HydroBASINS

Global watershed boundaries and sub-basin delineations derived from HydroSHEDS data at 15 second resolution

Technical Documentation Version 1.c (with and without inserted lakes)

prepared by Bernhard Lehner
(bernhard.lehner@mcgill.ca)
July, 2014

1. Overview and background

This documentation accompanies a series of polygon layers that depict watershed boundaries and sub-basin delineations at a global scale. The goal of this product is to provide a seamless global coverage of consistently sized and hierarchically nested sub-basins at different scales (from tens to millions of square kilometers), supported by a coding scheme that allows for analysis of watershed topology such as up- and downstream connectivity.

Watershed boundaries provide important geospatial units for many applications, but at a global scale there is a lack of high-quality mapping sources. The HydroSHEDS database (Hydrological data and maps based on SHuttle Elevation Derivatives; Lehner et al. 2008; for more information see http://www.hydrosheds.org) provides hydrographic data layers that allow for the derivation of watershed boundaries for any given location based on the near-global, high-resolution SRTM digital elevation model. Using this hydrographic information, watersheds were delineated in a consistent manner at different scales, and a hierarchical sub-basin breakdown was created following the topological concept of the Pfafstetter coding system (Verdin & Verdin 1999). The resulting polygon layers are termed HydroBASINS and represent a subset of the HydroSHEDS database.

The HydroBASINS product has been developed on behalf of World Wildlife Fund US (WWF), with support and in collaboration with the EU BioFresh project, Berlin, Germany; the International Union for Conservation of Nature (IUCN), Cambridge, UK; and McGill University, Montreal, Canada. Major funding for this project was provided to WWF by Sealed Air Corporation; additional funding was provided by BioFresh and McGill University.

Citations and acknowledgements of the HydroBASINS data should be made as follows:

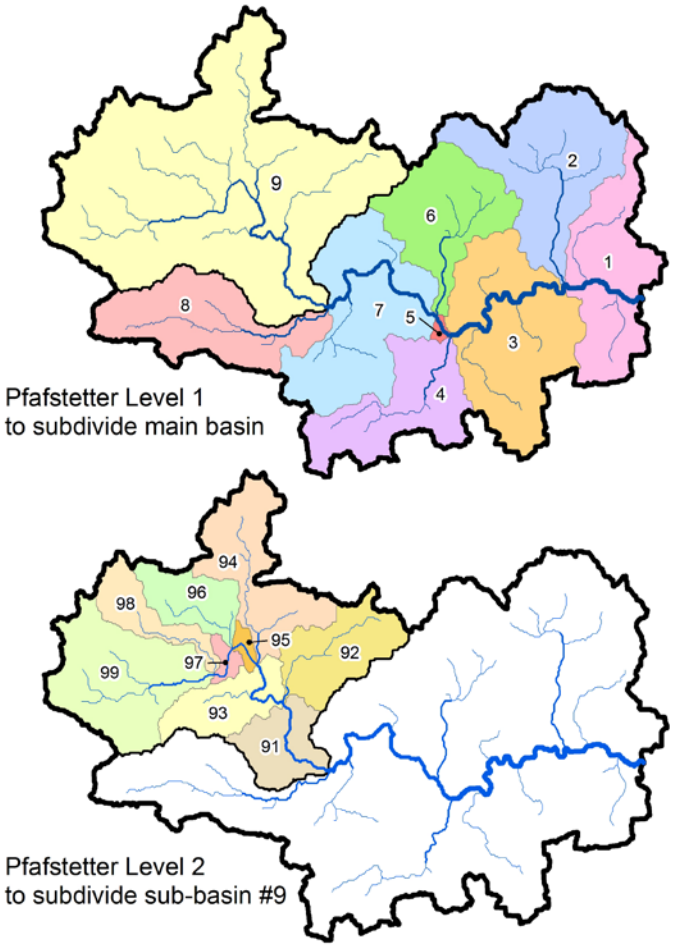
> *Lehner, B., Grill G. (2013): Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15): 2171–2186. Data is available at www.hydrosheds.org.*

2. Methods and data characteristics

All HydroBASINS layers were derived from World Wildlife Fund's HydroSHEDS data (Lehner et al. 2008; Lehner and Grill 2013) based on a grid resolution of 15 arc-seconds (approximately 500 m at the equator). For more information please refer to the Technical Documentation of HydroSHEDS at http://www.hydrosheds.org. It should be noted that the quality of the HydroSHEDS data is significantly lower for regions above 60 degrees northern latitude as there is no underlying SRTM elevation data available and thus a coarser scale DEM has been inserted (HYDRO1k provided by USGS; see http://gcmd.nasa.gov/records/GCMD_HYDRO1k.html).

2.1 Creation of sub-basin geometry

An important characteristic of any sub-basin delineation is the sub-basin breakdown, i.e. the decision of when and how to subdivide a larger basin into multiple tributary basins. Standard GIS tools offer the possibility to break out sub-watersheds at any confluence where the inflowing branches (i.e., a tributary and its main stem) exceed a certain size threshold, typically measured as the number of upstream pixels or the upstream catchment area. HydroBASINS follows the same concept and divides a basin into two sub-basins at every location where two river branches meet which each have an individual upstream area of at least 100 km^2. It should be noted that this concept still allows for smaller sub-basins to occur, namely the inter-basins between the tributaries (which can have any smaller size). Also, sub-basins can grow to sizes much larger than the 100 km^2 threshold if there is no tributary joining the main stem for a long distance. This inconsistency due to "oversized" sub-basins has been addressed and reduced in HydroBASINS by forcing additional subdivisions for every sub-basin larger than 250 km^2: these polygons are split into appropriately sized sub-basins by introducing break points along their main stem rivers.



2.2 Nested grouping and coding

A second critical feature of sub-basin delineations is the way the sub-basins are grouped or coded to allow for the breakout of nested sub-basins at different scales, or to navigate within the sub-basin network from up- to downstream. One of the easiest methods for navigation is to provide the ID of the next downstream object, which allows for moving from object to object in order to traverse the network. As for nesting and topological concepts, the 'Pfafstetter' coding system is frequently used due to its

relative simplicity and ease of application. The basic principle of the Pfafstetter coding is that a larger basin is sequentially subdivided into 9 smaller units (the 4 largest tributaries, coded with even numbers, and the 5 inter-basins, coded with odd numbers). Thus, the next finer resolution of a sub-basin delineation is achieved at the next Pfafstetter level by adding one digit to the code of the previous level. A more detailed description of the Pfafstetter coding is provided in literature (e.g., Verdin and Verdin 1999). The concept has successfully been applied both globally (e.g., HYDRO1k; USGS 2000) and regionally (e.g., Australia; Stein et al. 2014).

The HydroBASINS product follows the Pfafstetter concept and provides levels 1 to 12 globally. There are several aspects of the Pfafstetter coding, however, which have not been designed for global applications, thus the following modifications and updates were implemented:

- The first 3 levels of Pfafstetter codes for HydroBASINS were assigned manually. Level 1 distinguishes 9 continents (1 = Africa; 2 = Europe; 3 = Siberia; 4 = Asia; 5 = Australia; 6 = South America; 7 = North America; 8 = Arctic of North America; 9 = Greenland); Level 2 splits each continent into up to 9 large sub-units; and at Level 3 the largest river basins of each continent start to break out. From Level 4 onwards, the breakdown follows the traditional Pfafstetter coding (with further modifications as explained below).

- To provide a unique Pfafstetter code for every basin worldwide, the seeding of first-level coding numbers, as well as the successive seeding of all basins as they start to break out at higher Pfafstetter levels, starts at the north-eastern tip of the African continent and continues counter-clockwise around the continental coasts (Europe, Siberia, Asia, followed by Australia and the Americas). Islands are inserted into the sequence at appropriate (nearest) locations.

- The original Pfafstetter scheme was not designed to provide nested coding for islands. To incorporate islands as sub-units of continents, they have been grouped manually in HydroBASINS following the hierarchical Pfafstetter concept. For example, at Level 1 New Zealand is part of the Australian continent, at Level 2 both islands of New Zealand break out as one unit, and at Level 3 the North and South Islands are separated into their own units.

- Similarly to islands, there is no provision in the Pfafstetter scheme regarding the integration of endorheic basins (i.e., inland sinks that do not drain to the ocean). In HydroBASINS, endorheic basins have been grouped and then seeded with initial Pfafstetter codes manually (similar to islands) to provide a hierarchical nesting. The nesting may start with lumped groups of endorheic basins at lower levels and ends with the individual endorheic basins at higher levels. Once an endorheic basin is isolated, the standard Pfafstetter coding is used to continue the subdivision of the endorheic basin into smaller sub-basins.

- A second issue regarding endorheic basins—in particular for smaller ones—is that at the lower Pfafstetter levels only large river basins are broken out. Many large river basins, however, contain small endorheic basins inside them or adjacent to their watershed divide. For example, the Nile Basin contains many small endorheic basins in the dry middle region which are not connected to the main river via surface water flows yet are typically considered to be part of the overall basin (as they may be hydrologically connected via groundwater). In order to allow for these endorheic basins to be lumped with the larger river basin, and to

enable topological queries in which the endorheic discontinuities can be eliminated (traversed) to create contiguous regions, some endorheic sinks have been assigned a 'virtual' connection to an appropriate downstream polygon. These virtual connections can be identified in the attribute table (see details for attribute table below), and the user can decide whether or not to terminate the routing at an endorheic sink of this type.

- Another structural problem of the Pfafstetter code is apparent for very small coastal basins that drain into the ocean in between larger river basins. If left untreated, there would be a very large number of very small coastal basins globally, and the Pfafstetter coding would start to subdivide them into even smaller units. To avoid this spatial inconsistency, small coastal basins have been lumped in HydroBASINS to form their own coastal drainage units between larger river basins. The lumping was performed in a semi-automated procedure with some manual decisions about their grouping, and small islands close to the coast were included into the coastal basins where appropriate. Lumped coastal basins were allowed to reach a maximum individual size of up to 500 km^2 (or 700 km^2 for the lower quality areas north of 60ºN).

- Finally, an inherent issue of the Pfafstetter coding is the requirement to break out exactly 9 sub-basins at the next higher level. This can lead to highly inconsistent (and randomly distributed) sub-basin sizes at the same level. For example, there may be 4 large tributaries that form 4 large sub-basins, yet the 5 inter-basins may be very small. At the next Pfafstetter level, the 9 sub-basins each are subdivided again into 9 nested sub-unit, thus the size discrepancy can be passed along or even amplified (if the small inter-basins in turn contain more small inter-basins) from level to level. To reduce this inconsistency in sub-basin size, the breakdown in HydroBASINS was modified (in an automated procedure) to be guided by level-dependent tolerance ranges of sub-basin areas below which small sub-basins are allowed to skip a subdivision at the next higher Pfafstetter level. These skips are indicated with a value of 0 in the Pfafstetter coding digit of the respective level.

2.3 Two formats of HydroBASINS: standard (without lakes), and customized (with lakes)

Version 1.c of HydroBASINS was developed in two formats:

a) In **Format 1 (standard; without lakes)**, all sub-basins follow the standard concept of nested watersheds and are strictly derived from the underlying flow direction grids of HydroSHEDS by applying the area thresholds and modification rules as described above. This means that sub-basins are also created and coded in areas which (in reality) are covered by lakes (e.g., inside Lake Victoria). The shape of these particular sub-basins may not reflect true sub-basin boundaries, but their topology (i.e., interconnectedness) remains valid throughout the lake. Format 1 has a cleaner coding with a simpler structure than Format 2, and users can visually overlay additional lake layers to mask the affected sub-basins if needed. Format 1 consists of 12 individual polygon layers representing the 12 Pfafstetter levels. Additionally, an extra layer (Level 0) is provided which includes all sub-basins at their smallest breakdown with the full coding sequence of levels 1-12 in the attribute table. Users can derive any Pfafstetter level directly from the 'Level 0' layer by dissolving the sub-basin polygons based on the desired level code.

b) **Format 2 (customized; with inserted lakes)** was designed to specifically accommodate the special requirements of the BioFresh project and was produced in collaboration with IUCN. This dedicated version follows the same concept as the standard format, yet a global lake layer was integrated into the data. Lake polygons were taken from the Global Lakes and Wetlands Database (GLWD; Lehner and Döll 2004) and were clipped into the sub-basin polygons of HydroBASINS.

IMPORTANT NOTE: While every effort was made to fully integrate the lakes into the topological coding structure of HydroBASINS, some specific problems and artifacts were introduced by the lake features which necessitated a modification from the original coding scheme. For example, lakes may overlap and thus link multiple sub-basins from different tributaries; artificial reservoirs may link formerly independent watersheds; or lagoons may extend into the ocean and/or connect neighboring river basins. Strict up- and downstream topology is not always possible for these special circumstances, and connectivity may thus be inconsistent or was actively broken for certain constellations to avoid loops in the search algorithms. Another problem is that the GLWD lake polygons and the sub-basin polygons of HydroBASINS are not always well aligned to each other but may expose a spatial shift (in these cases, typically the local accuracy of the GLWD data is of lower quality). Due to this shift, some lakes (in particular small ones) may not be correctly registered to their surrounding sub-basins but may be recognized as 'downstream' instead of 'inside', or 'besides' instead of 'upstream', etc. Finally, some connectivity rules are more complex in Format 2. For example, there may be one or multiple lake polygons entirely within a sub-basin at a certain Pfafstetter level. In this case, the next downstream polygon of each lake is defined to be the surrounding sub-basin. This means in turn, however, that the lakes are assigned to be 'upstream' of the surrounding sub-basin, although this is only correct for parts of the sub-basin, and they have no up- or downstream connectivity among each other, even if in reality they may be situated along the same river.

Two additional modifications were introduced in the lake version of HydroBASINS upon request by BioFresh/IUCN:

- Lakes start to appear at certain Pfafstetter levels based on their size. This is to prevent small lakes from appearing as individual polygons already on low Pfafstetter levels. Lakes $\geq$ 1000 km^2 appear at level 4 (and higher); lakes $\geq$ 250 km^2 at level 5 (and higher); lakes $\geq$ 10 km^2 at level 8 (and higher); and lakes $\geq$ 2 km^2 appear at level 11 (and higher).

- If a lake covers the outlet of a sub-basin, the sub-basin is split into a left and a right part along its main river. This is to accommodate data being assigned to the left or right side of a lake. This modification is inconsistent with the traditional Pfafstetter scheme in which both sides have to keep the same Pfafstetter code because they belong to the same sub-basin. In HydroBASINS, the two sub-basin sides are distinguished by an additional attribute ('Side'), which can be L, R, or M (for left, right, or merged if there is no split). The according side is also reflected in the last digit of the sub-basin ID. Topologically, the two sides are neither up- nor downstream from each other. However, to keep connectivity and topological searches consistent, the following rule has been introduced in HydroBASINS: every right part of a split sub-basin has to first flow into its corresponding left part, then into the contained lake, and finally to right (or merged) part of the next downstream sub-basin polygon.

3. Data format and distribution

3.1 File name syntax

HydroBASINS data are provided as regional tiles in individual polygon shapefiles, one for each region and each Pfafstetter level. File names follow the syntax:

Hybas_XX_levYY_v1c.shp (standard format; without lakes)

or *Hybas_lake_XX_levYY_v1c.shp* (customized BioFresh/IUCN format; with inserted lakes)

where *XX* indicates the region and *YY* indicates the Pfafstetter level (01-12). The regional extents are defined by a two-digit identifier:

Identifier	*Region*
af	Africa
ar	North American Arctic
as	Central and South-East Asia
au	Australia and Oceania
eu	Europe and Middle East
gr	Greenland
na	North America and Caribbean
sa	South America
si	Siberia

3.2 Attribute table

Each HydroBASINS shapefile contains an attribute table with the following column structure and information:

Column	***Description***
Hybas_id	Unique basin identifier. The code consists of 10 digits: • First 1 digit represents the region: 1 = Africa; 2 = Europe; 3 = Siberia; 4 = Asia; 5 = Australia; 6 = South America; 7 = North America; 8 = Arctic (North America); 9 = Greenland • Next 2 digits define the Pfafstetter level (01-12). The value '00' is used for the 'Level 0' layer that contains all original sub-basins and all Pfafstetter codes (at all levels); 'Level 0' only exists in the standard format of HydroBASINS (without lakes). • Next 6 digits represent a unique identifier within the HydroSHEDS network; values larger than 900,000 represent lakes and only occur in the customized format (with lakes) • Last 1 digit indicates the side of a sub-basin in relation to the river network (0 = noSide; 1 = Left; 2 = Right). Sides are only defined for the customized format (with lakes).

Next_down	Hybas_id of the next downstream polygon. This field can be used for navigation (up- and downstream) within the river network. The value '0' indicates a polygon with no downstream connection. Note that small endorheic sinks may have a 'virtual' connection to an appropriate downstream polygon to allow for topological queries in larger river basins where discontinuities should be eliminated (e.g., the larger Nile Basin contains smaller endorheic basins that are virtually connected to the larger basin). Virtual connections can be identified as they carry a value of '2' in the 'Endo' field AND a value larger than '0' in the 'Next_down' field. Users can thus decide whether or not to terminate the routing at endorheic sinks.
Next_sink	Hybas_id of the next downstream sink. This field indicates either the ID of the next downstream endorheic sink polygon (if there is one) or the most downstream polygon of the river basin (if there is no endorheic sink in between). This field can be used to identify the entire, fully connected watershed that a polygon belongs to.
Main_bas	Hybas_id of the most downstream sink, i.e. the outlet of the main river basin. This field indicates the ID of the most downstream polygon of the river basin and can be used to identify the entire river basin that a polygon belongs to, including all associated endorheic basins. Note: small endorheic parts are typically lumped (via virtual connections) with their corresponding larger basin, while large endorheic watersheds can form their own basins.
Dist_sink	Distance from polygon outlet to the next downstream sink along the river network, in kilometers. This distance is measured to the next downstream endorheic sink (if there is one) or (if there is none) to the most downstream sink (i.e. the ocean).
Dist_main	Distance from polygon outlet to the most downstream sink, i.e. the outlet of the main river basin along the river network, in kilometers. The most downstream sink or outlet is that of the larger basin (to which smaller endorheic sub-basins may be virtually connected), i.e. either the outlet at the ocean, or the final sink of a large endorheic watershed which forms its own basin. Note that when small endorheic basins are lumped with a larger river basin, the virtual linkages are not measured as true distances but are calculated as direct (zero distance) connections.
Sub_area	Area of the individual polygon (i.e. sub-basin), in square kilometers.
Up_area	Total upstream area, in square kilometers, calculated from the headwaters to the polygon location (including the polygon). The upstream area only comprises the directly connected watershed area, i.e. it does not include endorheic regions that may be part of the larger basin through virtual connections.
Pfaf_id	The Pfafstetter code. For general description see literature (e.g., Verdin and Verdin 1999). The Pfafstetter code uses as many digits as the level it represents. This field can be used to cluster or subdivide sub-basins into nested regions. This field is only available for levels 1-12 (i.e. not for the 'Level 0' layer of the standard format).
Side	Indicates the side of a sub-basin in relation to the river network: L = Left; R = Right; M = Merged (direction defined looking downstream). This index enables a distinction between the two sides along lake shorelines (see text for more explanation). Polygons have only been split into left and right parts where lakes exist. This field is only available in the customized format (with lakes).

Lake	Indicator for lake types: 0 = no Lake; 1 = Lake; 2 = Reservoir; 3 = Lagoon. This field is only available in the customized format (with lakes).
Endo	Indicator for endorheic (inland) basins without surface flow connection to the ocean: 0 = not part of an endorheic basin; 1 = part of an endorheic basin; 2 = sink (i.e. most downstream polygon) of an endorheic basin.
Coast	Indicator for lumped coastal basins: 0 = no; 1 = yes. Coastal basins represent conglomerates of small coastal watersheds that drain into the ocean between larger river basins.
Order	Indicator of river order (classical ordering system): order 1 represents the main stem river from sink to source; order 2 represents all tributaries that flow into a 1st order river; order 3 represents all tributaries that flow into a 2nd order river; etc.; order 0 is used for conglomerates of small coastal watersheds.
Sort	Indicator showing the record number (sequence) in which the original polygons are stored in the shapefile (i.e. counting upwards from 1 in the original shapefile). The original polygons are sorted from downstream to upstream. This field can be used to sort the polygons back to their original sequence or to perform topological searches.
Pfaf_1 to Pfaf_12	Pfafstetter codes for all levels (1 to 12). For general description see literature (e.g., Verdin and Verdin 1999). The Pfafstetter code uses as many digits as the level it represents. These fields can be used to create sub-basins at all Pfafstetter levels by dissolving the polygons accordingly. These fields are only available for the 'Level 0' layer of the standard format (without lakes).

3.3 Vector data format and projection

The polygon data sets of HydroBASINS are distributed in ESRI 'shapefile' format (ESRI 1998). Each HydroBASINS shapefile consists of five main files (.dbf, .sbn, .sbx, .shp, .shx). Additionally, basic metadata information is provided in XML format (.xml). Projection information is provided in an ASCII text file (.prj). All shapefiles are in geographic (latitude/longitude) projection, referenced to datum WGS84.

3.4 Data distribution

HydroSHEDS data is available electronically in compressed zip file format from http://www.hydrosheds.org. *[Please note that the former data download site at the EROS Data Center of USGS at http://hydrosheds.cr.usgs.gov is now discontinued.]* To use the data files, the zip files must first be decompressed. Each zip file includes a copy of the HydroBASINS Technical Documentation.

4. Disclaimer and acknowledgement

4.1 License agreement

HydroBASINS is covered by the same License Agreement as the HydroSHEDS database, which is available at http://www.hydrosheds.org. HydroBASINS data (as defined in the License Agreement) are free for non-commercial and commercial use. For all regulations regarding license grants, copyright, redistribution restrictions, required attributions, disclaimer of warranty, indemnification, liability, waiver of damages, and a precise definition of licensed materials, please refer to the License Agreement.

By downloading and using the data the user agrees to the terms and conditions of the License Agreement.

4.2 Acknowledgement and citation

We kindly ask users to cite HydroBASINS in any published material produced using the data. If possible, online links to the HydroSHEDS website (http://www.hydrosheds.org) should be provided.

Citations and acknowledgements of the HydroBASINS data should be made as follows:

> *Lehner, B., Grill G. (2013): Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15): 2171–2186. Data is available at www.hydrosheds.org.*

5. References

ESRI – Environmental Systems Research Institute (1998): ESRI Shapefile Technical Description - An ESRI white paper. Available at http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

Lehner, B., Döll, P. (2004): Development and validation of a global database of lakes, reservoirs and wetlands. Journal of Hydrology 296(1-4): 1-22.

Lehner, B., Grill G. (2013): Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15): 2171–2186.

Lehner, B., Verdin, K., Jarvis, A. (2008): New global hydrography derived from spaceborne elevation data. Eos, Transactions, AGU, 89(10): 93-94.

Stein, J.L., Hutchinson, M.F., Stein J.A. (2014): A new stream and nested catchment framework for Australia. Hydrology and Earth System Science 18: 1917-1933.

USGS – U.S. Geological Survey (2000): HYDRO1k Elevation Derivative Database. USGS EROS Data Center, Sioux Falls, SD.

Verdin, K.L., Verdin, J.P. (1999): A topological system for delineation and codification of the Earth's river basins. Journal of Hydrology 218 (1-2): 1-12.