

SCIENTIFIC DATA

OPEN Data Descriptor: FLO1K, global maps of mean, maximum and minimum annual streamflow at 1 km resolution from 1960 through 2015

Received: 3 October 2017

Accepted: 8 February 2018

Published: 27 March 2018

Valerio Barbarossa¹, Mark A.J. Huijbregts^{1,2}, Arthur H.W. Beusen³, Hylke E. Beck⁴, Henry King⁵ & Aafke M. Schipper^{1,2}

Streamflow data is highly relevant for a variety of socio-economic as well as ecological analyses or applications, but a high-resolution global streamflow dataset is yet lacking. We created FLO1K, a consistent streamflow dataset at a resolution of 30 arc seconds (~1 km) and global coverage. FLO1K comprises mean, maximum and minimum annual flow for each year in the period 1960–2015, provided as spatially continuous gridded layers. We mapped streamflow by means of artificial neural networks (ANNs) regression. An ensemble of ANNs were fitted on monthly streamflow observations from 6600 monitoring stations worldwide, i.e., minimum and maximum annual flows represent the lowest and highest mean monthly flows for a given year. As covariates we used the upstream-catchment physiography (area, surface slope, elevation) and year-specific climatic variables (precipitation, temperature, potential evapotranspiration, aridity index and seasonality indices). Confronting the maps with independent data indicated good agreement (R^2 values up to 91%). FLO1K delivers essential data for freshwater ecology and water resources analyses at a global scale and yet high spatial resolution.

Design Type(s)	data integration objective • modeling and simulation objective • source-based data analysis objective
Measurement Type(s)	water flow process
Technology Type(s)	Neural Network
Factor Type(s)	
Sample Characteristic(s)	Earth (Planet) • water flow process • elevation • slope • hydrological precipitation process • temperature of air • evapotranspiration • aridity

¹Radboud University, Institute for Water and Wetland Research, Department of Environmental Science, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands. ²PBL Netherlands Environmental Assessment Agency, Department of Nature and Rural Areas, P.O. Box 30314, 2500 GH The Hague, The Netherlands. ³PBL Netherlands Environmental Assessment Agency, Department of Information, Data and Methodology, P.O. Box 30314, 2500 GH The Hague, The Netherlands. ⁴Princeton University, Department of Civil and Environmental Engineering, Princeton, New Jersey, USA. ⁵Unilever R&D, Safety and Environmental Assurance Centre, Colworth Science Park, Sharnbrook, Bedfordshire MK44 1LQ, UK. Correspondence and requests for materials should be addressed to V.B. (email: vbarbarossa@science.ru.nl).

Background & Summary

Quantifying streamflow is critical to a variety of socio-economic and ecological analyses and applications^{1–3}. Examples include the study of freshwater biodiversity patterns^{4–7}, assessments of global water resources^{8,9}, for example irrigation supply, hydropower or water footprinting^{10–12}, analyses of the fate of pollutants¹³ and quantification of sediment fluxes^{14,15}. Most of the stream reaches in the world are poorly or not monitored at all^{16,17}, due to the inaccessibility of most headwaters and a lack of financial and human resources¹⁸, highlighted by a substantial decline in monitoring since the mid-1980s^{17–19}. Streamflow is commonly quantified with process-driven global hydrological models (GHMs) and land surface models (LSMs)^{20–24}. GHMs/LSMs are typically run at coarse spatial resolutions (~10 to 50 km), due to computational constraints, and consequently are unable to provide reasonable streamflow estimates for small rivers (defined here by Strahler stream order < 5), which comprise 94.6 % of the total stream length and riparian interface on the planet²⁵. Streamflow data at higher spatial resolution would be highly beneficial for ecological applications and water resources assessment, for example understanding/modelling freshwater species distributions or modelling the fate and effects of pollutants in the aquatic environment^{13,26–29}.

Compared to process-based models, data-driven models like regression equations and neural networks are more suited for generating high-resolution streamflow data with large spatial extent, thanks to their computational efficiency and relatively quick parameterization³⁰. Data-driven models typically quantify streamflow based on upstream catchment characteristics related to topography, climate, land cover, and soils^{30–33}. Data-driven approaches have been mostly employed at a local scale³⁴. Recent studies demonstrated, however, the feasibility of applying a data-driven approach at a global scale, resulting in streamflow estimates that may have greater accuracy than the output of GHMs/LSMs^{31,32}. Despite these encouraging results, consistent high-resolution global streamflow maps are not yet available.

Here we present FLO1K: a consistent dataset of global annual streamflow maps at 1 km resolution for each year in the period 1960–2015. Annual flow (AF) metrics include mean annual flow as well as minimum and maximum monthly flow for a given year. We produced the maps with feed-forward Artificial Neural Networks (ANNs) trained on yearly AF metric values from 6600 monitoring stations worldwide, using catchment-averaged covariates representing topography and climate. We delineated the upstream catchments based on the 1-km HydroSHEDS (www.hydrosheds.org) hydrography³⁵, extended with Hydro1k (<https://lta.cr.usgs.gov/HYDRO1K>) for latitudes above 60°N not covered by HydroSHEDS, thereby achieving a global coverage (excluding Antarctica). For the training of the ANNs, we used 10 yearly values of mean, minimum and maximum AF per monitoring station and climate covariates for the corresponding years. We then constructed the AF metric maps by first computing for each year and each 30 arc seconds grid cell the upstream catchment-averaged covariates (which varied from year to year for climate), and then applying the trained ANNs. The streamflow is calculated for each terrain grid cell, i.e., it represents the potential in-channel discharge that would occur in the presence of a natural watercourse. The flow maps have a resolution 10 to 50 times higher than those typically produced using state-of-the-art GHMs/LSMs^{36,37} and global data-driven approaches³². For each of the three AF metrics, 56 yearly layers (1960–2015) are available packed in the NetCDF-4 format CF-compliant. In addition, we provide the FLO1K layers upscaled to 5 and 30 arc minutes resolutions for coarser-grain applications, including comparisons with GHMs/LSMs outputs. The FLO1K database can be downloaded from <http://geoservice.pbl.nl/download/opendata/FLO1K> and figshare (Data Citation 1).

Methods

General approach and streamflow network

The procedure to generate the maps consisted of (i) model fitting, including observed streamflow data preparation, extraction of covariates, and training of the ANNs, and (ii) application of the ANNs to generate the global AF maps. Figure 1 provides a general outline of the procedure. We used the 30 arc seconds (~1 km) version of HydroSHEDS³⁵ extended with Hydro1k for latitudes above 60°N to retrieve the drainage direction network and delineate the upstream catchment of each grid cell^{38,39}. The HydroSHEDS hydrography is based on the National Aeronautics and Space Administration (NASA) Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM)⁴⁰, which covers the entire terrestrial land surface from latitudes 56°S to 60°N. To achieve a global spatial coverage, we extended HydroSHEDS with Hydro1k^{38,39}, the latter being a United States Geological Survey (USGS) product derived from the GTOPO30 Digital Elevation Model (DEM) (<https://lta.cr.usgs.gov/GTOPO30>). The resulting drainage direction network is available at <http://files.nts.g.umd.edu/data/DRT/>.

Streamflow observations

We derived mean, maximum and minimum AF values from flow records in the Global Runoff Data Centre (GRDC) database (www.bafg.de/GRDC)⁴¹. The GRDC comprises daily and monthly streamflow records from 9252 monitoring stations worldwide. The GRDC monitoring stations are not directly referenced on the hydrography employed in this study. This means that mismatched monitoring stations might encompass the wrong upstream catchment basin, which in turn may lead to errors when training the ANNs. As the GRDC dataset includes the estimated catchment area upstream of each monitoring station, we geo-referenced each station in order to match the most similar upstream area on the 30 arc seconds stream network, following the procedure previously used to allocate GRDC stations on the

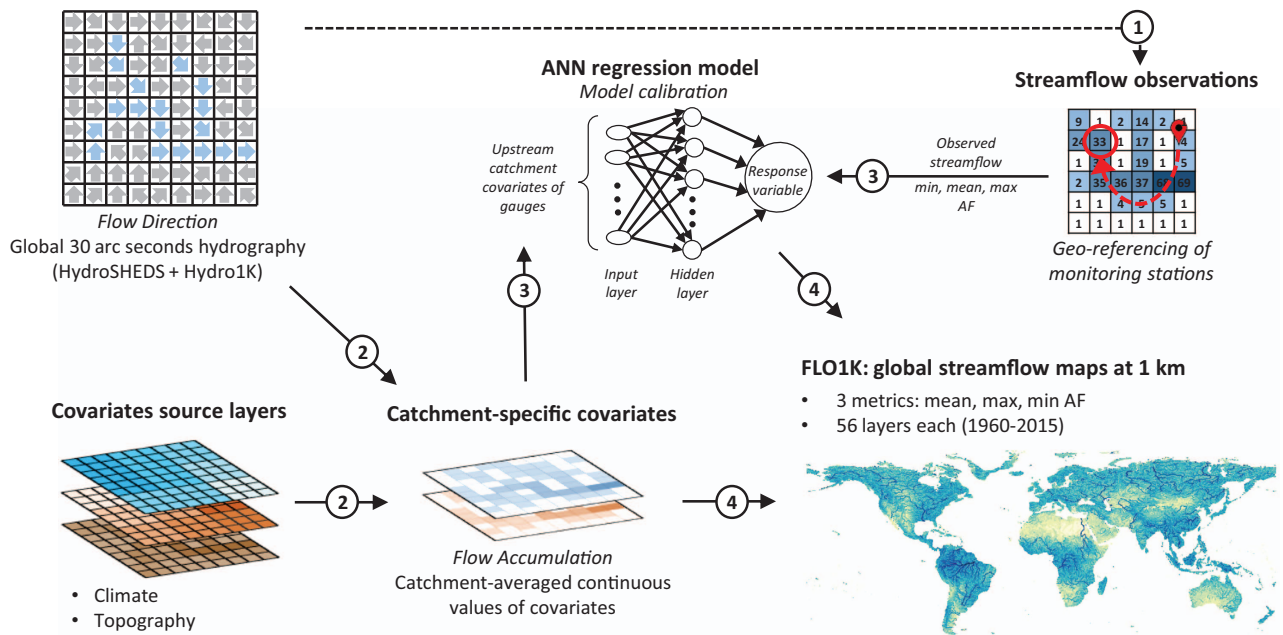


Figure 1. Schematic overview of the streamflow mapping procedure. The procedure consisted of four main steps: 1) monitoring stations (gauges) are geo-referenced based on the global hydrography, 2) catchment-specific covariates are compiled by aggregating climatic and physiographic variables over the upstream catchment of each cell, 3) ANNs are trained on monitoring data of AF metrics and covariates of the corresponding upstream catchment, 4) the trained ANNs are applied to the spatially-continuous covariates to create the global streamflow maps.

HydroSHEDS 15 arc seconds hydrography⁴². For each station, a new location is selected that minimizes discrepancies in catchment area and distance from the original location, within a 5 grid cells (~5 km) search radius. Out of the original 9252 monitoring stations, 285 were excluded as they did not report coordinates. Of the remaining 8967, 746 (~8%) were excluded because there was no matching catchment area within the search radius (based on a threshold of maximum 50% difference⁴²). Out of the remaining 8221, 65% reported an area difference smaller than 5%, 15% had an area difference between 5% and 10%, and 20% had an area difference between 10% and 50%.

We used the monthly records provided by the GRDC to calculate AF metrics for the period 1960–2015. We computed the mean AF for each year by averaging the 12 monthly values, and retrieved maximum and minimum AF by selecting the highest and lowest monthly values for each year, respectively. We considered only those years with a complete 12 months record and selected monitoring stations with at least 10 years of data from 1960 through 2015. The remaining set of stations totaled 6600 and were globally distributed as shown in Figure 2.

Catchment-specific covariates

As covariates of the flow metrics we used topography and climate, which we retrieved from publicly available spatially explicit sources and then aggregated to the upstream catchment of each grid cell. The choice of the covariates set and source data was based on previous studies^{30–34,43,44}, expert knowledge and data availability. A list of the covariates and related source databases is provided in Table 1.

We calculated the area of the upstream catchment of each cell by summing the areas of the upstream grid cells. We derived the upstream catchment-averaged elevation from the SRTM DEM⁴⁰ resampled at 30 arc seconds as provided by HydroSHEDS³⁵, supplemented with the GTOPO30 DEM for areas lacking SRTM coverage, i.e., latitudes above 60°N. We transformed the elevation values by adding a constant value of 500 m to avoid negative values, the lowest being represented by the shores of the Dead Sea at 430 m below sea level. We employed the USGS slope map developed for the Prompt Assessment of Global Earthquakes for Response (PAGER) system⁴⁵ to calculate upstream catchment-averaged surface slope values. This map is based on the same SRTM+GTOPO30 DEM and has been corrected for the discrepancy between ground units (arc degrees) and elevation units (meters)⁴⁵.

We derived the upstream catchment-averaged values for annual mean, maximum and minimum air temperature (T_{air}) and precipitation (P), as well as potential evapotranspiration (PET), aridity index (AI) and seasonality index for P and PET, for every year over the period 1960–2015. For air temperature, we employed the Climate Research Unit (CRU) Time Series (TS) dataset⁴⁶ (version 3.24.01; monthly temporal and 0.5° spatial resolution). For precipitation, we used the Multi-Source Weighted-Ensemble

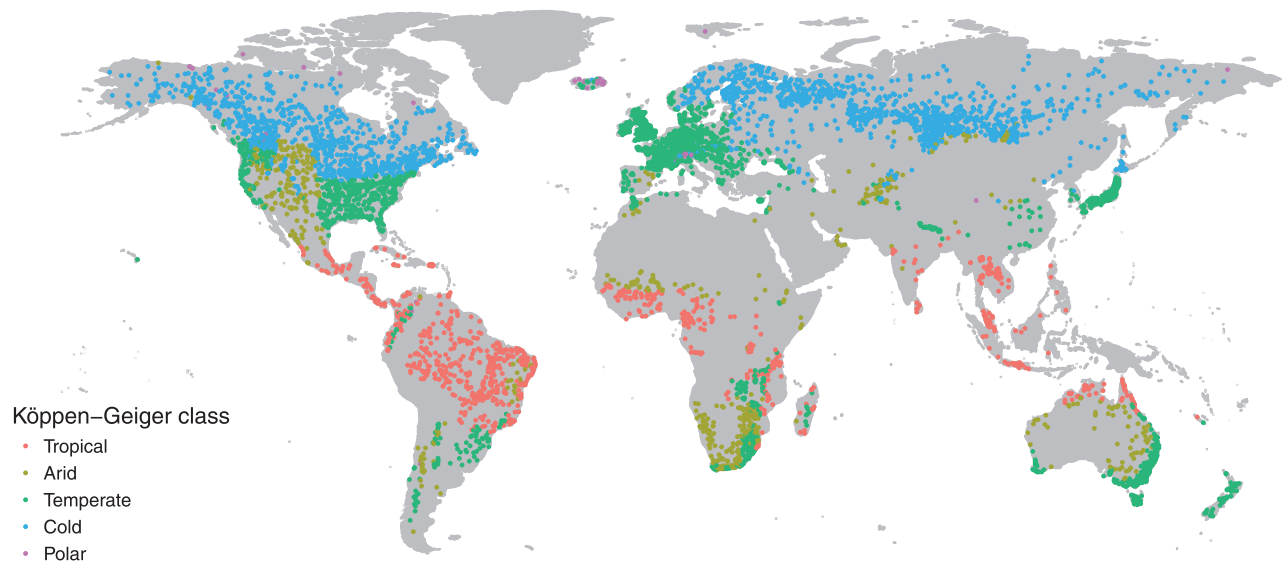


Figure 2. Distribution of the 6,600 GRDC stations monitored for at least 10 years in the period 1960–2015. Stations are coloured according to the Köppen–Geiger climate classification⁶⁷.

Category	Variable description	Source	Annual metric	Unit	No. layers	Spatial resolution	Temporal coverage
Topography	Upstream catchment area	This study	-	km ²	1	~ 1 km	-
	Elevation	SRTM ⁴⁰ + GTOPO30	-	m	1	~ 1 km	-
	Surface slope	USGS ⁴⁵	-	°	1	~ 1 km	-
Climate	Precipitation	MSWEP ⁴⁷ + GPCC ⁴⁸	Mean	mm · month ⁻¹	56 × 4	~ 25 km	1960–2015
			Minimum	mm · month ⁻¹			
			Maximum	mm · month ⁻¹			
			Seasonality index	-			
	Air temperature	CRU TS 3.24.01 ⁴⁶	Mean	K	56 × 3	~ 50 km	1960–2015
			Minimum	K			
			Maximum	K			
	Potential evapotranspiration	This study	Mean	mm · month ⁻¹	56 × 2	~ 50 km	1960–2015
			Seasonality index	-			
Aridity index	This study	-	-	56	~ 25 km	1960–2015	

Table 1. Comparison of R^2 values for streamflow metrics based on calendar vs US water year. The comparison is based on 2,484 stations north of 40N latitude, monitored for at least 30 years in the period 1960–2015. The R^2 was calculated from log-transformed values. LT: long term; YR: yearly.

Precipitation (MSWEP) dataset⁴⁷ (version 1.2; 3-hourly temporal and 0.25° spatial resolution; 1979–2015) supplemented with the Global Precipitation Climatology Centre (GPCC) Full Data Reanalysis⁴⁸ (version 7; monthly temporal and 0.5° spatial resolution) prior to 1979. MSWEP merges a wide range of gauge, satellite, and reanalysis datasets to achieve precipitation estimates with greater accuracy than any other global dataset⁴⁷. To combine the GPCC and MSWEP datasets, we rescaled the GPCC estimates such that the 1979–2013 mean of GPCC matched that of MSWEP. For each year and grid cell, we retrieved the mean annual value of T_{air} and P as the mean over the 12 monthly layers, and the minimum and maximum as the lowest and highest monthly values, respectively. We computed mean annual potential evapotranspiration from monthly T_{air} values following the temperature-based approach of Hargreaves et al.⁴⁹ and employing the same CRU TS v. 3.24.01 source data for temperature. Similarly, we calculated seasonality index layers for P and PET as $X_{si} = X_{yr}^{-1} \sum |X_m - X_{yr}/12|$, where si , yr and m stand for seasonality index, yearly and monthly values, respectively⁵⁰. We downscaled the raster layers for the climate-related covariates to match the 30 arc seconds resolution of the hydrography using nearest-neighbour resampling. In addition, we calculated the aridity index for each year as PET/P , using mean annual P and PET .

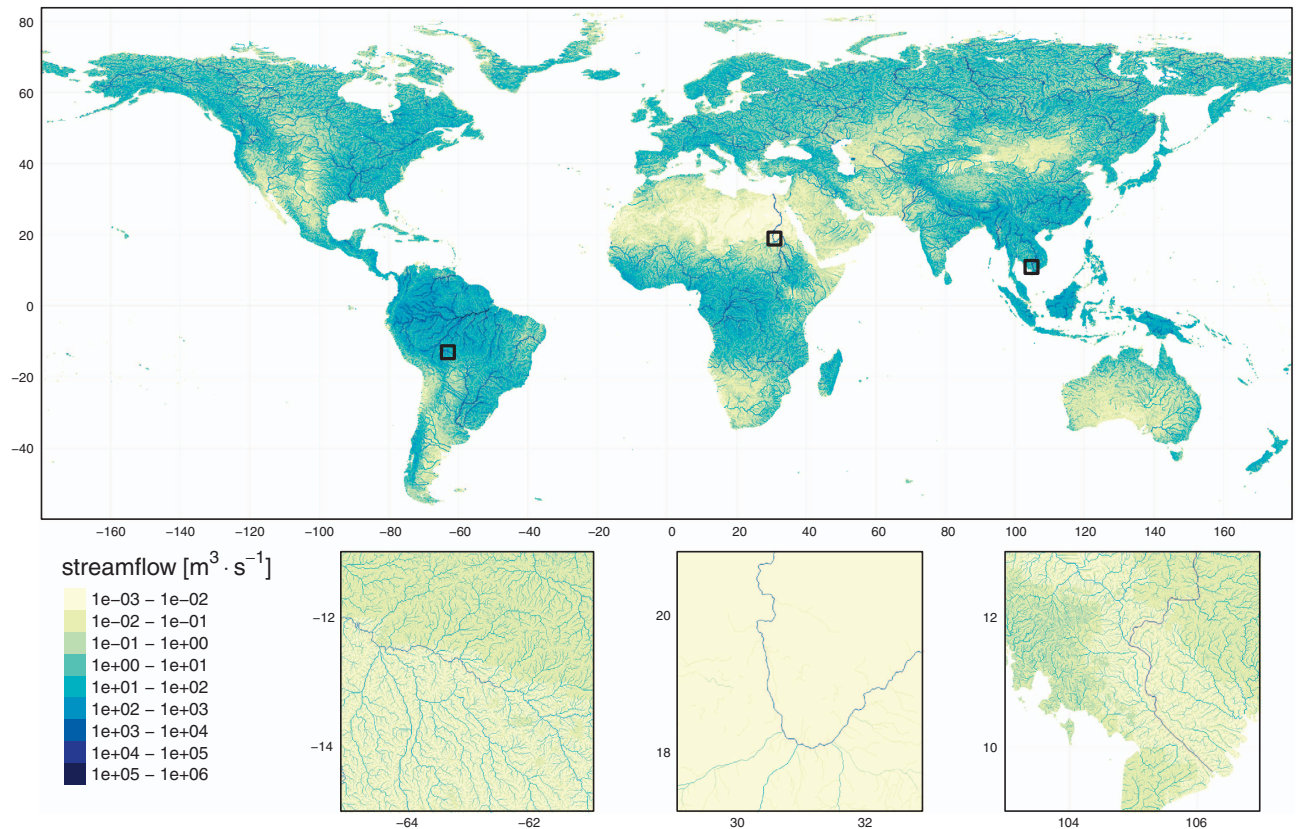


Figure 3. Long-term mean annual flow map overview. The long-term mean annual flow represents the average of the year-specific FLO1K maps for mean AF over the period 1960–2015. The global map has been upscaled using maximum-value resampling by a factor of 20 for clarity of visualization. Insets show the original 1 km resolution. Location of each inset is marked on the global map by a black square.

To calculate the upstream catchment-average values of the covariates, we employed the TauDEM software (Terrain Analysis Using Digital Elevation Models, <http://hydrology.usu.edu/taudem>). TauDEM is an open-source C++ software explicitly designed to implement the flow algebra for large datasets, employing a Message Passing Interface (MPI, <http://mpi-forum.org>) to implement highly parallelized processing algorithms^{51–53}. We extracted the covariates for the upstream catchment of each cell of the global hydrological network via the so-called flow accumulation technique (‘AreaD8’ in TauDEM). This technique considers each grid cell as a pour point and subsequently calculates the number of upstream grid cells or the sum of the attribute values of these upstream grid cells, using the flow direction map to delineate the watershed boundaries of the upstream catchment. To derive continuous upstream catchment-averaged values for the predictor variables, we divided the sum of the upstream covariate values by the total number of upstream grid cells at each pour point. To speed-up the calculations, we split the global flow direction layer into six continents (North America, South and Central America, Europe, Africa, Asia, Oceania). Adjacent continents (e.g., Europe and Asia) were separated along watershed boundaries.

Training of Artificial Neural Networks

We quantified the relationships between the flow metrics and the covariates using artificial neural networks (ANNs), which have been widely used for hydrological modelling from local⁵⁴ to global^{32,33} scales. We employed the feed-forward ANN algorithm based on the multi-layer perceptron structure with one hidden layer^{55,56} (Figure 1). We trained the ANNs based on year-specific values of mean, minimum and maximum AF, using the upstream-catchment topography and year-specific climate as covariates (Table 1). We applied a Box-Cox transformation to normalize the distributions of each variable (response and covariates)⁵⁷. In addition, we standardized each distribution to zero mean and unit standard deviation, as required for the ANNs⁵⁶. To avoid possible bias due to differences in monitoring intensity among the stations, we randomly picked 10 yearly values from those stations monitored at least 10 years across the 1960–2015 period. We then iterated the ANNs training 20 times, sampling different years from those stations having a record longer than 10 years. Prior to the training, we tuned the number of neurons of the hidden layer of the ANNs and the weights decay value to regularize the ANNs cost function, and therefore control for overfitting. To this end, we used 10-fold cross-validation (CV) whose

folders were based on excluded monitoring stations, and identified the number of neurons and weights decay value that maximized the median coefficient of determination (R^2) and minimized the median Root Mean Square Error (RMSE) of the testing set. As a result, we employed 20 neurons for the ANNs hidden layer and a weights decay value of 0.01.

Generating mean, maximum and minimum AF global maps

We applied the ANNs model to produce 30 arc seconds maps with mean, maximum and minimum annual flow from 1960 through 2015 (Data Citation 1). For each grid cell, we computed the AF metrics as the median across the outputs of 20 trained ANNs and back-transformed the values to $\text{m}^3 \cdot \text{s}^{-1}$.

We upscaled the 30 arc seconds layers to 5 and 30 arc minutes resolutions, in order to serve potential coarser-grain applications. We based the upscaled output on the 5 and 30 arc minutes flow direction grids produced by applying the dominant river tracing (DRT) algorithm to the same 30 arc seconds flow direction layer used in this study^{38,39}. The 5 and 30 arc minutes flow direction grids are freely available for download at <http://files.nts.g.umt.edu/data/DRT/>. We upscaled the 30 arc seconds streamflow values by choosing the value of the cell that minimized the differences in upstream-drainage area between the native 30 arc seconds and the coarser resolution grid cell. For the 5 arc minutes grids it was necessary to employ a one-cell search radius to avoid losing connectivity.

Code availability

The code used to generate the covariate data, geo-reference the monitoring stations, train the ANNs and generate the flow maps (Data Citation 1) was written and run in R version 3.3.2. TauDEM tools⁵² were used to produce the catchment-specific covariate layers and GDAL library⁵⁸ functions were employed to handle the analyses on large raster data. The scripts are available on request.

The ensemble of trained ANNs are available as R objects (.rds) and as Portable Model Markup Language (PMML) objects for cross-platform compatibility (.pmml, <http://dmg.org>). The parameters used for the Box-Cox transformation and standardization of the variables employed by the ANNs are also available in CSV format.

Data Records

The FLO1K dataset is a set of gridded layers packed as NetCDF-4 files freely available for download (Data Citation 1). For each of the three AF metrics, 56 yearly layers are available from 1960 through 2015, yielding a total of 168 layers. Each non-null cell represents the potential streamflow in $\text{m}^3 \cdot \text{s}^{-1}$, stored as 32-bit floating point. Layers are in the WGS84 coordinate system with a cell size of 30 arc seconds (~ 1 km) and a global extent, including all continents except for Antarctica (90°N to 90°S latitude and 180°W to 180°E longitude). In addition, upscaled data are available at 5 and 30 arc minutes.

Technical Validation

To evaluate the quality of the FLO1K maps, we run a 10-fold cross-validation for each of the 20 ANN runs, such that each observation was included in the test set once and by splitting the folds by stations. We assessed the overall map quality with R^2 and RMSE calculated based on log-transformed values to evaluate the performance across the full spectrum of streamflow values (10^{-3} - 10^5 $\text{m}^3 \cdot \text{s}^{-1}$). Cross-validation results showed high agreement between training (90%) and independent testing (10%) data, with negligible variation among the replicates (Table 2).

We assessed the uncertainty per grid cell resulting from the sub-sampling of the monitoring stations, by computing the coefficient of variation (CoV) over the 20 replicates. Uncertainty was very low (CoV < 0.5) for the main river stems globally and smaller reaches in wet regions (Fig. 4). We found higher uncertainty (higher CoV values) for low streamflow values in dry areas, e.g., the upper basin of the Nile (central inset of Fig. 4). These higher CoV values likely reflect the lower number of streamflow observations available for calibrating the ANNs in these areas. The highest CoV values (> 3.5) were found in grid cells with a low number of upstream grid cells (typically < 5) in dry areas. In these grid cells, most of the ANN replicates yielded zero-flow values whereas one or few replicates yielded close-to-zero values, resulting in a low mean yet large CoV across the 20 replicates.

We checked for potential bias in streamflow estimates in the northern hemisphere due to snowmelt delays, e.g., the contributing effect of snowfall in November-December of the previous year on the streamflow in May-June. To this end, we generated streamflow maps based on the US water year (November-October) for stations north of 40°N and compared their performance to the original (calendar year-based) FLO1K maps. We tuned the ANNs ensemble and computed the streamflow fields adopting the US water year for both the streamflow data and the climate input variables. Differences in R^2 between models based on calendar versus US water year were smaller than 0.01 and therefore considered negligible (Table 3).

Usage notes

The FLO1K dataset reports the potential streamflow in $\text{m}^3 \cdot \text{s}^{-1}$ in each grid cell, i.e., the discharge that would occur if there were a natural watercourse. To avoid confusion, we emphasize that the estimates represent volumetric streamflow rather than specific runoff. As such, the estimates cannot directly be compared with outputs from climate or land surface models without a streamflow routing component.

AF metric	R ²		RMSE	
	Training	Testing	Training	Testing
Mean	0.92 (0.001)	0.91 (0.002)	0.32 (0.002)	0.34 (0.004)
Maximum	0.91 (0.001)	0.90 (0.002)	0.33 (0.002)	0.34 (0.005)
Minimum	0.85 (0.001)	0.83 (0.003)	0.48 (0.002)	0.51 (0.005)

Table 2. Model performance statistics. The R² and RMSE values represent medians (with standard deviations in brackets) of the 10-fold cross-validation of 200 replicates. Both R² and RMSE were calculated from log-transformed values, therefore the RMSE is unitless.

	AF metric					
	Mean (LT)	Mean (YR)	Max (LT)	Max (YR)	Min (LT)	Min (YR)
Calendar year	0.975	0.953	0.969	0.936	0.933	0.885
US water year	0.975	0.955	0.969	0.942	0.928	0.879

Table 3. Description of the predictor variables used as input for the modelling of AF. The spatial resolution refers to the source data; for the analysis all variables were resampled to ~1 km.

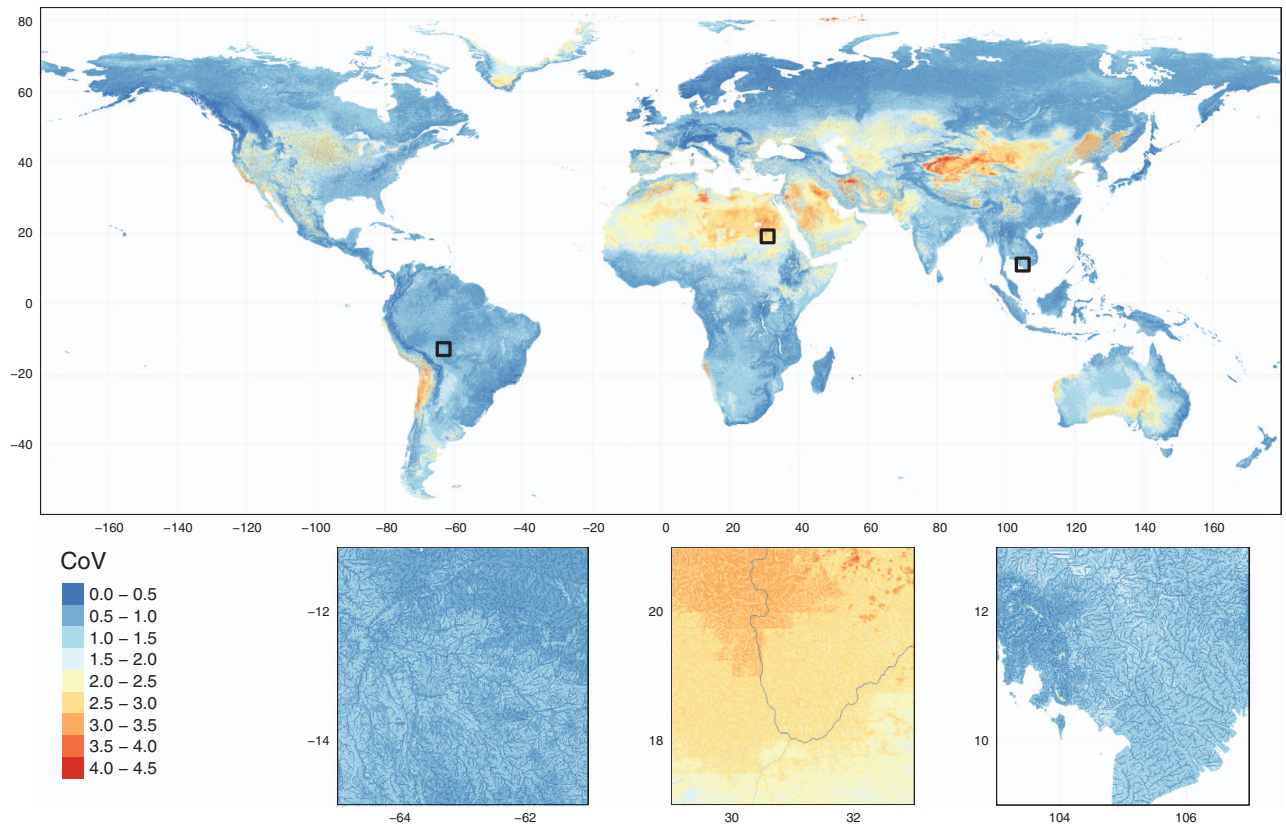


Figure 4. Uncertainty in mean AF due to differences in sub-sampling among the 20 ANN replicates. Uncertainty is expressed as coefficient of variation (CoV) averaged across the overall period 1960–2015. The global map has been upscaled using mode resampling by a factor of 5 for clarity of visualization. Insets show the original 1 km resolution. Location of each inset is marked on the global map by a black square.

We refrained from filtering the output to the actual stream network because there are multiple methods for stream network delineation^{59–66}, which users of FLO1k may want to select or refine according to their needs. For global-scale analyses one might adopt an arbitrary upstream catchment area threshold in order to delineate the network (e.g., 25 upstream grid cells as in Hydrosheds³⁵), as to our knowledge more refined methods have not yet been developed/tested.

The estimated maximum and minimum flow values for a given year reflect the highest and the lowest monthly values of that year. This does not give an indication about which months of the year belong to the maximum or minimum flow. The corresponding months might change from year to year based on the yearly distribution of the precipitation.

Users of the upscaled streamflow grids should keep in mind that these are contingent on the respective DRT flow direction layers^{38,39}. Further, the accuracy of the upscaled grids has not been evaluated.

References

- Vörösmarty, C. J. *et al.* Global threats to human water security and river biodiversity. *Nature* **467**, 555–561 (2010).
- WWAP (United Nations World Water Assessment Programme). *The United Nations world water development report 2015: water for a sustainable world. 1*. UNESCO Publishing, (2015).
- Bakker, K. Water Security: Research Challenges and Opportunities. *Science* (80-) **337**, 914–915 (2012).
- Iwasaki, Y., Ryo, M., Sui, P. & Yoshimura, C. Evaluating the relationship between basin-scale fish species richness and ecologically relevant flow characteristics in rivers worldwide. *Freshw. Biol* **57**, 2173–2180 (2012).
- Xenopoulos, M. A. & Lodge, D. M. Going with the Flow : Using Species-Discharge Relationships to Forecast Losses in Fish Biodiversity Published by : Ecological Society of America content in a trusted digital archive. *We use information technology and tools to increase productivity and fa* **87**, 1907–1914 (2014).
- Oberdorff, T. *et al.* Global and regional patterns in riverine fish species richness: A review. *Int. J. Ecol.* **2011** (2011).
- Poff, N. L. & Zimmerman, J. K. H. Ecological responses to altered flow regimes: A literature review to inform the science and management of environmental flows. *Freshw. Biol* **55**, 194–205 (2010).
- Dai, A. in *Terrestrial Water Cycle and Climate Change* 17–37 John Wiley & Sons, Inc., (2016).
- Dai, A., Qian, T., Trenberth, K. E. & Milliman, J. D. Changes in Continental Freshwater Discharge from 1948 to 2004. *J. Clim* **22**, 2773–2792 (2009).
- Hanafiah, M. M., Xenopoulos, M. A., Pfister, S., Leuven, R. S. E. W. & Huijbregts, M. A. J. Characterization Factors for Water Consumption and Greenhouse Gas Emissions Based on Freshwater Fish Species Extinction. *Environ. Sci. Technol.* **45**, 5272–5278 (2011).
- Tendall, D. M., Hellweg, S., Pfister, S., Huijbregts, M. A. J. & Gaillard, G. Impacts of river water consumption on aquatic biodiversity in life cycle assessment—a proposed method, and a case study for Europe. *Environ. Sci. Technol.* **48**, 3236–3244 (2014).
- Hoekstra, A. Y., Chapagain, A. K., Aldaya, M. M. & Mekonnen, M. M. *The Water Footprint Assessment Manual. Febrero 2011* (2011).
- Grill, G., Khan, U., Lehner, B., Nicell, J. & Ariwi, J. Risk assessment of down-the-drain chemicals at large spatial scales: Model development and application to contaminants originating from urban areas in the Saint Lawrence River Basin. *Sci. Total Environ.* **541**, 825–838 (2016).
- Syvitski, J. P., Peckham, S. D., Hilberman, R. & Mulder, T. Predicting the terrestrial flux of sediment to the global ocean: a planetary perspective. *Sediment. Geol.* **162**, 5–24 (2003).
- Syvitski, J. P. M. Impact of Humans on the Flux of Terrestrial Sediment to the Global Coastal Ocean. *Science* (80-) **308**, 376–380 (2005).
- Sivapalan, M. Prediction in ungauged basins: a grand challenge for theoretical hydrology. *Hydrol. Process.* **17**, 3163–3170 (2003).
- Fekete, B. M. & Vörösmarty, C. J. The current status of global river discharge monitoring and potential new technologies complementing traditional discharge measurements. *Predict. Ungauged Basins PUB Kick-off (Proceedings PUB Kick-off Meet. held Bras. Novemb. 2002)*, IAHS Publ. no. 309 **309**, 129–136 (2007).
- Shiklomanov, A. I., Lammers, R. B. & Vörösmarty, C. J. Widespread decline in hydrological monitoring threatens Pan-Arctic Research. *Eos, Trans. Am. Geophys. Union* **83**, 13–17 (2002).
- Hannah, D. M. *et al.* Large-scale river flow archives: importance, current status and future needs. *Hydrol. Process.* **25**, 1191–1200 (2011).
- Bierkens, M. F. P. Global hydrology 2015: State, trends, and directions. *Water Resour. Res.* **51**, 4923–4947 (2015).
- Haddeland, I. *et al.* Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results. *J. Hydrometeorol.* **12**, 869–884 (2011).
- Haddeland, I. *et al.* Global water resources affected by human interventions and climate change. *Proc. Natl. Acad. Sci* **111**, 3251–3256 (2014).
- Schellekens, J. *et al.* A global water resources ensemble of hydrological models: The earth2Observe Tier-1 dataset. *Earth Syst. Sci. Data* **9**, 389–413 (2017).
- Beck, H. E. *et al.* Global evaluation of runoff from ten state-of-the-art hydrological models. *Hydrol. Earth Syst. Sci.* **21**, 2881–2903 (2017).
- Downing, J. A. *et al.* Global abundance and size distribution of streams and rivers. *Int. Waters* **2**, 229–236 (2012).
- Labay, B. J. *et al.* Can Species Distribution Models Aid Bioassessment when Reference Sites are Lacking? Tests Based on Freshwater Fishes. *Environ. Manage.* **56**, 835–846 (2015).
- Domisch, S., Amatulli, G. & Jetz, W. Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution. *Sci. Data* **2**, 150073 (2015).
- Domisch, S., Jähnig, S. C., Simaika, J. P., Kuemmerlen, M. & Stoll, S. Application of species distribution models in stream ecosystems: the challenges of spatial and temporal scale, environmental predictors and species occurrence data. *Fundam. Appl. Limnol* **186**, 1–2 (2015).
- Thorp, J. H., Thoms, M. C. & DeLong, M. D. The riverine ecosystem synthesis: biocomplexity in river networks across space and time. *River Res. Appl.* **22**, 123–147 (2006).
- Verdin, K. L. & Worstell, B. A fully distributed implementation of mean annual streamflow regional regression equations. *J. Am. Water Resour. Assoc.* **44**, 1537–1547 (2008).
- Barbarossa, V. *et al.* Developing and testing a global-scale regression model to quantify mean annual streamflow. *J. Hydrol.* **544**, 479–487 (2017).
- Beck, H. E., de Roo, A. & van Dijk, A. I. J. M Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments*. *J. Hydrometeorol.* **16**, 1478–1501 (2015).
- Beck, H. E. *et al.* Global patterns in base flow index and recession based on streamflow observations from 3394 catchments. *Water Resour. Res.* **49**, 7843–7863 (2013).
- Razavi, T. & Coulibaly, P. Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. *J. Hydrol. Eng.* **18**, 958–975 (2013).
- Lehner, B., Verdin, K. & Jarvis, A. New global hydrography derived from spaceborne elevation data. *Eos (Washington. DC)* **89**, 93–94 (2008).
- Verzano, K. *et al.* Modeling variable river flow velocity on continental scale: Current situation and climate change impacts in Europe. *J. Hydrol.* **424–425**, 238–251 (2012).

37. Wada, Y., Wisser, D. & Bierkens, M. F. P. Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth Syst. Dynam* **5**, 15–40 (2014).
38. Wu, H. *et al.* A new global river network database for macroscale hydrologic modeling. *Water Resour. Res.* **48** (2012).
39. Wu, H., Kimball, J. S., Mantua, N. & Stanford, J. Automated upscaling of river networks for macroscale hydrological modeling. *Water Resour. Res.* **47** (2011).
40. Farr, T. G. *et al.* The Shuttle Radar Topography Mission. *Rev. Geophys.* **45**, RG2004 (2007).
41. GRDC. Long-Term Mean Monthly Discharges and Annual Characteristics of GRDC Stations / Online provided by the Global Runoff Data Centre of WMO (2017).
42. GRDC. Watershed Boundaries of GRDC Stations / Global Runoff Data Centre (2011).
43. Vogel, R. M., Wilson, I. & Daly, C. Regional Regression Models of Annual Streamflow for the United States. *J. Irrig. Drain. Eng* **125**, 148–157 (1999).
44. Farmer, W. H. & Vogel, R. M. Performance-weighted methods for estimating monthly streamflow at ungauged sites. *J. Hydrol.* **477**, 240–250 (2013).
45. Verdin, K. L. *et al.* Development of a Global Slope Dataset for Estimation of Landslide Occurrence Resulting from Earthquakes. Colorado: U.S. Geological Survey, Open-File Report **1188** (2007).
46. Harris, I., Jones, P. D., Osborn, T. J. & Lister, D. H. Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *Int. J. Climatol.* **34**, 623–642 (2014).
47. Beck, H. E. *et al.* MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrol. Earth Syst. Sci.* **21**, 589–615 (2017).
48. Schneider, U. *et al.* GPCC Full Data Reanalysis Version 7.0 at 0.5°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data (2015).
49. Hargreaves, G. L., Hargreaves, G. H. & Riley, J. P. Irrigation Water Requirements for Senegal River Basin. *J. Irrig. Drain. Eng* **111**, 265–275 (1985).
50. Walsh, R. P. D. & Lawler, D. M. Rainfall seasonality: description, spatial patterns and change through time. *Weather* **36**, 201–208 (1981).
51. Tesfa, T. K. *et al.* Extraction of hydrological proximity measures from DEMs using parallel processing. *Environ. Model. Softw* **26**, 1696–1709 (2011).
52. Tarboton, D. G. Terrain Analysis Using Digital Elevation Models (Taudem) (2008).
53. Tarboton, D. G., Schreuders, K. A. T., Watson, D. W. & Baker, M. E. Generalized terrain-based flow analysis of digital elevation models. in *The 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Cairns, Australia from 13–17 July 2009*. **1**, 2377–2383 (2009).
54. Yaseen, Z. M., El-shafie, A., Jaafar, O., Afan, H. A. & Sayl, K. N. Artificial intelligence based models for stream-flow forecasting: 2000–2015. *Journal of Hydrology* **530**, 829–844 (2015).
55. Bishop, C. M. & M. C. *Neural networks for pattern recognition*. Clarendon Press, (1995).
56. Haykin, S. S. *Neural networks: a comprehensive foundation*. Macmillan, (1994).
57. Box, G. E. P. & Cox, D. R. An analysis of transformations. in *Journal of the Royal Statistical Society. Series B (Methodological)* **26**, 211–252 (1964).
58. GDAL Development Team. GDAL - Geospatial Data Abstraction Library, Version 2.2.0 (2017).
59. Tarboton, D. G., Bras, R. L. & Rodriguez-Iturbe, I. On the extraction of channel networks from digital elevation data. *Hydrol. Process.* **5**, 81–100 (1991).
60. Tarboton, D. G., Bras, R. L. & Rodriguez-iturbe, I. A physical basis for drainage density.pdf. *Geomorphology* **5**, 59–76 (1992).
61. Hancock, G. R. The use of digital elevation models in the identification and characterization of catchments over different grid scales. *Hydrol. Process.* **19**, 1727–1749 (2005).
62. Thompson, J. A., Bell, J. C. & Butler, C. A. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma* **100**, 67–89 (2001).
63. Istanbuluoglu, E., Tarboton, D. G., Pack, R. T. & Luce, C. A probabilistic approach for channel initiation. *Water Resour. Res.* **38**, 61-1–61-14 (2002).
64. Russell, P. P., Gale, S. M., Muñoz, B., Dorney, J. R. & Rubino, M. J. A Spatially Explicit Model for Mapping Headwater Streams. *JAWRA J. Am. Water Resour. Assoc* **51**, 226–239 (2015).
65. Sangireddy, H., Stark, C. P., Kladzyk, A. & Passalacqua, P. GeoNet: An open source software for the automatic and objective extraction of channel heads, channel network, and channel morphology from high resolution topography data. *Environ. Model. Softw* **83**, 58–73 (2016).
66. Avcioglu, B., Anderson, C. J. & Kalin, L. Evaluating the Slope-Area Method to Accurately Identify Stream Channel Heads in Three Physiographic Regions. *JAWRA J. Am. Water Resour. Assoc* **53**, 562–575 (2017).
67. Kotttek, M., Grieser, J., Beck, C., Rudolf, B. & Rubel, F. World Map of the Köppen-Geiger climate classification updated. *Meteorol. Zeitschrift* **15**, 259–263 (2006).

Data Citation

1. Barbarossa, V *et al.* *figshare* <https://dx.doi.org/10.6084/m9.figshare.c.3890224> (2018).

Acknowledgements

The authors would like to thank the Global Runoff Data Center for providing the streamflow data. The authors would also like to thank Huan Wu and Kristine L. Verdin for providing the global hydrography base layer and the global surface slope data (PAGER), respectively. This project has received funding from the Europeans Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 641459.

Author Contributions

V.B. and A.M.S. conceived the idea and set up the study. V.B. performed the analyses, produced the FLO1K database and wrote the paper. H.B. provided the precipitation data. All authors contributed to discussions about the approach and results and commented on the paper.

Additional information

Competing interests: The authors declare no competing interests.

How to cite this article: Barbarossa, V *et al.* FLO1K, Global Maps of Mean, Maximum and Minimum Annual Streamflow at 1 km Resolution From 1960 Through 2015. *Sci. Data* 5:180052 doi: 10.1038/sdata.2018.52 (2018).

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



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

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files made available in this article.

© The Author(s) 2018