's curves with application to coronal heating simulations. *Nonlinear Process Geophys* 10:323–333
- Popat E, Kuleshov A, Kronenberg R, Bernhofer C (2020) Data-driven discharge analysis: a case study for the Wernersbach catchment, Germany. *Meteorol Hydrol Water Manage* 8(1):54–62
- Rousseeuw PJ (1984) Least median of squares regression. *J Am Stat Assoc* 79(388):871–880
- Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York
- Salinas JL, Castellarin A, Kohnová S, Kjeldsen TR (2014) Regional parent flood frequency distributions in Europe-part 2: climate and scale controls. *Hydrol Earth Syst Sci* 18(11):4391–4401
- Sharghi E, Nourani V, Soleimani S, Sadikoglu F (2018) Application of different clustering approaches to hydroclimatological catchment regionalization in mountainous regions, a case study in Utah State. *J Mt Sci* 15:461–484
- Simková T (2017) Statistical inference based on l-moments. *Statistika: Stat Econ J* 97:44–58
- Singh VP (2017) Handbook of applied hydrology, Second Ed. McGraw-Hill Education, New York, Chicago, San Francisco, Athens, London, Madrid, Mexico City, Milan, New Delhi, Singapore, Sydney, Toronto
- Smith A, Sampson C, Bates P (2015) Regional flood frequency analysis at the global scale. *Water Resour Res* 51(1):539–553
- Tallaksen LM, Van Lanen HAJ (2004) Hydrological drought. Processes and estimation methods for streamflow and groundwater. *Developments in Water Science*, 48. Elsevier Science B.V., Amsterdam, Netherlands
- Vargo E, Pasupathy R, Leemis L (2010) Moment-ratio diagrams for univariate distributions. *J Qual Technol* 42(3):1–11
- Wang F, Huang G, Li Y, Xu J, Wang G, Zhang J, Duan R, Ren J (2021) A statistical hydrological model for Yangtze river watershed based on stepwise cluster analysis. *Front Earth Sci* 9:853

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.