



POTSDAM-INSTITUT FÜR
KLIMAFOLGENFORSCHUNG

Originally published as:

Piniewski, M. (2017): Classification of natural flow regimes in Poland. - River Research and Applications, 33, 7, 1205-1218

DOI: [10.1002/rra.3153](https://doi.org/10.1002/rra.3153)

Classification of natural flow regimes in Poland

M. Piniewski* †

Abstract

Hydrological classifications are aimed at simplifying spatio-temporal variability of flow regimes, and secondly, at supporting environmental flow management. The objective of this study was to perform classification of natural flow regimes in Poland using an inductive approach based on a set of hydrological metrics (HMs), and to develop a model for prediction of class membership based on a set of environmental variables (EVs). A set of 147 gauges with relatively unmodified flow regimes was identified and, for each gauge, values of 73 HMs and 28 EVs were computed. Classification was performed using k-means and k-medoids techniques (both produced equivalent results), based on four principal components explaining 73.4% of variability in HMs. Out of seven distinguished classes, four (P1-P4) were spread across the Polish Plain, one (U5) was restricted to uplands and two (M6 and M7) to mountains. The between-class differences in HMs and EVs were generally high, although some classes (P2, P4, U5, M6, M7) were more easily distinguishable than others (P1, P3). Mean predictive accuracy of the developed random forest model was 79% which is high compared to other studies of this type. The lowest accuracies (0 and 50%) were achieved by two classes with the lowest counts. Variables representing diverse aspects: hydrography, climate, topography and geology had the highest importance in the RF model. Future research can benefit from the database of selected gauges with computed HMs, EVs and assigned classes, freely available through a long-lasting data repository. With this study, the first step towards application of the Ecological Limits of Hydrological Alteration (ELOHA) framework for environmental flow management at regional scale has been achieved.

Keywords hydrologic classification, random forest, hydrologic metric, Indicators of Hydrologic Alteration, catchment properties, ELOHA, environmental flow

1 Introduction

River flow regimes vary in space and time on an extraordinary scale. The temporal scale spans from minutes (flash floods) to years (supra-seasonal droughts), while the spatial variability is controlled by catchment properties such as climate, topography, land cover, soils and geology. River scientists have for a long time tried to simplify this huge spatio-temporal variability by performing hydrological classifications. First classifications were more descriptive in nature

*Warsaw University of Life Sciences, Poland. Corresponding author: mpiniewski@levis.sggw.pl

†Potsdam Institute for Climate Impact Research, Germany

and focused on monthly variability, the number of peaks during an annual cycle and their dominant feeding mechanism (Pardé, 1933; Lvovich, 1938). Further developments (Gottschalk *et al.*, 1979) also focused on flow genetic sources at monthly scale, but reduced the subjectivity of the previous methods by introducing strict discrimination criteria.

On top of the natural factors, human influences such as water abstraction, dams or diversions may be another driver shaping the flow regimes. In the 1990s the awareness rose that humans altered flow regimes on unprecedented scale worldwide, which caused the extensive ecological degradation and the loss of biodiversity (Poff *et al.*, 1997). This has led to development of biologically-relevant flow metrics, based on daily data and describing different aspects of flow regime: magnitude, timing, duration, frequency and rate of change (Richter *et al.*, 1996). These new metrics, called Indicators of Hydrologic Alteration (IHA), together with their various offsprings (Olden, Poff, 2003) have since then been increasingly used for deriving hydrological classifications. This new context of classifications was partly driven by the new notion of environmental flows (Acreman, Dunbar, 2004) and the need to assign the natural flow regime class to perturbed sites. Most notably, flow regime classification was designed as one of key steps in the Ecological Limits for Hydrologic Alteration (ELOHA; Poff *et al.*, 2010) framework for assessing ecological effects of streamflow alteration at regional scales.

The ELOHA framework directly embraces flow-ecology relationships that are usually region-specific. Hence, it is of vital importance whether such relationships have been established for Polish rivers. The answer is mixed. On the one hand, there are no studies, to my best knowledge, that quantify such relationships from the flow regime classes perspective. On the other hand, there is a solid base to establish such relationships due to a wealth of studies reporting ecological data on aquatic or river-dependent organisms in a hydrological context. This kind of research in Poland included macrophytes (Jusik *et al.*, 2015; Szoszkiewicz *et al.*, 2010), riparian wetlands (Oświt *et al.*, 1996; Pusłowska-Tyszewska *et al.*, 2014; Szewczyk *et al.*, 2005), zooplankton (Czerniawski, Domagała, 2010), aquatic invertebrates (Dumnicka, Koszałka, 2005; Grzybkowska, Witczak, 1990; Grzybkowska *et al.*, 1996; Kajzer-Bonk *et al.*, 2013; Skalski *et al.*, 2012; Wrzesiński, 2014; Wyzga *et al.*, 2012), fish (Wyzga *et al.*, 2009; Bischoff, Wolter, 2001), birds (Kajtoch, Figarski, 2013; Kloskowski *et al.*, 2015; Tryjanowski *et al.*, 2009) and mammals (Wuczyński, Jakubiec, 2013). The majority of these studies, though, dealt with the effects of high river flows. The character of reported relationships between flow and ecology was diverse across different studies: a positive role of floods on floodplain vegetation (Oświt *et al.*, 1996), fish diversity (Bischoff, Wolter, 2001) and aquatic warbler density (Tryjanowski *et al.*, 2009) was reported, but negative effects of floods on certain species, e.g. chironomids (Grzybkowska *et al.*, 1996), white stork (Tryjanowski *et al.*, 2009) were also identified. These studies were typically restricted to one or several sites on a river or on a group of neighbouring rivers, which brings about a question of their applicability on a wider scale, relevant for environmental flow management. First attempts to integrate flow-ecology-relationships for different groups of communities (wetland vegetation, fish) in a river-basin scale have been made in near-pristine NE part of Poland (Piniewski *et al.*, 2011; Piniewski *et al.*, 2014). More recently, the first national-scale study was performed, trying to upscale flow-ecology relationships from several low-modified, representative

sites in Poland across the entire country for the sake of the environmental flow method definition (as a part of future water management legislation) (KZGW, 2015). The developed method is based on the combination of a Hydrology-based Environmental Flow Regime (HEFR; Opdyke *et al.*, 2014) methodology and the habitat simulation model MesoHABSIM (Parasiewicz, 2007). Hence, an extensive natural flow regime classification, not performed in Poland to date, may have a practical importance in further enhancements of this new method.

Natural flow regime classifications were performed over medium and large scales in numerous locations worldwide. The typical examples include country- or state-wide classifications: Australia (Kennard *et al.*, 2010b), France (Snelder *et al.*, 2009), USA (McManamay *et al.*, 2014; Archfield *et al.*, 2014), New Zealand (Snelder, Booker, 2013) and Iran (Tavassoli *et al.*, 2014). Less frequent were classifications applied at geographical region level, usually within one country, e.g. Ozark–Ouachita Interior Highlands (Leasure *et al.*, 2016), south-east Queensland (Mackay *et al.*, 2014), Washington (Liermann *et al.*, 2012) and the Mediterranean region (Oueslati *et al.*, 2015) or large river basins, e.g. the Ebro (Solans, Poff, 2013) or the Segura (Belmar *et al.*, 2011). A natural extension of numerous classifications is application of decision tree models for class membership prediction based on landscape characteristics. For example, Classification and Regression Trees (CART) have been used (Liermann *et al.*, 2012; Kennard *et al.*, 2010b), as well as random forests (McManamay *et al.*, 2014; Snelder, Booker, 2013; Leasure *et al.*, 2016; Mackay *et al.*, 2014; Liermann *et al.*, 2012) and boosted regression trees (Snelder *et al.*, 2009).

In this study the overarching methodological framework for hydrological classifications developed by Olden *et al.* (2012) is followed. The most commonly applied pathway of inductive reasoning is pursued, in which similarities among rivers are characterised according to a set of diagnostic hydrological metrics that vary spatially across the landscape. In contrast, deductive approaches use indirect environmental surrogates for hydrology to establish flow regime classifications. Inductive approach consists of several steps (Olden *et al.*, 2012): (1) acquisition and evaluation of hydrological data; (2) selection of hydrological metrics; (3) computation of hydrological metrics; (4) conducting the hydrological classification and (5) interpretation and (optionally) spatial modelling of the developed classification.

The objective of this study is to perform a hydrological classification of natural streamflow regimes in Poland. More specifically, the purpose is to: (1) identify a set of relatively unimpaired and representative flow gauges for the area of Poland; (2) compute hydrological metrics for each gauge and identify which metrics are the most influential; (3) compute environmental variables characterising each gauged catchment and identify which variables are the most influential; (4) perform flow regime classification using several statistical methods, select the final classification and interpret the results; (5) develop a decision tree using the random forest technique in order to predict class membership for ungauged streams.

2 Materials and Methods

2.1 Study area

Poland lies in Central and Eastern Europe with a total area of 311,888 km² (Fig. 1). It is predominantly a lowland country: its largest part belongs to the Central European Plain and Eastern Baltic Plain (region names after Kondracki, 2002). These two regions were formed by glacial erosion in the Pleistocene ice age. The southern part of this plain characterised by a flat relief was shaped by older glaciations, while the northern part is full of lakes and low hills formed by younger glaciations. Elevation gradient increases towards the south. The Czech Massif (consisting of the Sudetes mountains and their foothills) stretches along the south-western part of the Polish border, while in the eastern part of the country the Central European Plain transforms first into the Polish Upland, and then into Carpathians and their Foothills in the south-eastern corner (Fig. 1). Poland has a temperate climate with cold winters and warm summers, which is influenced by air masses from all directions: maritime air from the west, cold polar air from Russia and Scandinavia, as well as sub-tropical air from the Atlantic Ocean, Mediterranean, and Black Sea. Land cover (assessed using CORINE Land Cover 2006) is predominantly agriculture (63%) and forests (32%).

The complex interplay of climatic and physiographic factors has a direct effect on streamflow of Poland's rivers. The classic work of Dynowska (1971) reported five types of flow regimes occurring in Poland (Supplementary Material Figure S1) that can be distinguished by simple rules ($Q_{3,4}$ - the highest of the mean March and April flows; $Q_{6,7,8}$ - the highest of the mean June, July and August flows; Q_{mean} - mean annual flow):

1. Weak nival ($Q_{3,4} < 1.3 \cdot Q_{mean}$);
2. Strong nival ($Q_{3,4} > 1.8 \cdot Q_{mean}$);
3. Moderate nival ($Q_{3,4} \in [1.3 \cdot Q_{mean}, 1.8 \cdot Q_{mean}]$);
4. Nivo-pluvial ($Q_{3,4} \in [1.3 \cdot Q_{mean}, 1.8 \cdot Q_{mean}]$ and $Q_{6,7,8} > 1.1 \cdot Q_{mean}$);
5. Pluvio-nival ($Q_{6,7,8} \in [1.3 \cdot Q_{mean}, 1.8 \cdot Q_{mean}]$).

Hence, the classification scheme of Dynowska, 1971 is based on two parameters referring exclusively to the time of occurrence (spring or summer) and standardized magnitude of peak monthly streamflow. The naming system suggests that different classes are characterised by different origin of flood waters: nival (snow melt) and pluvial (rain).

2.2 Streamflow data

In this study mean daily measured discharge data from the Institute of Meteorology and Water Management - National Research Institute (IMGW-PIB) were used. Gauge selection process has to follow certain criteria that are fit-for-purpose with respect to the study objective. In our case the most important criteria were: (1) good geographical coverage taking into account spatial climatic and physiographic variability; (2) good temporal record availability and

(3) the minimum level of human disturbance manifested by nearly unimpaired flow regime. The temporal window for all analyses was set to the 30-year period 1984-2013. The minimum record availability was set to 80% (24 years). This is much more than the threshold of 15 years and the minimum overlap across records of 50% recommended by Kennard *et al.* (2010b), so the uncertainty of this classification originating from different lengths of records and data overlap should not be high.

For assessing disturbance level previous experience from large-scale hydrological modelling of natural streamflow in Poland was used (Piniewski *et al.*, submitted), in which a benchmark catchment dataset was constructed. First, gauges with catchment area of more than 10,000 km² were excluded, as at this scale flow alteration is almost unavoidable (Leasure *et al.*, 2016). For the remaining gauges, a wide array of GIS layers related to water management and land cover was collected. All catchments containing reservoirs with capacity larger than 5 million m³ were excluded, while for all remaining catchments the relative importance of existing small reservoirs with respect to the mean flow was assessed on the case-by-case basis. Catchments containing cities with population above 200,000 inhabitants were all excluded. Only one catchment with population above 100,000 inhabitants (Kielce) was present but its relative impact on streamflow was assessed as low due to large catchment size. Other reasons for exclusion were: presence of major water abstractions or point sources, cross-basin water transfers, mining and large amount of fish ponds. In the next step, all streamflow hydrographs were screened and in case of dubious patterns they were compared to the hydrographs from the nearby catchments. In such cases, selected hydrological metrics that are sensitive to some specific types of flow alteration were calculated and gauges that were clear outliers were excluded. After performing all these analyses a set of 147 gauges shown in Figure 1 and listed in the Supplementary Material Table S1 was identified. Geographical coverage is good, albeit some regions known for their lower anthropogenic pressure (e.g. north-eastern Poland) are better represented than others. The minimum, mean and maximum catchment area was equal to 6, 856 and 6800 km², respectively.

The missing records were handled in the following way. The gaps shorter than 30 days (extremely rare) were filled in by interpolation. Whenever there was a gap longer than 30 days, data from the whole year were being removed. 73% of gauges had the complete data record, while 12% had only one year missing. For 16 out of 30 years flow data availability was above 99% of the whole gauge data set, while the lowest availability was achieved in 2013 with 87% of active records. It is acknowledged that some level of anthropogenic flow alteration is still present in most of selected catchments, however the work done represents our best efforts to select the subset of least human-impacted gauges.

2.3 Hydrological metrics

Due to a general ecological rationale of this classification, a suite of hydrological metrics (HMs) that describe the totality of the flow regime was selected (Olden *et al.*, 2012)). The Time Series Analysis module of the River Analysis Package (Marsh, 2004) was used to calculate the majority (70 of 73) of the metrics (Table 1). Scale dependence was removed by standardizing flow data with respect to mean annual flow. The only exception were three area-specific runoff metrics (Masr, Med_MinAsr and Med_MaxAsr), which refer to the mean, median of the

minimum and median of the maximum annual flows, respectively, divided by the catchment area. Metrics accounted for magnitude and variability of average and extreme flows, frequency and duration of high and low pulses, rate and frequency of flow changes and timing of extreme and seasonal flows.

2.4 Environmental variables

Environmental variables (EVs) include climatic and physiographic catchment descriptors that are expected to be correlated with some of the HMs. A wide range of EVs (Table 2) was selected representing different categories such as the size, geography, climate, topography, geology, soil, land cover and hydrography. Selection was a trade-off between a review of studies that also used EVs for similar purposes (Kennard *et al.*, 2010b; Liermann *et al.*, 2012; McManamay *et al.*, 2014; Snelder *et al.*, 2009; Snelder, Booker, 2013; Leasure *et al.*, 2016; Mackay *et al.*, 2014) and the availability of respective GIS data in Poland (Table 2). In order to compute these characteristics for each of 147 gauges, first each gauge was associated with its upstream catchment. Next, the gridded characteristics (climate and topography) were calculated for each catchment as the statistics (mean, minimum or maximum) of intersecting grid cells, while vector characteristics (geology, soil, land cover and hydrography) were calculated based on intersection of respective layers within catchment boundaries and performing spatial summaries.

2.5 Flow regime classification approach

2.5.1 Ordination analysis

Due to high redundancy within the HM dataset, it is not recommended to use all metrics for classification. In this study the Principal Component Analysis (PCA) was performed on the whole set of HMs in order to minimize their multi-collinearity by identifying a parsimonious sub-set of non-correlated synthetic indices (principal components, PCs) that represent several dimensions of hydrological variability. The number of non-trivial components (those with highest factor loadings) was determined using the broken stick rule. Five HMs having the highest correlation with selected PCs were used for interpretation of the synthetic indices. Projection of cases on two-dimensional factor plain was done in order to further interpret the results.

2.5.2 Cluster analysis

The statistical techniques used for organising rivers into similar classes are numerous and vary in the output (Olden *et al.*, 2012). The most commonly applied hard clustering methods assume that gauges can be divided into non-overlapping clusters with well-defined boundaries. Two main sub-types of hard clustering are (agglomerative) hierarchical and partitional clustering methods. The first group is based on a bottom-up approach in which each gauge starts in its own cluster, and pairs of clusters are merged upon moving up the hierarchy. The latter group seeks to identify clusters of equal distinction. Partitional clustering, on the other hand, attempts to directly decompose the data set into a set of disjoint clusters, each of which is characterised by a central vector. Finally, in soft clustering

techniques, stream gauges can belong to multiple classes with a certain degree of membership.

In this study we applied two partitional methods: k-means and k-medoids.

The common feature of these two techniques is that they both divide N observations with P dimensions (variables) into k clusters (pre-defined value) so that the sum of distances between observations and the points designated as cluster centres is minimised. In the k-means technique, the cluster centre is the mean of observations in the cluster and the distance measure is calculated as the within-cluster sum of squares. In the k-medoids algorithm, the cluster centre is one of its members, called a medoid. The k-medoids technique is known to be more robust to outlier data than the k-means.

Both k-means and k-medoids were run for different values of k between 2 and 15. In order to overcome the problem of finding the local minimum, 40 random starting cluster configurations were used for each k and the best solution (for each k) was being retained. As in the study of Snelder, Booker (2013), for k-means, the optimum number of clusters was selected based on analysing the plot of k values against the percent of variation (ratio of within-cluster sum of squares for a given k to the whole-dataset sum of squares). This value decreases with increasing k and the candidate optimal point is the one for which the rate of decrease goes down. For k-medoids partitioning, the optimal number of clusters was the one with the highest mean silhouette value, s (Kaufman, Rousseeuw, 1990). The higher s , the better cluster assignment of a given observation. If s is close to 0, it means that the observation falls in between two different clusters.

2.6 Random forest

A random forest (RF) classification tree was used as an empirical model of the class membership prediction as a function of EVs. The RF consists of a collection of simple CART-type tree predictors, each capable of associating a specified set of independent predictor values with one of the categories present in the dependent variable. Multiple trees built using random subsets of the data are then combined to produce one prediction for each observation. For each observation, the class with the highest proportion of trees voting for membership in this class wins. The advantage of RFs over traditional approaches is an increased classification accuracy and robust approaches to estimating variable importance based on decrease in classification accuracy caused by random permutations of each variable (Breiman, 2001).

In order to select variables to be used as potential predictors, the PCA was first performed on the set of EVs from Table 2 using the same approach as described in section 2.5.1 for HMs. The EVs representing different categories (Table 2) and having highest correlation with four first PCs were selected. The optimal size of the subset of predictor variables was defined as \sqrt{k} , where k is the number of predictors. The maximum level and number of nodes of an individual tree were set to 8 and 30, respectively. The maximum allowed number of trees to grow was set to 200. The percentage of sample used for training was 65%, while the remaining 35% were used to calculate mis-classification rates in cross-validation procedure. Another measure of classification quality was Cohen's κ coefficient of agreement assessing the predictive performance of the classification tree compared to random expectations.

3 Results

3.1 Flow regime classification

The first four PCs had eigenvalues greater than those determined by the corresponding broken stick distribution, explaining 73.5% of the variation in 73 hydrological variables. Table 3 specifies the variation explained by each PC and up to five HMs with the highest loadings. Interpretation of high values of PCs is also provided. In summary, each of the four PCs represents truly different aspects of the flow regime: (1) low flows and predictability, (2) high and low spells, (3) monthly flow magnitude and (4) contingency and high flows.

Both criteria used for optimal number of clusters selection (percent of variation for k-means algorithm and average silhouette for k-medoids algorithm) suggested seven classes as the most appropriate number. Furthermore, both techniques led to nearly the same classification, with only one out of 147 gauges assigned to different classes, so they can be regarded as equivalent. For clarity, in the further text, the output of k-means classification shall be used as the final one, due to a better interpretability of the class assignment of the ambiguous gauge.

Most classification algorithms force a grouped structure to what may otherwise be a continuously varying distribution (Olden *et al.*, 2012). Placement of different class members in the four-dimensional ordination space (projections on the PC1-PC2 and PC3-PC4 plains) shows that although a grouping structure can be distinguished, the borders between some of the classes are vague (Fig. 2). The majority of points are located in the proximity of their respective cluster centres, but each class has members whose location in the ordination space suggests that assignment to another class could also be possible.

Figure 3 illustrates geographical location of seven natural flow regime classes across Poland. The main geographical distinction can be made with respect to prevailing location of class members to either plain (P), upland (U) or mountain (M) areas. For this reason, first four classes were named P1-P4, the fifth class was named U5 and the last two classes were named M6-M7. While classes P2, P4, U5, M6 and M7 (the last one, only when elevation is included) can be distinguished from each other basing solely on geographical location, there are two classes (P1 and P3) that are widely scattered around the Polish Plain. This suggests that other factors than location play an important role. Silhouette values (Fig. 3) are usually the highest in the proximity of class geographic centres, which suggests that some of the gauges located far from the centres may be mis-classified. This effect is particularly strong for classes P1 and P4.

Table 4 presents the most important features distinguishing hydrological classes, including already discussed location (cf. Fig. 3), their monthly flow regime (Fig. 4), other hydrological metrics and environmental variables (Fig. 5). The only common feature of all monthly hydrographs across seven classes is that the mean monthly flow is the highest in March or April, and in the next months it gradually decreases until the end of summer/beginning of autumn (all classes apart from M7) or winter (M7). Otherwise, classes differ with respect to the magnitude and variability of monthly flows, but some classes (e.g. P1 and U5) show similar temporal patterns, which suggests that other factors might play a distinguishing role.

In most cases, between-group variability in HMs exceeds within-group vari-

ability (cf. Fig. 5 showing 10 selected HMs with high correlations with first four PCs). In consequence, it is often possible to distinguish a class using just one or more HMs, as presented in Table 4. For example, a gradient can be observed for some HMs when analysed across four "plain stream" classes (P1-P4): base-flow index, low flow magnitude and flow predictability decrease when moving from class P1 to P4, while high flow magnitude and daily coefficient of variation increase when moving in the same direction. This is also well reflected in the seasonal stability of flow regimes depicted in Figure 4, *i.e. standardised monthly flow values are the closest to the value of 1 for the class P1, and they keep diverging from 1 for classes P2 through P4. The size of the boxes and whiskers representing inter-annual variability, also grows from P1 to P4* (Fig. 4). Classes P1 and U5 are characterised by similar monthly flow regime and are not distinguishable by any of the aforementioned HMs, but class P1 has considerably higher mean annual specific runoff (Masr) and lower variability in rise rate (VarAnnMRateRise) than class U5 (Fig. 5). Two classes of mountainous streams, M6 and M7, also have a lot of similar hydrological features, but their main discriminating factors are the median Julian date of the annual minimum flow occurrence (JDMinMed) and Masr (both much higher values for M7 than M6).

Since this classification was based on PCs derived from the full set of HMs, it is natural to expect some variability in HMs across different classes. However, it is less obvious whether EVs allow for discriminating hydrological classes. Figure 5 also includes 10 EVs (those representing different environmental aspects and having high correlations with first four PCs derived from EV-based PCA), while Table 4 contains an indication of EVs that allow for discriminating certain classes.

The easiest way to *distinguish* classes M6 and M7 is by elevation, slope or precipitation. All EVs representing these categories differ fundamentally between these two classes and the remaining five. Summer (TMP_JJA) and winter temperature (TMP_DJF) as well as percentage of agricultural land (LU_agr) allow to discriminate between M6 and M7 reasonably well. In contrast, class U5 can be distinguished from others based on geological properties: presence of loess or limestone bedrock. Class P2 has significantly higher percentage of lakes (HYD_lake) and wetlands (HYD_wet) than any other class. Among other factors, class P4 can be discriminated by relatively low effective infiltration into the uppermost aquifer (GEO_inf), which is connected to low summer precipitation (PCP_JJA) and high percentage of till (*i.e.* predominantly impermeable) bedrock (GEO_till).

In contrast, there is no single EV that makes it possible to discriminate classes P1 and P3. In fact, the most abundant class P3 seems to be the most problematic class in this respect, as it cannot be distinguished by location (Fig. 3), any single HM and any single EV (Table 4). Its monthly flow regime is "in between" stable classes P1-P2 and highly variable class P4 (Fig. 4). In fact, this difficulty with class P3 should not be surprising in the light of Figure 2 showing that P3 is situated directly in the middle of PC1-PC2 space, neighbouring with all classes apart from M7. Hence, simple rules cannot work in this case and more advanced approaches are needed to predict class membership.

3.2 Prediction of class membership

Seventeen EVs selected for the random forest model are marked with * symbol in Table 2. Thus, the number of variables in a single tree was set to $\sqrt{17} \approx 4$. Predictive accuracy of the RF model on the training dataset ($N = 95$) was equal to 94%. Six classes had the accuracy above 93% and only M7 class (with only four counts for the training dataset) had the accuracy of 50%. Predictor importance analysis (cf. Table 2 for variable definitions) showed that stream density (HYD_str) had the highest importance among selected 17 EVs, followed by summer temperature (TMP_JJA), mean slope (TOP_slp), effective infiltration (GEO_inf) and percentage of wetlands (HYD_wet). The importance of these EVs was higher than 90%. Two EVs with the lowest importance were: limestone and loess bedrock percentage (GEO_lim and GEO_loess, with the accuracy of 45 and 53%, respectively). All remaining variables had the accuracy in the range of 70-90%.

The RF model did not perform as well for the testing dataset as for the training dataset (Fig. 6). Predictive accuracy was equal to 0.79 and Cohen's κ coefficient of agreement to 0.73. The accuracy was highly variable across classes. For classes P1-P4 it ranged between 0.67 and 0.86, for class U5 it was equal to zero, while for classes M6 and M7 to 1 and 0.5, respectively. The results for both U5 and M7 should be treated with caution, as they are based on only 2 cases. It can be observed that even though the most abundant class P3 had a high accuracy (0.86), it was the only incorrectly allocated class for P1, P2 and P4 (Fig. 6). The mean silhouette value for gauges with correctly predicted class was higher than the corresponding value for gauges with incorrectly allocated classes (0.3 vs. 0.21 for training and 0.36 vs. 0.31 for validation). This suggests that incorrect predictions happen more likely for streams whose assignment to hydrological class is more uncertain.

4 Discussion

Classification of natural flow regimes in Poland performed in this study is, to our knowledge, the first such classification in Central-Eastern Europe. Its outcome differs substantially from the classification of Polish river regimes of Dynowska (1971), mainly due to differences in methodology (a large set of HMs representing the totality of the daily flow regime vs. an approach based on flood seasonality and interpretation in terms of dominant feeding mechanism) and input data. Nevertheless, some similarities exist, e.g. the weak and moderate nival classes partly correspond to the P2 and P1 classes, respectively (Supplementary Material Figure S1). The pluvio-nival regime of Dynowska (1971) partly overlaps with the M7 class, which is associated with the highest elevation. The main feature distinguishing two approaches is that our classification shows that some regions are inhomogeneous and there can be three-four different classes close to each other (e.g. north-east or south-east of Poland; cf. Fig. 3), like in the study of Kennard *et al.* (2010b). Such a short-distance variability is most likely controlled by other factors than climatic ones. It is noteworthy that a wide array of different EVs appeared to have a strong influence on hydrological properties and the derived classification (Fig. 5B), which was also reflected by the fact that top five EVs with the highest variable importance in the RF model represented

hydrography, climate, topography and geology. The fact that climate was not a single dominating factor like in some other classifications can be explained by the regional (i.e. not continental) scale of this application (McManamay *et al.*, 2014).

Predictive accuracy of the developed random forest (79%) was relatively high compared to other similar studies: 73% (Leasure *et al.*, 2016), 75% (Liermann *et al.*, 2012), 63% (Snelder *et al.*, 2009), 67% for the best configuration (Snelder, Booker, 2013) and 76% for the best configuration (McManamay *et al.*, 2014). In contrast to Snelder *et al.* (2009) who reported that higher class predictability was achieved for more hydrologically distinctive classes and those with strong relationships between HMs and EVs, it was found in this study, similar to Liermann *et al.* (2012), that low accuracy was exclusively associated with classes that had very few members (U5 and M7). These two classes had highly distinctive regimes and catchment properties, in contrast to class P3 that had a transitional (and not very distinctive) regime and properties (cf. Table 4), but the most abundant P3 had a much higher predictive accuracy than the least abundant U5 and M7. This suggests that in the future updates of the RF model, its error rate could be decreased by increasing gauge counts in classes U5 and M7.

A direct application of the results of this study depends on two aspects, (1) whether a gauged or ungauged river is considered; (2) whether it has a modified or unmodified flow regime. Firstly, for unmodified, gauged, sites having a sufficient amount of continuous flow records (here, a minimum of 25 years was used; a shorter time series can also be used, but it would increase the uncertainty, cf. Kennard *et al.*, 2010a, and below the catchment area of 10,000 km², flow regime classes can be attributed on the basis of direct calculations of hydrological indices. Secondly, for gauged sites that are suspected to have a modified flow regime, two approaches can be applied simultaneously: flow regime classification and prediction based on EVs. If the same class is derived using both approaches, it may be a hint that the flow modification is not so strong, whereas the opposite case would be a partial evidence for a considerable modification. Thirdly, for ungauged sites, regardless whether modified or not, only the RF model can be applied in order to predict the natural flow regime class. This yields a higher uncertainty than assigning a class based on hydrological indices, but is still the best available option.

In the context of environmental flow management, a direct application is somewhat limited due to the fact that establishing environmental flow standards should not rely solely on hydrological classes, but also on hydromorphology and flow-ecology relationships (Poff *et al.*, 2010). Although none of these issues was addressed in this paper, the developed class membership prediction model can be applied across the entire network of Polish rivers, generating a high-resolution, national map of flow regime types, that would be a valuable asset in itself (cf. Snelder *et al.*, 2009; Liermann *et al.*, 2012; Leasure *et al.*, 2016). If, for example, published evidence on flow-ecology relationships in Polish rivers is collated into a comprehensive database (such as the one of McManamay *et al.*, 2013, for the South Atlantic Region of the U.S.), these relationships could then be stratified by flow regime classes and statistically analysed. Another opportunity arises from future enhancements of the newly developed method for environmental flow assessment in Poland (KZGW, 2015). Ultimately, environmental flow standards used in national water legislation could also be class-dependent.

In the view of ongoing climate change that is projected to considerably affect natural streamflow in Poland (Piniewski *et al.*, 2014), future work should also attempt to understand how the designed classes could change (e.g. shift between existing classes, new classes, etc.) in response to this phenomenon. **To this end, different approaches can be used: e.g. driving a class membership prediction model with downscaled climate projections data (Dhungel *et al.*, 2016) or performing independent classifications for current and future flows simulated by a hydrological model (Zhang *et al.*, 2015).**

Finally, in order to increase the availability of the results of this study, the spreadsheet database of reference Polish gauges storing computed hydrological metrics, environmental variables, principal components and assigned flow regime classes (Supplementary Table S1) is also available in a free, long-term research data repository (DOI: 10.4121/uuid:3cd28c08-8c30-4686-8ff9-d1ecb0a5a29c). Potential uses of this unique dataset include: (1) creation of one's own classification, e.g. using other techniques or focused on specific aspects of the flow regime, (2) examining relationships between catchment properties and hydrological metrics, and (3) development of new predictive models for class membership.

5 Acknowledgements

Support of the Alexander von Humboldt Foundation, the Ministry of Science and Higher Education of the Republic of Poland and the Climate Change Impact on Selected Sectors in Poland (CHASE-PL) project funded within the Polish-Norwegian Research Programme under the Norwegian Financial Mechanism 2009-2014 (Project Contract No. Pol Nor/200799/90/2014) is gratefully acknowledged. The Institute of Meteorology and Water Management - National Research Institute (IMGW-PIB) is acknowledged for providing the streamflow data. **Piotr Parasiewicz and one anonymous reviewer provided valuable comments on the manuscript.** Special thanks are addressed to Prof. Zbigniew W. Kundzewicz for fruitful discussions and useful comments on early drafts of this manuscript, and to Mateusz Szcześniak for technical support.

References

- Acreman, MC, Dunbar, MJ. 2004. Defining environmental river flow requirements: a review. *Hydrology and Earth System Sciences*. **8**: pp. 861–876.
- Archfield, SA, Kennen, JG, Carlisle, DM, Wolock, DM. 2014. An objective and parsimonious approach for classifying natural flow regimes at a continental scale. *River Research and Applications*. **30**: pp. 1166–1183.
- Belmar, O, Velasco, J, Martinez-Capel, F. 2011. Hydrological Classification of Natural Flow Regimes to Support Environmental Flow Assessments in Intensively Regulated Mediterranean Rivers, Segura River Basin (Spain). *Environmental Management*. **47**: pp. 992–1004.
- Berezowski, T, Szcześniak, M, Kardel, I, Michałowski, R, Okruszko, T, Mezghani, A, Piniewski, M. 2016. CPLFD-GDPT5: High-resolution gridded daily precipitation and temperature dataset for two largest Polish river basins. *Earth Syst. Sci. Data*. **8**: pp. 127–139.

- Bischoff, A, Wolter, C. 2001. The flood of the century on the River Oder: Effects on the 0+ fish community and implications for floodplain restoration. *Regulated rivers - Research & Management*. **17**: 171–190.
- Breiman, L. 2001. Random Forests. *Machine Learning*. **45**: pp. 5–32.
- Czerniawski, R, Domagała, J. 2010. Similarities in zooplankton community between River Drawa and its two tributaries (Polish part of River Odra). *Hydrobiologia*. **638**: pp. 137–149.
- Dhungel, S, Tarboton, DG, Jin, J, Hawkins, CP. 2016. Potential Effects of Climate Change on Ecologically Relevant Streamflow Regimes. *River Research and Applications*. **32**. RRA-16-0052: pp. 1827–1840.
- Dumnicka, E, Koszałka, J. 2005. The effect of drought on Oligochaeta communities in small woodland streams. *Biologia, Bratislava*. **60**: pp. 143–150.
- Dynowska, I. 1971. *Typy reżimów rzecznych w Polsce (in English: River regime types in Poland)*. Zeszyty Naukowe UJ. Kraków: Prace Inst. Geogr.
- Gottschalk, L, Jensen, JL, Lundquist, D, Solantie, R, Tollan, A. 1979. Hydrologic regions in the Nordic countries. *Hydrology Research*. **10**: pp. 273–286.
- Grzybewska, M, Temech, A, Dukowska, M. 1996. Impact of long-term alternations of discharge and spate on the chironomid community in the lowland Widawka River (Central Poland). *Hydrobiologia*. **324**: 107–115.
- Grzybewska, M, Witczak, J. 1990. Distribution and production of Chironomidae (Diptera) in the lower course of the Grabia River (Central Poland). *Freshwater Biology*. **24**: 519–531.
- Jusik, S, Szoszkiewicz, K, Kupiec, JM, Lewin, I, Samecka-Cymerman, A. 2015. Development of comprehensive river typology based on macrophytes in the mountain-lowland gradient of different Central European ecoregions. *Hydrobiologia*. **745**: pp. 241–262.
- Kajtoch, Ł, Figarski, T. 2013. Short-term restoration of riverine bird assemblages after a severe flood. *Bird Study*. **60**: pp. 327–334.
- Kajzer-Bonk, J, Nowicki, P, Bonk, M, Skórka, P, Witek, M, Woyciechowski, M. 2013. Local populations of endangered Maculinea (Phengaris) butterflies are flood resistant. *Journal of Insect Conservation*. **17**: 1105–1112.
- Kaufman, F, Rousseeuw, PJ. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Zeszyty Naukowe UJ. John Wiley & Sons.
- Kennard, MJ, Mackay, SJ, Pusey, BJ, Olden, JD, Marsh, N. 2010a. Quantifying uncertainty in estimation of hydrologic metrics for ecohydrological studies. *River Research and Applications*. **26**: pp. 137–156.
- Kennard, MJ, Pusey, BJ, Olden, JD, Mackay, SJ, Stein, JL, Marsh, N. 2010b. Classification of natural flow regimes in Australia to support environmental flow management. *Freshwater Biology*. **55**: pp. 171–193.
- Kloskowski, J, Tanneberger, F, Marczakiewicz, P, Wiśniewska, A, Choynowska, A. 2015. Optimal habitat conditions for the globally threatened Aquatic Warbler *Acrocephalus paludicola* in eastern Poland and their implications for fen management. *Ibis*. **157**: pp. 406–412.
- Kondracki, J. 2002. *Geografia Regionalna Polski (in English: Regional Geography of Poland)*. Warszawa: PWN.
- KZGW. 2015. *Identification of the ecological flow method for Poland. Stage II Final Report - Synthesis*.
- Leasure, DR, Magoulick, DD, Longing, SD. 2016. Natural Flow Regimes of the Ozark–Ouachita Interior Highlands Region. *River Research and Applications*. **32**: pp. 18–35.

- Liermann, CAR, Olden, JD, Beechie, TJ, Kennard, MJ, Skidmore, PB, Konrad, CP, Imaki, H. 2012. Hydrogeomorphic classification of Washington state rivers to support emerging environmental flow strategies. *River Research and Applications*. **28**: pp. 1340–1358.
- Lvovich, M. 1938. *Opyt Klassifikatsii Rek SSSR (Experience of In English: Classification of Rivers of the USSR)*. Leningrad: Trudy GGI6.
- Mackay, SJ, Arthington, AH, James, CS. 2014. Classification and comparison of natural and altered flow regimes to support an Australian trial of the Ecological Limits of Hydrologic Alteration framework. *Ecohydrology*. **7**: pp. 1485–1507.
- Marsh, N. 2004. *Time Series Analysis Module: River Analysis Package*. Tech. rep. Cooperative Research Centre for Catchment Hydrology, Monash University, Melbourne Australia.
- McManamay, RA, Bevelhimer, MS, Kao, SC. 2014. Updating the US hydrologic classification: an approach to clustering and stratifying ecohydrologic data. *Ecohydrology*. **7**: pp. 903–926.
- McManamay, RA, Orth, DJ, Kauffman, J, Davis, MM. 2013. A database and meta-analysis of ecological responses to stream flow in the South Atlantic Region. *Southeastern Naturalist*. **12**: 1–36.
- Olden, JD, Kennard, MJ, Pusey, BJ. 2012. A framework for hydrologic classification with a review of methodologies and applications in ecohydrology. *Ecohydrology*. **5**: pp. 503–518.
- Olden, JD, Poff, NL. 2003. Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. *River Research and Applications*. **19**: pp. 101–121.
- Opdyke, DR, Oborny, EL, Vaughn, SK, Mayes, KB. 2014. Texas environmental flow standards and the hydrology-based environmental flow regime methodology. *Hydrological Sciences Journal*. **59**: pp. 820–830.
- Oświt, J, Dembek, W, Wszola, A. 1996. Natural and landscape values of the Upper Narew valley between Siemianówka reservoir and Suraż and the basis for their protection. *Advances in Agricultural Sciences Problem Issues*. **428**. (in Polish): pp. 103–111.
- Oueslati, O, De Girolamo, AM, Abouabdillah, A, Kjeldsen, TR, Lo Porto, A. 2015. Classifying the flow regimes of Mediterranean streams using multivariate analysis. *Hydrological Processes*. **29**: pp. 4666–4682.
- Parasiewicz, P. 2007. The MesoHABSIM model revisited. *River Research and Applications*. **23**: pp. 893–903.
- Pardé, M. 1933. *Fleuves et Rivières*. Paris: Collection Armand Collin.
- Piniewski, M, Acreman, M, Stratford, C, Okruszko, T, Gielczewski, M, Teodorowicz, M, Rycharski, M, Oświecimska-Piasko, Z. 2011. Estimation of environmental flows in semi-natural lowland rivers - the Narew basin case study. *Polish Journal of Environmental Studies*. **20**: pp. 1281–1293.
- Piniewski, M, Okruszko, T, Acreman, MC. 2014. Environmental water quantity projections under market-driven and sustainability-driven future scenarios in the Narew basin, Poland. *Hydrological Sciences Journal*. **59**: pp. 916–934.
- Piniewski, M, Szcześniak, M, Kardel, I, Berezowski, T, Okruszko, T, Srinivasan, R, Vikharnar Shuler, D, Kundzewicz, Z. Submitted. Hydrological modelling of the Vistula and Odra river basins using SWAT. *Hydrological Sciences Journal*.

- Poff, NL, Allan, JD, Bain, MB, Karr, JR, Prestegard, KL, Richter, BD, Sparks, RE, Stromberg, JC. 1997. The Natural Flow Regime. *Bioscience*. **47**: 769–784.
- Poff, NL, Richter, BD, Arthington, AH, Bunn, SE, Naiman, RJ, Kendy, E, Acreman, M, Apse, C, Bledsoe, BP, Freeman, MC, Henriksen, J, Jacobson, RB, Kennen, JG, Merritt, DM, O’Keeffe, JH, Olden, JD, Rogers, K, Tharme, RE, Warner, A. 2010. The ecological limits of hydrologic alteration (ELOHA): a new framework for developing regional environmental flow standards. *Freshwater Biology*. **55**: pp. 147–170.
- Pusłowska-Tyszewska, D, Jaroszewicz, B, Chormański, J, Piróznikow, E, Okruszko, T. 2014. Hydrologic conditions in riparian habitats of small lowland river valleys: the Narewka valley case study. *Sylvan*. **158**: pp. 132–142.
- Richter, BD, Baumgartner, JV, Powell, J, Braun, DP. 1996. A Method for Assessing Hydrologic Alteration within Ecosystems. *Conservation Biology*. **10**: pp. 1163–1174.
- Skalski, T, Kędzior, R, Radecki-Pawlik, A. 2012. Riverine ground beetles as indicators of inundation frequency of mountain stream: a case study of the Ochotnica Stream, Southern Poland. *Baltic J. Coleopterol.* **12**: pp. 117–126.
- Snelder, TH, Booker, D. 2013. Natural flow regime classifications are sensitive to definition procedures. *River Research and Applications*. **29**: pp. 822–838.
- Snelder, TH, Lamouroux, N, Leathwick, JR, Pella, H, Sauquet, E, Shankar, U. 2009. Predictive mapping of the natural flow regimes of France. *Journal of Hydrology*. **373**: pp. 57–67.
- Solans, MA, Poff, NL. 2013. Classification of the natural flow regime in the Ebro basin (Spain) by using a wide range of hydrologic parameters. *River Research and Applications*. **29**: pp. 1147–1163.
- Szewczyk, M, Dembek, W, Kamocki, A. 2005. Response of the riparian vegetation to the decrease of flooding: Narew National Park, Poland. *Ecohydrology & Hydrobiology*. **5**: pp. 79–86.
- Szoszkiewicz, K, Jusik, S, Lawniczak, AE, Zgola, T. 2010. Macrophyte development in unimpacted lowland rivers in Poland. *Hydrobiologia*. **656**: pp. 117–131.
- Tavassoli, HR, Tahershamsi, A, Acreman, M. 2014. Classification of natural flow regimes in Iran to support environmental flow management. *Hydrological Sciences Journal*. **59**: pp. 517–529.
- Tryjanowski, P, Sparks, TH, Profus, P. 2009. Severe flooding causes a crash in production of white stork (*Ciconia ciconia*) chicks across Central and Eastern Europe. *Basic and Applied Ecology*. **10**: pp. 387–392.
- Wrzesiński, D. 2014. “Detection of changes in the flow regime of rivers in Poland”. *Woda w mieście (in English: Water in the city)*. Ed. by T Ciupa, R Suligowski. Vol. 2. Monografie Komisji Hydrologicznej PTG. Instytut Geografii UJK w Kielcach: pp. 319–330.
- Wuczyński, A, Jakubiec, Z. 2013. Mortality of game mammals caused by an extreme flooding event in south-western Poland. *Natural Hazards*. **69**: pp. 85–97.
- Wyżga, B, Amirowicz, A, Radecki-Pawlik, A, Zawiejska, J. 2009. Hydromorphological conditions, potential fish habitats and the fish community in a mountain river subjected to variable human impacts, the Czarny Dunajec, Polish Carpathians. *River Research and Applications*. **25**: pp. 517–536.

- Wyźga, B, Oglęcki, P, Radecki-Pawlik, A, Skalski, T, Zawiejska, J. 2012. Hydro-morphological complexity as a driver of the diversity of benthic invertebrate communities in the Czarny Dunajec River, Polish Carpathians. *Hydrobiologia*. **696**: pp. 29–46.
- Zhang, Y, Fu, G, Sun, B, Zhang, S, Men, B. 2015. Simulation and classification of the impacts of projected climate change on flow regimes in the arid Hexi Corridor of Northwest China. *Journal of Geophysical Research: Atmospheres*. **120**. 2015JD023294: pp. 7429–7453.

Table 1: List of hydrological metrics used in this study (adapted from Marsh, 2004).

Group ¹	Code	Definition ²
1	Masr	Mean annual specific runoff ³
	Med11Mean	Median of the November flow means
	Med12Mean	Median of the December flow means
	Med01Mean	Median of the January flow means
	Med02Mean	Median of the February flow means
	Med03Mean	Median of the March flow means
	Med04Mean	Median of the April flow means
	Med05Mean	Median of the May flow means
	Med06Mean	Median of the June flow means
	Med07Mean	Median of the July flow means
	Med08Mean	Median of the August flow means
	Med09Mean	Median of the September flow means
	Med10Mean	Median of the October flow means
	CV11Mean	Variability of the November flow means
	CV12Mean	Variability of the December flow means
	CV01Mean	Variability of the January flow means
	CV02Mean	Variability of the February flow means
	CV03Mean	Variability of the March flow means
	CV04Mean	Variability of the April flow means
	CV05Mean	Variability of the May flow means
	CV06Mean	Variability of the June flow means
	CV07Mean	Variability of the July flow means
	CV08Mean	Variability of the August flow means
	CV09Mean	Variability of the September flow means
	CV10Mean	Variability of the October flow means
	Med	Median of daily flows
	CV	Coefficient variation of daily flows
Skw	Skewness of daily flows	
Var	Variability (range to median ratio) of daily flows	
S.Lg	Standard deviation of the log of daily flows	
MedAnnBFI	Median of all years Baseflow Index (calculated using Lyn-Hollick digital filter)	
VarAnnBFI	Variability of all years Baseflow Index	
2	Med_MinAsr	Median of the minimum annual specific runoff ³
	Med_MaxAsr	Median of the maximum annual specific runoff ³
	MedAnnMin	Median of all years Minimum
	MedAnnMax	Median of all years Maximum
	MedAnnP 10	Median of all years Percentile 10
	MedAnnP 90	Median of all years Percentile 90
	VarAnnMin	Variability of all years Minimum
	VarAnnMax	Variability of all years Maximum
	VarAnnP 10	Variability of all years Percentile 10
	VarAnnP 90	Variability of all years Percentile 90
	PS1YrARI	Partial series 1 Yr ARI (Average Return Interval)
PS10YrARI	Partial series 10 Yr ARI	
3	MedAnnHSNum	Median of all years Number of High Spell
	MedAnnHSLong	Median of all years Longest High Spell
	MedAnnHSPeak	Median of all years Mean of High Spell Peaks
	MedAnnHSMeanDur	Median of all years Mean Duration of High Spell
	VarAnnHSNum	Variability of all years Number of High Spell
	VarAnnHSLong	Variability of all years Longest High Spell
	VarAnnHSPeak	Variability of all years Mean of High Spell Peaks
	VarAnnHSMeanDur	Variability of all years Mean Duration of High Spell
	MedAnnLSNum	Median of all years Number of Low Spell
	MedAnnLSLong	Median of all years Longest Low Spell
	MedAnnLSPeak	Median of all years Mean of Low Spell troughs
	MedAnnLSMeanDur	Median of all years Mean Duration of Low Spell
	VarAnnLSNum	Variability of all years Number of Low Spell
	VarAnnLSLong	Variability of all years Longest Low Spell
VarAnnLSPeak	Variability of all years Mean of Low Spell troughs	
VarAnnLSMeanDur	Variability of all years Mean Duration of Low Spell	
4	MedAnnNumRise	Median of all years Number of Rises
	MedAnnMRateRise	Median of all years Mean rate of Rise
	MedAnnNumFall	Median of all years Number of Falls
	MedAnnMRateFall	Median of all years Mean rate of Fall
	VarAnnNumRise	Variability of all years Number of Rises
	VarAnnMRateRise	Variability of all years Mean rate of Rise
	VarAnnNumFall	Variability of all years Number of Falls

Continued on next page

Table 1 – Continued from previous page

Group ¹	Code	Definition ²
5	VarAnnMRateFall	Variability of all years Mean rate of Fall
	P_MDFM	Predictability based on monthly mean daily flow
	C_MDFM	Constancy based on monthly mean daily flow
	M_MDFM	Contingency based on monthly mean daily flow
	JDMaxMed	Median of all years Julian dates of maximum flows
	JDMinMed	Median of all years Julian dates of minimum flows

Note: ¹ Code: 1 - Mean flow magnitude and variability; 2 - Extreme flow magnitude and variability; 3 - Frequency and duration of high and low pulses; 4 - Rate and frequency of flow changes; 5 - Timing of extreme and seasonal flows. ² All variability metrics were calculated as inter-percentile range divided by the median. ³ These metrics are expressed in $m^3 \cdot s^{-1} \cdot km^{-2}$ and were calculated outside the River Analysis Package.

Table 2: Environmental variables description.

Code	Indicator	Category	Source
AREA	Upstream catchment area [km ²]	Size	Calculations in GIS
LOC_long*	Longitude [decimal degrees]	Geography	Calculations in GIS
LOC_lat	Latitude [decimal degrees]	Geography	Calculations in GIS
LOC_coast*	Distance to coast [km]	Geography	Calculations in GIS
PCP_ann	Mean annual precipitation [mm]	Climate	CPLFD-GDPT5
PCP_DJF*	Mean DJF precipitation [mm]	Climate	CPLFD-GDPT5
PCP_JJA*	Mean JJA precipitation [mm]	Climate	CPLFD-GDPT5
TMP_ann	Mean annual temperature [deg. C]	Climate	CPLFD-GDPT5
TMP_DJF*	Mean DJF temperature [deg. C]	Climate	CPLFD-GDPT5
TMP_JJA*	Mean JJA temperature [deg. C]	Climate	CPLFD-GDPT5
TOP_elev*	Mean elevation [m a.s.l.]	Topography	CODGiK
TOP_slp*	Mean catchment slope [-]	Topography	CODGiK
GEO_inf*	Mean annual effective infiltration into the uppermost aquifer [mm]	Geology	PIG-PIB
GEO_sand	Percentage of sandy bedrock [-]	Geology	PIG-PIB
GEO_lime*	Percentage of limestone bedrock [-]	Geology	PIG-PIB
GEO_loess*	Percentage of loess bedrock [-]	Geology	PIG-PIB
GEO_till*	Percentage of till bedrock [-]	Geology	PIG-PIB
GEO_outsand	Percentage of <i>sandr</i> bedrock [-]	Geology	PIG-PIB
SOL_coar*	Percentage of coarse-textured soils [-]	Soil	IUNG-PIB
SOL_fin	Percentage of fine-textured soils [-]	Soil	IUNG-PIB
SOL_org	Percentage of organic soils [-]	Soil	IUNG-PIB
LU_for*	Percentage of forest land [-]	Land cover	Corine Land Cover 2006
LU_agr	Percentage of agricultural land [-]	Land cover	Corine Land Cover 2006
HYD_str*	Mean stream density [km/km2]	Hydrography	IMGW-PIB
HYD_lake*	Percentage of lakes [-]	Hydrography	IMGW-PIB
HYD_sink	Percentage of sinks [-]	Hydrography	IMGW-PIB
HYD_wet*	Percentage of wetlands with water table depth above -0.5 m. [-]	Hydrography	KZGW

Abbreviations: CPLFD-GDPT5: CHASE-PL Forcing Data - Gridded Daily Precipitation and Temperature Dataset (Berezowski *et al.*, 2016), CODGiK - Central Agency for Geodetic and Cartographic Documentation (provider of the Digital Elevation Model); PIG-the effective infiltration and geological units map); IUNG-PIB - Institute of Soil Science and Plant Cultivation - National Research Institute (provider of the soil map); IMGW-PIB - Institute of Meteorology and Water Management - National Research Institute (provider of the Map of Hydrographic Division of Poland, MPHP); KZGW - National Water Management Authority (provider of the GIS layer "Groundwater-dependent land ecosystems") . * symbol next to variable code denotes that this variable has been selected for the random forest model

Table 3: Interpretation of PCA results for hydrological metrics and environmental variables.

PCA Axis ¹	Metrics/variable name ²	Interpretation (of a high value)
Hydrological metrics		
PC1 (40.7)	Med (0.97), P_MDFM (0.96), MedAnnBFI (0.95), S_Lg (-0.95), MedAnnP 10 (0.94)	High values of low flows, base-flow contribution and predictability
PC2 (17.3)	MedAnnHSNum (0.81), MedAnnLSNum (0.72), MedAnnLSLong (-0.72), Masr (0.7), Med02Mean (-0.7)	High frequency and low duration of high and low spells, high mean specific runoff and low February flow
PC3 (7.7)	Med01Mean (0.62), Med05Mean (-0.61), Med12Mean (0.58), JD-MinMed (-0.58), Med06Mean (-0.54)	High December/January flows, low May/June flows, early occurrence of minimum flow
PC4 (6.3)	M_MDFM (0.62), Med04Mean (0.57), MedAnnP 90 (0.52), VarAnnMRateRise (-0.53), VarAnnHSLong (-0.57)	High contingency, high April flows, high values of high flows, low variability in rise rate and in high spell duration
Environmental variables		
PC1 (44.1)	TOP_elevmean (0.97), PCP_JJA (0.97), TOP_elevmax (0.97), TOP_slp (0.95), PCP_ann (0.94)	High elevation and slopes, high precipitation, low temperature
PC2 (13.4)	GEO_loess (0.64), LU_agr (0.63), LU_for (-0.61), LOC_DistCoast (0.56), LOC_Lat (-0.52)	High percentage of loess bedrock, high percentage of agriculture (and low of forest), high distance to coast and low latitude
PC3 (8.4)	TMP_DJF (-0.8), LOC_Long (0.78), TMP_ann (-0.55)	Low winter and annual temperature, high longitude (east)
PC4 (7.5)	GEO_till (0.63), SOL_coars (-0.52)	High percentage of till bedrock and low percentage of coarse-textured soils

¹ Percentage of variation in hydrological metrics explained by a given PC is shown in brackets. ² Up to five metrics/variables with the highest absolute correlations (higher than 0.5) are given. Correlation coefficients of a given metric with a respective PC are shown in brackets.

Table 4: Descriptive characteristics of the distinguished flow regime classes.

Class	Monthly flow regime (cf. Fig. X)	Other hydrological metrics (cf. Fig. X)	Geography (cf. Fig. X)	Environmental variables (cf. Fig. X)
P1	Very stable regime with a peak in March and summer low flows; usually low inter-annual variability in all seasons	High predictability (due to high contingency)	Scattered across northern and eastern part of PL, with one homogeneous part clustered close to the coast	No single distinguishing feature
P2	Stable regime with a peak in March-April and summer low flows; usually low inter-annual variability in all seasons	High predictability (due to high contingency), low magnitude of 10-year flood, low mean fall and rise rate	Exclusively between 52.5 and 54 latitude N, although not as close to the coast as many gauges in class P1	High percentage of lakes and wetlands
P3	Variable regime with a peak in March and summer low flows; usually medium inter-annual variability with occasionally higher variability in summer	High value of PC3; no single metric clearly distinguishes this class (the most appropriate is Predictability)	Scattered across the whole country with exception of the far south (mountains) and far north (coast); highest frequency in the NE quarter	No single distinguishing feature
P4	Highly variable regime with a peak in March and low flows in summer and autumn; high inter-annual variability in all seasons (slightly lower in spring)	Low predictability, low specific runoff	Predominantly in the centre of PL, south of latitude 53 N	Low summer precipitation, low catchment slopes, low percentage of forests, hardly no wetlands
U5	Very stable regime with a peak in March and low flows in summer and autumn; low inter-annual variability in all seasons apart from summer (medium)	Low periodicity, low high flow magnitude	Central part of the southern PL (with exception of the far south - mountains)	High percentage of loess and limestone bedrock, low percentage of forests
M6	Variable regime with a peak in March-April and low flows in late summer and autumn; medium inter-annual variability in winter and spring and high in summer and autumn	High daily flow CV, high magnitude of 10-year flood	South of latitude 51 N, mountains and foothills	Max. elevation (lower than in M7 and higher than in remaining classes)
M7	Variable regime with a peak in April and low flows in autumn and winter; low inter-annual variability in all seasons apart from summer (medium)	High specific runoff, late minimum and maximum flow occurrence	South of latitude 51 N, highest mountain ranges	High mean/max. elevation, summer precipitation, low summer temperature

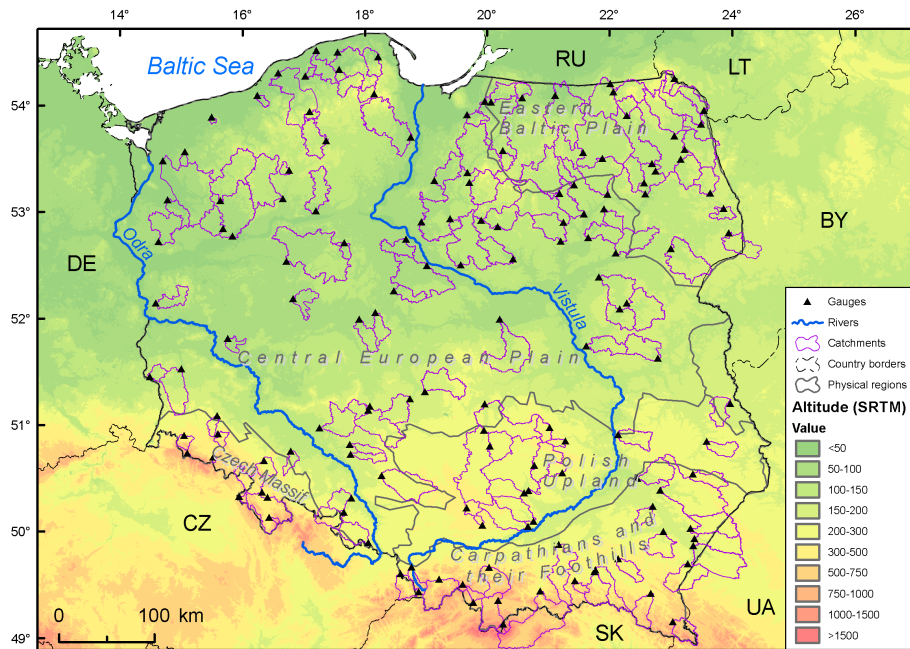


Figure 1: Study area and location of selected flow gauging stations.

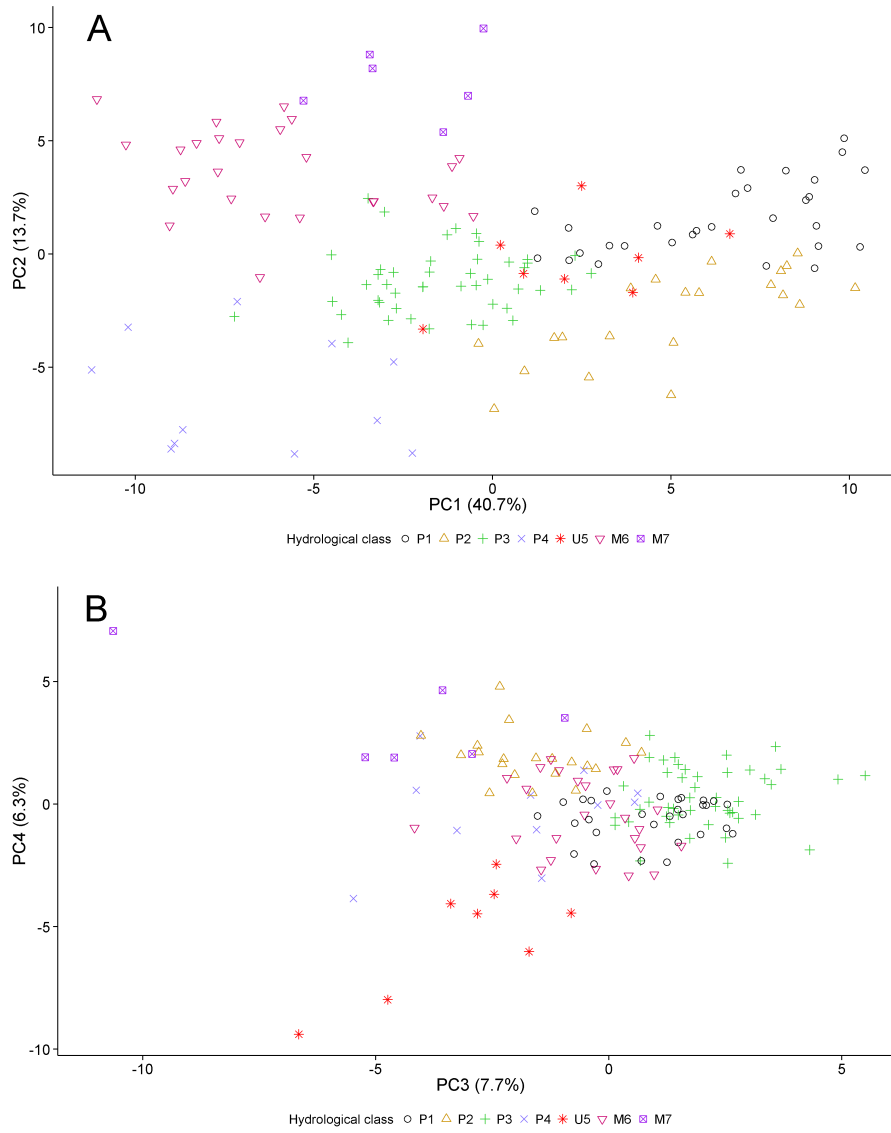


Figure 2: Projection of the cases (gauges) on 2D factor plain of two first PCs with highest loadings. Percentage of variation in hydrological metrics explained by a given PC is shown in brackets.

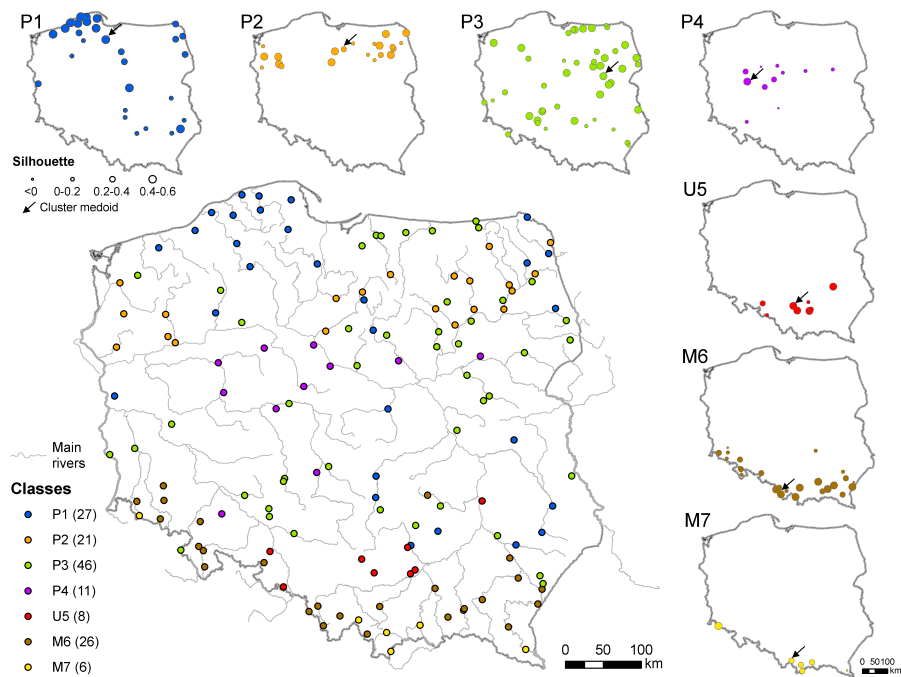


Figure 3: Spatial distribution of flow regime classes in Poland. Class counts are given in the main legend in brackets. Dot diameter in small maps is proportional to the silhouette value: the greater this value, the better given gauge is classified.

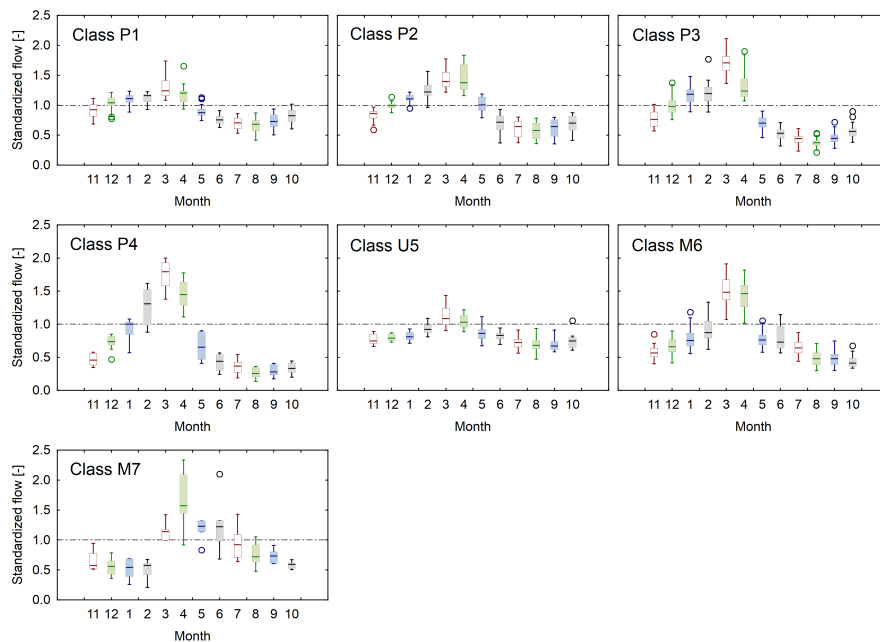


Figure 4: Standardized mean monthly flows for seven distinguished classes. Box plot variability for each month was calculated across class members.

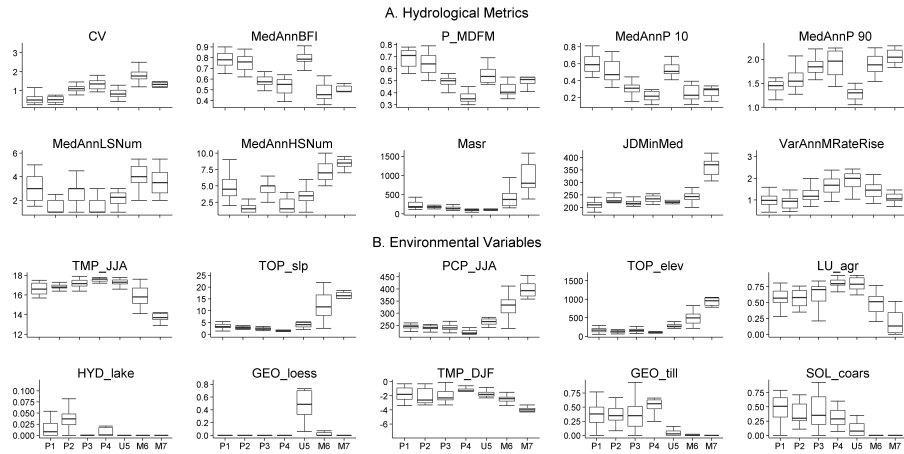


Figure 5: Box plots of selected hydrological metrics (A) and environmental variables (B) categorised by flow regime classes.

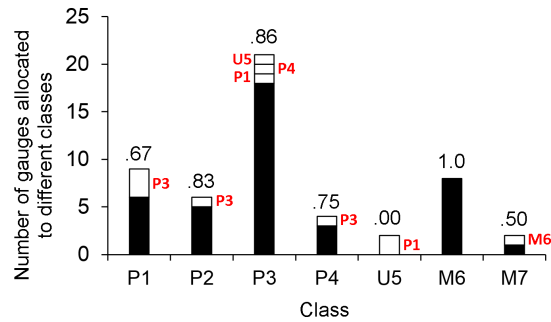


Figure 6: Predictive accuracy of the random forest model across different flow regime classes. Black and white bars refer to the number of correctly and incorrectly allocated classes, respectively. Red labels next to white bars refer to incorrectly allocated class names and the numbers above each bar denote mean accuracy for each class.