https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Earth System Science Data
Open Access
Discussions

MERIT-FullBasin: A global 90-meter basin dataset with a per-pixel upstream index and morphometric attributes

Lulu Jiang[1,2,3,4], Huan Wu[1,2*], Weitian Chen[5], Zhijun Huang[3,4], Ting Yang[6]

[1] State Key Laboratory of Climate System Prediction and Risk Management/Key Laboratory of Hydrometeorological Disaster Mechanism and Warning of Ministry of Water Resources/Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China

[2] School of Hydrology and Water Resources, Nanjing University of Information Science and Technology, Nanjing 210044, China

[3] Southern Marine Science and Engineering Laboratory (Zhuhai), and School of Atmospheric Sciences, Sun Yat-sen University, Zhuhai 519082, China

[4] Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, Sun Yat-sen University, Zhuhai 519082, China

[5] Future Water Laboratory, Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing 314102, China

[6] CAS Engineering Laboratory for Yellow River Delta Modern Agriculture, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

Correspondence to: Huan Wu (wuhuan3@mail.sysu.edu.cn, https://orcid.org/0000-0003-2920-8860)

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

Abstract:

Fine-resolution hydrographic data are essential for flood prediction and risk mapping, large-sample hydrology, and machine-learning streamflow models. Yet no existing global product simultaneously provides basin identity, upstream-catchment topology, and morphometric attributes at per-pixel resolution. The 90 m MERIT-Hydro flow-direction and flow-accumulation rasters are the input, but computing per-pixel attributes from this 75-billion-pixel global grid exceeds the memory limits of standard geographic information system (GIS) software, and tile-based workflows fragment basins across boundaries. Here we present MERIT-FullBasin, a globally seamless dataset that delivers all three layers at per-pixel 90 m resolution. The first tier delineates 340,991 basins above 1 km² in upstream drainage area, together covering over 99.5% of global land and partitioning the MERIT-Hydro land mask without gaps. The second tier provides a per-pixel depth-first search interval index (dfs_in, dfs_out); a single interval-containment comparison returns the complete upstream catchment of any target pixel, and the index losslessly encodes the underlying D8 flow-direction raster. The third tier delivers sixteen per-pixel upstream morphometric attribute rasters (eight terrain statistics and eight basin-shape metrics) at pixels with an upstream drainage area of at least 10 km². The full pipeline completes in about 21 hours on a single 13-thread, terabyte-memory high-performance computing node. Each tier is verified independently: tier 1 by exact pixel-count match against the flow-accumulation reference, tier 2 by exact reconstruction of the D8 raster from the index alone, and tier 3 by perimeter benchmarking against three independent reference implementations. Products are distributed as GeoTIFFs co-registered with MERIT-Hydro together with per-basin attribute tables. Per-pixel queries reduce to a single GeoTIFF read or a single table lookup. Among published global hydrographic products, MERIT-FullBasin is the first dataset distributed as a per-pixel queryable flow-tree index. The dataset is openly available at https://doi.org/10.5281/zenodo.20344113 (Jiang et al., 2026).

Keywords: Global basin delineation; MERIT-Hydro; High-resolution hydrography; Depth-first-search flow index; Per-pixel upstream catchment; Catchment morphometry

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

1. Introduction

Sub-kilometer global basin data are a hard requirement of contemporary hydrology. A watershed is the closed land area whose surface drainage converges at a single outlet, coastal or endorheic, and is the basic unit at which surface runoff, sediment, and pollutant load are analyzed. Climate-driven intensification of extreme precipitation (Tabari, 2020) has pushed flood forecasting, water-resources management, and Earth-system modeling onto sub-kilometer hydrography (Wu et al., 2014). At this resolution, the spatial accuracy and topological consistency of the basin layer set the ceiling on simulated runoff concentration and streamflow prediction skill (Jiang et al., 2025).

MERIT-Hydro (Yamazaki et al., 2017, 2019) supplies a globally consistent flow-direction map at 3 arcsec (approximately 90 m), with artificial sinks and broken flow paths in earlier digital elevation model (DEM) products removed through multi-source error correction and hydrography conditioning. A flow-direction raster encodes only neighbor-to-neighbor connectivity, and three downstream products are needed to make hydrography directly usable at per-pixel resolution: a closed basin partition for every land pixel, a per-pixel upstream-catchment query that returns the full upstream area of any chosen outlet, and per-pixel upstream morphometric attributes for those catchments. Producing all three across the global 3 arcsec grid (75.17 billion grid cells, 22.26 billion valid land pixels) is the gap among current global hydrography products.

The first gap concerns global basin delineation, and two strategies have been pursued. One is algorithm-centric: parallel watershed-delineation algorithms designed for continental scope, exploiting multi-core central processing unit (CPU) parallelism, graphics processing unit (GPU) acceleration, and out-of-core memory schemes (Kotyra, 2023; Cho, 2025; Zhao et al., 2025). These have been demonstrated at domains of 2 to 15 billion grid cells, but no continuous global delineation at the full 75.17 billion-cell scale has yet been reported. They deliver algorithm implementations but no released global watershed dataset.

The other is engineering-centric: scripted tiled workflows in mature geographic information system (GIS) software underlie most global hydrographic datasets of the past decade, including the Hydrologic Derivatives for Modeling and Applications database (HDMA; ArcGIS; Verdin, 2017), MERIT-Basins (TauDEM; Lin et al., 2021), Hydrography90m (GRASS GIS; Amatulli et al., 2022), and Basin90m (MATLAB TopoToolbox; He et al., 2024). These workflows simplify the global delineation problem to fit single-machine memory and the INT32 indexable limit (approximately 2.1×10^9 pixels) through two complementary moves. First, raster delineation is converted to vector polygons that render basin topology explicit, supporting cross-tile reconciliation and Pfafstetter-style hierarchical coding (Verdin and Verdin, 1999). Second, area thresholds remove small basins (Basin90m $\geq$ 50 km^2; HDMA $\geq$ 250 km^2; MERIT-

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Basins ≥ 25 km²; Hydrography90m does not impose one) to a tractable count. Both moves cost information: vectorization discards per-pixel raster addressability, and area thresholds discard the headwater catchments most relevant to ecohydrological and flood-process studies (Alexander et al., 2007). Tile boundaries within the workflow introduce basin fragmentation, cross-tile misclassification, and edge omission. A viable global watershed product therefore needs to address both fronts at once: the algorithmic capacity to process tens of billions of grid cells into a topologically closed delineation, and the engineering goal of releasing this delineation directly as a user-ready raster dataset.

The second gap concerns per-pixel upstream-catchment query: any land pixel can serve as an outlet, and large-sample hydrology, contaminant source tracing, and dense sampling along stream networks all require the full upstream catchment of arbitrary pixels rather than predefined basin outlets. Two approaches exist. Raster tools such as TauDEM (Tarboton, 1997) and the ArcGIS Pro Watershed tool retain pixel-level precision but reconstruct each catchment by traversing the flow-direction grid pixel by pixel, with per-query cost growing linearly with upstream area. Vector tools such as NHDPlus (Moore et al., 2019), the HydroBASINS/HydroRIVERS vector layers of HydroSHEDS (Lehner et al., 2008; Lehner and Grill, 2013), and ArcGIS Create Watersheds operate on a pre-simplified set of stream segments and run orders of magnitude faster, but the simplification drops the smallest basins and requires arbitrary query pixels to be snapped to the nearest segment, sacrificing pixel-level addressability. No global product currently offers pixel-level upstream-catchment query in a directly indexable form. Interval-based indexing of the flow-direction tree (Tarjan, 1972; Haag et al., 2018; Haag and Shokoufandeh, 2019) closes this gap by replacing per-pixel tracing with a single global pass and a per-query interval-containment test.

The third gap concerns per-pixel upstream morphometric attributes for those catchments: perimeter, basin length, basin width, convexity, and dimensionless shape ratios, alongside elevation- and slope-based aggregates. These attributes feed flood-hydrograph and unit-hydrograph modeling, donor-catchment selection in the Prediction in Ungauged Basins framework (Sivapalan et al., 2003), and feature engineering for long short-term memory (LSTM) streamflow models (Kratzert et al., 2019). Local terrain descriptors (slope, aspect, curvature) are computed from each pixel's 3 × 3 neighborhood and are routinely provided as global 90 m products (Geomorpho90m, Amatulli et al., 2020). Upstream-aggregated statistics such as mean upstream elevation or upstream relief require one upstream sweep per pixel and remain computationally tractable, yet no global product has distributed them at every pixel rather than at the sub-basin level (HydroATLAS, Linke et al., 2019). Morphological descriptors of the upstream catchment depend on its boundary, which must be reconstructed and traced for every pixel: Shen et al. (2016, 2017) delivered the first global per-pixel boundary-walk dataset on the HydroSHEDS 1 km grid, with nine attributes taking approximately one week of single-

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

core compute. At 3 arcsec the same procedure faces approximately 100 times more pixels, each with a proportionally longer boundary, for a 100- to 1,000-fold increase in total tracing work, making direct extension impractical. The interval-based indexing introduced above, paired with terabyte-memory high-performance computing (HPC) nodes, drops boundary-walk cost back to linear in pixel count and makes a global 90 m per-pixel attribute layer feasible.

Here we present MERIT-FullBasin, a 3 arcsec global watershed dataset derived from MERIT-Hydro by processing the full global domain (75.17 billion grid cells, 22.26 billion valid land pixels) as a single continuous entity on a terabyte-memory HPC node. The release contains three tiers. The first tier holds 24,587,290 outlets, of which 340,991 are basins above 1 km² in upstream drainage area, organized into 96 hydrographic regions and produced by a single seamless traversal of MERIT-Hydro. The second tier is a per-pixel interval index, derived from a single depth-first traversal of the flow-direction network at the native 3 arcsec resolution; it reduces upstream aggregation over any co-registered raster to a single interval-containment comparison per query, in place of explicit upstream tracing. The third tier holds sixteen per-pixel upstream terrain and shape attributes, populated at the approximately 440 million pixels (about 2% of all land pixels) with upstream drainage area ≥ 10 km² (an empirical noise floor below which D8 flow-direction artifacts dominate per-pixel shape metrics), distributed as per-region tiles.

2. Input Data and Computing Environment

The MERIT-Hydro dataset (Yamazaki et al., 2017, 2019) provides a flow direction raster (DIR) built on a bias-corrected, canopy-height-removed DEM and hydrologically conditioned against global river and water-body reference datasets. Two further layers are derived from DIR: the upstream pixel count (UPG) and the upstream drainage area (UPA; in km²). All three layers are at the native 3 arcsec (approximately 90 m) resolution. The original 5° × 5° tiles are mosaicked into globally seamless GeoTIFFs. The resulting rasters span 180° W to 180° E and 85° N to 60° S on a 432,000 × 174,000 (columns × rows) grid totaling 75.17 billion cells, of which 22.26 billion (29.6%) are valid land pixels.

All computations ran on an HPC node at Nanjing University of Information Science and Technology, comprising four Intel Xeon Gold 6416H CPUs (72 cores in total) and 4 TB of RAM. The 4 TB capacity allows the global rasters to remain resident throughout the pipeline, with peak resident memory of approximately 500 GB. Parallelism is applied within each raster operation (e.g., row-wise scans and per-cell updates), distributing the work across 13 threads.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

3. Building the Global Basin Dataset

The dataset is built in four stages: indexing every basin outlet (Sect. 3.1), delineating the watershed upstream of each outlet (Sect. 3.2), grouping basins into 96 regions (Sect. 3.3), and linking each basin's global and local identifiers with paired bounding boxes (Sect. 3.4). The outputs are four products (Table 1): two rasters (the basin-ID raster, BSN; the region-ID raster, RGN) and their attribute tables (BAT, RAT). Each is released both as a global mosaic and as per-region tiles. Fig. 1 shows the overall workflow.

3.1 Outlet Indexing

The Basin Attribute Table (BAT) is an outlet-indexed inventory of every basin. An outlet is the most-downstream pixel of a basin, marked in the MERIT-Hydro flow-direction raster (DIR; Table 1) by code 0 at the coast (exorheic) or code 255 at an interior continental sink (endorheic). Each BAT record stores the outlet's grid coordinates (outlet_row, outlet_col), its upstream pixel count (basin_cell_count, read from UPG), and its upstream drainage area (basin_area_km2, read from UPA). Records are ranked by basin_area_km2 in descending order, and basin_id is assigned sequentially from 1. The largest basin (basin_id = 1) is the Amazon; basin_id values run from 1 to 24,587,290 (Table 2).

$$\text{Basin\_ID} = \text{Table\_Index} + 1 \tag{1}$$

Endorheic share rises with basin size (Table 2): below 0.4% at the 0.05 km² floor and 35.8% (443 of 1,238) above 10,000 km², a class that includes the Caspian and Aral systems. The 0.05 km² floor is a Table 2 presentation cutoff only. The released BAT retains all 24,587,290 outlet records: 13.3 million single-pixel coastal entries that reflect pixel-scale artifacts in the MERIT-Hydro flow-direction grid along coastlines and fall below the 0.05 km² Table 2 cutoff, 340,991 basins above 1 km², and the remainder between 0.05 km² and 1 km².

3.2 Watershed Delineation

Delineation produces a single basin-ID raster (BSN) covering all land pixels, with every basin closed against its outlet record in BAT. The procedure is a five-step window sweep (Fig. 1). Step 1 initializes BSN from the outlet seeds in BAT, writing the basin_id from Sect. 3.1 at each outlet pixel under the convention of Eq. (1), so that each labeled pixel maps directly to its BAT row. Step 2 sweeps the global domain with six sequential 36°×36° window passes (43,200×43,200 = 1.87 billion pixels per window, about 13% below INT32_MAX). Pass 1 centers windows on the 30 largest basins, seeding each into BSN before Passes 2–6. Passes 2–5 step by 18° along longitude, latitude, and the diagonal, so every basin no larger than 36°×36° lies inside at least one window. Pass 6 wraps the antimeridian. Basins larger than 36° in either

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

dimension, such as the Amazon and the Nile, are completed over multiple passes by accumulating labels in BSN. Step 3 enforces closure per pass. Let labeled_count[Basin_ID−1] denote the labeled pixel count for basin Basin_ID in BSN, accumulated outlet-to-source by the window sweep, and let BAT[Basin_ID−1].basin_cell_count denote its reference cell count stored in BAT (accumulated source-to-outlet from UPG at the outlet; Sect. 3.1). Before each window pass the two counts are compared and a per-basin status mask MSK records the result according to Eq. (2). Basins flagged closed (MSK = 2) drop out of subsequent passes; open basins keep receiving labels. Step 4 restricts upstream tracing within each window to open basins and writes new labels back into the global BSN. Step 5 scans the completed BSN and stores each basin's row and column extents in BAT.

$$\text{labeled\_count}[\text{Basin\_ID}-1] = \text{BAT}[\text{Basin\_ID}-1].\text{basin\_cell\_count} \tag{2}$$

Fig. 2 shows the paired products. BSN labels every land pixel with a basin_id (Fig. 2a); the 30 largest basins (basin_id 1–30) are highlighted in the panel. The same map shows antimeridian-spanning basins (e.g., basin_id = 495) and the endorheic interior, including the Caspian Basin (basin_id = 3) and the smaller Aral Basin. BAT pairs each label with its outlet record from Sect. 3.1 and the row-column bounding box written by Step 5 (Fig. 2b); the bounding box defines the window for any per-basin raster read on the MERIT grid.

3.3 Region Partitioning

Region partitioning reuses the manually delineated upper levels of HydroBASINS (Lehner and Grill, 2013), keeping regional boundaries aligned with the Level-02/03 hydrographic divides and compatible with datasets already indexed on this layer. Each basin is first assigned a preliminary region_id by pixel-area-weighted overlap between its BSN mask and the Level-03 polygons, with ties broken by the lower region_id. Level-03 children sharing the same Level-02 parent are then merged into that parent, but only when both the merged pixel count and twice its valid-pixel count stay below the INT32 limit (Eq. (3); see Sect. 4 for the per-pixel storage). Parents that fail this test keep their Level-03 children as separate regions. Basins crossing the antimeridian and isolated island basins are kept outside the Level-02 merge and grouped on their own.

$$\begin{aligned} region\_nrow \times region\_ncol &< \texttt{INT32\_MAX} \\ 2 \times \text{region\_cell\_count} &< \texttt{INT32\_MAX} \end{aligned} \tag{3}$$

The partition divides the global mask into 96 regions: 65 continental regions tracking HydroBASINS Level-02 (IDs 11–92), two antimeridian groups (IDs 10001–10002), and 29 island groups (IDs 20001–20029), with RGN labeling each land pixel (Fig. 3a) and RAT

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

carrying the bounding-box and cell-count attributes per region (Fig. 3b). The integer-degree alignment of each box (multiples of 1,200 cells at 3 arcsec) allows regional tiles to be read directly against the global MERIT-Hydro rasters without resampling, and the INT32 headroom enforced by Eq. (3) keeps every per-pixel field in Sects. 4 and 5 within a single INT32 raster per region.

3.4 Linking Global and Local Indices

Each basin in BAT is given a triple key (global_basin_id, region_id, local_basin_id) and paired global and local bounding boxes. The global_basin_id is the basin_id from Sect. 3.1, unique across the global domain. The local_basin_id restarts from 1 inside each region and preserves the global basin_id ordering within the region, and the local bounding box is the global one shifted by the region's RAT origin (e.g., local_row_min = global_row_min − region_row_min; Fig. 3b). The triple key is shared by every basin-level layer in Table 1: global_basin_id indexes a basin in the global rasters, and (region_id, local_basin_id) indexes the same basin inside its regional tile.

4. Depth-First Search Interval: A Per-pixel Flow-Tree Index

The 24,587,290 outlets indexed in Sect. 3 partition the global land surface, but the intra-basin flow topology remains unresolved: the D8 flow-direction raster records only each pixel's immediate downstream neighbor, so any upstream–downstream relationship must be reconstructed by recursive traversal. At 3 arcsec resolution, an upstream trace from a single query pixel in a large basin can visit tens of millions of cells. Repeated on-demand tracing is intractable for pixel-level queries at global scale. The flow-tree index of MERIT-FullBasin replaces this traversal with a per-pixel depth-first search (DFS) interval encoding that makes the intra-basin topology directly queryable: upstream membership reduces to two integer comparisons.

A depth-first traversal of the D8 flow tree assigns every pixel two integers: an entry time dfs_in recorded when the traversal first visits the pixel, and an exit time dfs_out recorded when it finally leaves. The dfs_in / dfs_out labeling is the standard preorder/postorder DFS timestamping (Tarjan, 1972), known in database literature as the Nested Set Model (Celko, 2004). Haag et al. (2018) and Haag and Shokoufandeh (2019) first applied this construction to hydrologic flow-tree indexing, demonstrating it on the Delaware River Basin.

Fig. 4 illustrates the labeling algorithm on a synthetic 8 × 8 grid containing two basins. Panel (a) gives the D8 flow-direction codes with arrows, and panel (b) gives the row-major linear pixel indices (i0–i63) used as a coordinate reference. A counter advances by one at each event; starting from each basin outlet, the traversal stamps dfs_in on entry, recurses into the upstream

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

children in UPA-descending order, and stamps dfs_out on return. This UPA-descending sibling order is a design choice rather than a generic DFS feature; it underlies the eldest-child-chain property shown in Fig. 5(f) and Eq. (6). Panel (c) shows a mid-traversal snapshot before the first tributary backtrack. Panel (d) shows the completed traversal, with the two basins receiving disjoint dfs_in ranges. Panels (e) and (f) give the dfs_in and dfs_out rasters, respectively.

Fig. 5 illustrates the role of the DFS interval index on the same 8 × 8 example. Panels (a)–(c) show that the [dfs_in, dfs_out] pair is a lossless encoding of the per-pixel flow-tree structure: any per-pixel quantity derivable from the underlying D8 flow direction can be recovered from the pair alone, with no auxiliary raster. Three reconstructions illustrate this: (a) The D8 flow direction itself, recovered at each pixel from the [dfs_in, dfs_out] pair alone, by inspecting the timestamps of its eight neighbors. (b) The distance to the outlet in cell counts, equal to the count of [dfs_in, dfs_out] intervals enclosing each pixel (each ancestor in the flow tree contributes one enclosing interval). (c) The upstream cell count per pixel, equal to (dfs_out − dfs_in + 1)/2.

Panels (d)–(f) give three applications, each reducing to interval comparisons rather than recursive flow-direction traces. (d) The upstream catchment of a target pixel T_1 is the set of all pixels T_2 whose dfs_in lies within T_1's interval (Eq. (4)).

$$\mathrm{Catchment}(T_1) = \{\, T_2 : \mathrm{dfs\_in}[T_1] \le \mathrm{dfs\_in}[T_2] \le \mathrm{dfs\_out}[T_1] \,\} \tag{4}$$

(e) The catchment boundary of the same target is the set of catchment pixels T_2 with at least one eight-neighbor T_3 outside the catchment (Eq. (5)), where $N(T_2)$ is the eight-neighborhood of T_2 and off-grid or NoData neighbors count as outside.

$$\mathrm{Boundary}(T_1) = \{\, T_2 \in \mathrm{Catchment}(T_1) : \exists\, T_3 \in N(T_2), T_3 \notin \mathrm{Catchment}(T_1) \,\} \tag{5}$$

(f) Under the UPA-descending sibling order, any maximal run of consecutive dfs_in values traces an eldest-child chain: the path traced by repeatedly following the largest-UPA child from a starting node v (Eq. (6)). L(v) is the chain length in pixels.

$$\mathrm{Chain}(v) = \{\, p : \mathrm{dfs\_in}[v] \le \mathrm{dfs\_in}[p] < \mathrm{dfs\_in}[v] + L(v) \,\} \tag{6}$$

The chain rooted at a basin outlet is the basin's main stem; chains rooted at interior nodes are tributary stems. Panel (f) marks four such chains: Basin A's main stem (blue), two of its tributary stems (teal, orange), and Basin B's main stem (green).

Fig. 6 shows the resulting index at global extent, together with a zoom into an example basin (basin 8595, southeastern Fujian, China; 729 km²). Panels (a) and (c) display the dfs_in and dfs_out rasters; values are assigned per region, with each region's indices occupying a contiguous range starting at zero (INT32-storable per region; cf. Sect. 3, Eq. (3)). The region

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

boundaries are overlaid as thin white lines on (a) and (c). The UPA-descending traversal makes major river networks visible in the dfs_in raster itself, with tributary confluences appearing as abrupt jumps between adjacent pixels. This is the same eldest-child-chain property illustrated on the 8 × 8 example in Fig. 5(f), now visible at continental scale. Panels (b) and (d) show this example basin at its native 3 arcsec resolution, together with its boundary and outlet marker; the dfs_in gradient within the basin makes the flow-tree topology directly readable: the main stem appears as a monotonic run of dfs_in values increasing from the outlet upstream.

5. Per-pixel Upstream Terrain and Shape Attributes

The sixteen per-pixel upstream terrain and shape attributes (Table 3) are defined at every 3 arcsec MERIT-Hydro pixel through the flow-tree index of Sect. 4 and released at approximately 440 million pixels with UPA ⩾ 10 km². The sixteen attributes fall into three algorithmic groups (sized 6, 4, and 6) reflecting how each is computed: six upstream-to-downstream aggregates of elevation and slope, four planform metrics traced from the catchment boundary, and six arithmetic derivatives. By scientific use these regroup into eight terrain statistics (the six elevation and slope aggregates plus the two elevation-based derivatives) and eight basin-shape metrics (the four planform metrics plus the four dimensionless shape indices), shown separately in Figs. 8 and 9. The first group comprises six aggregates: catchment-mean elevation, catchment-mean slope, minimum elevation, maximum elevation, elevation standard deviation, and slope standard deviation; all six are computed in a single upstream pass over the flow-tree index, with each basin pixel visited once and contributing to all its downstream targets, so cost scales linearly with the number of pixels.

The four planform metrics extracted from each target's upstream catchment boundary are perimeter (Vossepoel and Smeulders, 1982), basin length (Schumm, 1956), basin width, and convexity (Willemin, 2000); together they dominate the compute cost across the 440 million targets, and an independent boundary scan per target at 3 arcsec global resolution would be prohibitive. The 10 km² release floor is set here: below it, catchment boundaries are traced from too few pixels for raster-discretization noise to average out in the planform metrics. The flow-tree index of Sect. 4 makes the computation tractable (Fig. 7). Catchment membership is determined by a single interval test on [dfs_in, dfs_out] (Eq. (4)). The boundary set is then extracted by Eq. (5), and a single Moore-neighbor walk around it returns a closed 8-connected chain from which all four metrics are read off in one pass: perimeter from the step-length sum, basin length and basin width from the chain's principal-axis extents, and convexity from the enclosed area normalized by its convex hull (Fig. 7a). A pre-filter discards interior pixels with all eight neighbors already upstream, which cannot lie on any target's boundary. Adjacent targets on the same stream share most of their boundary (Fig. 7b), so each stream is walked once at its starting target and updated incrementally downstream; independent streams run in

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

parallel (Fig. 7c), and the main stem is split into equal-length segments to balance runtime (Fig. 7d). For the example basin (1,932 stream pixels at UPA $\geqslant$ 10 km^2), the scheme decomposes into 20 concurrent streams, with the main stem split into three segments (Fig. 7c, d).

The third group comprises six arithmetic derivatives, each computed at every pixel from quantities in the first two groups. Two combine the upstream elevation rasters of the first group: basin relief and the hypsometric integral (Strahler, 1952). The remaining four are dimensionless shape indices built from the planform metrics of the second group: elongation ratio (Schumm, 1956), Gravelius compactness (Gravelius, 1914), circularity ratio (Miller, 1953), and lemniscate ratio (Chorley et al., 1957). These six layers add no new traversal, so the sixteen attributes stay mutually consistent at every pixel; full definitions and units appear in Table 3.

Fig. 8 shows the eight terrain attributes on the example basin: the six elevation and slope aggregates and the two elevation-based arithmetic derivatives. Panel (a) is the input elevation field; the eight derived attributes are shown in panels (b–i). The six upstream-to-downstream aggregates (Fig. 8b–g) track the catchment's downstream evolution directly: minimum elevation decreases monotonically toward the outlet, since the most downstream pixel of any catchment is also its lowest, while maximum elevation grows as more headwater terrain is incorporated. Mean elevation and mean slope evolve as area-weighted integrals over the expanding catchment, and the elevation and slope standard deviations summarize the topographic heterogeneity upstream of every point. The two arithmetic derivatives reflect the same downstream evolution: basin relief increases steadily along the main stem (Fig. 8h), while the hypsometric integral decreases downstream (Fig. 8i), marking the transition from the steep dissected upland in the source region to the broad low coastal plain at the outlet.

Fig. 9 shows the eight shape attributes on the same basin: the four planform metrics of the second group and the four shape indices from the third group. Panel (a) is the upstream drainage area (logarithmic scale); the eight derived attributes are shown in panels (b–i). The boundary-walk metrics (Fig. 9b–e) trace each upstream catchment's outline: basin length, basin width, and perimeter (Fig. 9b–d) grow smoothly along the main stem as the catchment accumulates downstream area, whereas convexity (Fig. 9e) is highest along the lower main stem and lowest in the elongate headwater branches. The four shape indices in Fig. 9f–i (circularity ratio, Gravelius compactness, elongation ratio, and lemniscate ratio) capture the same transition from elongate headwater catchments to compact lower-basin catchments. Together with the BSN, BAT, RGN, RAT, and flow-tree layers of Sects. 3 and 4, these sixteen attribute layers complete the dataset described in Sect. 6.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

6. The MERIT-FullBasin Dataset

MERIT-FullBasin is distributed as GeoTIFF raster layers with CSV attribute tables, all co-registered to the native MERIT-Hydro 3 arcsec grid (approximately 90 m at the equator; 432,000 × 174,000 pixels). The complete inventory, data types, and NoData values are listed in Table 1; the DOI and repository location are given in the Data availability section. Two rasters cover the global land surface: bsn_global_90m.tif (UInt32, pixel value = global_basin_id) and rgn_global_90m.tif (UInt16, pixel value = region_id), paired with attribute tables bat_global_90m.csv (BAT, 24,587,290 rows) and rat_global_90m.csv (RAT, 96 rows). The triple key (global_basin_id, region_id, local_basin_id) defined in Sect. 3.4 carries through every basin-level layer and links these global products to the per-region tiles described below.

All other layers are distributed as per-region tiles across the 96-region partition of Sect. 3.3, sized so each region fits in memory under standard GIS or array tools. Each region carries its basin-ID raster in two forms. The first, bsn_region*_90m.tif (UInt32, keyed by global_basin_id), supports cross-region or global queries and joins against the global BAT. The second, bsn_region*_90m_remapped.tif (UInt32, keyed by local_basin_id), indexes the per-region BAT subset (bat_region*_90m.csv) directly: the pixel value at any location is the local_basin_id of that pixel's basin, pointing to the corresponding row in that region's BAT table.

The flow-tree index is distributed as two rasters per region, dfs_in_region*_90m.tif and dfs_out_region*_90m.tif (Int32, NoData = −1), recording the depth-first entry and exit timestamps of each pixel. The upstream catchment of any target pixel is the set of pixels within the same region whose entry timestamp falls between the target's entry and exit timestamps: a single interval test, without traversing the flow-direction raster. The region partition of Sect. 3.3 keeps each basin entirely within one region, so upstream queries never cross region boundaries.

The sixteen per-pixel upstream terrain and shape attribute layers are distributed as Float32 GeoTIFFs across the same 96-region partition. Definitions, units, and groupings are listed in Table 3, with algorithms described in Sect. 5. The layers are populated only at pixels with UPA $\geqslant$ 10 km² (approximately 440 million pixels globally, about 2% of all land pixels); all other locations carry the NoData value −9999, also written into the GeoTIFF NoData metadata so GDAL and rasterio load it as NaN.

All per-region GeoTIFFs are written with DEFLATE compression; global totals before and after compression are listed in Table 1. The basin and region ID layers each compress to about 1 GB globally, since long pixel runs within a basin or region share the same integer value. The sixteen Float32 upstream-attribute layers compress to roughly 3 GB each globally: the 98% of

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

pixels that fall below the UPA ⩾ 10 km² threshold are NoData and compress heavily. The two flow-tree index layers are the largest distributed components: dfs_in and dfs_out each occupy about 40 GB globally compressed, from roughly 280 GB uncompressed, since each valid pixel carries a unique timestamp, and only the NoData values compress well. Although DEFLATE raises wall-clock cost during dataset production, the full compressed dataset totals approximately 140 GB.

7. Technical Verification

MERIT-FullBasin is verified along three axes corresponding to its three product tiers: the basin and region geometry (BSN, RGN, BAT, RAT; Sect. 3), the flow-tree index (dfs_in, dfs_out; Sect. 4), and the sixteen per-pixel upstream terrain and shape attributes (Sect. 5). The basin geometry is verified by two independent checks. First, completeness: all basin outlets (including inland endorheic sinks) are enumerated and registered in the BAT before delineation begins, and every valid land pixel drains in a finite number of D8 steps to one of these outlets (Sect. 3.1), so no basin can be structurally omitted. Second, topological closure: each basin must yield the same pixel count whether traversed source-to-outlet (UPG accumulation) or outlet-to-source (basin delineation). These are two independent code paths over the same D8 network; their per-basin counts coincide only when the network is everywhere self-consistent (no leakage, no duplication, no topological break). Globally, all 22,257,225,902 valid land pixels are assigned to a basin that reaches closed status (Eq. (2)).

The flow-tree index is verified by showing that its encoding is lossless. Losslessness is the foundation of the index: all sixteen per-pixel attributes (Sect. 5) use the [dfs_in, dfs_out] interval to define their upstream pixel set, and any encoding error propagates to all sixteen. The losslessness follows from the nested-interval property of DFS preorder/postorder timestamping (Tarjan, 1972): every pixel's [dfs_in, dfs_out] interval is strictly contained within its parent's interval in the flow tree. Each pixel's downstream neighbor (its D8 parent) is therefore the unique pixel whose interval most tightly encloses it, and the full D8 flow-direction raster DIR is recoverable from [dfs_in, dfs_out] alone.

All sixteen per-pixel attributes derive from the same upstream pixel set defined by the [dfs_in, dfs_out] interval of each target. The lossless DFS encoding verified above guarantees that this set is computed correctly. Eight of the sixteen attributes (the six upstream-to-downstream aggregates and the two elevation-based arithmetic derivatives, basin relief and the hypsometric integral) follow directly from the set: they are simple aggregations (mean, min, max, std) and pixel-wise arithmetic over the upstream pixel set. The remaining eight attributes are the four boundary-walk metrics (perimeter, basin length, basin width, convexity) and the four shape ratios derived from them (circularity ratio, Gravelius compactness, elongation ratio, lemniscate

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

ratio); these depend on the 8-connected boundary of the upstream set, whose computation requires a Moore-neighbor walk and is substantially more error-prone. Among the boundary-walk metrics, perimeter is the sharpest diagnostic of boundary integrity. Let A and P denote the pixel counts of the target catchment's area and perimeter. The single-pixel boundary-error sensitivity scales as $4/P$, exceeding the $1/A$ sensitivity of area-integral statistics for any practical basin. Perimeter agreement is therefore a sufficient diagnostic for boundary correctness, which in turn validates all eight boundary-dependent attributes.

Perimeter was verified on a global validation set of approximately 150,000 upstream-catchment masks in two tiers. The outlet tier (approximately 95,000) is one mask per basin with UPA $\geqslant$ 10 km^2 globally; it verifies the standalone Moore–Vossepoel–Smeulders (Moore–VS) perimeter algorithm across the full global domain. The sub-mid tier (approximately 58,000) is one interior mask per basin with outlet UPA $\geqslant$ 20 km^2, testing the incremental boundary update along streams (Sect. 5): the sub-mid is taken as a tributary (or main-stem mid-point if no tributary qualifies) whose UPA lies between 30% and 70% of the outlet UPA; reaching the sub-mid from the outlet requires the released algorithm to apply tens to hundreds of incremental boundary updates, so the sub-mid perimeter directly probes update correctness. For every mask, perimeter was computed offline using three independent Python implementations (full descriptions and cross-algorithm statistics in Appendix B): Crofton 4-direction (skimage.measure.perimeter_crofton), marching squares (skimage.measure.find_contours), and GIS staircase (rasterio.shapes → shapely + pyproj). The released perimeters trace a fractal power-law scaling against catchment area, $P = 4.77 \cdot A^{0.558}$, $R^2 = 0.96$ (Fig. 10a); the exponent exceeds the Euclidean value of 0.5, as expected for the fractal character of drainage divides (Mandelbrot, 1967). On the same masks, the released Moore–VS perimeter has a median offset of +0.4% from the Crofton 4-direction reference, with a 2nd-to-98th percentile (P2–P98) band of ±1.8% (Fig. 10b), confirming consistency with the integral-geometry definition of curve length. Marching squares overestimates by about 5.6% (median), the expected magnitude of its staircase bias on irregular boundaries. The GIS staircase overestimates by about 28%, in quantitative agreement with its theoretical ceiling of $4/\pi \approx 1.273$ (i.e., +27.3%), which explains why raster-to-polygon GIS pipelines systematically over-report drainage-basin perimeter at sub-grid resolutions.

For external validation against an independent product, the Shen et al. (2016, 2017) dataset, built on HydroSHEDS 1 km grids, is the only published global per-pixel basin-morphometric dataset that overlaps MERIT-FullBasin. The two datasets share nine variables; eight of these show a median deviation $\leqslant$ 1% (Appendix B). The only systematic offset is in perimeter (−5.94%), tied to a method difference: MERIT-FullBasin applies Vossepoel–Smeulders (VS) minimum-bias weights on the Moore chain code, whereas Shen et al. (2016) use an unweighted variant. Disabling the VS weights brings the deviation below 1% (Appendix B).

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

8. Discussion

8.1 Computational Feasibility

MERIT-FullBasin covers the global 90 m grid (75.17 billion pixels, 22.26 billion on land) and delivers seamless basin delineation, a per-pixel DFS interval index (dfs_in, dfs_out), and sixteen per-pixel upstream attribute layers. The full workflow runs end-to-end on a 13-thread HPC node as a single C/Numba pipeline without manual GIS intervention. Numerical correctness is verified in Sect. 7; the production wall times reported here set the scaling basis for next-generation 30 m grids such as HydroSHEDS v2, which carry roughly nine times the present pixel count.

Earlier global delineations either scale to continental but not global extent (Cho, 2025; Zhao et al., 2025; Kotyra, 2023) or rely on tiled single-machine workflows that relax boundary fidelity (Lin et al., 2021; Amatulli et al., 2022). The MERIT-FullBasin pipeline (Fig. 1) traces every land pixel to its outlet through a six-pass window sweep; each pass opens with a per-basin closure test that drops fully labeled basins, so no pixel is retraced once its basin closes. Production wall time is approximately 100 min (Fig. A1).

The second data tier is the per-pixel DFS interval index (dfs_in, dfs_out). A depth-first traversal of each region's flow tree writes both timestamp fields; wall time across the 96 regions is approximately 1.76 h (Fig. A2).

The third data tier is the sixteen per-pixel attribute layers (Table 3); production wall time is approximately 17.7 h, of which the boundary walk takes 12 h (Fig. A2). The arithmetic kernel, six layers derived in place from preceding rasters, is compression-bound: per-pixel compute totals 0.08 h while compressed GeoTIFF read/write takes 2.44 h, with each layer compressed from approximately 280 GB to 3 GB by the release format (Float32 + DEFLATE). On matched single-core hardware on the HydroSHEDS 1 km grids, MERIT-FullBasin computes the nine attributes shared with Shen et al. (2016, 2017) in approximately 1 h, against the one week reported by Shen et al. (2016) (Fig. A3; see Appendix B for details).

8.2 Data Use

MERIT-FullBasin's foundational data layer, comprising BSN and RGN rasters, BAT and RAT lookup tables, and the per-pixel DFS interval index (dfs_in, dfs_out), is navigable from the global domain through region and basin to pixel.

The 96-region partition is paired with a global ↔ local index system: every basin carries a triple key (global_basin_id, region_id, local_basin_id) and paired global and region-local bounding boxes, with pixel rasters released in both global-keyed and local-keyed forms. The local-keyed per-region tiles fit within INT32 indexing for native handling in GDAL, QGIS, ArcGIS, and numpy with no further tiling. At the basin level, BSN pixel values are the global basin_id and BAT rows are sorted in ascending basin_id order per Eq. (1), with ID 1 reserved for the Amazon Basin; the expression BAT[BSN − 1] then retrieves the basin-level record (outlet coordinates, cell count, drainage area, bounding box) for every pixel as a pure positional index, vectorized across numpy, pandas, R, Julia, or any environment that supports positional indexing. The positional form avoids the multi-gigabyte memory footprint a hash table over 24.6 million basin keys would incur. Other global raster products distribute on fixed latitude–longitude grids (MERIT-Hydro at 5° × 5°, Hydrography90m at 20° × 20°; Amatulli et al., 2022), under which basins crossing tile boundaries are mosaicked on the client side before basin-level analysis; vector products such as MERIT-Basins (Lin et al., 2021) represent each basin as a polygon and lose per-pixel addressability in the conversion.

At the pixel level, the (dfs_in, dfs_out) index reduces any upstream query for a target pixel to a single interval-containment test (Eq. (4)). Raster traversal tools (e.g., the ArcGIS Pro Watershed tool) retrace the flow-direction grid for every query, at per-query cost that grows with upstream area; vector segment-based tools (e.g., the Create Watersheds service in ArcGIS Online) operate on a pre-simplified stream network at fixed per-query cost, but snap arbitrary query pixels to the nearest segment and lose headwater catchments below the segment resolution. MERIT-FullBasin retains pixel-level precision at the fixed per-query cost of the segment-based tools. Any external raster co-registered to the MERIT-Hydro 90 m grid (precipitation, land cover, population) is sorted by dfs_in and prefix-summed once, after which the upstream sum, mean, or variance at every pixel is returned at unit cost per pixel; the mechanism applies per time step to time-series products and to products at other resolutions after resampling to the MERIT-Hydro grid. These three input classes correspond to the hazard, surface-response, and exposure components of flood-disaster analysis, so the same mechanism supports basin-scale flood-risk workflows. Among published global hydrographic products, MERIT-FullBasin is the only dataset distributed as a pixel-level queryable flow-tree index. Fig. A4 illustrates three implementations of this index, namely the per-pixel D8 flow direction, distance to outlet, and upstream pixel count shown in Fig. 5(a–c), recovered from (dfs_in, dfs_out) alone. The companion Python reference implementation FlowDex (Jiang et al., 2026) exposes the data layer through interactive workflows (Fig. A5).

8.3 Hydrological Applications

The sixteen per-pixel attribute layers extend the dataset to scientific applications. The closest prior dataset, Shen et al. (2016, 2017), delivered nine boundary-derived attributes on HydroSHEDS 1 km grids; MERIT-FullBasin raises the resolution from 1 km to 90 m and extends the variable set from nine to sixteen. The resolution gain delivers three direct benefits:

(a) basins below 100 km² retain reliable shape, where a 1 km grid resolves only about 100 pixels per basin at this size and degenerates to a few pixels below 10 km²; (b) drainage-divide geometry is sharper, where a 1 km grid merges adjacent divides in steep or narrow terrain; (c) sub-kilometer descriptors are directly available for urban- and sub-basin-scale flood modeling and for headwater-catchment process studies.

Grouped by scientific use, the sixteen attributes fall into two families of eight. Basin shape exerts first-order control on flood-hydrograph peak timing, peak magnitude, and unit-hydrograph form; the eight shape attributes (perimeter, basin length, basin width, convexity, circularity ratio, Gravelius compactness, elongation ratio, lemniscate ratio) are the standard similarity descriptors used in regionalization and Prediction in Ungauged Basins (PUB; Sivapalan et al., 2003). The eight elevation- and slope-based attributes (six statistical aggregates over the upstream DEM and slope set, with basin relief and the hypsometric integral as their arithmetic derivatives) characterize the within-basin distribution of elevation and slope: Strahler (1952) defined the hypsometric integral as a landform-evolution diagnostic, with high values for the youthful (inequilibrium) stage, intermediate values for the mature (equilibrium) stage, and low values for the monadnock (old-age) stage; aggregates of relief and slope further control topography-driven runoff generation and the catchment-scale energy gradient. The full sixteen-attribute set augments the basin-descriptor tables of large-sample hydrology collections such as the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS; Addor et al., 2017) and Caravan (Kratzert et al., 2023), and supplies inputs for shape- and terrain-similarity-based parameter transfer in PUB and for feature engineering in LSTM streamflow prediction (Kratzert et al., 2019).

8.4 Limitations and Future Work

The watershed delineation in MERIT-FullBasin inherits the D8 flow-direction raster of MERIT-Hydro. DEM artifacts can redirect flow paths and shift basin boundaries in low-relief terrain (large endorheic basins in arid regions, extensive wetland plains such as the Sudd and the Pantanal, and high-latitude periglacial areas). These limitations are intrinsic to DEM-derived flow routing; users working in such regions should cross-reference basin boundaries with independent datasets or satellite-derived water-extent products.

MERIT-FullBasin does not bundle a stream-network extraction layer, because the choice of threshold depends on the application. Existing global products adopt fixed drainage-area thresholds that span more than two orders of magnitude: 0.05 km² (Hydrography90m; Amatulli et al., 2022), 5 km² (recommended for MERIT-Hydro; Yamazaki et al., 2019), 10 km² (HydroSHEDS; Lehner et al., 2008), and 25 km² (MERIT-Basins; Lin et al., 2021). Drainage density itself varies substantially with climate, relief, and soil–vegetation conditions (Lin et al., 2021), so a uniform threshold may systematically misrepresent stream-network morphology

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

across different landscapes. The released DFS interval index supports stream-network construction: once a stream mask is defined at the user's chosen threshold, drainage density and stream-network topology can be computed directly. A threshold-selection scheme that varies with local basin characteristics is developed in a companion paper, MERIT-FullRiver (in preparation).

9. Conclusions

We present MERIT-FullBasin, a globally seamless 90 m basin dataset derived from MERIT-Hydro across all 22.26 billion land pixels. The Basin Attribute Table records 24,587,290 outlets, of which 340,991 are basins above 1 km² in upstream area. Three tiers are released together: globally seamless basin-ID and region-ID rasters with their attribute tables; a per-pixel depth-first search interval (dfs_in, dfs_out) that encodes the global flow tree; and sixteen per-pixel upstream terrain and shape attributes at approximately 440 million pixels with upstream drainage area ⩾ 10 km². Each tier passes its own verification benchmark. The dataset supports Prediction in Ungauged Basins, flood-risk modeling, large-sample hydrology, and machine-learning streamflow prediction. Usage requires no hydrologic modeling: per-pixel queries reduce to a single GeoTIFF read or table lookup. All products are openly available; see Data availability.

Data availability

The MERIT-FullBasin dataset is openly available on Zenodo at https://doi.org/10.5281/zenodo.20344113 (Jiang et al., 2026), released as v1.0 under the CC BY 4.0 license. The release contains three tiers, following the structure of Sects. 3 to 5: the basin and region masks (BSN, RGN) paired with their attribute tables (BAT, RAT); the per-pixel depth-first search interval index (dfs_in, dfs_out); and sixteen per-pixel upstream-catchment attribute layers, comprising eight terrain and relief statistics and eight basin-shape metrics, populated where the upstream contributing area reaches at least 10 km². File names, formats, and grid specifications are listed in Table 1. The MERIT-Hydro flow-direction, upstream-pixel-count, and upstream-area rasters that underlie the workflow are not redistributed and should be downloaded from the MERIT-Hydro archive (Yamazaki et al., 2019).

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Code availability

The code accompanying this dataset is deposited together with the data archive at https://doi.org/10.5281/zenodo.20344113 (CC BY 4.0). The release contains three components: (i) FlowDex v0.1, a Python reference implementation of the per-pixel query functions (BAT/RAT lookup, DFS-interval upstream/downstream extraction, and attribute sampling on the sixteen upstream-catchment layers), with an interactive GUI for loading the published rasters; (ii) the validation scripts that reproduce the cross-comparison against Shen et al. (2017) reported in Sect. 7; and (iii) the figure scripts used to generate the figures in this manuscript.

Video supplement

A narrated walkthrough of the MERIT-FullBasin dataset is available at https://www.youtube.com/watch?v=wJwJT6D3Snc.

Author contributions

LJ: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. HW: Conceptualization, Methodology, Resources, Supervision, Funding acquisition, Writing – review & editing. WC: Methodology, Software, Validation. ZH: Investigation, Validation. TY: Validation, Writing – review & editing.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grants 42275019 and 42088101), the National Key R&D Program of China (Grant 2024YFC3013302), and the Open Fund Project for Heavy Rain of China Meteorological Administration (Grant BYKJ2024Z10). Partial support was also provided by the Hainan Provincial R&D Program (Grants CXFZ2022J074 and SCSF202203) and the Hainan Key Research and Development Project (Grant ZDYF2023SHFZ125).

We thank Dai Yamazaki and colleagues for the openly distributed MERIT-Hydro dataset, which provided the flow-direction and flow-accumulation grids underlying this work, and the

HydroSHEDS team for the manually delineated Levels 1, 2, and 3 region boundaries adopted as the regional partitioning.

Figures and Tables

Table 1. Summary of input datasets and output products. All native-resolution rasters share a global grid (432,000 × 174,000 pixels) and are distributed as DEFLATE-compressed GeoTIFF files. Tabular products are in CSV format. All basin-level tables share three linking keys (global_basin_id, region_id, local_basin_id) enabling direct cross-referencing between raster masks and attribute records. BAT and RAT column definitions are introduced where each field first appears in Sect. 3. File-name wildcard * denotes the region_id (e.g., bsn_region00011_90m.tif, bsn_region20029_90m.tif).

File Name	Variable	Type	NoData	File Size	Compressed
Input baseline topology (global coverage)					
dir_global_90m.tif	Flow direction	UInt8	247	~70 GB	~6.5 GB
upg_global_90m.tif	Flow accumulation in cells	UInt32	0	~280 GB	~18 GB
upa_global_90m.tif	Flow accumulation in km²	Float32	-9999	~280 GB	~36 GB
Basin and region masks with attribute tables (global coverage)					
bsn_global_90m.tif	Basin mask	UInt32	0	~280 GB	~0.8 GB
rgn_global_90m.tif	Region mask	UInt8	0	~70 GB	~0.1 GB
bat_global_90m.csv	Basin attribute table	—	—	—	—
rat_global_90m.csv	Region attribute table	—	—	—	—
Basin and region masks with attribute tables (per-region tiles)					
bat_region*_90m.csv	Basin attribute table	—	—	—	—
bsn_region*_90m.tif	Basin mask	UInt32	0	~280 GB	~0.85 GB
bsn_region*_90m_remapped.tif	Basin mask remapped using local basin id	UInt32	0	~280 GB	~0.85 GB
DFS interval index (per-region tiles)					
dfs_in_region*_90m.tif	DFS entry timestamp	Int32	-1	~280 GB	~40 GB
dfs_out_region*_90m.tif	DFS exit timestamp	Int32	-1	~280 GB	~40 GB
Upstream morphometric attributes (per-region tiles)					

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

mean_elev_upstream_region*_90m.tif	Mean upstream elevation	Float32	-9999	~280 GB	~3 GB
min_elev_upstream_region*_90m.tif	Minimum upstream elevation	Float32	-9999	~280 GB	~3 GB
max_elev_upstream_region*_90m.tif	Maximum upstream elevation	Float32	-9999	~280 GB	~3 GB
std_elev_upstream_region*_90m.tif	Standard deviation of upstream elevation	Float32	-9999	~280 GB	~3 GB
mean_slope_upstream_region*_90m.tif	Mean upstream slope	Float32	-9999	~280 GB	~3 GB
std_slope_upstream_region*_90m.tif	Standard deviation of upstream slope	Float32	-9999	~280 GB	~3 GB
perimeter_catchment_region*_90m.tif	Catchment perimeter	Float32	-9999	~280 GB	~3 GB
length_basin_region*_90m.tif	Basin length	Float32	-9999	~280 GB	~3 GB
width_basin_region*_90m.tif	Basin width	Float32	-9999	~280 GB	~3 GB
convexity_planform_region*_90m.tif	Convexity	Float32	-9999	~280 GB	~3 GB
relief_basin_region*_90m.tif	Basin relief	Float32	-9999	~280 GB	~3 GB
hypsometric_integral_region*_90m.tif	Hypsometric integral	Float32	-9999	~280 GB	~3 GB
elongation_ratio_region*_90m.tif	Elongation ratio	Float32	-9999	~280 GB	~3 GB
compactness_gravelius_region*_90m.tif	Gravelius compactness	Float32	-9999	~280 GB	~3 GB
circularity_ratio_region*_90m.tif	Circularity ratio	Float32	-9999	~280 GB	~3 GB

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

lemniscate_ratio_region*_90m.tif	Lemniscate ratio	Float32	-9999	~280 GB	~3 GB

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

Table 2. Global outlet counts fully stratified by drainage destination (exorheic and endorheic) across six area thresholds (≥ 0.05, ≥ 1, ≥ 25, ≥ 100, ≥ 1,000, and ≥ 10,000 km²).

	Drainage destination	⩾0.05 km²	⩾1 km²	⩾25 km²	⩾100 km²	⩾1,000 km²	⩾10,000 km²
This study	All outlets	2,569,927	340,991	57,339	26,970	6,985	1,238
	Exorheic	2,560,220	331,299	48,572	20,228	4,348	795
	Endorheic	9,707	9,692	8,767	6,742	2,637	443

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Table 3. Sixteen per-pixel upstream morphometric attributes of MERIT-FullBasin, delivered as GeoTIFF rasters at native 3 arcsec (approximately 90 m) resolution for targets with upstream drainage area ≥ 10 km². Attributes fall into three algorithmic classes: upstream aggregation and boundary walk are structural passes over the DFS-indexed flow tree, while arithmetic entries are pixel-wise combinations of preceding rasters. Classical references for each variable are listed in the rightmost column and discussed in Sect. 5. The corresponding raster filenames are listed in Table 1.

#	Variable	Symbol	Units	Category	Method / formula	Classical reference
1	Mean upstream elevation	$\bar{z}$	m	Upstream aggregation	$\bar{z} = \frac{1}{n}\sum_i z_i$	—
2	Minimum upstream elevation	z_{min}	m	Upstream aggregation	$z_{min} = \min_i z_i$	—
3	Maximum upstream elevation	z_{max}	m	Upstream aggregation	$z_{max} = \max_i z_i$	—
4	Upstream elevation standard deviation	σ_z	m	Upstream aggregation	$\sigma_z = \sqrt{\frac{1}{n}\sum_i (z_i - \bar{z})^2}$	—
5	Mean upstream slope	$\bar{S}$	deg	Upstream aggregation	$\bar{S} = \frac{1}{n}\sum_i S_i$	—
6	Upstream slope standard deviation	σ_S	deg	Upstream aggregation	$\sigma_S = \sqrt{\frac{1}{n}\sum_i \left(S_i - \bar{S}\right)^2}$	—
7	Catchment perimeter	P	km	Boundary walk	$P = \sum_i w_i \, l_i$	Vossepoel & Smeulders (1982)
8	Basin length	L_b	km	Boundary walk	$L_b = \max_b d_h\,(T, b)$	Schumm (1956)
9	Basin width	L_w	km	Boundary walk	$L_w = \max(\mathrm{proj}_W) - \min(\mathrm{proj}_W)$	Schumm (1956)

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

10	Convexity	C_v	—	Boundary walk	$C_v = \frac{A}{A_{\text{convex hull}}}$	Willemin (2000)
11	Basin relief	H	m	Arithmetic	$H = z_{max} - z_{min}$	Strahler (1952)
12	Hypsometric integral	HI	—	Arithmetic	$HI = \frac{\bar{z} - z_{min}}{H}$	Strahler (1952)
13	Elongation ratio	R_e	—	Arithmetic	$R_e = \frac{2}{L_b}\sqrt{\frac{A}{\pi}}$	Schumm (1956)
14	Gravelius compactness	K_G	—	Arithmetic	$K_G = \frac{P}{2\sqrt{\pi A}}$	Gravelius (1914)
15	Circularity ratio	R_c	—	Arithmetic	$R_c = \frac{4\pi A}{P^2}$	Miller (1953)
16	Lemniscate ratio	k	—	Arithmetic	$k = \frac{\pi L_b^2}{4A}$	Chorley et al. (1957)

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

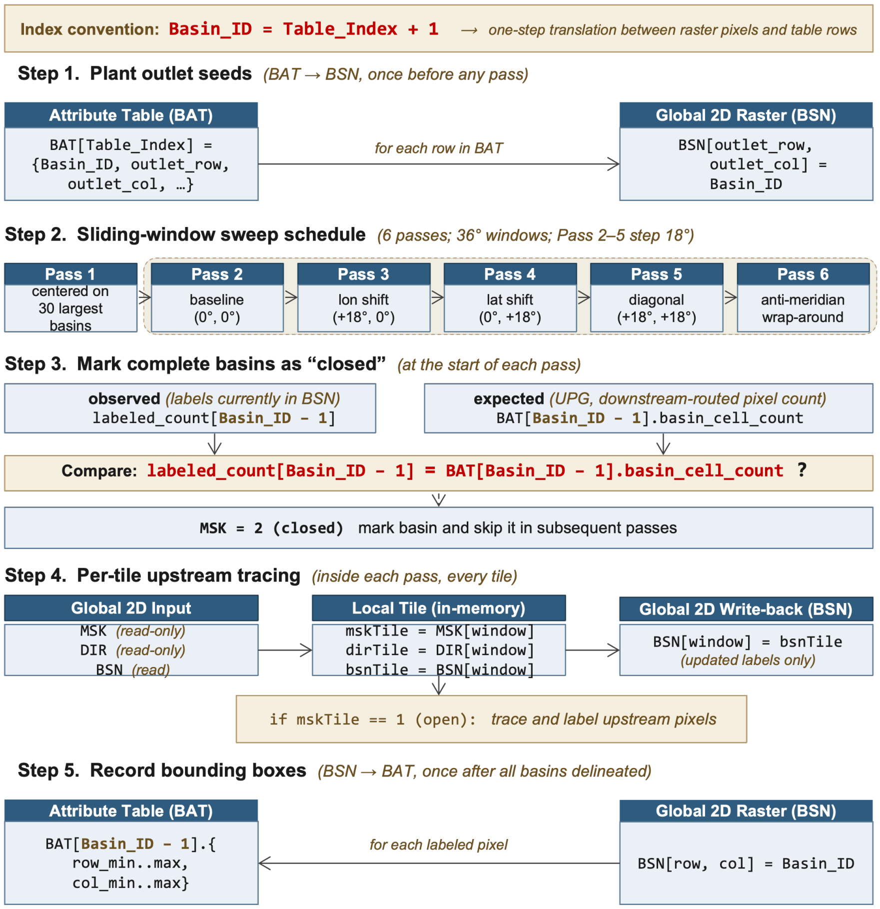
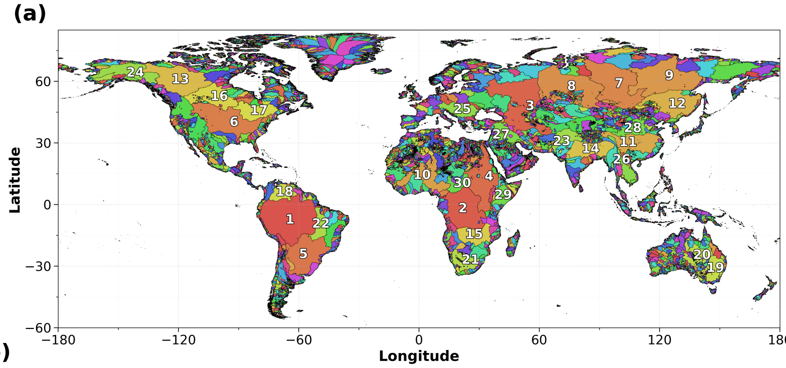


Figure 1. Schematic of the seamless global delineation framework, organized as five steps. Step 1 plants outlet seeds: for each row in the Basin Attribute Table (BAT), the outlet pixel in the global raster (BSN) is set to its Basin_ID under the index convention Basin_ID = Table_Index + 1 (Eq. (1)), and all other land pixels are set to NoData. Step 2 specifies a sliding-window sweep schedule of six 36° passes: Pass 1 centers windows on the 30 largest basins; Passes 2–5 step by 18° in longitude, latitude, and the diagonal from a (0°, 0°) baseline; Pass 6 wraps the antimeridian. Step 3 marks complete basins as closed at the start of each pass: the observed labeled_count[Basin_ID − 1] from BSN is compared against the expected BAT[Basin_ID − 1].basin_cell_count (the downstream-routed UPG cell count); on equality (Eq. (2)) the basin's mask is set to MSK = 2 (closed) and skipped in subsequent passes. Step 4 traces upstream per tile inside each window: read-only globals MSK, DIR, and BSN are buffered into local tiles mskTile, dirTile, bsnTile; for cells with mskTile == 1 (open), upstream pixels are traced and labeled; updated labels in bsnTile are written back into the global BSN. Step 5 records bounding boxes once all basins are delineated: BSN is scanned and per-basin row/column extents are written back to BAT as row_min..max, col_min..max.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

(a)



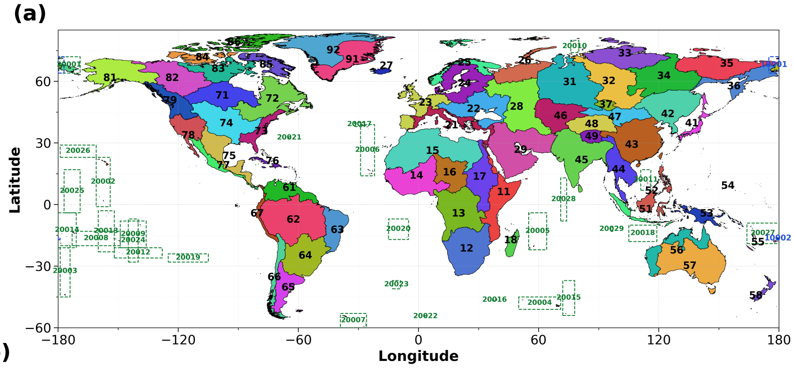
(b)

bat index	region id	local basin_id	global basin_id	...	outlet dir	basin cell_count	basin area_km2	basin nrow	basin ncol	global basin_col_min	...	local basin_col_min	...
0	62	1	1	...	0	698,585,522	5904430.00	30,935	36,062	120,460	...	460	...
1	13	1	2	...	0	432,249,569	3675760.00	27,266	26,708	230,116	...	8,116	...
2	28	1	3	...	255	587,886,873	3243760.00	33,942	41,754	254,360	...	1,160	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
494	10,001	1	495	...	0	8,434,837	27988.30	2,093	6,479	-3,349	...	2,651	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
928	20,001	1	929	...	0	4,202,217	14024.40	1,706	3,811	1,900	...	1,900	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8,594	43	163	8,595	...	0	93,432	729.01	292	614	356,884	...	32,884	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
340,990	51	11,094	340,991	...	0	117	1.00	17	15	339,708	...	10,908	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2,569,926	18	12,877	2,569,927	...	0	6	0.05	6	6	270,183	...	6,183	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 2. Global basin mask and Basin Attribute Table. (a) Global basin mask (BSN), where each basin is displayed using a distinct color. The 30 largest basins, ranked by drainage area, are labeled for reference. (b) Structure and example records of the Basin Attribute Table (BAT). Key fields include the triple linking key (global_basin_id, region_id, local_basin_id), outlet attributes (drainage destination, cell count, and upstream drainage area), basin footprint dimensions (number of rows and columns), and spatial bounding box indices on both the global and regional grids. The complete BAT contains 24,587,290 outlet records, of which 13,332,885 correspond to single-pixel coastal entries that reflect pixel-scale artifacts in the MERIT-Hydro flow-direction grid along coastlines (each pixel drains directly to the ocean with no upstream aggregation) and are retained in the BAT for completeness. For global_basin_id = 495, the global bounding box extends across the 180° E/W antimeridian. For global_basin_id = 3, the outlet_dir equals 255, indicating an endorheic (internally draining) basin (the Caspian Basin, approximately 3.24 million km²).

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

(a)



(b)

region id	...	region row_min	region row_max	region col_min	region col_max	...	region nrow	region ncol	...	region cell_count	...
11	...	64,800	124,800	252,000	282,000	...	60,000	30,000	...	381,369,152	...
12	...	111,600	144,000	229,200	260,400	...	32,400	31,200	...	587,084,897	...
...	...	...	...	...	...	...	...	...	...	...	...
43	...	50,400	84,000	324,000	363,600	...	33,600	39,600	...	609,079,064	...
...	...	...	...	...	...	...	...	...	...	...	...
92	...	1,200	22,800	128,400	201,600	...	21,600	73,200	...	527,035,091	...
10,001	...	15,600	25,200	-6,000	3,600	...	9,600	9,600	...	19,168,845	...
10,002	...	121,200	122,400	-1,200	1,200	...	1,200	2,400	...	9,650	...
20,001	...	15,600	25,200	0	13,200	...	9,600	13,200	...	28,616,669	...
...	...	...	...	...	...	...	...	...	...	...	...
20,029	...	115,200	117,600	331,200	332,400	...	2,400	1,200	...	2,792	...

Figure 3. Global processing regions and Regional Attribute Table. (a) Global region mask (RGN) showing the final set of 96 processing regions. The first 65 regions (IDs 11–92) correspond to continental landmasses and broadly follow the HydroBASINS Level-02 hierarchy; two antimeridian regions (IDs 10001–10002) group basins that cross the 180° E/W antimeridian; and 29 island-group regions (IDs 20001–20029) consolidate isolated island basins by spatial proximity. (b) Structure and example entries of the Regional Attribute Table (RAT), listing spatial bounding indices (region_row_min/max and region_col_min/max), region dimensions (region_nrow and region_ncol), and the number of valid land cells (region_cell_count) within each bounding box. For region 10001, the negative region_col_min reflects the bounding box extending west of the antimeridian. All regional bounding boxes are snapped to integer-degree boundaries (multiples of 1,200 cells at 3 arcsec resolution) to ensure exact alignment with global rasters.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

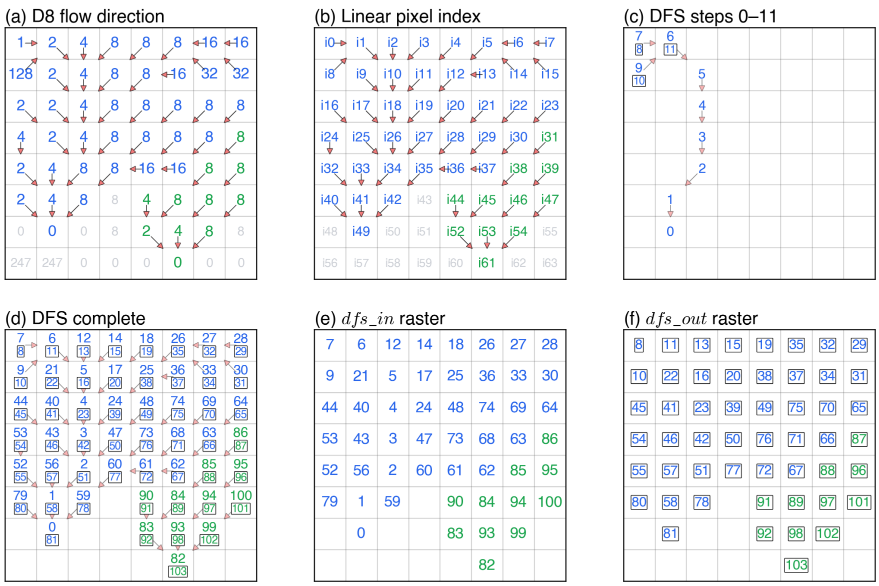


Figure 4. Schematic illustration of DFS interval labeling on an 8 × 8 D8 flow-direction raster containing two basins: Basin A (blue, 41 pixels) and Basin B (green, 11 pixels). Children are sorted by upstream drainage area (UPA) in descending order so that the main stem is traversed first. (a) D8 flow-direction codes with arrows indicating flow paths. (b) Linear pixel indices (i0–i63); gray cells denote invalid pixels. (c) A snapshot after DFS steps 0–11: arrows and timestamps appear only for pixels visited so far; plain numbers indicate the entry timestamp (dfs_in) and boxed numbers indicate the exit timestamp (dfs_out). (d) Completed DFS traversal showing both entry and exit timestamps for all 52 valid pixels. (e) The final dfs_in raster, recording the time each pixel is first visited. (f) The final dfs_out raster, recording the time each pixel's subtree is fully explored.

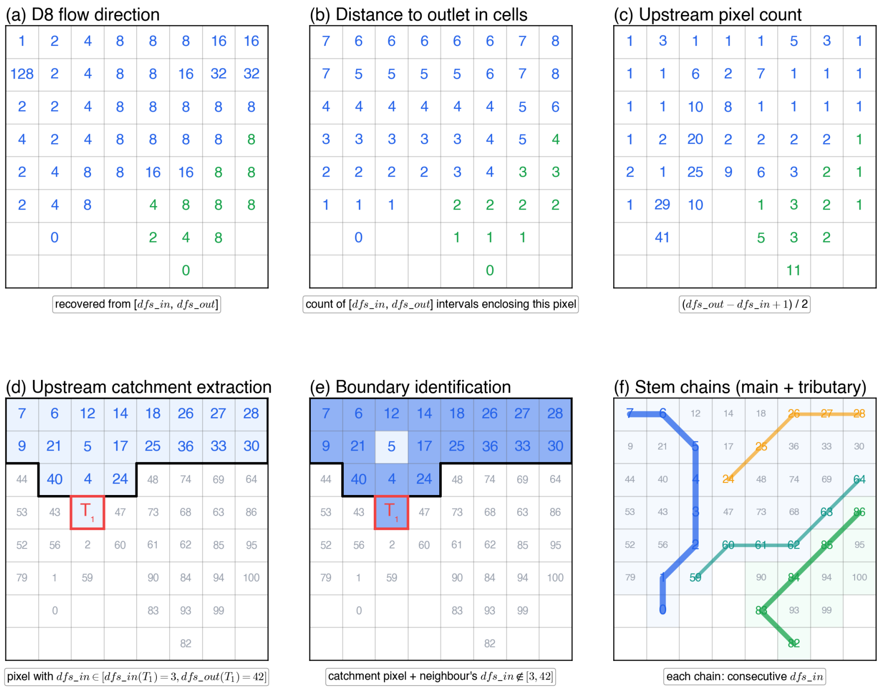


Figure 5. Capabilities of the DFS interval index, illustrated on the 8 × 8 example grid of Fig. 4. Panels (a)–(c) show three per-pixel quantities recoverable from the [dfs_in, dfs_out] pair alone: (a) the D8 flow direction at each pixel, recovered by reconstructing each pixel's parent from the dfs interval pair and encoding the (Δrow, Δcol) offset to a D8 power-of-two code; (b) the distance to the outlet in cells, equal to the count of [dfs_in, dfs_out] intervals enclosing each pixel; (c) the upstream pixel count per pixel, given by (dfs_out − dfs_in + 1)/2. Panels (d)–(f) demonstrate applications of the index, taking T_1 = pixel i26 in Basin A as the example target with [dfs_in[T_1], dfs_out[T_1]] = [3, 42]: (d) the upstream catchment of T_1, extracted as the set of pixels whose dfs_in lies in T_1's interval (interval-containment test), with the 4-connected geometric boundary drawn in black; (e) the catchment boundary, identified as catchment pixels with at least one neighbor whose dfs_in lies outside T_1's interval (highlighted in darker blue); (f) stem chains, identified as maximal runs of consecutive dfs_in values. The chain rooted at a basin outlet is the basin's main stem, and chains rooted at interior branching points are tributary stems (blue: Basin A's main stem; teal and orange: two of its tributary stems; green: Basin B's main stem). Cell values shown are: D8 codes in (a), distance-to-outlet values in (b), upstream-pixel counts in (c), and dfs_in values in (d)–(f).

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

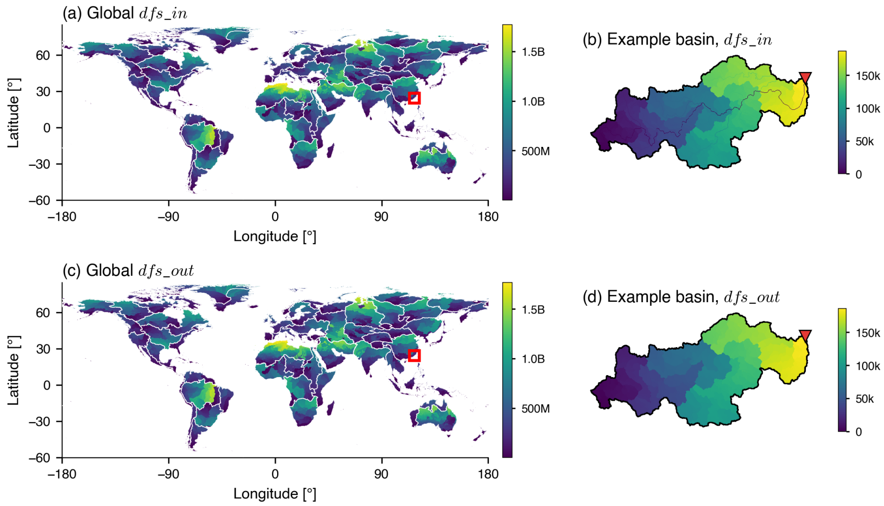


Figure 6. Global DFS interval index of MERIT-FullBasin with a zoom into the example basin. (a, c) Thin white lines trace the region boundaries, making the per-region DFS-index segmentation explicit. Because sibling branches are visited in descending order of upstream drainage area, consecutive dfs_in values trace each basin's main stem before branching outward into tributaries, and the spatial pattern of dfs_in visually recovers the river network. The red rectangle locates the example basin (basin 8595, 729 km², southeastern Fujian, China). (b, d) Zoom-in on the example basin at the native 3 arcsec resolution showing dfs_in and dfs_out, respectively; the black polygon traces the basin boundary and the red triangle (▼) marks the outlet (117.90° E, 24.40° N). Each panel uses an independent viridis color scale.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

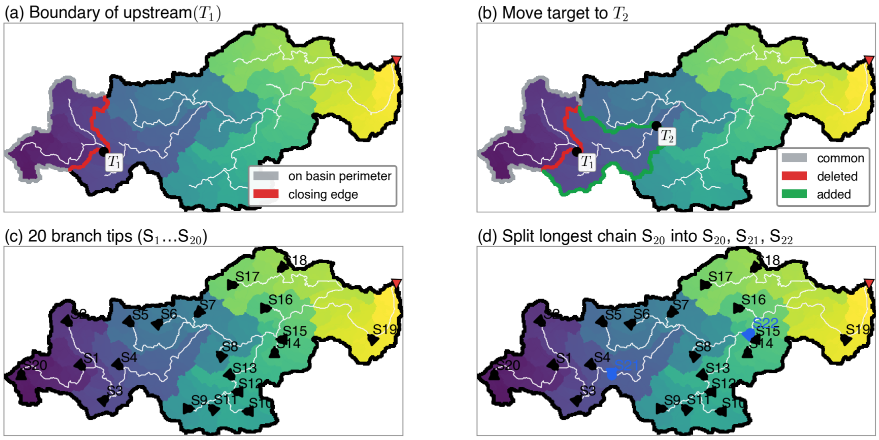


Figure 7. Boundary walk on the example basin (basin 8595, southeastern Fujian, China; 729 km², 3 arcsec resolution). All four panels share the same dfs_in color fill, basin perimeter (black), stream polylines (white; UPA $\geqslant$ 10 km²), and outlet marker (▼). (a) For a target T_1 on the main stem, the boundary of upstream(T_1) is identified by the eight-neighbor interval test, requiring one comparison per candidate; the segment of the basin perimeter that lies within upstream(T_1) is shown in gray, while the new closing edge interior to the basin appears in red. (b) Moving the target downstream to T_2, the new ring is derived from the previous ring by symmetric difference: pixels common to ring(T_1) and ring(T_2) (gray), pixels lost when $T_1 \rightarrow T_2$ (red, deleted), and pixels added at T_2 (green). (c) The N stream sources of the basin are labeled $S_1 \ldots$S_N at their channel heads, ordered upstream-first along their junction position on the main stem; the main stem itself is S_N (the longest chain). Black arrows indicate the downstream flow direction at each source. (d) Splitting the longest chain S_N at 33% and 67% of its length yields two additional sub-streams S_{N+1} and S_{N+2} (blue), demonstrating intra-basin parallelism for the boundary walk.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

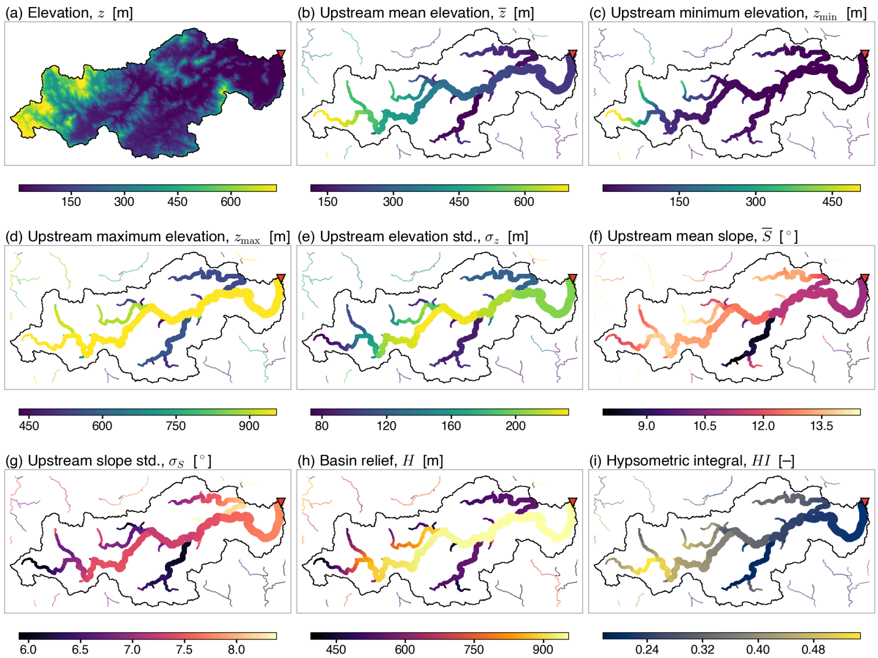


Figure 8. Per-pixel upstream terrain attributes of the example basin (global_basin_id = 8595; local_basin_id = 163; 729 km², southeastern Fujian, China): (a) basin elevation z (90 m DEM, shown for context), (b) upstream mean elevation $\bar{z}$, (c) upstream minimum elevation z_{min}, (d) upstream maximum elevation z_{max}, (e) upstream elevation standard deviation σ_z, (f) upstream mean slope $\bar{S}$, (g) upstream slope standard deviation σ_S, (h) basin relief H, and (i) hypsometric integral HI. Panels (b)–(g) are the six upstream aggregates in Table 3; (h) and (i) are the two arithmetic derivatives, computed pixel-wise from (b)–(d). Each target pixel (UPA ≥ 10 km²) is drawn as a filled circle whose radius scales with log10(UPA), so main-stem pixels appear larger. Pixels from neighboring basins sharing the bounding box are plotted at a uniform small radius with 50% opacity for geographic context. The black polygon traces the basin boundary; the red triangle (▼) marks the outlet.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

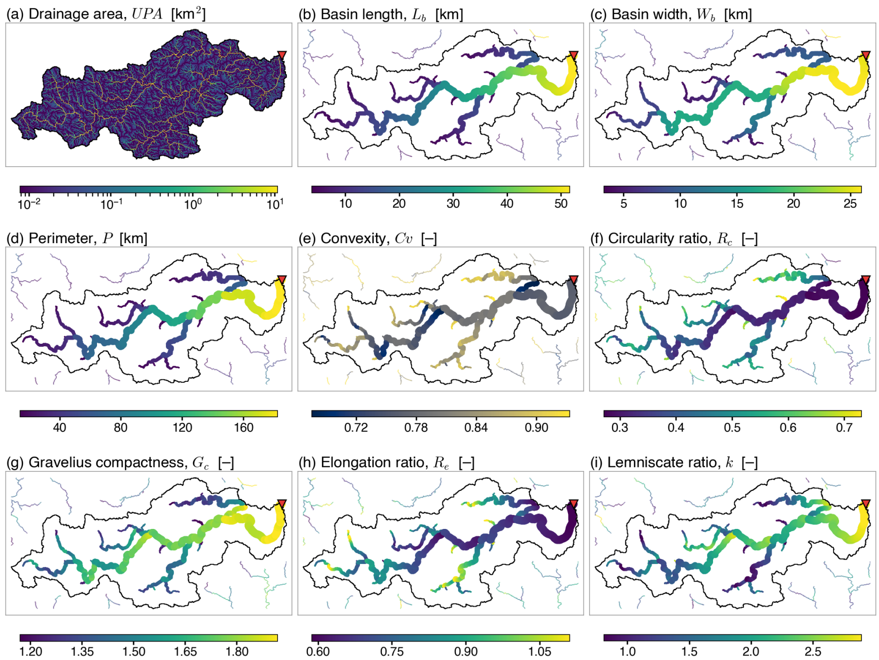


Figure 9. Per-pixel planform and shape attributes of the example basin (global_basin_id = 8595; local_basin_id = 163; 729 km², southeastern Fujian, China): (a) drainage area UPA at every pixel (shown for context), (b) basin length L_b, (c) basin width L_w, (d) perimeter P, (e) convexity C_v, (f) circularity ratio R_c, (g) Gravelius compactness K_G, (h) elongation ratio R_e, and (i) lemniscate ratio k. Panels (b)–(e) are the four boundary-walk metrics in Table 3, traced on each target's upstream-catchment outline; panels (f)–(i) are the four dimensionless shape ratios, computed pixel-wise from panels (a), (b), and (d). Each target pixel (UPA ⩾ 10 km²) is drawn as a filled circle whose radius scales with log10(UPA), so main-stem pixels appear larger. Pixels from neighboring basins sharing the bounding box are plotted at a uniform small radius with 50% opacity for geographic context. The black polygon traces the basin boundary; the red triangle (▼) marks the outlet.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

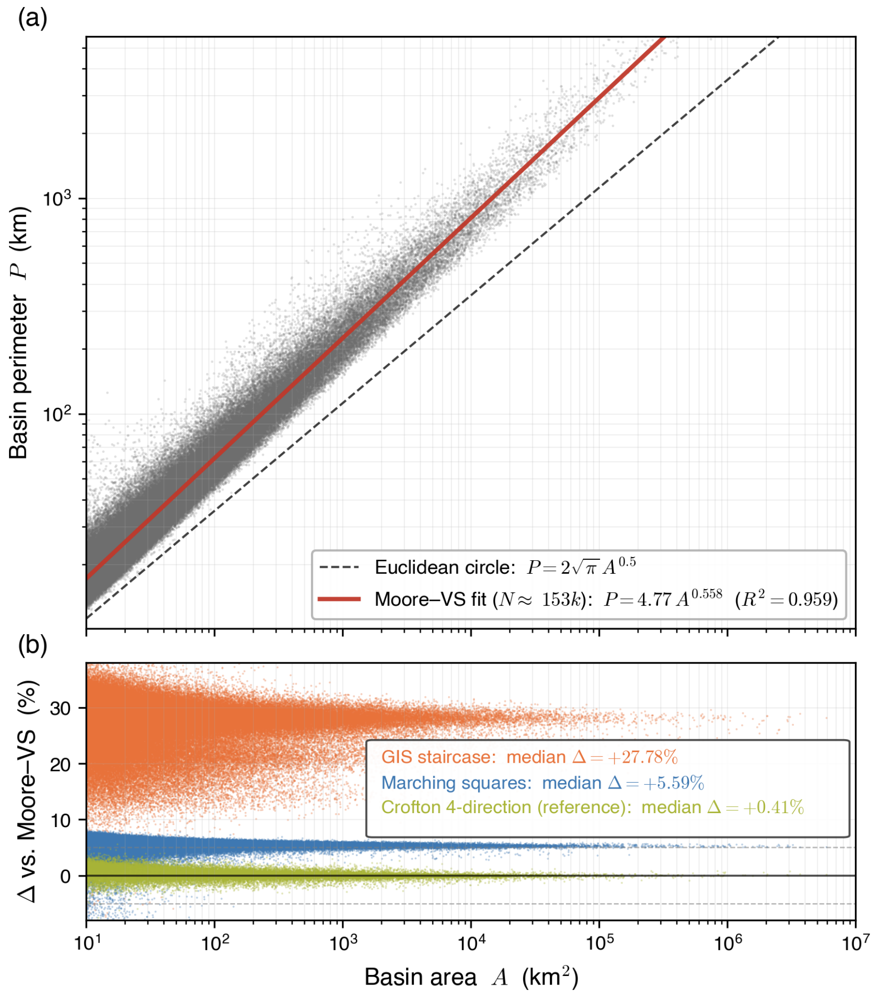


Figure 10. Perimeter–area scaling and cross-validation of the released Moore–VS perimeter on 153,219 global records (outlet + sub-mid, A ≥ 10 km²). (a) Log-log scatter of basin perimeter P against area A, with the Moore–VS power-law fit P = 4.77 · A^0.558 (R^2 = 0.96) and the Euclidean-circle reference α = 0.5 (dashed). The exponent exceeds 0.5, consistent with the fractal character of drainage divides (Mandelbrot, 1967). (b) Relative deviation Δ = 100 × (P_method − P_Moore–VS) / P_Moore–VS versus A for three alternative methods: GIS staircase, marching squares, and Crofton 4-direction (the integral-geometry reference, restricted to pixel aspect ≤ 1.2); reference lines mark Δ = 0 and ±5%. Appendix B gives full method descriptions and cross-algorithm statistics.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Appendix A

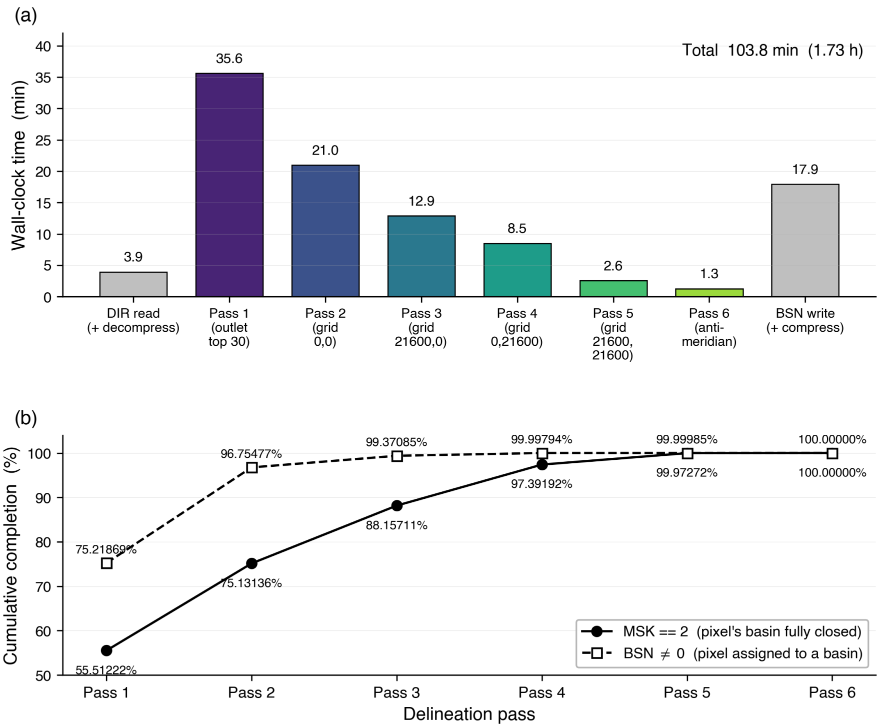


Figure A1. Wall-clock performance of the multi-pass global basin delineation on the MERIT-Hydro 3 arcsec flow-direction grid (432,000 × 174,000 pixels; 22.26 billion valid land pixels; 24.59 million outlets). (a) Per-stage time: DIR GeoTIFF read, the six delineation passes (one outlet-centered tile sweep over the 30 largest basins, four offset grid sweeps, and a final antimeridian-wrap pass), and BSN GeoTIFF write. Total end-to-end wall-clock is 103.8 min (1.73 h); both I/O bars include (de)compression, not only raw disk transfer. (b) Cumulative completion after each pass: MSK == 2 (a pixel's basin is fully closed, solid line) and BSN ≠ 0 (a pixel has been assigned a basin ID, dashed line). The two curves converge to 100% by Pass 6; the early gap between BSN ≠ 0 and MSK == 2 reflects pixels that carry a basin ID before their basin's traversal completes.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

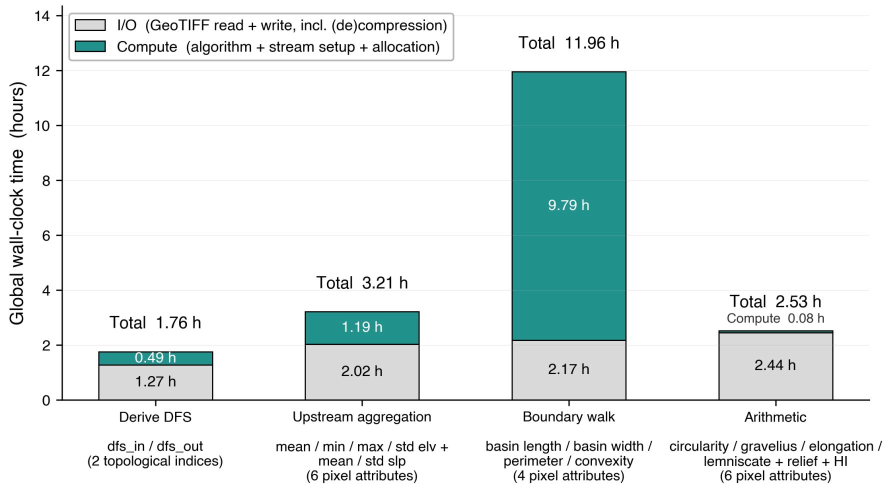


Figure A2. Global wall-clock time for each of the four computational kernels of the MERIT-FullBasin pipeline. Each bar is stacked into I/O (GeoTIFF read + write, including (de)compression) and Compute (algorithm core and per-tile setup). Kernel outputs are: Derive DFS: dfs_in, dfs_out (2 pixel layers); Upstream aggregation: mean, min, max, std elevation and mean, std slope (6 pixel attributes); Boundary walk: basin length, basin width, perimeter, convexity (4 pixel attributes); Arithmetic: circularity ratio, Gravelius compactness, elongation ratio, lemniscate ratio, basin relief, hypsometric integral (6 pixel attributes). Boundary walk dominates the per-pixel attribute wall-clock at 11.96 h (~68% of the 17.7 h spent on the 16 attribute layers), driven by compute rather than I/O; the other three kernels are I/O-bound and lie within a factor of two of each other.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

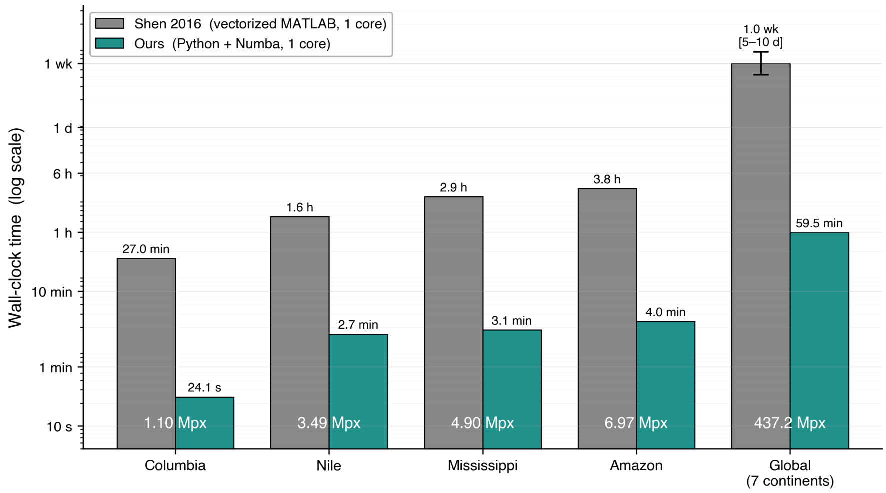


Figure A3. Single-core wall-clock reproduction of Shen et al. (2016, Table 2). The figure compares five cases: four reference basins (Columbia, Nile, Mississippi, Amazon) and the global 30 arcsec 7-continent grid (437.2 million pixels). Shen values come from the vectorized MATLAB implementation of Shen et al. (2016) on an Intel i7-4860HQ (2.4 GHz base, 3.6 GHz turbo; 1 core, 2013); MERIT-FullBasin values come from a Python-with-Numba reimplementation on an Apple M2 (1 core, 2022), reported as 3-run medians after warm-up. The y-axis is a logarithmic wall-clock scale; the error bar on the global Shen case brackets Shen et al. (2016, Sect. 5)'s "≈ one week of production time" as 5–10 d. The release provides full replication scripts, per-continent pixel-level accuracy tables, and a perimeter-method comparison table (see Code availability).

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Open Access Earth System Science Data Discussions

(a) DFS interval → D8 flow direction (DIR)

```
Input:  dfs_in, dfs_out
Output: DIR
S ← empty stack
for each i in argsort(dfs_in):                 ▷ d2u order
    while S ≠ ∅ and dfs_out[top(S)] < dfs_in[i]:
        pop(S)
    parent[i] ← top(S) if S ≠ ∅ else i          ▷ outlet → self
    push(S, i)
for each i:  DIR[i] ← encode_D8(parent[i] − i)  ▷ (Δrow, Δcol) → D8
```

(b) DFS interval → distance to outlet in cells (RNK)

```
Input:  dfs_in, dfs_out
Output: RNK
S ← empty stack
for each i in argsort(dfs_in):                 ▷ d2u order
    while S ≠ ∅ and dfs_out[top(S)] < dfs_in[i]:
        pop(S)
    RNK[i] ← height(S)                          ▷ = ancestor depth
    push(S, i)
```

(c) DFS interval → upstream pixel count (UPG)

```
Input:  dfs_in, dfs_out
Output: UPG
for each pixel i:
    UPG[i] ← (dfs_out[i] − dfs_in[i] + 1) / 2   ▷ pure arithmetic, no sweep
```

Figure A4. Pseudocode for recovering the three per-pixel scalars of Fig. 5(a–c) from the released [dfs_in, dfs_out] pair alone. (a) D8 flow direction. The algorithm reads pixels in order of increasing dfs_in (outlet to upstream) and pushes each pixel onto a stack on entry and pops it on exit. At any moment the pixel's parent in the flow tree is the one currently below it on the stack; the (Δrow, Δcol) offset to the parent gives the D8 code, and pixels with no parent become outlets. (b) Distance to the outlet, in cells. The same pass tracks the height of the stack at each pixel's entry; that height is the pixel's distance to the outlet, because every ancestor on the flow path is on the stack at that moment. Both (a) and (b) take a single linear pass over valid pixels. (c) Upstream pixel count. A subtree of n pixels generates 2n entry/exit events, so (dfs_out − dfs_in + 1) / 2 gives n directly.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

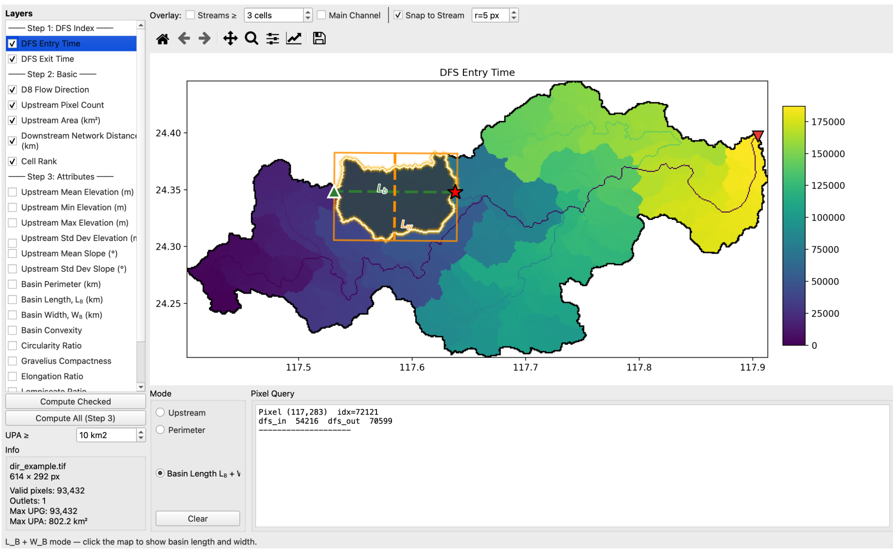


Figure A5. FlowDex: a Python reference implementation of MERIT-FullBasin query operations on the released rasters. A single click on any pixel returns the upstream catchment, the perimeter, or the basin length and basin width, each derived at query time from dfs_in, dfs_out, DIR, and UPA via direct interval containment on the DFS index (Sect. 4). No per-basin precomputation is required. The figure illustrates the queryable-index character of the data product rather than interface specifics; source ships with the dataset (see Code availability).

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Appendix B

Basin perimeter on a raster grid is method-dependent: every finite computation carries an algorithm-dependent discretization bias (Proffitt and Rosen, 1979), in the same way that coastline length depends on the ruler used (Mandelbrot, 1967). To bound this method-dependence, we measure perimeter with four methods based on distinct mathematical principles on the same set of basin masks. One method (Moore–VS) is the production algorithm released with the data product; the other three serve as independent cross-checks.

The validation sample comprises ~150 k binary basin masks: ~95 k outlet masks (one per basin with UPA $\geqslant$ 10 km², spanning the full global domain) and ~58 k interior sub-mid masks (one per basin with outlet UPA $\geqslant$ 20 km², seeded at a tributary or main-stem point whose UPA lies between 30% and 70% of the outlet UPA, a range that samples mid-network points away from both the outlet and the headwater extremes). The outlet sample tests the full-boundary Moore–VS computation on every basin; the sub-mid sample also tests the incremental boundary update along streams used by the production code (Sect. 5). All four methods are applied to the same masks.

The production method is an 8-connected (Moore) chain code with Vossepoel–Smeulders weights (Freeman, 1961; Vossepoel and Smeulders, 1982), hereafter Moore–VS. The algorithm traces the boundary once, starting from the top-most left-most interior pixel and walking counter-clockwise through the 8-connected neighborhood; each step contributes a length term to the running total. The naive Freeman sum (one pixel side per cardinal step, $\sqrt{2}$ pixel sides per diagonal step) over-estimates the length of straight lines of arbitrary orientation by several percent. Vossepoel and Smeulders (1982) derived bias-minimizing weights: each cardinal step contributes 0.980 pixel sides, each diagonal step contributes 1.406 pixel sides, and each transition between cardinal and diagonal steps subtracts a corner penalty of 0.091 times the mean pixel side length. The algorithm recomputes physical pixel side lengths at the step mid-latitude on the WGS84 authalic sphere (an equal-area sphere of radius 6,371,007 m), accounting for pixel anisotropy at high latitudes without any square-pixel approximation. The implementation is a C routine with OpenMP per-basin parallelism, and a pure-Python Numba reimplementation handles cross-language validation.

The first cross-check is the Crofton 4-direction formula (Crofton, 1868; Santaló, 1976), an integral-geometry estimator; we use it as the reference for comparing Moore–VS and marching squares. Rather than tracing a boundary, the method counts 0↔1 pixel transitions along four lattice directions (horizontal, vertical, and the two diagonals) and combines them into a weighted sum that approaches the true length as the number of sampling directions tends to infinity. We call scikit-image's perimeter_crofton with directions = 4 (van der Walt et al., 2014) and convert pixel units to physical units via the geometric mean of the north–south and east–

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

west pixel side lengths. Because this conversion assumes isotropic pixels, the method is restricted to basins whose mean pixel aspect ratio is $\leqslant$ 1.2; basins failing this threshold are excluded from the Crofton comparison but retained for the other three methods. The threshold limits the anisotropy-induced error in the geometric-mean conversion and corresponds approximately to the domain south of ~40° N; it admits essentially all tropical and subtropical basins.

The second cross-check is 4-connected vector polygonization, a staircase-polygon method commonly used in standard GIS workflows; this is the GIS staircase method in Fig. 10b. The mask is polygonized through 4-connected pixel edges, yielding an exterior ring measured as a geodesic polyline on the WGS84 ellipsoid. In Python we use rasterio.features.shapes (a thin wrapper around the GDAL OGRPolygonize primitive) followed by pyproj.Geod.geometry_length. Because every route ultimately calls the same GDAL primitive, the same result is reproducible in standard GIS software without Python: in QGIS, Raster → Raster to Polygon (gdal:polygonize) followed by length($geometry) on the resulting exterior under a geographic coordinate reference system with ellipsoid WGS84; in ArcGIS Pro, Raster to Polygon followed by Calculate Geometry → Shape_Length with Method = Geodesic; on the command line, gdal_polygonize.py followed by ST_Length(geometry). All three routes agree to meter-level numerical noise. For straight lines of uniform orientation, the staircase over-estimate has a theoretical upper bound of $4/\pi \approx 1.273$ (Proffitt and Rosen, 1979).

The third cross-check is marching squares, a sub-pixel iso-contour method (Lorensen and Cline, 1987, 2-D variant). Rather than snapping the boundary to pixel edges, the method traces the 0.5 iso-contour through each 2×2 sub-cell, with vertices at pixel-edge midpoints, forming a piecewise-linear boundary. We use scikit-image's find_contours with level = 0.5 on a once-padded mask, keep the longest closed contour, and measure its geodesic length by summing pyproj.Geod.inv distances over consecutive segments; the one-pixel frame of zeros guarantees closure for basins touching the raster edge. This class of methods is widely used for coastline and contour extraction and removes the staircase bias of the 4-connected polygonization.

On the ~150 k-mask sample, the three non-staircase estimates (Moore–VS, Crofton 4-direction, marching squares) agree to within a few percent of one another. Taking the Crofton 4-direction estimate as the integral-geometry reference, Moore–VS shows the tightest distribution against it, with a median offset of about −0.4% and a 2nd-to-98th-percentile (P2–P98) band of roughly ±1.8% (Fig. 10b). The 4-connected polygonization (GIS staircase in Fig. 10b), included as a control, has a median offset of about +28%, in quantitative agreement with its independent theoretical upper bound of $4/\pi \approx 1.273$ (i.e., +27.3%). The three-way agreement among the non-staircase methods indicates that the released perimeter does not depend on the implementation choice, and the match between the 4-connected over-estimate and its theoretical upper bound indicates that the four methods are independent in their sources of bias.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Basin perimeter is a standard descriptor in studies of basin shape and scale-dependent geometry (Breyer and Snow, 1992); among the four methods tested, Moore–VS provides the tightest finite approximation to the Crofton reference and ships with the data product.

As a further cross-check against an external reference dataset, we compare Moore–VS against the published P.tif perimeters of Shen et al. (2017), a global per-pixel basin-morphometric dataset at 1 km on HydroSHEDS. The Moore–VS perimeter raster is re-computed on Shen's North America grid (4,319 × 10,241 cells, HydroSHEDS 1 km aligned to Shen's NA/BL.tif) and differenced against Shen's NA P.tif at every pixel with Strahler order $\geqslant 1$ (2.872 million pixels, the domain over which Shen's P is defined). On this 2.872-million-pixel sample, Moore–VS agrees with Shen at 91.3% within 10% with a median signed offset of −5.94%, a mean signed −6.68%, and $\rho = 0.989$. The −6% offset is the signature of the Vossepoel–Smeulders correction: when the correction is disabled and the raw Moore chain code is written to a companion P8.tif raster, the residual against Shen drops to 91.9% within 5% and 98.7% within 10%, with median signed −0.54% and mean signed −1.59%, indicating that Shen's 8-neighbor chain-code method is algorithmically the uncorrected limit of the same walk. As a global cross-continent check, the same comparison is repeated at every basin outlet on HydroSHEDS (each pit or coastal-drainage pixel) with at least 100 upstream cells across the seven Shen continents, giving 31,171 outlets where Moore–VS and Shen perimeter are both defined; Moore–VS sits a median −5.95% below Shen with a P5–P95 band of [−9.08%, −4.23%] and a cross-continent spread in the median of only 0.18% (from −5.87% on Africa to −6.05% on North America). The match between the per-pixel North America residual (−5.94%) and the per-outlet global residual (−5.95%) shows that the difference between the released perimeter and Shen's reference is a stable method-definition gap and not a computational error.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Alexander, R. B., Boyer, E. W., Smith, R. A., Schwarz, G. E., and Moore, R. B.: The role of headwater streams in downstream water quality, J. Am. Water Resour. Assoc., 43, 41–59, https://doi.org/10.1111/j.1752-1688.2007.00005.x, 2007.

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

Amatulli, G., Garcia Marquez, J., Sethi, T., Kiesel, J., Grigoropoulou, A., Üblacker, M. M., Shen, L. Q., and Domisch, S.: Hydrography90m: a new high-resolution global hydrographic dataset, Earth Syst. Sci. Data, 14, 4525–4550, https://doi.org/10.5194/essd-14-4525-2022, 2022.

Breyer, S. P. and Snow, R. S.: Drainage basin perimeters: a fractal significance, Geomorphology, 5, 143–157, https://doi.org/10.1016/0169-555X(92)90062-S, 1992.

Celko, J.: Joe Celko's Trees and Hierarchies in SQL for Smarties, 1st edn., Morgan Kaufmann, San Francisco, CA, 2004.

Cho, H.: Avoid backtracking and burn your inputs: CONUS-scale watershed delineation using OpenMP, Environ. Model. Softw., 183, 106244, https://doi.org/10.1016/j.envsoft.2024.106244, 2025.

Chorley, R. J., Malm, D. E. G., and Pogorzelski, H. A.: A new standard for estimating drainage basin shape, Am. J. Sci., 255, 138–141, https://doi.org/10.2475/ajs.255.2.138, 1957.

Crofton, M. W.: On the theory of local probability, applied to straight lines drawn at random in a plane; the methods used being also extended to the proof of certain new theorems in the integral calculus, Philos. Trans. R. Soc. Lond., 158, 181–199, https://doi.org/10.1098/rstl.1868.0008, 1868.

Freeman, H.: On the encoding of arbitrary geometric configurations, IRE Trans. Electron. Comput., EC-10, 260–268, https://doi.org/10.1109/TEC.1961.5219197, 1961.

Gravelius, H.: Grundriss der gesamten Gewässerkunde, Band 1: Flusskunde (Compendium of Hydrology, Vol. I), Göschen, Berlin, 1914.

Haag, S., Shakibajahromi, B., and Shokoufandeh, A.: A new rapid watershed delineation algorithm for 2D flow direction grids, Environ. Model. Softw., 109, 420–428, https://doi.org/10.1016/j.envsoft.2018.08.026, 2018.

Haag, S. and Shokoufandeh, A.: Development of a data model to facilitate rapid watershed delineation, Environ. Model. Softw., 122, 103973, https://doi.org/10.1016/j.envsoft.2018.07.016, 2019.

He, C., Yang, C.-J., Turowski, J. M., Ott, R. F., Braun, J., Tang, H., Ghantous, S., Yuan, X., and Stucky de Quay, G.: A global dataset of the shape of drainage systems, Earth Syst. Sci. Data, 16, 1151–1166, https://doi.org/10.5194/essd-16-1151-2024, 2024.

Jiang, L., Wu, H., Yang, T., Qu, L., and Huang, Z.: Accelerating large-scale hydrological modeling with stepwise spatial–temporal multimember parallelization, Environ. Model. Softw., 191, 106495, https://doi.org/10.1016/j.envsoft.2025.106495, 2025.

Jiang, L., Wu, H., Chen, W., Huang, Z., and Yang, T.: MERIT-FullBasin: a topologically closed global watershed dataset at 3 arcsec (~90 m) resolution, with companion FlowDex Python reference implementation, Zenodo [data set and code], https://doi.org/10.5281/zenodo.20344113, 2026.

Kotyra, B.: High-performance watershed delineation algorithm for GPU using CUDA and OpenMP, Environ. Model. Softw., 160, 105613, https://doi.org/10.1016/j.envsoft.2022.105613, 2023.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviours via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan – A global community dataset for large-sample hydrology, Sci. Data, 10, 61, https://doi.org/10.1038/s41597-023-01975-w, 2023.

Lehner, B., Verdin, K., and Jarvis, A.: New global hydrography derived from spaceborne elevation data, Eos, Trans. AGU, 89, 93–94, https://doi.org/10.1029/2008EO100001, 2008.

Lehner, B. and Grill, G.: Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems, Hydrol. Process., 27, 2171–2186, https://doi.org/10.1002/hyp.9740, 2013.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Lin, P., Pan, M., Wood, E. F., Yamazaki, D., and Allen, G. H.: A new vector-based global river network dataset accounting for variable drainage density, Sci. Data, 8, 28, https://doi.org/10.1038/s41597-021-00819-9, 2021.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., and Thieme, M.: Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution, Sci. Data, 6, 283, https://doi.org/10.1038/s41597-019-0300-6, 2019.

Lorensen, W. E. and Cline, H. E.: Marching cubes: a high resolution 3D surface construction algorithm, ACM SIGGRAPH Comput. Graph., 21, 163–169, https://doi.org/10.1145/37402.37422, 1987.

Mandelbrot, B. B.: How long is the coast of Britain? Statistical self-similarity and fractional dimension, Science, 156, 636–638, https://doi.org/10.1126/science.156.3775.636, 1967.

Miller, V. C.: A Quantitative Geomorphic Study of Drainage Basin Characteristics in the Clinch Mountain Area, Virginia and Tennessee, Technical Report No. 3, Department of Geology, Columbia University, New York, 1953.

Moore, R. B., McKay, L. D., Rea, A. H., Bondelid, T. R., Price, C. V., Dewald, T. G., and Johnston, C. M.: User's guide for the National Hydrography Dataset Plus (NHDPlus) high resolution, U.S. Geological Survey Open-File Report 2019–1096, 66 pp., https://doi.org/10.3133/ofr20191096, 2019.

Proffitt, D. and Rosen, D.: Metrication errors and coding efficiency of chain-encoding schemes for the representation of lines and edges, Comput. Graph. Image Process., 10, 318–332, https://doi.org/10.1016/S0146-664X(79)80041-6, 1979.

Santaló, L. A.: Integral Geometry and Geometric Probability, Encyclopedia of Mathematics and Its Applications, Vol. 1, Addison-Wesley, Reading, MA, 1976.

Schumm, S. A.: Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey, Geol. Soc. Am. Bull., 67, 597–646, https://doi.org/10.1130/0016-7606(1956)67[597:EODSAS]2.0.CO;2, 1956.

Shen, X., Vergara, H. J., Nikolopoulos, E. I., Anagnostou, E. N., Hong, Y., Hao, Z., Zhang, K., and Mao, K.: GDBC: a tool for generating global-scale distributed basin morphometry, Environ. Model. Softw., 83, 212–223, https://doi.org/10.1016/j.envsoft.2016.05.012, 2016.

Shen, X., Anagnostou, E. N., Mei, Y., and Hong, Y.: A global distributed basin morphometric dataset, Sci. Data, 4, 160124, https://doi.org/10.1038/sdata.2016.124, 2017.

Sivapalan, M., Blöschl, G., Zhang, L., and Vertessy, R.: Downward approach to hydrological prediction, Hydrol. Process., 17, 2101–2111, https://doi.org/10.1002/hyp.1425, 2003.

Strahler, A. N.: Hypsometric (area-altitude) analysis of erosional topography, Geol. Soc. Am. Bull., 63, 1117–1142, https://doi.org/10.1130/0016-7606(1952)63[1117:HAAOET]2.0.CO;2, 1952.

Tabari, H.: Climate change impact on flood and extreme precipitation increases with water availability, Sci. Rep., 10, 13768, https://doi.org/10.1038/s41598-020-70816-2, 2020.

Tarboton, D. G.: A new method for the determination of flow directions and contributing areas in grid digital elevation models, Water Resour. Res., 33, 309–319, https://doi.org/10.1029/96WR03137, 1997.

Tarjan, R.: Depth-first search and linear graph algorithms, SIAM J. Comput., 1, 146–160, https://doi.org/10.1137/0201010, 1972.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the scikit-image contributors: scikit-image: image processing in Python, PeerJ, 2, e453, https://doi.org/10.7717/peerj.453, 2014.

Verdin, K. L.: Hydrologic Derivatives for Modeling and Applications (HDMA) database, U.S. Geological Survey [data release], https://doi.org/10.5066/F7S180ZP, 2017.

Verdin, K. L. and Verdin, J. P.: A topological system for delineation and codification of the Earth's river basins, J. Hydrol., 218, 1–12, https://doi.org/10.1016/S0022-1694(99)00011-6, 1999.

Vossepoel, A. M. and Smeulders, A. W. M.: Vector code probability and metrication error in the representation of straight lines of finite length, Comput. Graph. Image Process., 20, 347–364, https://doi.org/10.1016/0146-664X(82)90057-0, 1982.

Willemin, J. H.: Hack's law: sinuosity, convexity, elongation, Water Resour. Res., 36, 3365–3374, https://doi.org/10.1029/2000WR900229, 2000.

Wu, H., Adler, R. F., Tian, Y., Huffman, G. J., Li, H., and Wang, J.: Real-time global flood estimation using satellite-based precipitation and a coupled land surface and routing model, Water Resour. Res., 50, 2693–2717, https://doi.org/10.1002/2013WR014710, 2014.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and Bates, P. D.: A high-accuracy map of global terrain elevations, Geophys. Res. Lett., 44, 5844–5853, https://doi.org/10.1002/2017GL072874, 2017.

https://doi.org/10.5194/essd-2026-402
Preprint. Discussion started: 3 August 2026

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: a high-resolution global hydrography map based on latest topography dataset, Water Resour. Res., 55, 5053–5073, https://doi.org/10.1029/2019WR024873, 2019.

Zhao, F., Xiong, L., Wang, H., Luo, L., Na, J., Li, S., and Tang, G.: Parallel algorithm of watershed delineation by considering the topological relationship into information organisation, Environ. Model. Softw., 194, 106706, https://doi.org/10.1016/j.envsoft.2025.106706, 2025.