The Journal of Open Source Software

rabpro: global watershed boundaries, river elevation profiles, and catchment statistics

Jon Schwenk[1¶], **Tal Zussman**[2], and **Jemma Stachelek**[1]

1 Los Alamos National Laboratory, Division of Earth and Environmental Sciences **2** Columbia University, Department of Computer Science ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Chris Vernon

Reviewers:

- @thurber
- @atrcheema

Submitted: 28 February 2022
Published: unpublished

Summary

River and Basin Profiler (`rabpro`) is a Python package to delineate watersheds, extract river flowlines and elevation profiles, and compute watershed statistics for any location on the Earth's surface. As fundamental hydrologically-relevant units of surface area, watersheds are areas of land that drain via aboveground pathways to the same location. Delineations of watershed boundaries are typically performed on DEMs that represent surface elevations as gridded rasters. Depending on the resolution of the DEM and the size of the watershed, delineation may be very computationally expensive. With this in mind, we designed `rabpro` to provide user-friendly workflows to manage the complexity and computational expense of watershed calculations given an arbitrary coordinate pair. In addition to basic watershed delineation, `rabpro` will extract the elevation profile for a watershed's main-channel flowline. This enables the computation of river slope, which is a critical parameter in many hydrologic and geomorphologic models. Finally, `rabpro` provides a user-friendly wrapper around Google Earth Engine's (GEE) Python API to enable cloud-computing of zonal watershed statistics and/or time-varying forcing data from hundreds of available datasets. Altogether, `rabpro` provides the ability to automate or semi-automate complex watershed analysis workflows across broad spatial extents.

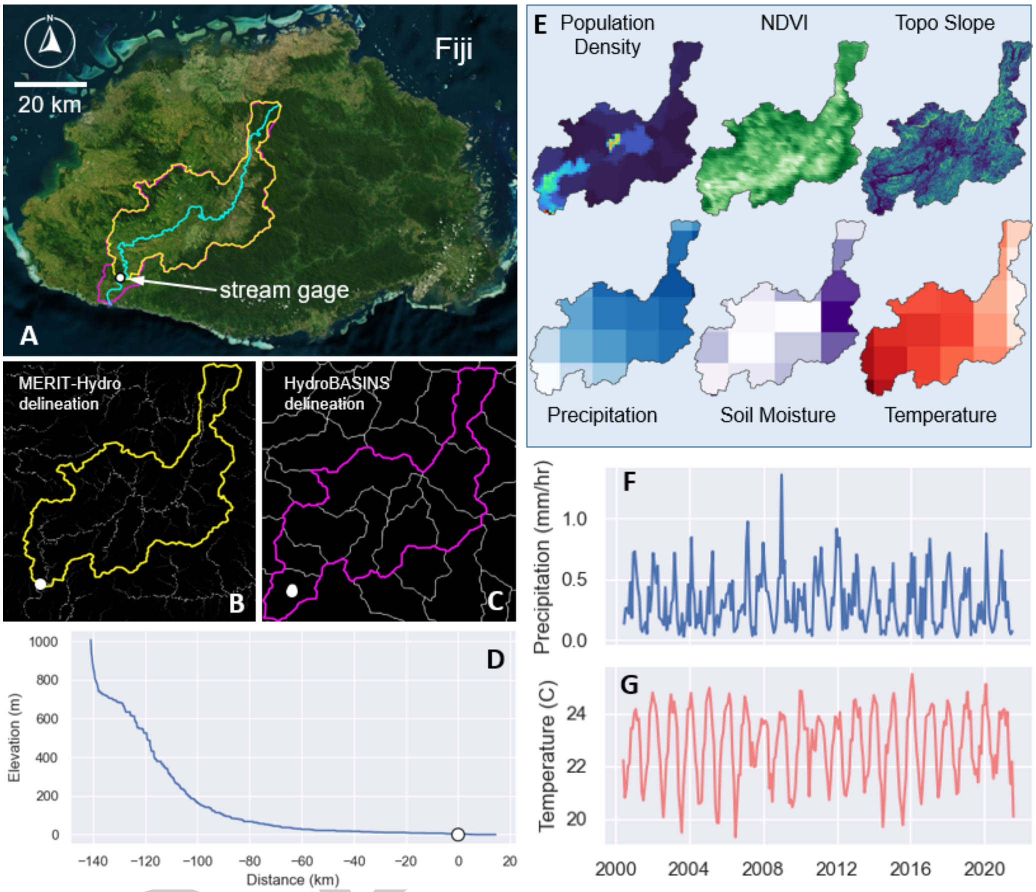

Figure 1: The core functionality of rabpro demonstrated on the Sigatoka River. (A) Study site with both MERIT and HydroBASINS delineations and river flowline extraction for a hypothetical gage station. Bing VirtualEarth base image. (B) MERIT-Hydro delineation with MERIT-Hydro flowlines underneath. (C) HydroBASINS delineation with level-12 HydroBASINS polygons as white outlines. (D) Extracted elevation profile with gage location denoted by white circle at Distance = 0. (E) Examples of time-averaged (where appropriate) basin characteristics retrieved by rabpro from Google Earth Engine. Sources: population (**center_for_international_earth_science_information_network-ciesin-columbia_university_gridded_2017?**), NDVI (Didan, 2015), topo slope (Amatulli et al., 2020), precipitation (**precipitation_processing_system_pps_at_nasa_gsfc_gpm_2019?**), soil moisture (**oneill__peggy_e_smap_2018?**), and temperature (Service, 2017). (F, G) Basin-averaged time-series data fetched by rabpro for the temperature and precipitation datasets in (E).

Statement of Need

Watersheds play a central and vital role in many scientific, engineering, and environmental management applications (See Brooks (2003) for a comprehensive overview). While `rabpro` can benefit any watershed-based research or analysis, it was designed to satisfy the needs of data-driven rainfall-runoff models. These models aim to predict streamflow (runoff) time series as a function of precipitation over upstream land area (i.e. the watershed). In addition to watershed delineations and precipitation estimates, they typically require data on both time-varying parameters (or forcing data) like temperature, humidity, soil moisture, and vegetation as well as static watershed properties like topography, soil type, or land use/land cover (Gauch et al., 2021; Kratzert et al., 2019, 2021; Nearing et al., 2021). The `rabpro` API enables users to manage the complete data pipeline necessary to drive such a model starting from the initial watershed delineation through the calculation of static and time-varying parameters. Some hydrologic and hydraulic models also require channel slope for routing streamflow (Boyle et al., 2001; Piccolroaz et al., 2016; Wilson et al., 2008), developing rating curves (Colby, 1956;

Fenton & Keller, 2001), or modeling local hydraulics (Schwenk et al., 2017, 2015; Schwenk & Foufoula-Georgiou, 2016).

The need for watershed-based data analysis tools is exemplified by the growing collection of published datasets that provide watershed boundaries, forcing data, and/or watershed attributes in precomputed form, including CAMELS (Addor et al., 2017), CAMELS-CL, -AUS, and -BR (Alvarez-Garreton et al., 2018; Chagas et al., 2020; Fowler et al., 2021), Hysets (Arsenault et al., 2020), and HydroAtlas (Linke et al., 2019). These datasets provide off-the-shelf options for building streamflow models, but they suffer from a degree of inflexibility. For example, someone desiring to add a watershed attribute, to use a new remotely-sensed data product, or to update the forcing data time-series to include the most recently available data must go through the arduous process of sampling it themselves. `rabpro` was designed to provide flexibility for both building a watershed dataset from scratch or appending to an existing one.

While we point to streamflow modeling as an example, many other applications exist. `rabpro` is currently being used to contextualize streamflow trends, build a data-driven model of riverbank erosion, and generate forcing data for a mosquito population dynamics model. `rabpro`'s focus is primarily on watersheds, but some users may also find `rabpro`'s Google Earth Engine wrapper convenient for sampling raster data over any geopolygon(s).

State of the field

The importance of watersheds, availability of DEMs, and growing computational power has led to the development of many excellent open-source terrain (DEM) analysis packages that provide watershed delineation tools, including TauDEM (Tarboton, 2005), pysheds (Bartos, 2020), [Whitebox Tools][https://github.com/jblindsay/whitebox-tools](Lindsay (2016)), SAGA (Conrad et al., 2015), among many others. Computing statistics and forcing data from geospatial rasters also has a rich history of development, and Google Earth Engine (Gorelick et al., 2017) has played an important role. Almost a decade has passed since Google Earth Engine has been available to developers, and the community has in-turn developed open-source packages to interface with its Python API in user-friendlier ways, including gee_tools (Principe, 2021), geemap (Wu, 2020), eemont (Montero, 2021), and restee (Markert, 2021)–each of which provides support for sampling zonal statistics and time series from geospatial polygons.

However, to our knowledge, `rabpro` is the only available package that provides efficient end-to-end delineation and characterization of watershed basins at scale. While a combination of the cited terrain analysis packages and GEE toolboxes can achieve `rabpro`'s functionality, `rabpro`'s blending of them enables simpler, less error-prone, and faster results.

One unique `rabpro` innovation is its automation of "hydrologically addressing" input coordinates. DEM watershed delineations require that the outlet pixel be precisely specified; in many `rabpro` use cases, this is simply a (latitude, longitude) coordinate that may not align with the underlying DEM. `rabpro` will attempt to "snap" the provided coordinate to a nearby flowline while minimizing the snapping distance and the difference in upstream drainage area (if provided by the user). Another unique `rabpro` feature is the ability to optimize the watershed delineation method according to basin size such that pixel-based (from MERIT-Hydro (Yamazaki et al., 2019)) delineations can be used for more accurate estimates and/or smaller basins, and coarser subbasin-based (from HydroBASINS (Lehner & Grill, 2014)) delineations can be used for rapid estimates of larger basins.

Functionality

`rabpro` executes watershed delineation based on either the MERIT-Hydro dataset, which provides a global, ~90 meter per pixel, hydrologically-processed DEM suite, or the HydroBASINS data product, which provides pre-delineated subbasins at approximately ~230 km^2 per subbasin. Conceptually, basin delineation is identical for both. The user-provided coordinate is hydrologically addressed by finding the downstream-most pixel (MERIT-Hydro) or subbasin

(HydroBASINS). The watershed is then delineated by finding all upstream pixels or subbasins that drain into the downstream pixel/subbasin and unioning these pixels/subbasins into a single polygon. A user must therefore download either the MERIT-Hydro tiles covering their study watershed or the appropriate HydroBASINS product; `rabpro` provides tooling to automate these downloads and create its expected data structure (See the Downloading data notebook).

There are three primary operations supported by `rabpro`: 1) basin delineation, 2) elevation profiling, and 3) subbasin (zonal) statistics. If operating on a single coordinate pair, the cleanest workflow would be instantiating an object of the `profiler` class and calling (in order) the `delineate_basins()`, `elev_profile()`, and `basin_stats()` methods (See the Basic Example notebook). If operating on multiple coordinate pairs, the workflow would loop through each coordinate pair while delineating each watershed (optionally calculating its elevation profile). As the loop runs, the user collects each basin polygon in a list, concatenates the list, and directly calls `subbasin_stats.compute()` on the resulting GeoDataFrame (See the Full Example notebook). More details on package functionality can be found in the documentation.

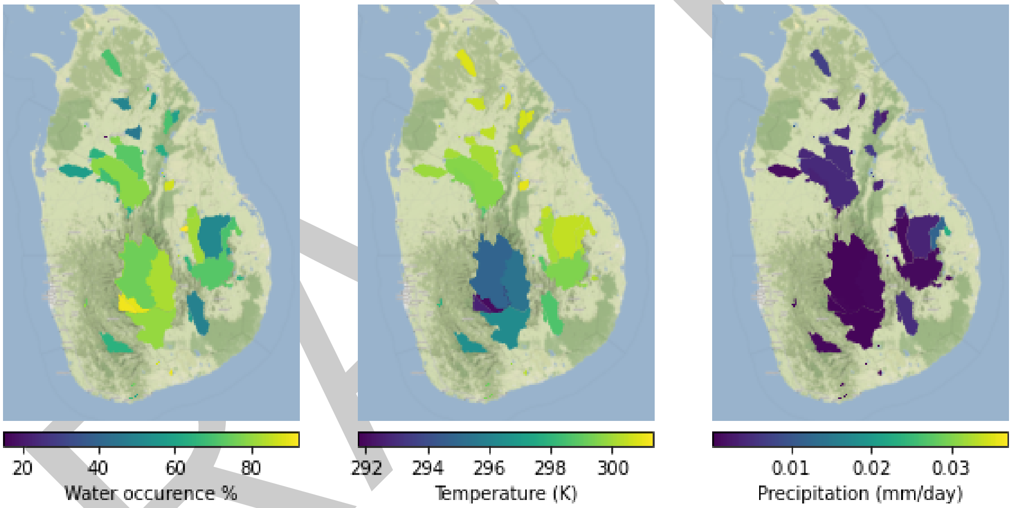

Figure 2: rabpro can return statistics for multiple polygons with a single call. Here, dam-associated (Prior et al., 2022) watersheds in Sri Lanka are delineated and zonal statistics are run for water occurrence, temperature, and precipitation.

Dependencies

`rabpro` relies on functionality from the following Python packages: GDAL (GDAL/OGR contributors, 2020), NumPy (Harris et al., 2020), GeoPandas (Jordahl et al., 2020), Shapely (Gillies et al., 2007), pyproj (Snow et al., 2021), scikit-image (Van der Walt et al., 2014), scipy (Virtanen et al., 2020), and earthengine-api (Gorelick et al., 2017). Use of the watershed statistics methods requires a free Google Earth Engine account. Required MERIT-Hydro and HydroBASINS data are freely available for download by visiting their websites or using `rabpro`'s download scripts; MERIT-Hydro requires users to first register to receive a username and password for access to downloads.

Acknowledgements

Jordan Muss, Joel Rowland, and Eiten Shelef envisioned and created a predecessor to `rabpro` and helped guide its early development. `rabpro` was developed with support from the Laboratory Directed Research and Development program of Los Alamos National Laboratory (Project numbers 20210213ER, 20220697PRD1) and as part of the Interdisciplinary Research for Arctic

Coastal Environments (InteRFACE) project through the Department of Energy, Office of Science, Biological and Environmental Research Earth and Environment Systems Sciences Division RGMA program, awarded under contract grant #9233218CNA000001 to Triad National Security, LLC ("Triad"). TZ was supported by funding from the Columbia Undergraduate Scholars Program Summer Enhancement Fellowship.

References

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R.others. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies–Chile dataset. *Hydrology and Earth System Sciences*, *22*(11), 5817–5846. https://doi.org/10.5194/hess-22-5817-2018

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., & Domisch, S. (2020). Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. *Scientific Data*, *7*(1), 162. https://doi.org/10.1038/s41597-020-0479-6

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., & Poulin, A. (2020). A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds. *Scientific Data*, *7*(1), 1–12. https://doi.org/10.1038/s41597-020-00583-2

Bartos, M. (2020). *Pysheds: Simple and fast watershed delineation in python*. https://doi.org/10.5281/zenodo.3822494

Boyle, D. P., Gupta, H. V., Sorooshian, S., Koren, V., Zhang, Z., & Smith, M. (2001). Toward improved streamflow forecasts: Value of semidistributed modeling. *Water Resources Research*, *37*(11), 2749–2759. https://doi.org/10.1029/2000WR000207

Brooks, K. N. (Ed.). (2003). *Hydrology and the management of watersheds* (3rd ed). Iowa State Press. ISBN: 978-0-8138-2985-2

Chagas, V. B., Chaffe, P. L., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C., & Siqueira, V. A. (2020). CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth System Science Data*, *12*(3), 2075–2096. https://doi.org/10.5194/essd-12-2075-2020

Colby, B. (1956). *Relationship of sediment discharge to streamflow*. US Dept. of the Interior, Geological Survey, Water Resources Division,. https://doi.org/10.3133/ofr5627

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., & Böhner, J. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*, *8*(7), 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015

Didan, K. (2015). *MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006*. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13A2.006

Fenton, J. D., & Keller, R. J. (2001). *The calculation of streamflow from measurements of stage*. Cooperative Research Centre for Catchment Hydrology.

Fowler, K. J., Acharya, S. C., Addor, N., Chou, C., & Peel, M. C. (2021). CAMELS-AUS: Hydrometeorological time series and landscape attributes for 222 catchments in Australia. *Earth System Science Data*, *13*(8), 3847–3867. https://doi.org/10.5194/essd-13-3847-2021

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021

GDAL/OGR contributors. (2020). *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation. https://gdal.org

Gillies, S.others. (2007). *Shapely: Manipulation and analysis of geometric objects*. toblerity.org. https://github.com/Toblerity/Shapely

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Jordahl, K., Bossche, J. V. den, Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., … Leblanc, F. (2020). *Geopandas/geopandas: v0.8.1*. Zenodo. https://doi.org/10.5281/zenodo.3946761

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, *25*(5), 2685–2703. https://doi.org/10.5194/hess-25-2685-2021

Lehner, B., & Grill, G. (2014). *HydroBASINS: Global watershed boundaries and sub-basin delineations derived from HydroSHEDS data at 15 second resolution—Technical documentation version 1. c* [Technical {Report}].

Lindsay, J. B. (2016). Whitebox GAT: A case study in geomorphometric analysis. *Computers & Geosciences*, *95*, 75–84. https://doi.org/10.1016/j.cageo.2016.07.003

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H.others. (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data*, *6*(1), 1–15. https://doi.org/10.1038/s41597-019-0300-6

Markert, K. (2021). *Restee*. GitHub. https://github.com/KMarkert/restee

Montero, D. (2021). Eemont: A Python package that extends Google Earth Engine. *Journal of Open Source Software*, *6*(62), 3168. https://doi.org/10.21105/joss.03168

Nearing, G. S., Klotz, D., Sampson, A. K., Kratzert, F., Gauch, M., Frame, J. M., Shalev, G., & Nevo, S. (2021). Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. *Hydrology and Earth System Sciences Discussions*, 1–25. https://doi.org/10.5194/hess-2021-515

Piccolroaz, S., Di Lazzaro, M., Zarlenga, A., Majone, B., Bellin, A., & Fiori, A. (2016). HYPERstream: A multi-scale framework for streamflow routing in large-scale hydrological model. *Hydrology and Earth System Sciences*, *20*(5), 2047–2061. https://doi.org/10.5194/hess-20-2047-2016

Principe, R. E. (2021). *Gee_tools*. GitHub. https://github.com/gee-community/gee_tools

Prior, E., Schwenk, J., & Rowland, J. (2022). *VotE-Dams: A compilation of global dams' locations and attributes (v1)*. Environmental System Science Data Infrastructure for a Virtual Ecosystem. https://doi.org/10.15485/1843541

Schwenk, J., & Foufoula-Georgiou, E. (2016). Meander cutoffs nonlocally accelerate upstream and downstream migration and channel widening. *Geophysical Research Letters*, *43*(24). https://doi.org/10.1002/2016GL071670

Schwenk, J., Khandelwal, A., Fratkin, M., Kumar, V., & Foufoula-Georgiou, E. (2017). High spatiotemporal resolution of river planform dynamics from Landsat: The RivMAP toolbox and results from the Ucayali River: Annual Planform Morphodynamics, Ucayali. *Earth and Space Science*, *4*(2), 46–75. https://doi.org/10.1002/2016EA000196

Schwenk, J., Lanzoni, S., & Foufoula-Georgiou, E. (2015). The life of a meander bend: Connecting shape and dynamics via analysis of a numerical model. *Journal of Geophysical Research: Earth Surface*, *120*(4), 690–710. https://doi.org/10.1002/2014JF003252

Service, C. C. C. (2017). *ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate . Copernicus Climate Change Service Climate Data Store (CDS)*. https://cds.climate.copernicus.eu/cdsapp#!/home

Snow, A. D., Whitaker, J., Cochran, M., Van Den Bossche, J., Mayo, C., Miara, I., De Kloe, J., Karney, C., Couwenberg, B., Lostis, G., Dearing, J., Ouzounoudis, G., Filipe, Jurd, B., Gohlke, C., Hoese, D., Itkin, M., May, R., Heitor, … Da Costa, M. A. (2021). *pyproj4/pyproj: 3.3.0 Release*. Zenodo. https://doi.org/10.5281/ZENODO.2592232

Tarboton, D. G. (2005). Terrain analysis using digital elevation models (TauDEM). *Utah State University, Logan*, *3012*, 2018.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-image: Image processing in Python. *PeerJ*, *2*, e453.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Walt, S. J. van der, Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wilson, J. P., Aggett, G., Yongxin, D., & Lam, C. S. (2008). Water in the Landscape: A Review of Contemporary Flow Routing Algorithms. In Q. Zhou, B. Lees, & G. Tang (Eds.), *Advances in Digital Terrain Analysis* (pp. 213–236). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77800-4_12

Wu, Q. (2020). Geemap: A Python package for interactive mapping with Google Earth Engine. *Journal of Open Source Software*, *5*(51), 2305. https://doi.org/10.21105/joss.02305

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset. *Water Resources Research*, *55*(6), 5053–5073. https://doi.org/10.1029/2019WR024873