RESEARCH PAPER

IDENTIFYING CLIMATE CHANGE IMPACTS ON HYDROLOGICAL BEHAVIOR ON LARGE-SCALE WITH MACHINE LEARNING ALGORITHMS

Aleksander M. Ivanov[1], Artem V. Gorbarenko[1], Maria B. Kireeva[1*], Elena S. Povalishnikova[1]
[1]Lomonosov Moscow State University, Leninskie Gory – ГСП-1, Moscow, 119991, Russia
***Corresponding author:** kireeva_mb@mail.ru
Received: May 3rd, 2022 / Accepted: August 8th, 2022 / Published: October 01st, 2022
https://DOI-10.24057/2071-9388-2022-087

ABSTRACT. The article presents the results of study of the application of machine learning methods to the problem of classification and identification of different river water regimes in a large region – the European territory of Russia. An accumulation of hydrological observation data for the 60 – 80 years makes it possible to create an information basis for such studies. The article uses information on the average monthly runoff at 351 hydrological gauges during the period from 1945 to 2018. The most widely used data clustering approaches were used as analysis methods – K-means, EM-method, agglomerative hierarchical clustering, DBSCAN algorithms and the application of gradient boosting methods (CATBUST). Clustering and classification algorithms were given eight parameters as a basis for prediction. It was found that the most distinct and stable clusters are formed with three parameters, and the highest silhouette coefficient (SS = 0,3-0,5) is obtained using the numbers for months of the maximum and minimum runoff and the ratio of the maximum to the minimum water flow. The best result gives DBSCAN (SS = 0,6 – 0,7). Supervised classification models also show high correspondence with the reference classification, with an accuracy of 87%. Both clustering methods and classification methods showed a shift of clusters representing southern water regimes. In the central region these regimes expanded by a 1000 km to the north. Furthermore, results demonstrate that currently available data already makes it possible to apply machine learning methods to the analysis of hydrological data. Clusters corresponding to different types of water regime can be obtained by utilizing contemporary clustering algorithms. The study shows that over the past 40 years, the southern types of water regimes have noticeably shifted to the north.

KEYWORDS: Hydrological behavior, machine learning, climate change, East European Plain

CITATION: Ivanov A. M., Gorbarenko A. V., Kireeva M. B., Povalishnikova E. S. (2022). Identifying Climate Change Impacts On Hydrological Behavior On Large-Scale With Machine Learning Algorithms. Geography, Environment, Sustainability. 3(15), 80-87
https://DOI-10.24057/2071-9388-2022-087

ACKNOWLEDGEMENTS: This study was supported by the Russian Science Foundation, project no. 19-77-10032 in terms of calculations and analyzes. This research has been supported by the Interdisciplinary Scientific and Educational School of M.V.Lomonosov Moscow State University» Future Planet and Global Environmental Change» in methodology and climate change context.

Conflict of interests: The authors reported no potential conflict of interest.

INTRODUCTION

The hydrological regime of a river represents a specific pattern of changes in the state of the water body, unique for each territory (Frolova et al. 2021, 2022; Gelfan et al. 2021). The main characteristics of the hydrological regime of rivers are the character of inflow components, morphodynamic and climatic conditions (Blöschl et al. 2017; 2019; Hall and Blöschl 2018; Frolova et al. 2021; Frolova et al. 2020; Kireeva et al. 2019). The local unevenness of these values has led to the development of various methods for zoning rivers according to the types of water regime (Water regime… 2001; Frolova et al. 2021; Ayzel 2021). The study of geographic dependencies in the formation of the water regime, the analysis of the impact of economic activity – all of this is necessary to improve the existing methods of hydrological calculations. Climate change is also an important factor that affects rivers and leads to transformations and shifts in the water regime types.

One of the most important tasks of modern society is to develop resistance mechanisms and adapt to these changes (Frolova et al. 2021, 2022; Djamalov et al. 2014, 2015). In many regions, climate change has a negative impact on the quality and quantity of water resources, water temperature and the state of related ecosystems, leading to an increase in the scale and frequency of extreme natural events such as floods and droughts (Georgievsky and Shalygin 2012; Long-term fluctuations… 2021). All this, in turn, negatively affects many sectors of the economy, including agriculture, energy, fisheries, tourism and healthcare.

In addition, it is necessary to study the transformation of the water regime due to decreasing performance by currently existing methods and classifications. Most of the systematic studies devoted to the classification of the water regime of the ETR rivers were published decades ago and as of now have undergone significant changes that require a detailed analysis.

The goal of this work is to create a new model for automatic classification of ETR rivers by water regime type and assess the impact of climate change on changes in the water regime and its classes.

The abundance of up-to-date hydrometeorological information, automation of calculations, the development of different technologies and machine learning methods have made it possible to move away from general geographical patterns of classification to more modern quantitative methods. Modern approaches to the analysis of the water regime reduce the possibility of a subjective assessment of the analysis processes and allow to infor accurate numerical indicators of zoning and additional informative visualization.

Over the past 35 years, a lot of work has been done in the context of application of numerical methods to the task of water regime clustering. The pioneering work on the global data scale in this area was a 1988 paper (Haines et al. 1988). The authors set themselves the task of identifying different regions of the water regime and climatic zones algorithmically and exclusively on the basis of data. This work became possible precisely during this period due to accumulation of data on a significant scale and existence of reliable algorithms to process it. In their work, the authors used hierarchical clustering methods, and the Ochiai coefficient was taken as a distance measure. As a result of the work, the authors obtained the first map of water regimes in Western practice based only on the characteristics of the river runoff and made algorithmically. In the 2000s, a lot of new regional works appeared (Harris et al, 2000; Tavassoliet et al. 2014; Olden and Poff 2003; Kingston 2011; Brunner et al. 2018) where the authors have used algorithmic tools for assessing and classifying the water regime of rivers.

For example in (Harris et al. 2000) authors analyze the water regimes of rivers in the UK using multivariate analysis methods. Average monthly water discharges and temperature regime were separately classified in accordance with the form of their intra-annual distribution. As a result, the authors obtained 3 classes of rivers according to the shape of the hydrograph (peaks in November, December-January, March); these classes were also divided into 2 subclasses by years with different water levels: dry and high-water. In (Kidanewold et al. 2015) Ethiopian scientists have classified national rivers using daily average data and a multivariate (hierarchical) classification method. In general terms, this method is similar to that used in (Haines et al. 1988). The variables used were: average daily runoff modulus (flow water divided by the catchment area), the ratio of the average daily discharge to the average basis flow, dispersion of average daily water discharge, frequency and magnitude of abrupt changes in water discharge, average day of maximum annual discharge, and average number of days when the river dries up. As a result, 208 hydrological gauges were grouped into 3 clusters: «ephemeral» rivers that flow only after precipitation, «seasonal» rivers that flow only at certain times of the year, and «permanent» ones.

In «Classification of natural flow regimes in Iran to support environmental flow management» (Tavassoli et al. 2014), Iranian scientists have already attempted to clusterize water regimes using data from 539 stations over a period of 47 years. The data used were inferred from average daily discharges by converting it into 66 metrics used in the work (Olden and Poff 2003). Metrics were divided into the following groups: monthly water discharge and its statistical characteristics, magnitude and duration of annual maxima and minima of water discharge, dates of extreme discharges (start and end of maximum and minimum), frequency and duration of periods with the difference in average discharge by no less than a standard deviation, rate and frequency of changes in water discharge. As a clustering method, the authors chose the Bayesian mixture of distributions (Webb et al. 2007). This method works by choosing a most likely classification option out of several selected based on existing data. Each of the metrics above was modeled by the authors with a continuous normal distribution. As a result, 12 classes of rivers were obtained, while more than 90% of all stations were unambiguously assigned to any of the classes.

A radically different approach was tried by a team of authors in the article «Identification of Flood Reactivity Regions via the Functional Clustering of Hydrographs» (Brunner et al. 2018). In this work, the classification of hydrographs is carried out by the methods of functional analysis. Unlike other works described in this section, the authors are engaged in the classification of flood hydrographs. The authors propose to decompose the hydrograph into smooth functions and reduce the modeling problem to the identification of appropriate functions and parameters for them, which in total will give the actual values of the hydrograph. The authors use this approach for flood sample values from data collected in 163 Swiss watersheds. From the result obtained, the authors were able to derive three reference hydrographs corresponding to the average values of the clusters. The difference between the hydrographs was in the time that the flood lasts and the intensity of its growth.

The next qualitative leap occurred by using neural networks in this field of hydrology. In the work (Kratzert et al. 2019), the authors use a neural network with a long short-term memory (LSTM network). The method is based on a neural network of a special architecture. In the architecture of the LSTM network there are elements that are able to remember the previous state of a network node and transfer the values between layers distant from each other, which makes it possible to avoid signal blur. Among other things, the authors developed a special version of the LSTM network for solving the prediction problem, which differs in that the incoming signal is fed separately to each layer of the network. The authors were able to teach the network using the CAMELS dataset, which contains daily average precipitation, temperature, humidity, soil composition, and snowfall information for 531 gauging stations across the United States. As a result of training, the neural network was able to form an internal representation and identify two options for clustering the water regime for these gauging stations (with 5 and 6 clusters, respectively). As a result, the authors obtained clusters of water regimes geographically corresponding to: the US Northwest (Oregon and Washington), the Rocky Mountains and California, the Great Plains, the East Coast Southeast, and, finally, the East Coast Northeast and the Appalachians. The key characteristics influencing water regimes were: altitude, aridity, average daily precipitation, catchment area, presence of forests and other vegetation, and the average annual difference in the amount of green vegetation.

This model was successfully applied by the authors to the problems of flow forecasting using an input signal containing precipitation values and other meteorological variables. Finally, the team of authors from (Kratzert et al. 2019) presented the article «Accurate Hydrologic Modeling Using Less Information» (Shalev et al. 2019), in which they showed that a neural LSTM network pre-trained on the CAMELS dataset can learn to predict and classify rivers by water regime only on the basis of averaged data on discharge, temperature, and precipitation, and without taking into account average daily information on precipitation, temperature, humidity, and soil composition and snow cover. The method was applied to the data from Indian rivers. An experiment comparing the performance of models with available data on the static characteristics of the watershed (size, soil type, etc.) with a model without them demonstrated that it is possible to achieve comparable model quality without static characteristics of a watershed

It's rare to find a study with the use of machine learning methods in the problems of classifying the water regime for Russian rivers. In fact, the study (Ayzel 2021) is a unique work. A map of water regime types in the USSR is used as a class reference (Water regime... 2001). In this work the problem of reproducing the types of water regime in 1990 for the North-West of the European territory of Russia is solved by using the «random forest» class of methods on the basis of data on climatic runoff for the period 1979–2016. Then its transformation is estimated based on the calculated values of Future runoff projections (R5CH, 2006–2099) according to the three emission scenarios of the respective RCPs (RCP2.6, RCP6.0, RCP8.5). The calculations are carried out on a regular grid. As a result, the author obtained a very high classification accuracy – 91.6%, the calculations showed that by the end of the 21st century, the water regime of the rivers of the north-west of the ETR will change significantly: low periods of relatively stable water flow will become more intermittent or due to emerging rain floods or due to thaws. The second important aspect will be the transformation of the snowmelt flood – it will become significantly lower and will be observed at an earlier date.

In addition to regional studies, large-scale continental generalizations have begun to appear in recent years, using machine learning methods for problems of hydrological classification. In the work «Spatial patterns and characteristics of flood seasonality in Europe» (Hall and Blöschl 2018), a more general classification of the characteristics of the maximum runoff on the scale of the European continent was carried out. The authors took data from 4,105 measuring stations and used it to extract the maximum flow rates for each year. Then, for each station, a vector of 12 variables was constructed, where each variable corresponds to the frequency with which flood peaks occurred in that month. The K-means algorithm was chosen as the clustering algorithm, and the silhouette coefficient was chosen as the cluster quality assessment metric. The authors considered three options for the number of clusters: 4, 6, 7. As a result of the work, the authors identified 6 main clusters according to the peaks of floods, localized in geographically common subregions.

Another similar work – In «Regional classification, variability, and trends of northern North Atlantic river flow» (Kingston et al. 2011), the team of authors from (Harris et al. 2000) extended the problem of classifying water regime types to the North Atlantic. This time, instead of modeling statistical distributions, the authors used full-fledged clustering algorithms, in particular, several methods were tested:

Agglomerative hierarchical clustering with average pairwise distance metric;

Agglomerative hierarchical clustering with Ward's algorithm;

k-means method;

Agglomerative hierarchical clustering with subsequent application of the K-means method;

Using Principal Components and then Hierarchical Clustering.

The collective of authors came to the conclusion that the second approach is the most optimal in terms of the quality of the obtained clusters. The physical result of the work was the identification of seven different types of hydrographs in the region.

MATERIALS AND METHODS

Selected watersheds

Average monthly water discharges for 351 hydrological gauges located on the European territory of Russia (ETR) were used as a data source for this research. Catchment area size varied from 1000 to 200 000 square kilometers therefore both medium and large rivers were studied. The gauges were selected to cover the entire region of interest from the Far North to the arid south, including the natural zones of the tundra and forest-tundra, taiga, mixed and broad-leaved forests, forest-steppe and steppe.

Hydrological data and observation periods

Average values for monthly runoff of different rivers were used as data for this study. This dataset was created by converting the following publications into a digital format. Data for the period from 1985 to 2007 were purchased from the State Fund VNIIGMI-WDC (http://meteo.ru/). Data for the period from 2007 to 2019 were available online from the AIS GMVO (https://gmvo.skniivh.ru/). The selected parameters were calculated for each year and then averaged over two periods 1945–1977 and 1978–2019. The choice of periods is based on literary analysis, as according to the most modern studies (Long-term fluctuations... 2021; Frolova et al. 2022) in the period from 1978 and up to today hydrological systems start to display different behavior compared to historical period in response to changes in climate.

Feature selection

To carry out the analysis for each year of observation and for each river, the following hydrological characteristics were calculated:

Month number for the maximum average monthly flow (nMax) – the month in which the maximum value of water discharge was observed during the calendar year

Month number for the maximum average minimum flow (nMin) – the month in which the minimum value of water discharge was observed for the calendar year

The share of runoff volume during the spring season (dP) – was determined as the ratio of the sum of runoff volume for March, April, May to the sum of total runoff volume for the entire calendar year

Maximum average monthly discharge per year (Qmax)

Minimum average monthly discharge per year (Qmin)

The ratio of the maximum discharge to the average annual discharge (Qmax / Qyear) – the ratio of the maximum average monthly discharge for a calendar year to the average annual flow rate.

The ratio of the maximum flow to the minimum flow (Qmax / Qmin) – the ratio of the maximum average monthly flow to the minimum average monthly flow for a calendar year.

Coefficient of natural regulation (Phi) – was calculated as the ratio of the sum of the base annual runoff to the total total runoff for the year, where the base runoff is the sum of all discharge values that are less than the average. If the flow rate is greater than the average, then the average flow is used during summation instead.

During the aggregation of values for periods of 1945–1977 and 1978–2019, numerical values were averaged, and for categorical ones (i. e. nMax) a mode (most frequent value) was used.

Clusterization methods

Several algorithms were used to cluster data samples for two previously described periods by types of water regime. The K-means algorithm was first described in 1957 and has been one of most famous algorithms due to its widespread (Xu and Tian 2015). Modern versions of the algorithm

optimize its computational complexity to some extent or try to take advantage of various distance metrics. An important feature of the algorithm is the lack of guarantee to find an optimal solution in the global sense; it only finds a local one. Another disadvantage of the algorithm is the requirement to specify the number of clusters into which the data should be partitioned. Therefore this number should be inferred beforehand.

The next clustering algorithm that was used in the work is the EM-algorithm (Expectation–maximization algorithm) (Dempster et al. 1977). In general, this algorithm works similarly to the K-means algorithm. The main difference between them is that the EM-algorithm does not calculate the distance from points to centroids, but instead uses the probability that a point belongs to a particular cluster.

Hierarchical clustering, just like the K-means method, requires choosing a distance metric (usually, Euclidean one is used), but unlike the previous method, it is not that sensitive to changes in this metric. The idea of the algorithm is that a tree of elements is built and for each step of the algorithm the nearest clusters are glued together until only a single set remains. The choice of cutoff at which to stop gluing is left to the discretion of the researcher. The option where individual elements are combined into one is called agglomerative hierarchical clustering (Sasirekha and Baby 2013). The algorithm for determining the distance between the merged nodes also remains at the choice of the researcher, as a rule, the Ward criterion is used. Therefore an algorithm tries to minimize the total value of the variance within each cluster. The main advantage of this class of algorithms is the relative ease of use, which could have influenced their comparative popularity in the works of the 2000s.

Another interesting approach to perform data clustering is the DBSCAN algorithm proposed in (Schubert et al. 2017). Unlike previous algorithms, DBSCAN groups points into clusters according to the density of their distribution in space, and not according to the distances between them. Also, DBSCAN does not require a beforehand knowledge of the number of clusters that the researcher intends to obtain. The method is described in detail in (Schubert et al. 2017).

For all the methods described above their implementations in Python 3 programming language were used. Specifically, the Scikit-Learn machine learning library was used. Other libraries used in data analysis and transformation were Pandas and Numpy. Matplotlib was used as a visualization library. Data preparation consisted of analysis of the parameter variability and its limits, and structuring the data in a way appropriate for drawing maps.

The silhouette coefficient was chosen as a metric for assessing the quality of clustering, similar to (Haines et al. 1988; Hall and Bloshl 2018). The value of the silhouette coefficient S shows how similar the object is to its cluster compared to other clusters, which is described in detail in (Rousseeuw 1987).

The value of the coefficient lies between -1 and 1. The closer the score to 1 the more it indicates that the object is close to the objects in the cluster it was assigned to, and doesn't have much similarity with the objects from «foreign» clusters. If the majority of objects have a high value of this metric, then we can consider the clustering result to be of sufficient quality. If a large number of objects have low or negative silhouette coefficients, then there may be too many clusters, too few clusters, or the data simply isn't structured in a way that could be clusterized.

Classification using Gradient Boosting

During the course of work another approach was tried, gradient boosting algorithms were used to classify the types of water regimes. Unlike clustering algorithms, gradient boosting algorithms are from a family of supervised learning algorithms. First, the algorithm is trained on a labeled piece of data, and then the inferred underlying law is applied to the new data. This family of algorithms (boosting) was chosen because as of now it is a kind of an industry standard. Their implementations are:

Microsoft: LGBM algorithm;

Yandex: CatBoost algorithm;

XGboost algorithm implemented in the open package Sklearn is also widely used.

The popularity of this family of algorithms is the result of their fast speed of work and a relative ease of choosing input parameters. In fact, this algorithm is a special case of an ensemble of decision trees (i.e., a large number of decision trees are built and their average result value is taken). At each step of the algorithm, a temporary intermediate model is created and the residuals of this model are calculated (i.e., the difference between the actual value at the point and the value the algorithm returned). After that, a new ensemble of trees is created that models these residuals and the resulting model is added to the previous solution. This process goes on until the criteria specified at the start of a classification process are met (usually a set number of steps is specified). For a classification of water regime types authors chose the implementation of gradient boosting from Yandex (Prokhorenkova et al. 2019). The available sample was divided into training and test sets to assess the quality of the model. The training data set for 1945–1978 was labeled according to the Water Regime Types map (Water regime… 2001). MultiClass classification metric was chosen as the function to be optimized, i.e. a function that predicts the class of a point among several options and an overall accuracy of the model is calculated as a number of correct predictions divided by the number of datapoints. At the beginning, the authors made an attempt to simply build a model on 2000 steps, but later other parameters had to be adjusted.

The main feature of this algorithm is that it can work on relatively small amounts of data, which is a very useful feature given the amount of data used in this work. The difference between CatBoost and other gradient boosting algorithms also lies in the system for constructing decision trees. CatBoost uses absolutely symmetrical tree construction. To split a tree into branches, a certain metric is needed. In CatBoost, however, the value of the split depends on its ability to approximate the gradient vector. The splitting value is the value that is as close as possible to the gradient. According to the results of testing by the Yandex team, it was found that this mechanism really improves the quality of the algorithm. In the CatBoost algorithm, as in other algorithms, the calculation of the quality of the result is implemented for each split of the tree. The value with the best quality score in the end will be the split point of the tree. However, Yandex developers came up with the idea of adding a certain value to each quality result. This value will depend on the number of iterations passed and on the length of the gradient vector. The use of the CatBoost algorithm in this work took place in several stages, which gave different results at the output.

The basis of the metrics for assessing the quality of classification is the contingency matrix (Townsend 1971). The most common metric is accuracy, which was also used as a metric in the work (Ayzel 2021). This metric was used to evaluate the classification result, despite the fact that it has a significant drawback. It lies in the fact that it assigns the same weight to all classes, regardless of how many points fall into a particular class. However, it is the most common and frequently used metric for assessing the quality of a classification.

RESULTS AND DISCUSSION

Various combinations of parameters were tested by authors in an attempt to cluster water regime types. Three parameters out of the entire dataset were chosen as the optimal number of features to use. Authors were able to identify clusters using only Nmax, Nmin, and Qmax/Qmin. Additional characteristics did not improve the quality of clustering, but increased the instability. Among all clustering methods, K-means and DBSCAN, had the highest silhouette coefficient. By using K-means and setting the number of clusters to 8, authors acquired clusters with a silhouette score of 0.478 for the first period and 0.498 for the second one. The DBSCAN method performed much better. An algorithm found 9 clusters, 5 of significant size and 4 small ones. For the sample of data up to the year of 1978 parameters eps = 1, minPts = 3 were and the resulting silhouette coefficient was 0.610, for the sample after 1978 and parameters eps = 0.6, minPts = 3 the score was 0.720.

The distribution of points across clusters is uneven, 80% of all points fall into the three main clusters, the remaining ones account for less than 20%. At the same time, the clusters were very well localized in the geographical space, despite the fact that zonal characteristics (vegetation, soil, meteorological parameters) did not participate in any way in clustering. As a result, maps of localization of clusters obtained by the DBSCAN method for the period before and after 1977 were built, which are presented in (Fig. 1). The resulting clusters are associated with the types of water regime (TWR) on the map «Water regime of the rivers of Russia and adjacent territories» (Water regime… 2001). The dark blue cluster corresponds to the types numbered 15 and 2. The green cluster can be interpreted as types №14 and 3, and the yellow as 16. The algorithm also singled out the red cluster, which is intermediate between 14 and 16 TWR. The remaining clusters account for less than 20% of the points, of which the orange cluster corresponds to the 21st type of water regime on the map (Water regime… 2001), covering the Kuban basin, and the dark purple cluster corresponds to the 12th type, covering most of the Terek basin.

Table 1. Silhouette Score (SS) for different methods and dataframes for parameters N_{max}, N_{min}, Q_{max}/Q_{min}

Algorithm	Period	Silhouette score (SS) for N κclusters			
		N = 5	N = 6	N = 7	N = 8
K-means	Before 1978	0.438	0.463	0.468	0.478
	After 1978	0.464	0.469	0.482	0.498
EM-method	Before 1978	0.181	0.213	0.083	0.041
	After 1978	0.162	0.233	0.15	0.017
Agglomerative hierarchical clustering	Before 1978	0.423	0.428	0.432	0.437
	After 1978	0.446	0.447	0.455	0.463
		Number of clusters determined by an algorithm		Parameters	
DBSCAN	Before 1978	0.61		eps=1, minPts=3	
	After 1978	0.72		eps=0.6, minPts=3	

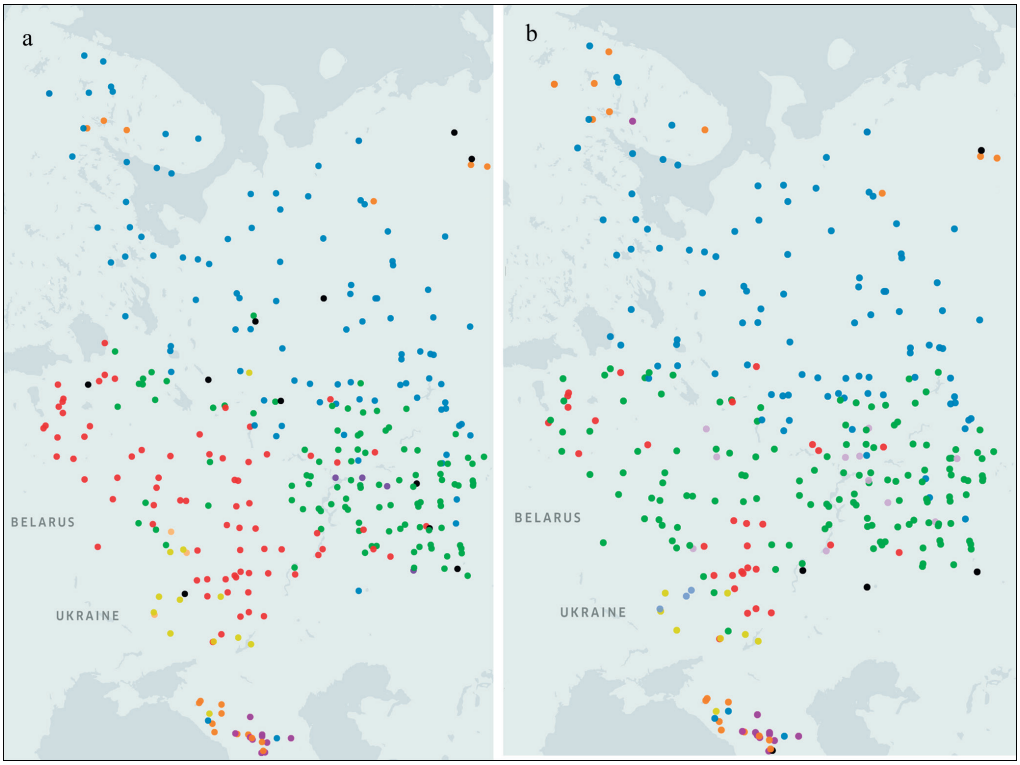
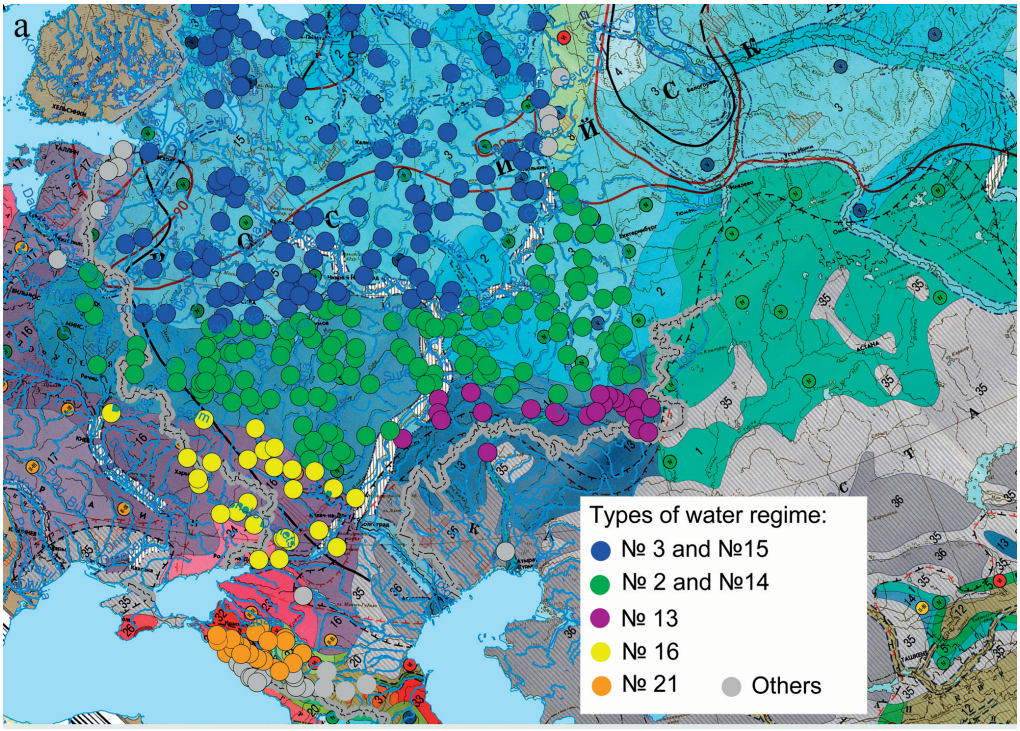


Fig. 1. Scheme of clusters for period 1945–1977 (a) and 19782019 (b) created with DBSCAN algorithm for three parameters: N_{min}, N_{max}, Q_{min}/Q_{max}

The result of clusterization largely corresponds to the map of water regime types created in (Frolova et al. 2021). An algorithm could not identify fractional clusters that differ in continentality conditions in the Central part of the Russian Plain. Figure 1 shows that for the second period there is a noticeable shift of the southern clusters; they are expanding to the north. For example, during the first period the yellow cluster mainly included points within the Seversky Donets basin, right-bank tributaries of the Lower Don. At the present stage, the yellow cluster corresponds to type 16 on the TVR map (Water regime… 2001) and already covers some tributaries of the Middle and Upper Don. The most noticeable changes affected the central zone – the red, intermediate cluster moved north by more than 1000 km, covering most of the Oka and Upper Volga basins, as well as the entire central and eastern part of the Don basin, while the initially dominant green cluster 14-15 TBP has been preserved only to the east of the Volga – in the Kama basin and partly on the Upper Volga. This result corresponds to the data obtained earlier in the work (Frolova et al. 2020), where estimates of the water regime transformation coefficient were given, and it was shown that this calculated coefficient is maximal in this region. At the same time, there is practically no shift of the green cluster to the north compared to the others, which indicates the relative stability of the water regime of the northern regions of the EPR.

The similarity of the obtained results compared to the existing map of water regime types (Water regime… 2001) suggested the possibility of using it to train the supervised model, with the aim of subsequent reproduction on a modern data set. The primary analysis of the «predicted classes» showed a low quality of classification compared to the existing map (about 0.68%). The reason behind this was an inability of the algorithm to recognize relatively similar water regime types: 2 and 14, 3 and 15, as there are relatively few data points in the sample to infer differences between them. As a result, it was decided to combine each pair into one class. After that, on the test part of the data set (1945–1977) with the parameters set to default, the accuracy of determining the type of water regime raised to 78%. This is a very good result, given the volume and quality of the data used by the algorithm. To improve the obtained values, manual selection of parameters of the CatBoost algorithm was carried out. In addition to this selection, a dynamic visualization from a CatBoost package was used to display the process of training the model. With its help, the point at which overfitting began was determined, which in turn made it possible to select the appropriate regularization parameters in order to avoid it. The quality of the algorithm reached 87% in terms of accuracy. In the field of application of machine learning, the result of metrics of 80+% is often considered good. All methods of improving the quality of the algorithm were tried: cross-validation, K-fold validation, One Hot Encoding, regularization, bagging, stacking, normalization and standardization. Subsequent improvements to the algorithm are possible only with the addition of the initial hydrological data. According to the results obtained by using the CatBoost algorithm, a water regime classification map was also built for the past and present periods (Fig. 2). Similar to the clustering, a classification algorithm wasn't able to distinguish between the water regimes for western and eastern regions of ETR. A border between rivers of «northern» (in a relative sense) «central» regions of ETR lies further to the south compared with the existing map and approximately corresponds to the Oka macrovalley. Supervised classification confirmed a noticeable shift of the more «southern» type of water regime (corresponding to No. 16 on the TBP map) to the north, but the shift of the 14th TBP to the north in the case of supervised classification was not detected as supervised training initially sets the classes to match the reference division . On the other hand, class 13 was separately identified, localized in the Ural basin, which compared to a historical period broadened to a larger area.



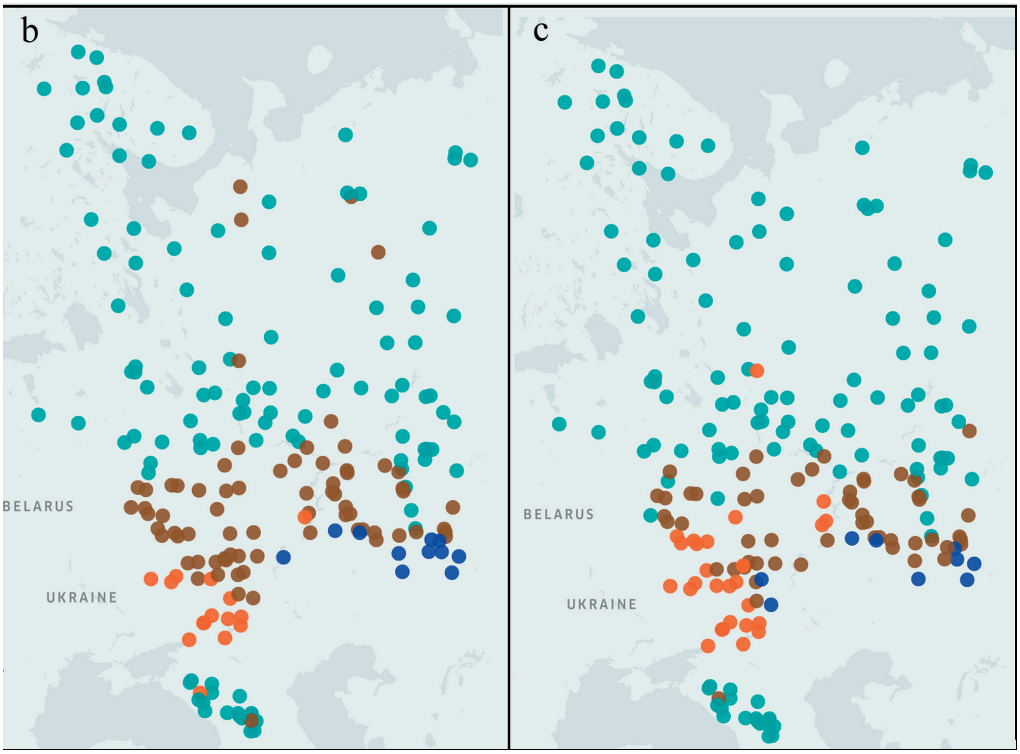


Fig. 2. Classes according to the map of Types of water regime of the rivers of the USSR with points of hydrological gauges (a), classes obtained by training on the test set 1945–1977 (b) and classes obtained using the trained model on modern data for 1978–2018 (c)

CONCLUSIONS

The results obtained allow us to formulate the following main conclusions:

The accumulated volume of hydrological data allows the use of machine learning methods in the problems of classifying water regime types.

The simplest class of methods – clustering methods shows that by selecting a combination of parameters and using data series with a length of 6070 years it is possible to obtain good results. The clustering performed by the DBSCAN method showed a high silhouette coefficient and good localization of clusters in space.

By using clustering methods, it is possible to assess the transformation of the water regime types over the past 40 years by dividing the sample into two periods.

Supervised classification models also show high correspondence with the reference classification, with an accuracy of 87%. However, the initial selection of clusters may not reveal the transitional types that are revealed by using unsupervised methods.

Both clustering methods and classification methods showed a shift of clusters representing southern water regimes. In the central region these regimes expanded by a 1000 km to the north.

REFERENCES

Ayzel G. (2021). Machine Learning Reveals a Significant Shift in Water Regime Types Due to Projected Climate Change. ISPRS Int. J. Geo-Inf., 10, 660, DOI: 10.3390/ijgi10100660.

Blöschl G., Hall J., Parajka J. et al. (2017). Changing climate shifts timing of European floods. Science, 357, 588-590, DOI: 10.1126/science.aan2506.

Blöschl G., Hall J., Viglione A., Perdigão R.A.P., Parajka J., Merz B., Lun D., Arheimer B., Aronica G.T., Bilibashi A., Boháč M., Bonacci O., Borga M., Čanjevac I., Castellarin A., Chirico G.B., Claps P., Frolova N., Ganora D., Gorbachova L., Gül A., Hannaford J., Harrigan S., Kireeva M., Kiss A., Kjeldsen T.R., Kohnová S., Koskela J.J., Ledvinka O., Macdonald N., Mavrova-Guirguinova M., Mediero L., Merz R., Molnar P., Montanari A., Murphy C., Osuch M., Ovcharuk V., Radevski I., Salinas J.L., Sauquet E., Šraj M., Szolgay J., Volpi E., Wilson D., Zaimi K., and Živković N. (2019). Changing climate both increases and decreases European river floods. Nature, 573, 10-111, DOI: 10.1038/s41586-019-1495-6.

Brunner M.I., Viviroli D. Furrer R., Seibert J., and Favre A.C. (2018). Identification of Flood Reactivity Regions via the Functional Clustering of Hydrographs. Water Resources Research, 54(3), 1852-1867, DOI: 10.1002/2017WR021650.

Dempster P., Laird N.M. and Rubin D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 1-38.

Djamalov R.G., Frolova N.L., Bugrov A.A., Grigoriev V.Yu., Igonina M.I., Kireeva M.B., Krichevets G.N., Rets E.P., Safronova T .I., Telegina A.A., Telegina E.A., and Fathi M.O. (2014). Renewable Water Resources of the European Part of Russia. Atlas. Moscow: Water Problem Institute of RAN. (In Russian).

Djamalov R.G., Frolova N.L., Kireeva M.B., Rets E.P., Safronova T.I., Bugrov A.A., Telegina A.A., Telegina E.A. (2015). Modern resources of underground and surface waters of the European part of Russia: formation, distribution, use. Moscow: GEOS. (In Russian).

Frolova N.L., Kireeva M.B., Kharlamov M.A., Samsonov T.E., Entin A.L., Lurie I.K. (2020). Mapping the current state and transformation of the water regime of the rivers of the European territory of Russia. Geodesy and Cartography, 7, 14-26, DOI: 10.22389/0016-7126-2020-961-7-14-26. (In Russian).

Frolova N.L., Magritsky D.V., Kireeva M.B., Grigoriev V.Yu., Gelfan A.N., Sazonov A.A., Shevchenko A.I. (2022). Runoff of Russian rivers under ongoing and predicted climate changes: a review of publications. 1. Assessment of changes in the water regime of Russian rivers based on observational data. Water Resources, 49(3), 251-269, DOI: 10.31857/S032105962203004X. (In Russian).

Frolova N.L., Povalishnikova E.S., Kireeva M.B. (2021). Classification and zoning of rivers by their water regime: History, methodology, and perspectives. Water Resources, 48(2), 169-181, DOI: 10.1134/s0097807821020056.

Gelfan A.N., Frolova N.L., Magritsky D.V., Kireeva M.B., Grigoriev V.Yu., Motovilov Yu.G., Gusev E.M. (2021). Influence of climate change on the annual and maximum runoff of rivers in Russia: assessment and forecast Fundamental and applied climatology, 7(1), 36-79, DOI: 10.21513/2410-8758-2021-1-36-79 (In Russian).

Georgievsky V.Yu., and Shalygin A.L. (2012). Hydrological regime and water resources. Methods for assessing the consequences of climate change for physical and biological systems, Moscow: Rosgidromet, 53-86. (In Russian).

Haines A.T., Finlayson B.L., McMahon T.A. (1988). A global classification of river regimes. Applied Geography, 8(4), 255-272, DOI: 10.1016/0143-6228(88)90035-5.

Hall J. and Blöschl G. (2018). Spatial patterns and characteristics of flood seasonality in Europe. Hydrol. Earth Syst. Sci., 22, 3883-3901, DOI: 10.5194/hess-22-3883-2018.

Harris N.M., Gurnell A.M., Hannah D.M., and Petts G.E. (2000). Classification of river regimes: a context for hydroecology. Hydrological Processes, 14(16-17), 2831-2848, DOI: 10.1002/1099-1085(200011/12)14:16/17<2831::AID-HYP122>3.0.CO;2-O.

Kidanewold B.B., Seleshi Y., Demissie S., and Melesse A.M. (2015). Flow Regime Classification and Hydrological Characterization: A Case Study of Ethiopian Rivers Water, 7, 3149-3165, DOI: 10.3390/w7063149.

Kingston D.G., Hannah D.M., Lawler D.M., and McGregor G.R. (2011). Regional classification, variability, and trends of northern North Atlantic river flow. Hydrol. Processes, 25(7), 1021-1033.

Kireeva M., Frolova N., Rets E., Samsonov T., Entin A., Kharlamov M., Telegina E., Povalishnikova E. (2019). Evaluating climate and water regime transformation in the European Part of Russia using observation and reanalysis data for the 1945-2015 period. International Journal of River Basin Management, 18(4), 1-12, DOI: 10.1080/15715124.2019.1695258.

Kratzert F., Klotz D., Shalev G., Klambauer G., Hochreiter S. and Nearing G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci., 23, 5089-5110, DOI: 10.5194/hess-23-5089-2019.

Long-term fluctuations and variability of water resources and the main characteristics of the flow of rivers in the Russian Federation. Scientific and Applied Handbook. (2021). St. Petersburg: LLC «RIAL».

Olden J.D., and Poff N.L. (2003). Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River Research and Applications, 19, 101-121, DOI: 10.1002/rra.700.

Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. (2019). CatBoost: unbiased boosting with categorical features. Available at: https://arxiv.org/pdf/1706.09516.pdf%20https://tech.yandex.ru/catboost/ [Accessed Apr. 2020].

Rousseeuw J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 5365, DOI: 10.1016/0377-0427(87)90125-7.

Sasirekha K. and Baby P. (2013). Agglomerative hierarchical clustering algorithm. A review. International Journal of Scientific and Research Publications, 83, 83.

Schubert E., Sander J., Ester M., Kriegel H. P., and Xu X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 1-21, DOI: 10.1145/3068335.

Shalev G., El-Yaniv R., Klotz D., Kratzert F., Metzger A., Nevo S. (2019). Accurate Hydrologic Modeling Using Less Information. Mathematics, Computer Science, Available at: https://arxiv.org/pdf/1911.09427.pdf.

Tavassoli H.R., Tahershamsi A. and Acreman M. (2014). Classification of natural flow regimes in Iran to support environmental flow management. Hydrological Sciences Journal, 59(3-4), 517-529, DOI: 10.1080/02626667.2014.890285.

Townsend J.T. (1971). Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics, 9(1), 40-50.

Water regime of the rivers of Russia and adjacent territories. Map for higher educational institutions. (2001). Scale 1:8 000 000, V.M. Evstigneev, N.V. Shenberg, N.V. Anisimova, A.A. Zaitsev (Eds.), Novosibirsk, Roskartografia Cartographic Factory. (In Russian).

Webb J.A., Bond N.R., Wealands S.R., Nally R.M., Quinn G.P., Vesk P.A., and Grace M.R. (2007). Bayesian clustering with AutoClass explicitly recognises uncertainties in landscape classification. Ecography, 30, 526-536, DOI: 10.1111/j.0906-7590.2007.05002.x.

Xu D., and Tian Y. (2015). A Comprehensive Survey of Clustering Algorithms. Ann. Data. Sci., 2, 165-193, DOI: 10.1007/s40745-015-0040-1.