SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Freshwater ecology

» Biogeography

» Macroecology

» Biodiversity

Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution

Sami Domisch[1], Giuseppe Amatulli[1] & Walter Jetz[1]

Received: 15 June 2015

Accepted: 05 November 2015

Published: 08 December 2015

The lack of freshwater-specific environmental information at sufficiently fine spatial grain hampers broad-scale analyses in aquatic biology, biogeography, conservation, and ecology. Here we present a near-global, spatially continuous, and freshwater-specific set of environmental variables in a standardized 1 km grid. We delineate the sub-catchment for each grid cell along the HydroSHEDS river network and summarize the upstream climate, topography, land cover, surface geology and soil to each grid cell using various metrics (average, minimum, maximum, range, sum, inverse distance-weighted average and sum). All variables were subsequently averaged across single lakes and reservoirs of the Global lakes and Wetlands Database that are connected to the river network. Monthly climate variables were summarized into 19 long-term climatic variables following the 'bioclim' framework. This new set of variables provides a basis for spatial ecological and biodiversity analyses in freshwater ecosystems at near global extent, yet fine spatial grain. To facilitate the generation of freshwater variables for custom study areas and spatial grains, we provide the '*r.stream.watersheds*' and '*r.stream.variables*' add-ons for the GRASS GIS software.

Design Type(s)	time series design ● observation design
Measurement Type(s)	climate ● topography ● land cover ● geology ● soil
Technology Type(s)	hydrography
Factor Type(s)	Collection Data Type
Sample Characteristic(s)	freshwater environment ● North America ● South America ● Europe ● Africa ● Asia ● Australasia

[1] Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06511, USA. Correspondence and requests for materials should be addressed to S.D. (email: sami.domisch@yale.edu).

Background & Summary

Freshwater habitats cover only 0.1% of Earth's surface, yet they provide habitat for ~10% of all animal species[1]. Many streams and lakes are considered biodiversity hotspots[2], and understanding the biogeography of freshwater organisms is key for biodiversity conservation and management, especially in the context of ongoing and projected climate and land cover change[3].

In addition to the macroevolutionary and biogeographic history of clades and past conditions, current-day species distributions are strongly determined by contemporary environmental factors. A variety of modelling tools, often summarized under the term of 'environmental niche' or 'species distribution models' have been developed to assess species ~ environment relationships and to predict species geographic distributions. Here, the species geographic occurrences are related to the environmental conditions at those locations, yielding a model in environmental space that can then be projected and extrapolated into geographic space[4]. For successful modelling, in addition to reliable and ideally environmentally and geographically representative species occurrence records[5], range-wide environmental data is needed. This usually comes in a spatially gridded format, where each grid cell is characterized by a continuous (e.g., temperature) or discrete (land cover class) variable. Over the past years a number of global high-resolution 1 km data sets have been created to assist the use of spatial modelling, such as gridded climate data from interpolated weather stations (WorldClim[6]), remote sensing products such as topography (SHuttle Elevation Derivatives at multiple Scales, SRTM[7]), or Moderate Resolution Imaging Spectroradiometer (MODIS) -derived products such as land cover[8], and derivatives of those.

Appropriately quantifying species associations with their (abiotic) environment requires consideration of the full spatial extent of their occurrence[9]. While the inclusion of range-wide environmental data in models is becoming common place in the terrestrial and marine realm, it poses a real challenge in the freshwater realm[5], since (i) the gridded site-level (terrestrial) environmental variables do not translate well into the directional freshwater ecosystems without accounting for the down-stream connectivity, and (ii) even if freshwater-specific environmental data are available, they are mostly restricted to single basins and watersheds, or political borders[5,10]. Likewise, the lack of range-wide freshwater-specific environmental information hampers comparable freshwater biodiversity and ecosystem analyses in general, such as metapopulation and -community models, ecosystem or ecoregion delineations, and functional biodiversity assessments[11].

To facilitate a more geographically inclusive and comparable studies in freshwater biodiversity science, we developed near-global 1 km gridded freshwater-specific information comprised of multiple, complementary environmental variables of known relevance for species distributions: climatic (monthly air temperature and precipitation), topographic (elevation and slope), land cover, surface lithological and soil variables (Data Citation 1). This newly developed information is based on the 1 km HydroSHEDS hydrography[12] (**Hydro**logical data and maps based on **SH**uttle **E**levation **D**erivatives at multiple **S**cales, www.hydrosheds.org) and accounts for the upstream connectivity within the stream network. Each environmental variable is computed for each single 1 km stream gird cell individually along the stream network[10], by (i) delineating the upstream sub-catchment for each grid cell, and (ii) summarizing and relating the upstream environmental conditions of each variable to each stream grid cell along various metrics (upstream minimum, maximum, average, sum, weighted average, sum and weighted sum, Table 1). This procedure integrates the connectivity, and consequently the upstream environmental conditions can be traced along the stream network (for instance, percent upstream forest cover). In addition, (iii) we extended the data to lakes and reservoirs of the 1 km gridded Global Lake and Wetlands Database[13] and unified the stream and lake data layers for each variable.

The newly developed layers facilitate—for the first time—large-scale models of the distribution and community characteristics of freshwater biota. While the layers (Data Citation 1) can be used 'as is', we also developed the GRASS-GIS[14] add-ons '*r.stream.watersheds*' and '*r.stream.variables*' that allow the re-calculation of specific environmental variables for a given study area or different spatial grain in an automated and parallelized manner. A subset of the layers can also be visualized online at www.earthenv.org/streams (Data Citation 1).

Methods

Base layer preparation

We used the HydroSHEDS 30 arc sec (hereafter referred to as 1 km) hydrography[12] as a basis for all computations. HydroSHEDS is based on the SRTM[7] Digital Elevation Model (DEM), and consists of a gridded drainage direction layer at 15 arc sec spatial grain (~ 500 m) and a vectorised near-global stream network (with a minimum of 100 upstream cells), and the upscaled products on a 1 km spatial grain[15]. The advantage of using the 1 km stream network for our analyses (as opposed to 500 m) was that it allowed aligning its spatial grain with the environmental source layers to create the freshwater-specific variables, avoiding uncertainties resulting from downscaling data at multiple resolutions[16].

Category	Source	Variable name	Number of source layers	Unit	Variable naming convention	Metric of upstream environment (number of layers)						
						Min	Max	Range	Average	Sum	Distance weighted average	Distance weighted sum
Climate	WorldClim[6]	Minimum monthly air temperature	12	[°C] *10	monthly_tmin_*.nc	12	12	12	12	–	12	–
Climate	WorldClim[6]	Maximum monthly air temperature	12	[°C] *10	monthly_tmax_*.nc	12	12	12	12	–	12	–
Climate	WorldClim[6]	Monthly sum of precipitation	12	[mm]	monthly_prec_*.nc	–	–	–	–	12	–	12
Climate	WorldClim[6]	Long-term hydroclimatic variables	36	[°C] *10 and [mm]	hydroclim_*.nc	–	–	–	11	8	11	8
Topography	HydroSHEDS[12]	Elevation	1	[m]	elevation.nc	1	1	1	1	–	–	–
Topography	HydroSHEDS[12]	Slope	1	[°] * 100	slope.nc	1	1	1	1	–	–	–
Topography	HydroSHEDS[12]	Flow length (upstream cells)	1	count of grid cells	flow_acc.nc	–	–	–	–	1	–	–
Topography	HydroSHEDS[12]	Flow accumulation (watershed size)	1	count of grid cells	flow_acc.nc	–	–	–	–	1	–	–
Land cover	Consensus land-cover[8]	Land cover	12	Percent cover	landcover_*.nc	12	12	12	12	–	12	–
Surface geology	USGS[20]	Geological age	92	count of grid cells	geology_weighted_sum.nc	–	–	–	–	–	–	92
Soil	ISRIC [21]	Soil type	10	%, cmol/kg, kg/m3, cm	soil_*.nc	10	10	10	10	–	10	–

Table 1. Overview of the variable categories, the source and names of the variables, the number of source layers in each category, the unit of measurement, the naming convention of the netCDF files (Data Citation 1), and the metrics calculated for each variable. Here, the numbers correspond to the number of calculated variables (Data Citation 1), the '–' indicates that no variables were created for the given metric.

As the stream network was originally computed in ESRI ArcGIS[17], we started the computation using this software and merged single continents regarding the vectorised stream network and the void-filled Digital Elevation Model (DEM) to single near-global layers (note that the SRTM[7] and hence the HydroSHEDS data does currently not exceed 60°N northern latitude[12]). Here, the use of the ESRI ArcGIS software ensured the alignment of the DEM and streams, avoiding pixel mismatching and geographic projection issues. The stream network was then transformed from 1 km vector data to 1 km grids to facilitate the subsequent computation, while being inclusive as no 'corner cutting' was performed (where diagonal cells could have an additional adjacent cell due to possible inaccuracies in the 1 km DEM[12]). We then used the flow direction layer from HydroSHEDS and identified the Strahler stream order[18] of each stream grid cell (Spatial analyst toolbox, Hydrology, Stream Order[17]). The DEM and stream order layer for the entire stream network were then imported into GRASS GIS 7.0 (ref. 19) under the Linux environment for all subsequent processing steps.

To ensure that the water is flowing downstream *in* the stream channels, we carved the stream network into the DEM by a depth of 22 m. This carving depth was chosen after a step-wise increase of the carving depth by 2 m and by checking for the downstream connectivity along the water courses (note that this carved elevation layer was not used for any elevation-based analyses later on). Running a 3 ×3 cell moving window analysis (*r.neighbors* function in GRASS[14]) along each stream order separately allowed us to remove coarse sinks and peaks, while possible remaining minor pits and peaks were smoothed using the *r.hydrodem*-function. We then calculated a new flow direction layer based on the carved DEM using the *r.watershed*-function, enabling water to flow into multiple down-stream cells (– MFD flag), forcing a positive flow accumulation for potential underestimates (– a flag) and emphasizing the flow accumulation in flat areas (– b flag). This layer needed to be created to avoid any incompatibilities between flow direction layers deriving from GRASS and the one originally created in ArcGIS.

All subsequent calculations were processed in parallel using the High Performance Computing facility at Yale University in compressed chunks of 10,000 files. This was done to maintain a fast I/O load and to avoid the overload of index nodes filling up the hard disk due to the high number of small files. To optimize the parallelization, we split the global stream network again into six continents (North and Central America, South America, Europe, Africa, Asia and Australasia, following the original divisions of[12]), and extracted the coordinates of the gridded stream network along a unique grid ID (hereafter 'gridID' layer). Each single continent and the IDs served as a template for creating the final layers later. In total, the entire near-global stream network consisted of 20,794,251 grid cells, and the spatial extent for all analyses ranged from 60°N to 5°S latitude, and 145°W to 180°E longitude (i.e., the spatial extent of HydroSHEDS).

Sub-catchment delineation

The flow direction layer served as a basis for delineating the upstream sub-catchment of each 1 km stream grid cell using the *r.water.outlet* -function in GRASS[14], where each stream grid cell along the stream network served as a pour point (see Fig. 2a). This is considered the most intense computation process because each sub-catchment needs to be delineated individually based on the flow direction layer, i.e., each outlet needs to 'find' its catchment across the near-global flow direction layer (we used the global layer here to avoid any truncation of sub-catchments located between continents). Besides storing the sub-catchment for each grid cell as a gridded GeoTIFF layer, we also extracted the stream network (water courses) within each sub-catchment separately. For each of these two files—sub-catchment and upstream water courses—we calculated the distance from the outlet to each grid cell within the specific sub-catchment with the grid cells as spatial units ('as the crow flies' and 'as the fish swims', respectively). These distances were used to compute an inverse distance weighting factor (weighting factor = 1/distance) to create distance-weighted averages of the variables (see below). Thus, the environment in and in the immediate vicinity of a given outlet grid cell is considered to have the most influence on the focal grid cell, and this influence decreases with an increasing distance from the given grid cell (Fig. 2a). These three layers (sub-catchment, and the two distance layers) served as a basis for the calculation of the stream variables for each grid cell.

Source layers of the stream variables

To create the stream variables we used globally available and spatially continuous gridded environmental data at a native spatial grain of 1 km as source layers (Table 1). These were (1) interpolated climate from the WorldClim[6] data set (monthly minimum and maximum air temperature, and the monthly sum of precipitation), (2) topography based on the HydroSHEDS void-filled DEM[12] (elevation, slope, flow accumulation and flow length as the sum of contributing grid cells), (3) land cover derived from the consensus land-cover product[8] (representing the percent cover for each of the 12 classes in a given grid cell), (4) surface geology[20] (where each of the 92 discrete classes represents the approximate geological age), and (5) 10 soil classes from the SoilGrids1km data base through ISRIC/WDC-Soils[21].

Because the spatial extent of the climate layers at the coasts were slightly smaller than the spatial extent of the HydroSHEDS stream network due to the upscaling procedure[15], we extended the climate layers by 15 grid cells into the oceans (*r.grow* function) to cover all stream cells and to avoid gaps and truncated coastal streams in the original stream network. The originally discrete surface geology layer was transformed into 92 binary 0–1 variables to facilitate the computation. Predictions regarding the original soil types in the SoilGrids1km database have been produced only for areas with vegetation cover and urban areas. No estimate is provided for shifting sands/deserts and permanent ice areas[21]. For the sake of consistency regarding the extent and number of grid cells (and NoData cells) we replaced the missing values in these areas with zero.

Derived metrics from the source layers

For each sub-catchment, we first overlaid each source layer and clipped all areas not covered by the sub-catchment. We then calculated various metrics of the remaining portion of each source layer that covered the sub-catchment (Table 1), including the minimum, maximum, average, sum, and distance-weighted average and distance-weighted sum of the upstream values using the *r.univar* function and custom code. All units were kept from the source layers (note that temperature values and slope values are multiplied by 10 and 100, respectively, to keep integers and to reduce the file size, see Table 1 and Table 2 (available online only).

All derived metrics were stored for each cell along the stream network, and then merged into near-global GeoTIFF layers using the gridID layer (see Fig. 1). Finally, the upstream averaged and weighted averaged climatic layers were further processed into 19 long-term 'hydro-climatic' variables following the bioclim framework[22] using monthly temperature and precipitation.

Extension to lakes and reservoirs

The HydroSHEDS data set does not contain lakes and reservoirs *per se*, although the Global Lake and Wetlands Database[13] (GLWD, 1 km spatial grain) was integrated during the computation of the hydrography[12]. In other words, the GLWD data set has been used to create the hydrography[12] in terms of the geographic location (thus the lakes and reservoirs spatially match with the stream network), however they are not marked as such in the stream network. The lakes and reservoirs can be identified by a 'fish-bone' structure (Fig. 2b,c) due to zero slope where the flow accumulation algorithm is forced to work properly.

We extracted the lakes and reservoirs with a surface area >0.1 km^2 from GLWD to extend the newly developed stream variables also to lentic areas along the river network. We first identified all lakes and reservoirs that represent single spatial units (*r.clump*—function), and then overlaid and averaged the newly developed variables (Data Citation 1) over each unit to mimic the more static environmental condition in standing waters (*r.stats.zonal* -function). Possibly steep drops at the interface of stream and lake/reservoir grid cells were smoothed by averaging the values in these grid cells at the interface within a 3 × 3 cell neighbourhood at these locations.

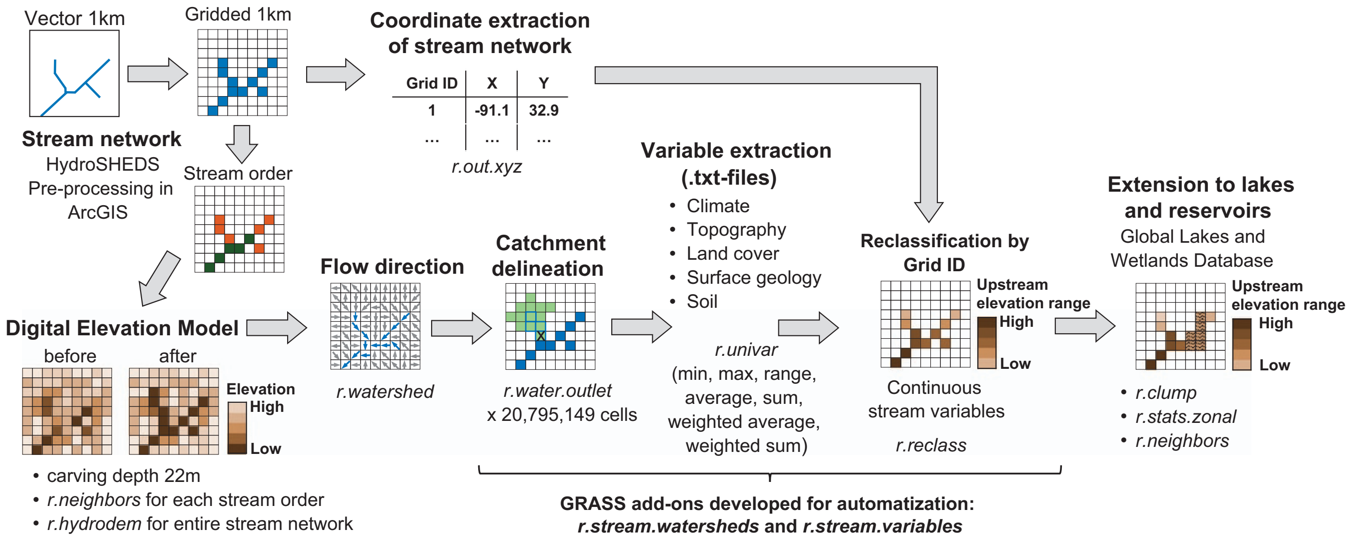

Figure 1. Schematic overview of the main steps for creating the freshwater variables and the *GRASS GIS*[14] *functions* used. First the vectorised HydroSHEDS[12] stream network was transformed into grids and the stream order computed, and this layer was used to recondition the Digital Elevation Model[12] (DEM). The corrected DEM was then used to calculate the flow direction and to delineate the upstream sub-catchment for each cell along the stream network. Various metrics (min, max, range, average, sum, weighted average and weighted sum) were extracted as text-files from existing climate[6], topography[12], land cover[8], surface geology[20], and soil[21] data sets, where each text-file contained the different metrics of a given variable. Once all catchments were processed, the text-files were merged and reclassified into a spatial grid, representing continuous upstream variables. Finally, lakes and reservoirs were extracted from the 1 km gridded Global Lakes and Wetlands Database[13] and the variables were averaged across each single lake and reservoir entity that intersected with the stream network.

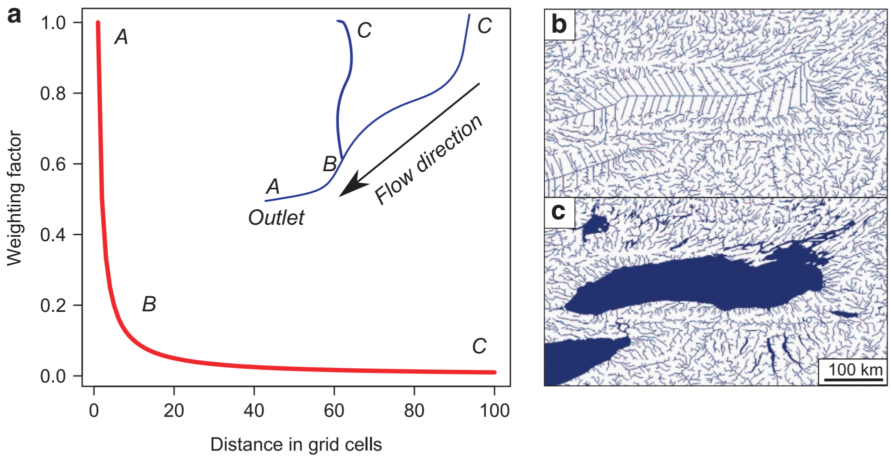

Figure 2. (**a**–**c**) Inverse-distance weighting and integration of lakes and reservoirs. (**a**) Scheme for obtaining the inverse distance weighting factor for calculating the weighted average and sum of the variables. The position of the letters in the stream network (inset) correspond to their approximate position on the curve. (**b**,**c**) Illustration of the modified HydroSHEDS[12] stream network before (**b**) and after (**c**) integrating the lakes and reservoirs of the Global Lakes and Reservoirs Database[13].

In summary, the entire procedure resulted in the computation of stream-specific variables with a continuous surface along the river continuum[23] (Fig. 3), and the variables were subsequently extended to lakes and reservoirs by averaging the upstream environmental conditions for each lake/reservoir connected to the stream network.

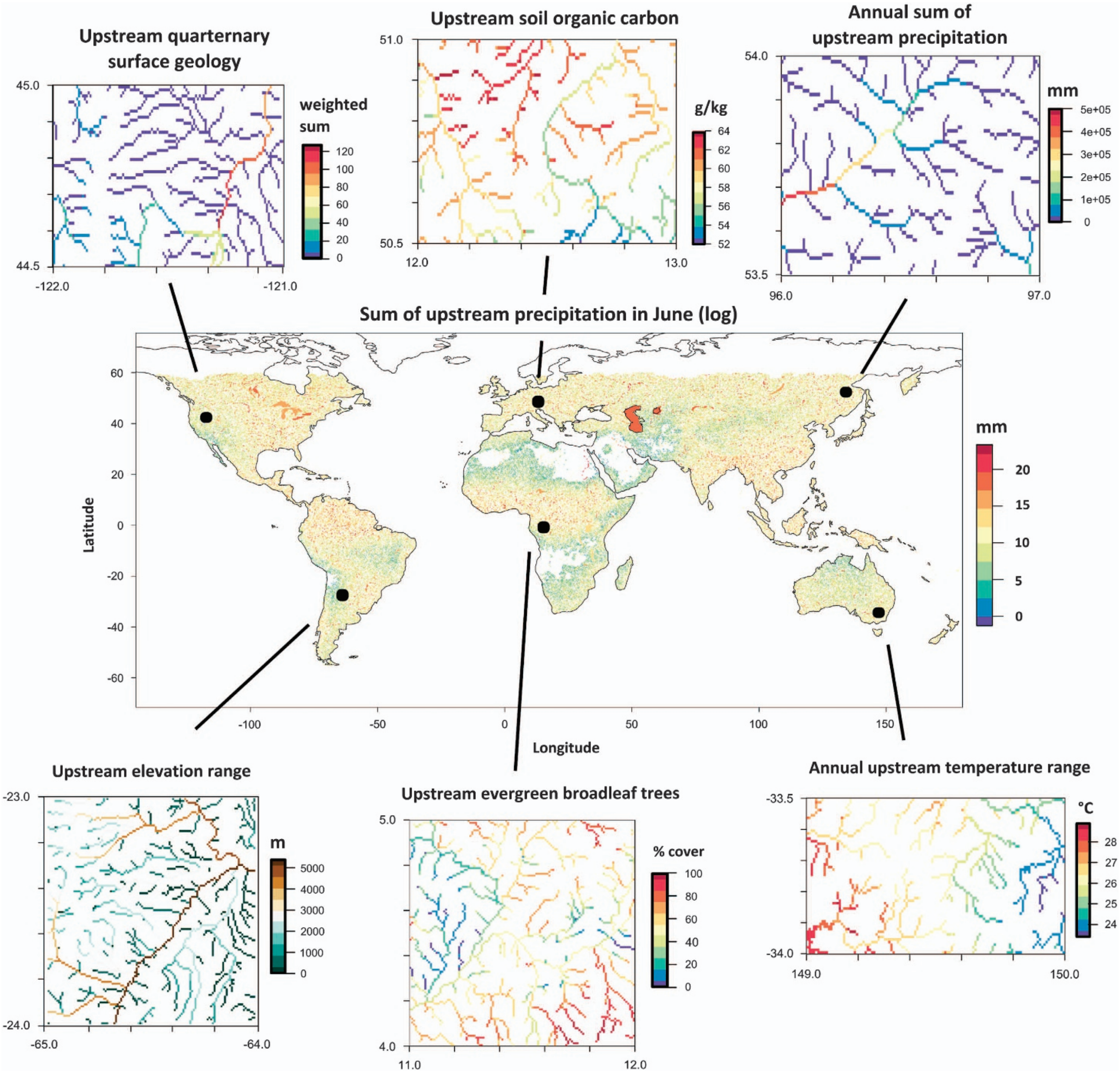

Figure 3. Example maps of the newly developed freshwater variables (Data Citation 1). A global overview of the natural log-transformed sum of precipitation in June (grid cells are aggregated by factor 4 for a better visualisation), and insets representing climate, topography, land cover, surface geology and soil (at the original units and 1 km spatial grain).

Code availability and GRASS GIS add-ons

We developed the '*r.stream.watersheds*' and '*r.stream.variables*' add-ons for GRASS GIS[19] for users who wish to calculate additional variables or metrics, and to apply the method for a specific study area on a different spatial grain. The add-ons automatize and parallelize the main processing chain with a user-specified number of cores, and can be downloaded from the GRASS repository (http://grass.osgeo.org/download/addons/). See the Supporting Information for exemplary code and the tutorial on spatial-ecology.net.

The '*r.stream.watersheds*' add-on delineates the sub-catchment for each stream grid cell and saves the single sub-watershed, and the stream sections within each sub-watershed as GeoTIFF raster files in zipped folders on the hard disk. Users need to provide (i) a flow direction layer generated in GRASS GIS (*r.watershed*) and (ii) a gridded stream network layer. Once this add-on has finished, the second '*r.stream.variables*' add-on can be used: it takes the output from '*r.stream.watersheds*' and overlays the single sub-watersheds with environmental variables to calculate various user-specified metrics simultaneously (number of contributing cells, minimum, maximum, range, average, sum, standard deviation, coefficient of variation). The final output for each metric is a contiguous GeoTIFF layer that is ready to use in a GIS software or e.g. species distribution modeling application. Note that all output layers are stored as integers (Int32 datatype) to reduce the file size, and users can set a scale factor to avoid the truncation of decimals.

The add-ons take advantage of multiple processors to speed up the calculation. For instance, using 8 processors on a 2.66 GHz PC for a stream network that consists of ~90,000 grid cells takes ~1.5 h for the watershed delineation, and ~30 min for each input layer (yielding up to eight output layers).

Data Records

All newly developed 1 km variables are available as compressed netCDF-4 layers on a near-global extent at www.earthenv.org/streams (Data Citation 1) and the Dryad Digital Repository (Data Citation 2). In addition, a visualisation of the layers is given online at www.earthenv.org/streams (Data Citation 1), where users can browse a subset of variables in each category at the stream level and in 1 km resolution.

We provide a variety of different metrics of the upstream environment along the stream network, resulting in a total of 324 layers (Table 1 and Table 2 (available online only)).

Each variable comes in the netCDF-4 format (network Common Data Form version 4) in a cell size of 0.0083333333° (30 arc-seconds, i.e., 30/3,600 of a degree) in the WGS84 coordinate system with an extent of 60°N to 5°S latitude and 145°W to 180°E longitude. All variables consists of 39,000 columns and 13,920 rows. To reduce the file size for download, each netCDF file contains one variable (e.g., upstream average land cover) where the single layers (e.g., 12 landcover classes) are stacked as single bands using the software NCO (ref. 24). The pixel type of the layers ranges from Byte (upstream land cover) to Float64 (upstream precipitation).

All variables were computed based on i) average or sum, or ii) inverse distance-weighted average or distance-weighted sum over the upstream network (except for surface geology where we only computed the distance-weighted sum under the assumption that the impact is dampened along the river continuum, opposed to the continuous accumulation along the river continuum). Regarding the topography, land cover and soil, we also computed the minimum, maximum and the range of the upstream values. Please see Table 2 (available online only) for a full description of all available layers.

Climatic variables consist of (i) 12 monthly minimum and maximum temperature and precipitation variables, and (ii) 19 long-term hydroclimatic variables following the bioclim framework[22]. For each, we computed the average and weighted average for temperature variables (units in °C * 10), and the sum and inverse distance weighted sum for precipitation variables (mm). Likewise, the hydroclimatic variables consist of the monthly average and sum, and the monthly inverse distance weighted average and sum for the aggregations (e.g., upstream sum of precipitation during the warmest quarter). The climate data derives from the WorldClim data base[6], consisting of monthly gridded climate data interpolated between point locations across the globe, and averaged across the years 1950–2000 on a 30-arc-second spatial grain.

The **stream topographic variables** consist of elevation (m), slope (° * 100), flow length and flow accumulation. Elevation and slope data were aggregated using the upstream minimum, maximum, range and average aggregation techniques. Flow accumulation and flow length are computed as the sum of contributing (1 km) grid cells for the entire sub-catchment and only regarding the stream network within the sub-catchment, respectively. The source data for these variables derives from HydoSHEDS[12] (30-arc-second spatial grain) which in turn is based on SRTM[7] data from the year 2000.

Landcover variables consist of 12 classes and depict the upstream percent coverage of a given landcover class: Evergreen/Deciduous Needleleaf Trees, Evergreen Broadleaf Trees, Deciduous Broadleaf Trees, Mixed/Other Trees, Shrubs, Herbaceous Vegetation, Cultivated and Managed Vegetation, Regularly Flooded Vegetation, Urban/Built-up, Snow/Ice, Barren, Open Water). For each class we computed the upstream minimum, maximum, range, average and inverse distance weighted average percent cover. The source data for these variables is the Consensus Landcover dataset[8] with a temporal coverage of 14 years (1992–2006) on a 30-arc-second spatial grain.

Surface geological variables consist of 92 variables indicating the geological age of a given sub-catchment based on the surface geology. Here, the data aggregation type used was the inverse distance-weighted sum of grid cells of a given surface geology type. The source data is derived from USGS[20] via the worldgrids.org portal on a 30-arc-second spatial grain, and the data acquisition period of the source data covers the timeframe from 1960–2000.

The **soil variables** consist of the upstream minimum, maximum, range, average, and inverse distance-weighted average of ten soil variables that were predicted within the standard depth of 2.5 cm (0–5 cm standard thickness): soil organic carbon (g/kg), soil pH (pH * 10), sand content mass fraction (%), silt content mass fraction (%), clay content mass fraction (%), coarse fragments (>2 mm fraction) volumetric (%), cation exchange capacity (cmol/kg), bulk density of the fine earth fraction (kg/m^3), depth to bedrock (R horizon) up to maximum 240 cm (cm), predicted probability of occurence (0–100%) of R horizon across sub-catchment. These variables derive originally from the soilgrid.org database (ISRIC)[21], where the spatially contiguous soil types are derived from predictions. The raw data for the predictions consist of observations at point locations from a 55 year time period (1950–2005). Note that the predictions on a

30-arc-second spatial grain have been produced only for areas with vegetation cover and urban areas, and no estimates are provided for shifting sands/deserts and permanent ice areas[21] (see http://www.isric.org/content/technical-specifications-soilgrids). For the sake of consistency regarding the extent and number of grid cells (and NoData cells) we replaced the missing values in these areas with zero.

Technical Validation

Quality control of the sub-catchment delineation

We checked for errors in the variables (Data Citation 1) that may arise from incorrect downstream routing of the flow accumulation prior to the extension to lakes and reservoirs. Possible reasons for an incorrect routing could be (i) remaining pits and peaks in the DEM (note that a more extensive 'cleaning' of the DEM based on a fixed threshold in one area can lead to an incorrect flattening in other areas); (ii) possible inaccuracies and intermittences in the underlying stream network[12,15] due to the 1 km spatial grain (since the HydroSHEDS stream network was upscaled from the original 90 m spatial grain), (iii) cells that are located at the 60°N latitude boundary may have a truncated catchment, (iv) coastal areas that have a very flat/constant elevation without any gradients. Moreover, (v) while the stream network layer was kept 'inclusive', i.e., all stream cells from the HydroSHEDS river network were kept for our analyses, this also required addressing to deal with these cells where the water flow could have been bypassed during the catchment delineation (no watershed was created). In addition, (vi) the differences in flow direction algorithms between GRASS and ArcGIS to compute the upstream watersheds can differ, leading to slightly different flow direction maps.

These cells were identified for each stream order separately through their lower than expected upstream flow accumulation given the stream order. For instance, a sub-catchment of a 2nd stream order needs to consist of at least three cells (two 1st order cells merging to a 2nd order cell). This scheme was applied to all stream orders, where the minimum number of upstream grid cells follows the hierarchy of Strahler's stream order[18] (with a minimum required number of 3, 7, 15, 31, 63, 127, 255, and 511 grid cells for 2nd to 9th order streams, respectively).

In total 314,084 grid cells were identified to have a truncated sub-catchment, which corresponds to 1.5% of all stream grid cells on the near-global extent (the layer 'missing_cells.nc' is provided in the 'quality_control.nc' file, Table 2 (available online only)). The values for each environmental variable in these cells were corrected by first assigning the maximum value of the surrounding 3×3 cell neighbourhood matrix for each stream order separately. This reduced the number of cells with incorrect values to 6,924, i.e., 97.8% of initially incorrect cells could be filled. However, as not all cells could be corrected, this also meant that their immediate neighbourhood was assigned incorrectly, and that the flow direction was routed incorrectly (note that these cells occurred mainly in flat areas such as in lakes or coastal areas). A second round of correction was undertaken by repeating this procedure iteratively in the neighbouring cells until all remaining incorrectly assigned cells were filled.

While this procedure forced calculated layer values (e.g., upstream average temperature) into stream reaches that would be otherwise underestimated, it left a counterpart of stream cells with overestimated values. To identify these river reaches, we overlaid the flow accumulation and stream order layers in Google Earth Engine with the global satellite topography imagery (TruEarth 15 m), and visually checked whether the upstream flow accumulation matches with the network topology in the satellite image and the stream order. In total, 3301 manual corrections were made globally, all near coasts, since these were the areas where the flow accumulation routing could fail due to flat areas. We decided to omit these stream cells from all layers due to high uncertainties in the flow direction, and provide a separate layer to users to identify these cells ('cells_removed.nc' in the 'quality_control.nc' file, Table 2 (available online only)).

Validation

Background. Among the most important predictors for delineating freshwater species distributions are water temperature and discharge[5].

Developing continuous stream layers directly addressing these variables would be ideal, but requires spatially and especially temporally continuous and detailed, high-quality information such infiltration rates, evapotranspiration rates, snow/ice cover and melt, and direct anthropogenic water abstraction and water use. These data are however only partially available on the required spatial and temporal resolution. Thus, our aim here is to validate the overall pattern of the variables we were able to produce, and to show how upstream processes influence those located more downstream over large spatial scales but fine grains via the transport characteristics of the stream network.

We therefore assessed how well the newly-developed monthly upstream temperature and precipitation layers may be suitable proxies for these variables. Specifically, we assessed how well local measurements of monthly water temperature and discharge[25–27] (Supplementary Fig. S1a,b) are predicted by the newly-developed freshwater variables (Data Citation 1) and how much this fit was improved when using upstream *average* or *distance-weighted average* variables instead. We plotted the relationships between observed data and the variables and used the R^2 from linear regressions (in log space) as a measure of goodness of fit (Table 3 and Figs 4 and 5, precipitation and discharge data were log-transformed for plotting). All calculations were done in the software R (ref. 28).

Month	Monthly minimum temperature		Monthly maximum temperature		Monthly discharge	
	Upstream average	Distance weighted average	Upstream average	Distance weighted average	Upstream sum	Distance weighted sum
January	**0.57**	0.13	0.43	**0.44**	**0.86**	0.03
February	**0.57**	0.22	0.77	**0.78**	**0.87**	0.03
March	**0.72**	0.35	0.83	**0.84**	**0.89**	0.03
April	0.65	**0.71**	0.74	**0.75**	**0.84**	0.02
May	0.68	**0.74**	**0.66**	0.63	**0.71**	0.01
June	0.70	**0.75**	**0.59**	0.55	**0.52**	0.00
July	0.64	**0.69**	**0.58**	0.55	**0.30**	0.00
August	0.62	**0.68**	**0.62**	0.60	**0.23**	0.00
September	0.64	**0.71**	**0.76**	**0.76**	**0.42**	0.00
October	0.57	**0.64**	**0.73**	**0.73**	**0.72**	0.01
November	**0.59**	0.55	0.76	**0.77**	**0.83**	0.01
December	**0.61**	0.27	0.82	**0.84**	**0.85**	0.02

Table 3. R^2 values derived from the linear regression between observed monthly minimum and maximum stream temperature, and the upstream average and distance weighted averaged air temperature within the stream network, as well as between the observed monthly discharge and upstream sum and distance weighted sum of precipitation across the sub-catchments. The monthly R^2 values correspond to the single plots in the Figs 4 and 5, respectively. Higher values among each pair (average/sum versus weighted average/sum) are in bold.

Monthly upstream temperature versus observed monthly water temperature. The observed stream temperature data was compiled from various sources (National Water Quality Monitoring Council[25], the GloRICH data base[26] and from the Environment Agency[27], Supplementary Fig. S1a) and aggregated to monthly minimum and maximum values to match the temporal resolution of the temperature variables (Data Citation 1). First, only observations from 1950–2000 were used as this matches the timeframe of the WorldClim[6] data set. Second, at least three observations per month for three years were required, resulting in a minimum of nine observations that were averaged for each site. This yielded 319 unique locations with 1740 observations that were moved to the closest grid cells in the stream network using a 'snapping' distance of 3 km using 'RasterTools' [29]. This distance was chosen due to possible spatial inaccuracies in the stream network while retaining as much of the data as possible for the validation (Supplementary Fig. S1a). We then extracted the upstream average and distance-weighted average of the temperature variables at those locations.

Goodness of fit (R^2) regarding the minimum monthly temperature was higher from November to March (ranging from 0.57 to 0.72) when the upstream average temperature was used (as opposed to the distance-weighted temperature), and the opposite pattern was found from May until October (R^2 ranges from 0.64 to 0.75, Table 3, Fig. 4).

For maximum monthly temperature the pattern was reversed, and the distance-weighted averaged values had in general a higher goodness of fit during the winter months (R^2 ranged 0.44–0.84) than the evenly averaged air temperature (Table 3). During the summer months, the averaged values indicated a better fit than the distance-weighted averaged vales (R^2 ranges 0.58–0.76, Table 3, Fig. 4).

Stream temperature is strongly related to air temperature, and stream/air temperature are generally highly correlated at longer timeframes (such as monthly data, ref. 30). Plotting the average minimum and maximum water and sub-catchment air temperatures over the course of the year (Supplementary Fig. S2) shows how the observed stream temperature lies mostly between the minimum and maximum upstream air temperature. However, due to the heat capacity of streams, a lag-effect regarding the adaptation of stream to air temperatures can be observed (Supplementary Fig. S2 and ref. 30), such that for certain months, the air temperature in locations more distant (upstream average) or in the immediate upstream vicinity (distance-weighted average) may be better proxies of stream temperature, respectively. In this regard, high resolution data such as regional climate datasets can give more insights into these patters using e.g., the diurnal range of stream temperatures, and could be calculated using the newly created add-ons *r.stream.watersheds* and *r.stream.variables* along with a high-resolution stream network and regional climate datasets.

Monthly upstream precipitation versus observed monthly discharge. Monthly observed discharge data was downloaded from the Global River Discharge v.1.1 data set[31] (Supplementary Fig. S1b). This data set contains monthly observed discharge in m^3/sec from 1807–1991 at 1018 stations across the globe. Only observations after 1950 and from sites deemed reliable were used (flagged as such in the data set).

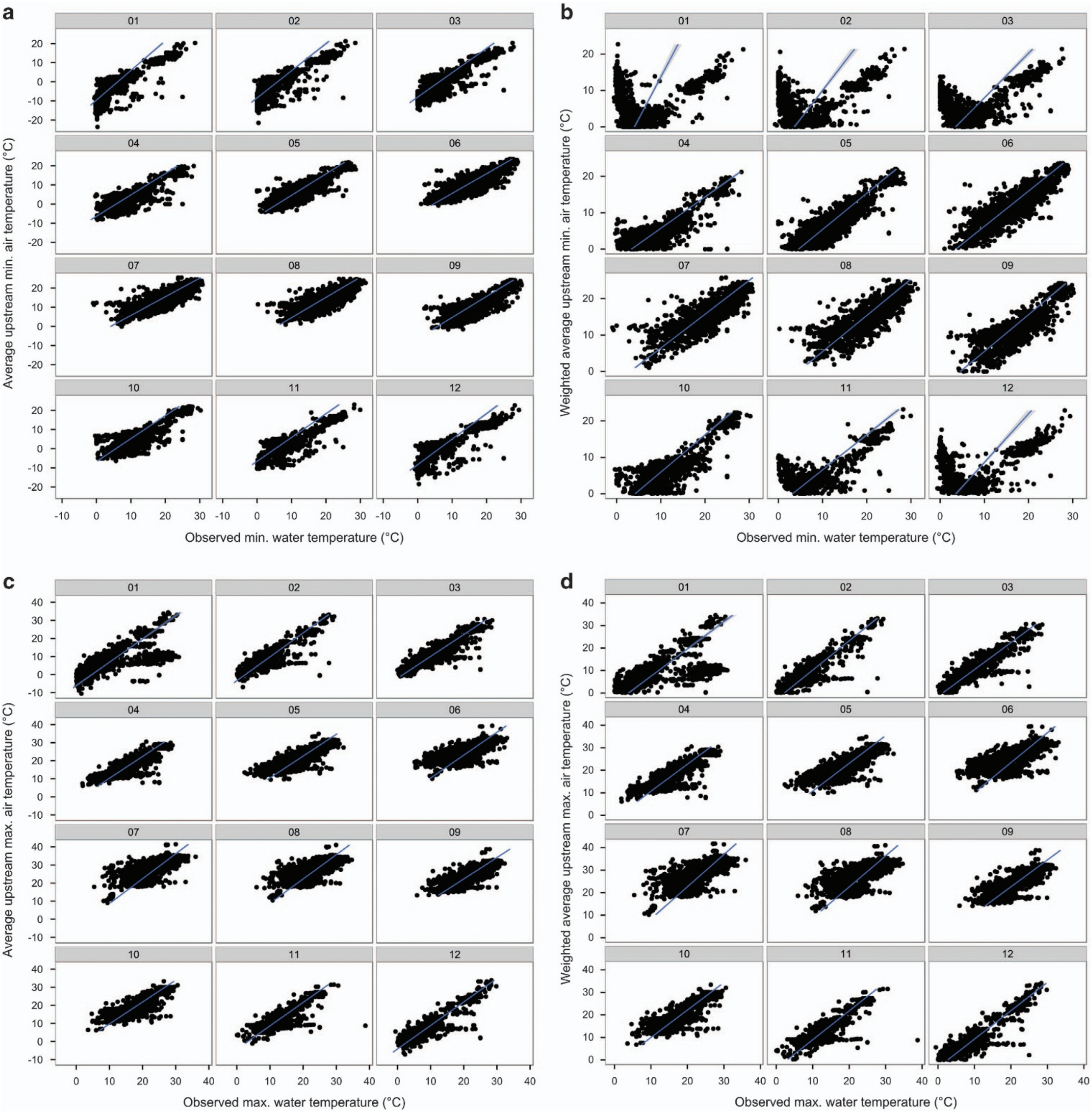

Figure 4. (**a**–**d**) Relationship between observed stream temperature and newly-developed temperature variables. Observed monthly minimum (**a**,**b**) and maximum (**c**,**d**) stream temperature plotted against the upstream average (**a**,**c**) and distance-weighted average (**b**,**d**) air-temperature for each month (plot headers correspond to January to December). Lines correspond to the fit of the linear regression with 95% confidence intervals.

We aggregated the monthly observations for each available year per station, where at least observations for three months needed to be available. A subset of 582 sites were marked reliable (71% of all stations) and were snapped to >= 3rd order streams (3 km tolerance), since most gauging stations are located at larger rivers, but could be incorrectly snapped to smaller nearby confluences, leaving 568 sites for our validation (69%, Supplementary Fig. S1b). We then extracted the upstream sum and distance-weighted sum of the precipitation variables (Data Citation 1) at those locations.

Accounting for the upstream sum of precipitation gives a better estimate of the observed discharge, and R^2 values ranged from 0.23 in August to 0.89 in March (Table 3, Fig. 5). It is therefore considered a more

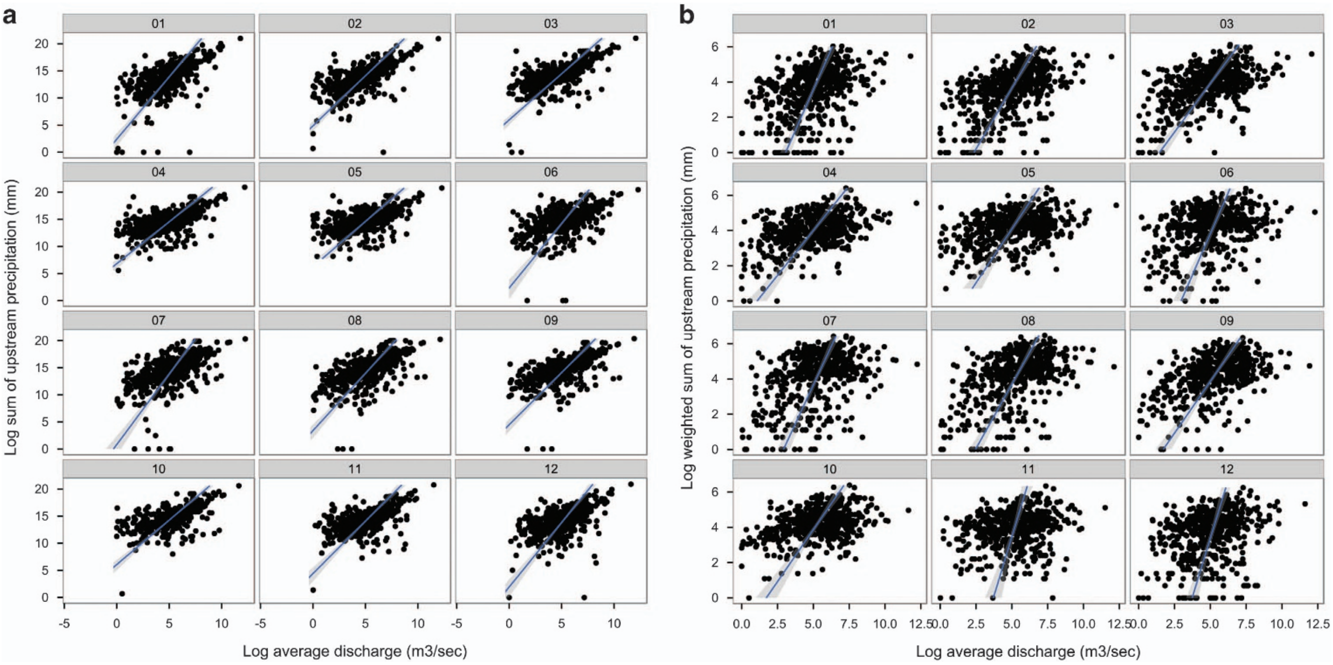

Figure 5. (**a**,**b**) Relationship between observed discharge and newly-developed precipitation variables. Natural log-transformed observed monthly average discharge plotted against the natural log-transformed sum (**a**) and distance-weighted sum (**b**) of upstream precipitation for each month (plot headers correspond to January to December). Lines correspond to the fit of the linear regression with 95% confidence intervals.

robust proxy for discharge than the distance weighted sum -metric, as the latter mentioned down weights the influence of precipitation at locations that are more distant to the discharge gauging station. Low goodness of fit scores of the upstream sum -metric are mostly due to low precipitation but high discharge patterns due to snow melt in winter/spring, and anthropogenic impact such as water management and release from reservoirs in arid regions during some months.

Usage Notes

The newly developed variables (Data Citation 1) have a variety of potential uses in freshwater ecology, conservation and spatial biodiversity science. They are suited for applications the modelling and mapping of the spatial variation in freshwater species and communities. For example, the variables (Data Citation 1) can be used to annotate community data or point occurrences of freshwater species to explore their position in multivariate environmental niche space, to ascertain and predict the environmental limits to their distribution, or map their potential distribution. For modelling freshwater species distributions and most other use cases, we encourage choosing complementary variables from different variable categories to limit collinearity. An example of a useful combination may be the elevation *range* combined with the *average* upstream temperature, the upstream *sum* of precipitation, the *maximum* upstream land or soil cover and *weighted sum* of the surface geology cover. In certain cases, the upstream range and maximum value of a land cover variable can be more useful than the average, as anthropogenic influences ('Urban/Built-up') can depict the retention of land cover effects in downstream freshwater habitats rather than a dampening/uptake (which is indicated by the average and weighted average).

We provide the distance-weighted precipitation variables to users who wish to explore the precipitation patterns in the immediate vicinity of a given location, however the validation showed that they are not suitable as a proxy for discharge.

Further layers regarding future climate and land use projections are currently under development, and we encourage potential users to visit www.earthenv.org/streams for updates on the layers.

We provide example code in the Supplementary Information to load and process the variables in the netCDF format in R, and to use the GRASS add-ons '*r.stream.watersheds*' and '*r.stream.variables*'.

References

1. Balian, E. V., Segers, H., Lévèque, C. & Martens, K. The Freshwater Animal Diversity Assessment—an overview of the results. *Hydrobiologia* **595,** 627–637 (2008).
2. Strayer, D. L. & Dudgeon, D. Freshwater biodiversity conservation: recent progress and future challenges. *J. N Am Benthol. Soc.* **29,** 344–358 (2010).
3. Assessment, M. E. *Ecosystems and human well-being: Biodiversity synthesis*. (World Resources Institute, 2005).

4. Elith, J. & Leathwick, J. R. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annu. Rev. Ecol. Evol. S* **40,** 677–697 (2009).
5. 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,** 45–61 (2015).
6. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25,** 1965–1978 (2005).
7. Jarvis, A., Reuter, H. I., Nelson, A. & Guevara, E. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90 m Database (http://srtm.csi.cgiar.org) (2008).
8. Tuanmu, M. N. & Jetz, W. A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. *Global Ecol. Biogeogr.* **23,** 1031–1045 (2014).
9. Barbet-Massin, M., Thuiller, W. & Jiguet, F. How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? *Ecography* **33,** 878–886 (2010).
10. Kuemmerlen, M. *et al.* Integrating catchment properties in small scale species distribution models of stream macroinvertebrates. *Ecol Model* **277,** 77–86 (2014).
11. Heino, J., Virkkala, R. & Toivonen, H. Climate change and freshwater biodiversity: detected patterns, future trends and adaptations in northern regions. *Biol. Rev. Camb. Philos. Soc.* **84,** 39–54 (2009).
12. Lehner, B., Verdin, K. & Jarvis, A. New global hydrography derived from spaceborne elevation data. *Eos, Transactions, AGU* **89,** 93–94 (2008).
13. Lehner, B. & Döll, P. Development and validation of a global database of lakes, reservoirs and wetlands. *J. Hydrol.* **296,** 1–22 (2004).
14. Neteler, M., Bowman, M. H., Landa, M. & Metz, M. GRASS GIS: a multi-purpose Open Source GIS. *Environ. Modell Softw.* **31,** 124–130 (2012).
15. Lehner, B. HydroSHEDS technical documentation, version 1.2. Conservation Science Program, World Wildlife Fund US Washington, DC 20037. Available from www.hydrosheds.org (2013).
16. Dendoncker, N., Bogaert, P. & Rounsevell, M. A statistical method to downscale aggregated land use data and scenarios. *Journal of Land Use Science* **1,** 63–82 (2006).
17. Esri. *ArcGIS Desktop: Release 10.2.* (Environmental Systems Research Institute, 2013).
18. Strahler, A. N. Quantitative analysis of watershed geomorphology. *Trans Am Geophys Union* **38,** 913–920 (1957).
19. Geographic Resources Analysis Support System (GRASS) Software, Version 7.0. Open Source Geospatial Foundation. http://grass.osgeo.org (2015).
20. Ribeiro, E. & Hengl, T. Resampled global surface geology available at http://worldgrids.org/doku.php?id=wiki:geaisg3&s[] =geology. Based on data from the International Surface Geology project http://certmapper.cr.usgs.gov/data/envision/index.html?widgets=geologymaps.
21. Hengl, T. *et al.* SoilGrids1km—Global Soil Information Based on Automated Mapping. *PLoS ONE* **9,** e105992 (2014).
22. Busby, J. R. BIOCLIM-a bioclimate analysis and prediction system. *Plant Protection Quarterly* **6,** 8–9 (1991).
23. Vannote, R. L., Minshall, G. W., Cummins, K. W., Sedell, J. R. & Cushing, C. E. River Continuum Concept. *Can J Fish Aquat Sci.* **37,** 130–137 (1980).
24. Zender, C. S. Analysis of Self-describing Gridded Geoscience Data with netCDF Operators (NCO). *Environ Modell Softw* **23,** 1338–1342 (2008).
25. National Water Quality Monitoring Council. Water quality data provided by USGS, EPA and USDA, available at http://waterqualitydata.us/.
26. Hartmann, J., Lauerwald, R. & Moosdorf, N. A Brief Overview of the GLObal RIver Chemistry Database, GLORICH. *Procedia Earth and Planetary Science* **10,** 23–27 (2014).
27. Environmental Agency: Surface Water Temperature Archive for England and Wales. available at http://www.geostore.com/environment-agency/WebStore?xml = environment-agency/xml/ogcDataDownload.xml.
28. R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, ISBN 3-900051-07-0, URL http://www.R-project.org (2015).
29. Verbruggen, H. RasterTools: moveCoordinatesToClosestDataPixel.jar version 1.03 http://www.phycoweb.net/software/rasterGIS/index.html (2012).
30. Caissie, D. The thermal regime of rivers: a review. *Freshw Biol.* **51,** 1389–1406 (2006).
31. Vorosmarty, C. J., Fekete, B. M. & Tucker, B. A. *Global River Discharge* 1807–1991, V. 1.1 (RivDIS) Data set. Available on-line [http://www.daac.ornl.gov]from Oak Ridge National Laboratory Distributed Active Archive Center, (1998).

Data Citations

1. Domisch, S., Amatulli, G. & Jetz, W. *EarthEnv* http://www.earthenv.org/streams (2015).
2. Domisch, S., Amatulli, G. & Jetz, W. *Dryad* http://dx.doi.org/10.5061/dryad.dv920 (2015).

Acknowledgements

We thank all primary data providers for providing the data: HydroSHEDS, Global Lakes and Wetlands Database (GLWD), the WorldClim database, the Consensus Land Cover database, US Geolgocical Survey (USGS), and the International Soil Reference and Information Centre (ISRIC) and soilgrids.org database. We wish to thank Jens Hartmann for sharing the GLORICH data, Jim Stagge and the Jetz Lab for valuable discussions and support, Steve Weston for help and advice regarding the cluster computation, Robert Hijmans for advice regarding the usage of the netCDF format in R, and Jeremy Malczyk for the online visualization of the layers. This study was funded by the Yale Climate & Energy Institute (YCEI), the Yale (YIBS) Program in Spatial Biodiversity Science and Conservation and supported in part by the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center. S.D. was funded by the German Research Foundation DFG (grant DO 1880/1-1). W.J. acknowledges funding from US National Science Foundation Grants DEB 1026764, DEB 1441737 and DBI 1262600, and NASA Biodiversity Grant NNX11AP72G.

Author Contributions

S.D. computed the data layers and wrote the manuscript; G.A. drafted the codes, contributed to the cluster computation and developed the GRASS GIS add-ons. S.D., G.A. and W.J. designed the study and all authors discussed the results and commented on the manuscript.

Additional Information

Table 2 is only available in the online version of this paper.

Supplementary Information accompanies this paper at http://www.nature.com/sdata

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Domisch, S. *et al.* Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution. *Sci. Data* 2:150073 doi: 10.1038/sdata.2015.73 (2015).