Towards the Unified Approach for Obtaining Hydro-Meteorological and Landscape Characteristics for River Catchments

Dmitriy Abramov[a], Georgy Ayzel[a,b] and Oleg Nikitin[c]

[a] Hydrological State Institute, Vasilyevsky Island, 2nd line, 23, St. Petersburg, 199004, Russian Federation
[b] Institute for Environmental Sciences and Geography, University of Potsdam, Potsdam 14476, Germany
[c] Computing Center of the Far Eastern Branch of the Russian Academy of Sciences, 65 Kim Yu Chena Ulitsa, Khabarovsk, 680000, Russian Federation

Abstract

A hydrological catchment is a complex product that is formed and evolves under the interaction of many processes. In general, these processes reflect and could be represented in a set of different geophysical parameters. Modern numerical hydrological models, both physically based and data-driven, benefit from the assimilation of such catchment parameters that allow them a closer representation of river runoff formation processes. However, no readily available tool allows us to obtain the same sets of geophysical parameters for any catchment across the globe. To fill this gap, here we present featureXtractor — an open, unified approach, and reproducible set of scripts for obtaining the large set of catchment properties [1]. It interacts with the open database HydroATLAS and aggregates different sets of hydrological, physiographic, climatic, land cover, soil, and anthropogenic parameters; then, it stores it in a user-defined format. Thus, any catchment across the globe could be represented with a consistent set of descriptors that opens a new way towards large-scale hydrological modeling and applications.

Keywords 1

Geophysical parameters extraction, HydroATLAS database, open source

1. Introduction

Hydrological processes are characterized by high spatio-temporal variability. From place to place, respective directions of transformation of precipitation into runoff occur in different ways. There are two primary sources of these differences: (1) the various dominant geophysical parameters that characterize the hydrological catchment and (2) meteorological forcing. Together, these factors determine the behavior of hydrological catchments in terms of specific runoff formation patterns and regions [2, 3, 4].

At the moment, several projects have been focused on collecting and aggregating universal sets of geophysical parameters and meteorological forcing for advancing large-scale hydrology and the respective development of hydrological models. Among the most well-known are CAMELS [5], LAMAH [6], MOPEX[7], CANOPEX [8]. However, each project operates at a specific region, uses a unique set of input data, thus utilizes different sets of catchment parameters (as well as tools to acquire them) and meteorological forcing. All these limit projects' comparability. Thus, after almost three decades of research in the field of large-scale hydrological modeling, there is no tool for obtaining a consistent set of catchment parameters that could be particularly beneficial for a research

VI International Conference Information Technologies and High-Performance Computing (ITHPC-2021),
September 14–16, 2021, Khabarovsk, Russia

EMAIL: dmbrmv96@gmail.com (A. 1); hydrogo@yandex.ru (A. 2); olegioner@gmail.com (A. 3);
ORCID: 0000-0003-2682-8722 (A. 1); 0000-0001-5608-9110 (A. 2); 0000-0001-9139-5180 (A. 3);

community. While obtaining meteorological data is generally solved by using reanalysis data, there is no such data introduced for obtaining catchment descriptors yet.

To fill the gap, we propose to use the HydroATLAS [9] database as the latest effort and the most up-to-date and state-of-the-art compilation of various geophysical datasets at different catchment levels.

This paper presents computational workflows and the open-source tool — featureXtractor — which allows aggregating different sets of geophysical parameters from the HydroATLAS database. As a case study, we demonstrate an application of the developed tool for deriving catchment properties for 1018 catchments included in the OpenForecast v2 system [10].

2. Data and Methods

2.1. HydroATLAS

HydroATLAS was chosen as a source database for several reasons:

1. First, based on the diversity of data that HydroATLAS database offers for users. There are 56 environmental variables that are partitioned into 281 individual attributes and organized into 6 categories: hydrology; physiography; climate; land cover & use; soils & geology; and anthropogenic influences.
2. The second reason is the global availability of data. HydroATLAS derives the hydro-environmental characteristics by aggregating and reformatting original data from well-established global digital maps, and by accumulating them along the drainage network from headwaters to ocean outlets. Hierarchically nested sub-basins are linked to attributes at multiple scales, as well as the individual river reaches, both extracted from the global HydroSHEDS database [hydrosheds] at 15 arc-second (~500 m) resolution. The sub-basin and river reach information is distributed in two companion datasets: BasinATLAS and RiverATLAS. The BasinATLAS dataset will be further utilized as a source dataset.
3. The third reason is the uniformity and consistency of data stored in companion datasets. BasinATLAS stores data in shapefiles that correspond to the individual sub-basin number of hydro-environmental characteristics. In this way, that allows us to automate the process of data extraction and preprocessing.

Environmental attributes from the HydroATLAS database are stored in six different categories: hydrology; physiography; climate; land cover & use; soils & geology; and anthropogenic influences. However, we reduce the number of considered attributes from 281 (in the original dataset) to 149. The reduction has been determined by the expert screening that defines the suitability of available characteristics for further use in hydrological modelling studies. The table with original BasinATLAS and expert-guided variants of datasets alongside auxiliary information is available on GitHub.

2.2. Research Catchments

For the test case study, we select research catchments of rivers across the Russian Federation with areas from 50 to 50 000 km². In total, our study includes 1018 catchments from the OpenForecast system [10]. The boundaries of the respective catchment are stored in shapefiles, allowing us to manipulate them in a programmatic way using specialized software libraries. Boundaries’ availability is the sole requirement for the developed computational scripts. Thus, the provided approach for feature extraction (Figure 1) allows us to prepare the unified and consistent set of catchment descriptors for any river catchment across the globe with the digitized boundary.

2.3. Computational workflow

Figure 1 illustrates the general concepts of the proposed computational workflow for deriving the unified set of catchment descriptors.

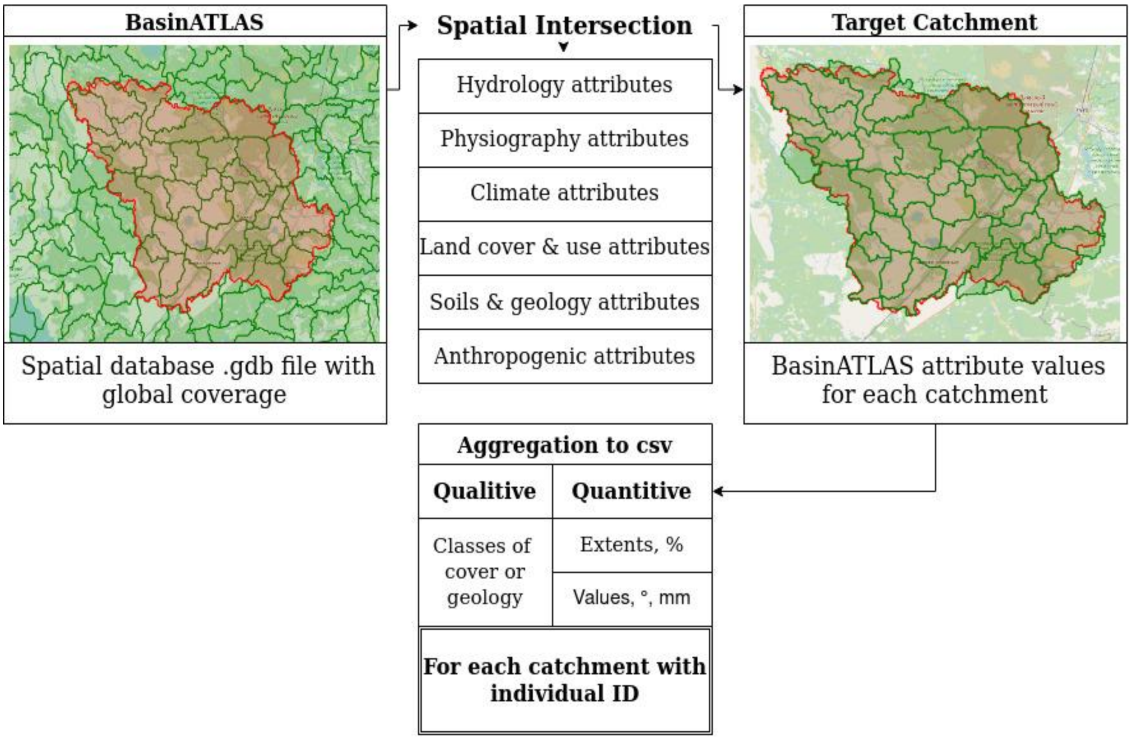
BasinATLAS
Spatial database .gdb file with global coverage
Spatial Intersection
Hydrology attributes
Physiography attributes
Climate attributes
Land cover & use attributes
Soils & geology attributes
Anthropogenic attributes
Target Catchment
BasinATLAS attribute values for each catchment
Aggregation to csv
Qualitive
Quantitive
Classes of cover or geology
Extents, %
Values, °, mm
For each catchment with individual ID

Figure 1: The illustration of the proposed computational workflow

To perform calculations, we need two inputs: (1) a shape boundary of the analyzed catchment and (2) a pre-downloaded BasinATLAS dataset [11]. Then, the workflow is as follows.

1. With the usage of Fiona [12] and GeoPandas [13] libraries, we read the shapefiles of boundaries and shapefiles of BasinATLAS datasets.
2. After reading the data, we start to perform intersection procedures based on the use of GeoPandas and shapely libraries' spatial functions. We use the sub-basin layer with the highest spatial resolution from the BasinATLAS dataset, where individual basin splits are approximately 50 sq. km.
3. To perform an intersection, there is one more step required. Before assigning sub-basin characteristics from BasinATLAS to the targeted basin, it is necessary to ensure that the sub-basin intersects enough with it. Thus, we calculate the fraction of the sub-basin and the intersection of it with the target basin. If the intersection share is more than 0.2, then the considered sub-basin could be included as characterizing.
4. After the intersection procedure, the next step is calculating aggregated values and splitting them to separate datasets based on their affiliation. The attributes, in general, can be divided into two types of data: qualitative (land cover, lithological classes) and quantitative (air temperature, extents of different characteristics). To aggregate quantitative attributes, a weighted mean was used. For the aggregation of the qualitative attributes, we use the spatial majority, i.e., we assign the most popular class from sub-basins as the descriptor of the whole target catchment.
5. After aggregation, the individual results are separated into separate sets describing: hydrology, physiography, climate, land cover & use, soils & geology, and anthropogenic characteristics.
6. To ensure linear speed up the calculations, the workflow has been parallelized using the standard multiprocessing library. The use of 8 threads (CPU: Intel 10700k) achieved a computational time of 1 hour for 1018 analyzed catchments.
7. After the main calculation procedure, the final results could be saved in any user-defined format available in standard pandas functionality (e.g., .csv, .tsv, .xls).

The resulting computational script — featureXtractor — is written in Python programming language[14] and is entirely based on open and freely distributed software packages: NumPy [15], pandas [16], geopandas [13] shapely [17]. It is available and ready to use in the GitHub repository [1] under the MIT license.

3. Results

The proposed script returns six files according to each category from the BasinATLAS dataset. Each file represents individual attributes as columns that simplify the further analysis. The first column of every file represents the unique basin ID. That is the anchor which builds a relation between files of different categories.

Figure 2 shows the spatial distribution of the OpenForecast basin dataset and the number of attributs obtained from the computation.

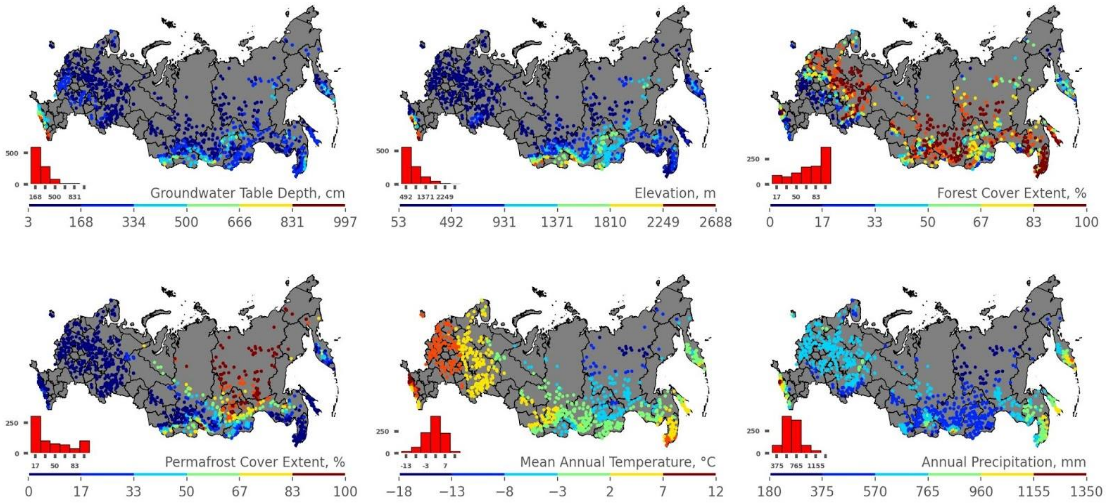

Figure 2: The distribution of values for characteristics of Groundwater Table Depth, Elevation, Forest Cover Extent, Permafrost cover extent, mean annual temperature, and annual precipitation. The histograms indicate the number of catchments (out of 1018) in each bin or category.

The distribution of the analyzed environmental variables (Figure 2) gives a reliable representation of features' spatial heterogeneity across the analyzed catchments. Also, the analyzed features correspond to specific landscapes and geographic regions. All obtained results and the code for their analysis and visualization are available in the GitHub repository [1].

4. Conclusion and Outlook

We introduced a universal tool and unified approach for obtaining an extensive, descriptive, and consistent set of hydro-climatic and landscape characteristics. The presented tool is a state-of-the-art and readily available swiss-knife for obtaining the set of catchment attributes for any river catchment across the globe. This tool was tested using 1018 river catchments on the territory of the Russian Federation and proved its efficiency for obtaining input data that is usually required for large-scale hydrological studies. The obtained wide range of geophysical characteristics opens new opportunities to quantitatively explore how the interplay between topography, climate, land cover, soil, and geology shapes hydrological behavior. Global coverage of the BasinATLAS dataset and open-source approach of the presented tool enables a possibility to test any hypothesis about the hydrological system functioning based on the consistent sample of catchment attributes available for any river catchment across the globe.

The field of hydrological modeling benefits from the introduced instrument. Modern data-driven models for runoff formation could assimilate the representation of catchment attributes while optimizing their parameters could lead to more reliable results [4]. Also, the vector of catchment attributes could provide deeper insights into hydrological processes that underlie runoff formation mechanisms.

Last but not least. We urge that research reproducibility brings benefits for a broad range of specialists. Thus, the developed tool makes hard-to-obtain data of catchment attributes easily accessible yet consistent and reliable. In this way, featureXtractor democratizes research in hydrological modeling, making one of the research-intensive procedures — data preparation — available for a broad community that wants to push forward citizen science.

5. Acknowledgements

The reported study was funded by RFBR, project number 19-35-60005.

The studies were carried out using the resources of the Center for Shared Use of Scientific Equipment "Center for Processing and Storage of Scientific Data of the Far Eastern Branch of the Russian Academy of Sciences", funded by the Russian Federation represented by the Ministry of Science and Higher Education of the Russian Federation under project No. 075-15-2021-663.

6. References

[1] Dmitriy A., Ayzel G., featureXtractor, (2021), GitHub repository, URL: https://github.com/dmbrmv/featureXtractor

[2] Glushkov V. G. : “Geographic-hydrological method.” Proc. of SHI, No. 57-58 (1933) [in Russian].

[3] Grigoriev A.A., Budyko M.I. “On the periodic law of geographic zoning” Reports of the USSR Academy of Sciences. 1956. vol. 110. № 1. p. 129–132

[4] Kratzert, F., Daniel K., Guy S., Günter K., Sepp H., and Grey N.. “Towards Learning Universal, Regional, and Local Hydrological Behaviors via Machine Learning Applied to Large-Sample Datasets.” Hydrology and Earth System Sciences 23, no. 12 (December 17, 2019): 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

[5] Addor, N., Andrew J. N., Naoki M., and Martyn P. C.. “The CAMELS Data Set: Catchment Attributes and Meteorology for Large-Sample Studies.” Hydrology and Earth System Sciences 21, no. 10 (October 20, 2017): 5293–5313. https://doi.org/10.5194/hess-21-5293-2017.

[6] Klingler, C, Karsten S., and Mathew H. “LamaH | *La*Rge-Sa*m*Ple D*a*Ta for *H*Ydrology and Environmental Sciences for Central Europe.” *Earth System Science Data Discussions*, March 18, 2021, 1–46. https://doi.org/10.5194/essd-2021-72.

[7] Duan, Q., J. Schaake, V. Andréassian, S. Franks, G. Goteti, H. V. Gupta, Y. M. Gusev, et al. “Model Parameter Estimation Experiment (MOPEX): An Overview of Science Strategy and Major Results from the Second and Third Workshops.” *Journal of Hydrology*, The model parameter estimation experiment, 320, no. 1 (March 30, 2006): 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031.

[8] Arsenault, R., Rachel B., Camille O. D., and François B.. “CANOPEX: A Canadian Hydrometeorological Watershed Database.” *Hydrological Processes* 30, no. 15 (2016): 2734–36. https://doi.org/10.1002/hyp.10880.

[9] Linke, S., Bernhard L., Camille O. D., Joseph A., Günther G., Mira A., Penny B., et al. “Global Hydro-Environmental Sub-Basin and River Reach Characteristics at High Spatial Resolution.” *Scientific Data* 6, no. 1 (December 9, 2019): 283. https://doi.org/10.1038/s41597-019-0300-6.

[10] Ayzel, G.. “OpenForecast v2: Development and Benchmarking of the First National-Scale Operational Runoff Forecasting System in Russia.” Hydrology 8, no. 1 (March 2021): 3. https://doi.org/10.3390/hydrology8010003.

[11] Lehner, B.; Linke, S.; Thieme, M. (2019): HydroATLAS version 1.0. figshare. Dataset. https://doi.org/10.6084/m9.figshare.9890531.v1

[12] Fiona is GDAL's neat and nimble vector API for Python programmers, 2021, URL: https://pypi.org/project/Fiona/

[13] Kelsey Jordahl, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, François Leblanc. (2020, July 15). geopandas/geopandas: v0.8.1 (Version v0.8.1). Zenodo. http://doi.org/10.5281/zenodo.3946761

[14] Python Core Team, Python Programming Language, 2021, URL: https://www.python.org/

[15] Harris, C. R., Jarrod M., Stéfan J. van der Walt, Gommers R.,Virtanen P., Cournapeau D., Wieser E., et al. "Array Programming with NumPy." Nature 585, no. 7825 (September 2020): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

[16] Reback, J.,McKinney W., Van den Bossche J., Augspurger T., Cloud P., et al. Pandas-Dev/Pandas: Pandas 1.0.3. Zenodo, 2020. https://doi.org/10.5281/zenodo.3715232.

[17] Gillies S. and others, toblerity.org, "Shapely: manipulation and analysis of geometric objects", 2021, URL: https://github.com/Toblerity/Shapely