AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE

10.1029/2021WR030054

Key Points:

- The SWOT River Database (SWORD) provides the foundation for SWOT river vector products including elevation, slope, width, and discharge
- SWORD combines multiple global river- and satellite-related data sets into a congruent product designed to integrate satellite observations
- SWORD contains 213,485 river reaches (~10 km in length) and 10.7 million nodes (~200 m spacing) with a consistent topological structure

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

E. H. Altenau,
ealtenau@unc.edu

Citation:

Altenau, E. H., Pavelsky, T. M., Durand, M. T., Yang, X., Frasson, R. P. d. M., & Bendezu, L. (2021). The surface water and ocean topography (SWOT) mission river database (SWORD): A global river network for satellite data products. *Water Resources Research*, *57*, e2021WR030054. https://doi.org/10.1029/2021WR030054

Received 24 MAR 2021
Accepted 23 JUN 2021

The Surface Water and Ocean Topography (SWOT) Mission River Database (SWORD): A Global River Network for Satellite Data Products

Elizabeth H. Altenau[1], Tamlin M. Pavelsky[1], Michael T. Durand[2], Xiao Yang[1], Renato Prata de Moraes Frasson[3], and Liam Bendezu[1]

[1]Department of Geological Sciences, University of North Carolina, Chapel Hill, NC, USA, [2]School of Earth Science and Byrd Polar and Climate Research Center, The Ohio State University, Columbus, OH, USA, [3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Abstract The upcoming Surface Water and Ocean Topography (SWOT) satellite mission, planned to launch in 2022, is the first mission to focus on measuring hydrological processes in Earth's surface water. As such, SWOT will vastly expand observations of global rivers ≥100 m wide. SWOT will provide a variety of data products, including a global vector river product containing water surface elevation (WSE), width, slope, and estimated discharge. Practical application and consistency of the SWOT vector products requires a prior global river network database divided into reaches. Here, we introduce the SWOT River Database (SWORD). SWORD will serve as the framework for the SWOT river vector products consisting of river reaches (~10 km long) and nodes (~200 m spacing). We generate SWORD by combining several global river- and satellite-related data sets into one congruent product. When defining river reaches, we incorporate natural and human-created river obstructions, basin boundaries, tributary junctions, and SWOT orbit track information. SWORD contains a total of 213,485 reaches and 10.7 million nodes. Globally, 77.3% of river reach lengths are between 10 and 20 km with a median reach length of 10.5 km. 95% of river reaches ≥10 km will have sufficient SWOT observations to provide discharge estimates at least once per orbit cycle. SWORD also contains many useful hydrologic and morphological attributes and is designed to be expandable in the future. Even before the launch of SWOT, it can serve as a framework for global hydrologic analyses using models, in situ measurements, and additional satellite observations.

Plain Language Summary The Surface Water and Ocean Topography (SWOT) satellite mission, planned to launch in 2022, is the first satellite with a specific aim to measure Earth's surface water fluctuations. SWOT will provide unprecedented observations of river water surface elevation, width, and slope. One product that will be provided is a vector-based data set designed for large regional-to-global scale analyses. Because rivers are dynamic features that change frequently, the vector product will allow scientists to analyze the data most effectively if the SWOT observations are attached to an existing database that is static in space and time. Here we introduce the SWOT River Database (SWORD), which will serve as the framework for the SWOT river vector products consisting of river reaches (~10 km long) and nodes (~200 m spacing within reaches). When defining river reaches, we consider natural and human-created boundaries as well as the boundaries of SWOT observation swaths. SWORD contains many useful hydrologic and morphological attributes, and it is designed to be expandable in the future. Even before the launch of SWOT, it can serve as a framework for modeling river flows at global scales and for conducting large-scale hydrologic analyses using ground measurements and/or additional satellite observations.

1. Introduction

Scheduled to launch in 2022, the Surface Water and Ocean Topography (SWOT) mission is the first satellite designed with a specific objective of observing earth's rivers (Biancamaria et al., 2016; Durand et al., 2010). During the 3 years that it will spend in its primary orbit, SWOT is designed to provide measurements of water surface elevation (WSE), width, and slope, along with estimates of discharge for global rivers wider than 100 m (Rodriguez et al., 2018). In addition to this science requirement, SWOT also has a science goal of providing these data for rivers as narrow as 50 m. Based on its orbit configuration, SWOT will observe

>95% of such rivers at least once and as many as 31 times every 21 days of the orbit cycle, with a median of 2 observations per cycle (Biancamaria et al., 2016). These data will allow useful measurements that can characterize river states at a wide range of relevant discharges (Nickles et al., 2019) and allow unprecedented remote sensing-based analysis of river morphology (Langhorst et al., 2019), streamflow (Durand et al., 2016), and flood behavior (Frasson, Schumann, et al., 2019).

In order to facilitate a wide range of new analyses with flexibility, the SWOT mission will provide a range of relevant data products. The most fundamental data product of broad use for river scientists is the so-called pixel cloud product (JPL Internal Document, 2020a), which provides measurements of inundation extent and WSE at relatively fine spatial resolution (10 m in the near swath range to 60 m in the far swath range) but with meter-scale elevation errors (Domeneghetti et al., 2018). This product will be useful for scientists who want to study a particular region using bespoke data analysis methods, but it may be unwieldy to work with at scales larger than regional or when conducting multitemporal analysis. More information regarding the SWOT pixel cloud product and its uncertainties can be found in Biancamaria et al. (2016) and Rodriguez et al. (2018). The SWOT mission will also provide a raster product, nominally at 100 and 250 m resolutions, that will aggregate the pixel cloud data onto a regularized grid. It will contain information about WSE and extent, but it will not include calculations of river slope, river discharge, or any characteristics of river network topology. As such, the SWOT mission will also provide river vector products stored in shapefile format for each SWOT overpass (JPL Internal Document, 2020b). These vector products will include aggregations of SWOT WSE and inundation extent approximately every 200 m along rivers (known as nodes) and also aggregations of WSE, inundation extent, slope, and estimated discharge along reaches approximately 10 km in length.

The SWOT vector data products will be most broadly useful if they allow multitemporal analysis of river nodes and reaches covering the same river areas. Doing so requires defining SWOT reaches and nodes a priori, so that SWOT data can be assigned to them. We recognize that such fixed a priori data sets will not be ideal for all analyses, such as in rivers that migrate rapidly or during major floods. We anticipate that the science community will start from the pixel cloud product to develop custom solutions for these more localized problems. In order to facilitate development of consistent global data products, however, we need a suitable global database of predefined river nodes and reaches. Existing global river databases such as HydroSHEDS (Lehner et al., 2008) and MERIT Hydro (Yamazaki et al., 2019) are primarily based on digital elevation models (DEMs). Though they provide very useful information about river characteristics such as drainage area and elevation, they do not always reflect the complex structure of river networks. They are also at a relatively coarse resolution (3 arcseconds, ~90 m at the equator) compared to the resolution of SWOT nodes. In contrast, global river representations based on satellite imagery such as the Global River Widths from Landsat (GRWL) database (Allen & Pavelsky, 2018) reflect river complexity at somewhat higher resolution (30 m) but without the necessary information on elevation and drainage area required to construct network topology. As such, neither a DEM-based nor a satellite-based product alone is sufficient to allow optimal construction of an a priori SWOT River Database.

Here, we describe development of the SWOT River Database (SWORD) (Altenau et al., 2021), which combines multiple global river- and satellite-related data sets to define the nodes and reaches that will constitute SWOT river vector data products. While SWORD is designed primarily for SWOT, it can also serve as a framework for other satellite-based river measurements, such as WSEs from nadir altimeters (Biancamaria et al., 2018) and river widths measured from optical and radar satellite images (e.g., Ishitsuka et al., 2020). Because it contains a great deal of information on river topology and network structure, it may also be of use to scientists seeking to analyze global river networks using numerical models. Additionally, once SWOT launches and the observations are attached to SWORD, scientists will be able to study regional to global scale river processes including flooding dynamics, water balance issues, and carbon cycle variations in unprecedented ways (Biancamaria et al., 2016). In the remainder of this paper, we describe the methods we use to develop SWORD (Section 2), the characteristics of the resulting data set (Section 3), and the implications for SWOT and global river analysis more broadly (Section 4).

Table 1
Summary of Data Sets Used in the Development of SWORD

Data set	Attribute contribution
Global River Widths from Landsat (GRWL) (Allen & Pavelsky, 2018)	Provides river centerline locations at 30-m resolution and associated width, water body type, and number of channels attributes.
MERIT Hydro (Yamazaki et al., 2019)	Provides elevation and flow accumulation at 3 arc-second resolution (~90 m at the equator).
HydroBASINS (Lehner & Grill, 2013)	Provides Pfafstetter nested basin codes up to level 6.
Global River Obstruction Database (GROD) (Whittemore et al., 2020)	Provides global locations of anthropogenic river obstructions along the GRWL river network.
Global Delta Maps (Tessler et al., 2015)	Provides the spatial extent of 48 of the world's largest river deltas.
SWOT Orbits (https://www.aviso.altimetry.fr/en/missions/future-missions/swot/orbit.html)	Provides polygons containing SWOT track coverage for each pass throughout the 21 days of cycle orbit.
HydroFALLS (http://wp.geog.mcgill.ca/hydrolab/hydrofalls/)	Provides global locations of waterfalls and natural river obstructions.

2. Methods

2.1. Input Data Sources

We generate SWORD by combining several global hydrography databases into one congruent product. This section describes the input data sources that we use in the development of SWORD. Table 1 provides a summary of the data sets and the attributes they contribute to the final product.

2.1.1. Global River Widths From Landsat (GRWL)

The primary data source for SWORD is the Global River Widths from Landsat (GRWL) database (Allen & Pavelsky, 2018). GRWL provides high-resolution (30 m) centerline locations for global rivers of 30 m wide and greater. It also contains corresponding channel width, number of channels, and water body type attributes associated with each centerline location. GRWL was developed by processing Landsat imagery at approximately mean annual flow, creating and cleaning the river masks, and generating centerlines along the final river masks. At each centerline point, the cross-sectional width and number of observed channels are provided. Water body type information is also attached to each centerline pixel, designating whether it is a river, lake/reservoir, canal, or tidally influenced river (Allen & Pavelsky, 2018). We use GRWL centerlines as the primary spatial data set for river centerline locations in SWORD. The GRWL centerline locations are preferred for the river locations over DEM-derived centerlines since they are derived from optical imagery and are more likely to agree with the river locations that SWOT will observe. In addition, DEM-derived centerlines such as those from HydroSHEDS (Lehner et al., 2008) and MERIT-Hydro (Yamazaki et al., 2019) do not always capture the complexity of river networks in places where topology is difficult (e.g., deltas, highly anabranching systems, etc.). River centerlines derived from DEMs are also subject to DEM noise and error in the elevations, which can result in centerlines that are offset from the true position of the river. DEM derived centerlines are also tricky to adjust. Burning in new centerlines is non-trivial and altering network flowlines after DEM-extraction will lose the coupling with the DEM. Additionally, it is inevitable that some rivers will migrate substantially over time requiring updates to the river centerlines, which cannot be done using DEM-derived methods alone since DEMs are often not multitemporal. The long temporal record and ongoing Landsat missions allow for efficient updates to these centerlines compared to DEM-derived centerline products. Finally, once SWOT launches, SWOT data may also be useful for updating river centerline locations.

2.1.2. MERIT Hydro

MERIT Hydro is a global hydrography database built using the MERIT DEM and several other inland water maps (G1WBM, GSWO, and OpenStreetMap) (Yamazaki et al., 2019). It contains high-resolution (3 arc-second) raster maps of flow direction, flow accumulation, hydrologically adjusted elevations, and river channel width. Elevation and flow accumulation are the primary products from MERIT Hydro used to develop the SWORD database. We use these attributes to help build the global topology structure in SWORD (Section 2.2.4) and provide a priori elevation values for the database.

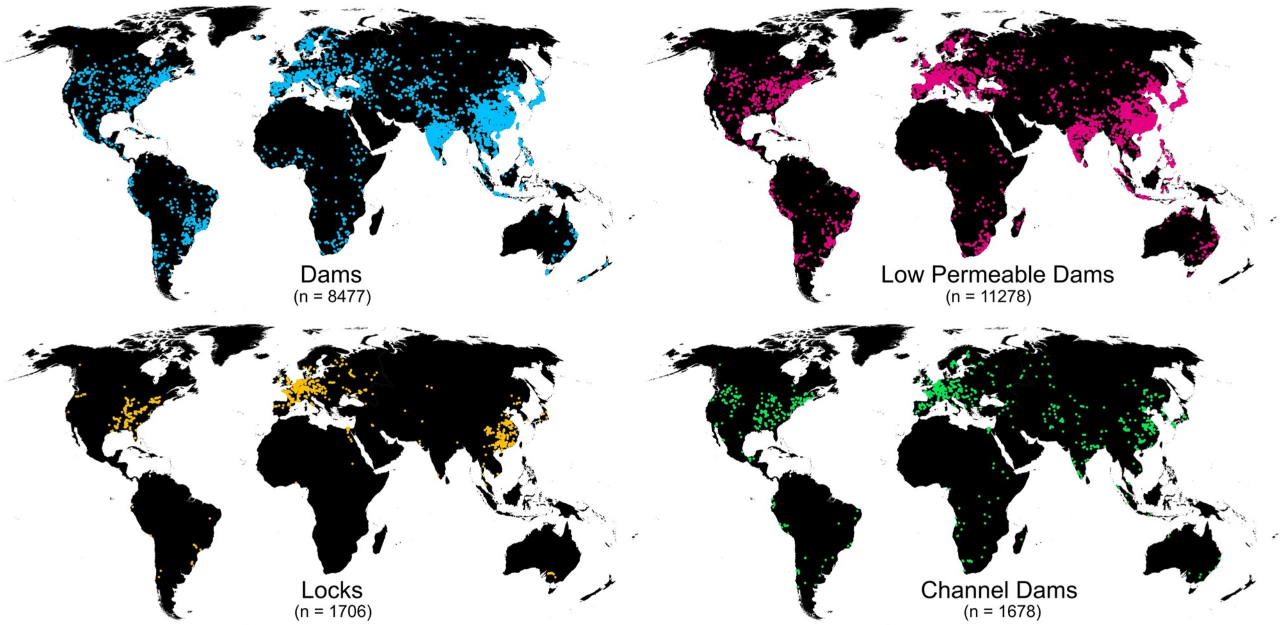

Figure 1. Global River Obstruction Database (GROD) features used in SWORD.

2.1.3. HydroBASINS

The HydroBASINS data product contains a series of watershed boundaries and sub-basin delineations at a global scale in polygon shapefile format (Lehner & Grill, 2013). A Pfafstetter coding system (Verdin & Verdin, 1999) is used to guide the basin nesting and topological organization. We use HydroBASINS's Pfafstetter basin system to provide the foundation for the SWORD river reach/node id structure and topology (Section 2.2.4). The HydroBASINS data set is globally consistent and commonly used throughout the hydrologic community, making it a suitable choice for building SWORD's topological structure.

2.1.4. Global River Obstruction Database (GROD)

The Global River Obstruction Database (GROD) is a new data set that includes all human-created river obstructions along the GRWL channel network (Whittemore et al., 2020). Obstructions were manually identified and categorized using a Google Earth Engine (GEE) application (Gorelick et al., 2017). GROD includes ∼30,000 obstructions that fall into six different categories: 1—Dams, 2—Channel Dams (dam that is only on one channel out of multiple channels), 3—Locks, 4—Low Permeable Dams, 5—Partial Dams ≥ 50 (dam that covers greater than or equal to 50% of the river channel), 6—Partial Dams < 50 (dam that covers less than 50% of the river channel) (Figure 1). We only include obstruction categories 1–4 in the SWORD database because they entirely cross a given river channel and are thus most likely to affect SWOT-observable river hydraulics. These categories account for 78% of GROD features. We use GROD river obstruction locations during the SWORD reach definition process (Section 2.2.3).

2.1.5. Global Delta Map

Tessler et al. (2015) built shapefiles defining delta extents for 48 of the world's largest river delta systems to help quantify changing flood risk in these vulnerable and dynamic environments. We use these delta maps to identify locations of river reaches in the SWORD database where it is difficult to establish proper topology (i.e., distinguishing flow direction).

2.1.6. HydroFALLS

HydroFALLS is a global database of validated waterfall points with quality ratings (http://wp.geog.mcgill.ca/hydrolab/hydrofalls/). This product is the first of its kind in terms of scale and spatial coverage at the global scale. The database was created by systematically merging, consolidating, and validating existing waterfall data sets to generate a waterfall point layer. Like GROD, we use the HydroFALLS features during

Figure 2. Example of GRWL centerline corrections along the Amazon River. Purple lines represent the GRWL centerlines before corrections, and orange lines represent the centerline updates. Upper right inset shows the selected location in South America.

the reach definition process. In addition to the human-created obstructions from GROD, we include natural "breaks" in river profiles such as waterfalls when creating river reaches. We want to avoid river reaches crossing dams or waterfalls because they create abrupt changes in elevation within a reach (Section 2.2.3). These abrupt changes in elevation and slope hinder the ability to estimate discharge along a reach (Frasson et al., 2017).

2.2. Database Development

2.2.1. GRWL Updates

Before combining GRWL with the other data sets, improvements to GRWL's centerline connectivity were needed because it contains a range of discontinuities associated with the way it was generated. To correct these discontinuities, we use a dedicated Google Earth Engine application to manually identify points along the GRWL centerlines that require connection. Pairs of points that need to be connected to each other are exported with latitude and longitude locations in csv file format. After we identify the points, we run the point locations through an algorithm implemented in the Interactive Data Language (IDL) that automatically connects these pairs of points at 30 m spatial resolution to match the existing GRWL centerlines. The algorithm connects the points linearly or based on the shortest path along the GRWL river masks between the two points. We assign a default value of one to the width and number of channels attributes at the corrected locations in the database. Most of the GRWL corrections occur at channel junctions or in complex river environments where width and number of channels are difficult to calculate automatically. Overall, we identified and fixed >7,000 discontinuities in the GRWL centerline network. These corrections improve continuity in the river network (Figure 2), which is essential for defining reaches and topology in SWORD, as well as estimating discharge across large scales.

2.2.2. Merging Databases

Once we update the GRWL centerlines, we attach the other data set attributes to the GRWL river network. First, we merge the MERIT Hydro elevation and flow accumulation values onto the GRWL centerlines. To do this, we need to identify the river pixel locations in MERIT Hydro that are associated with the GRWL centerline network. Since MERIT Hydro is derived from a DEM, the river network is much more extensive than the GRWL river network, which only contains rivers ≥30 m wide. If we include the much smaller headwater streams from MERIT Hydro, the smaller channels could interfere in the merging process and

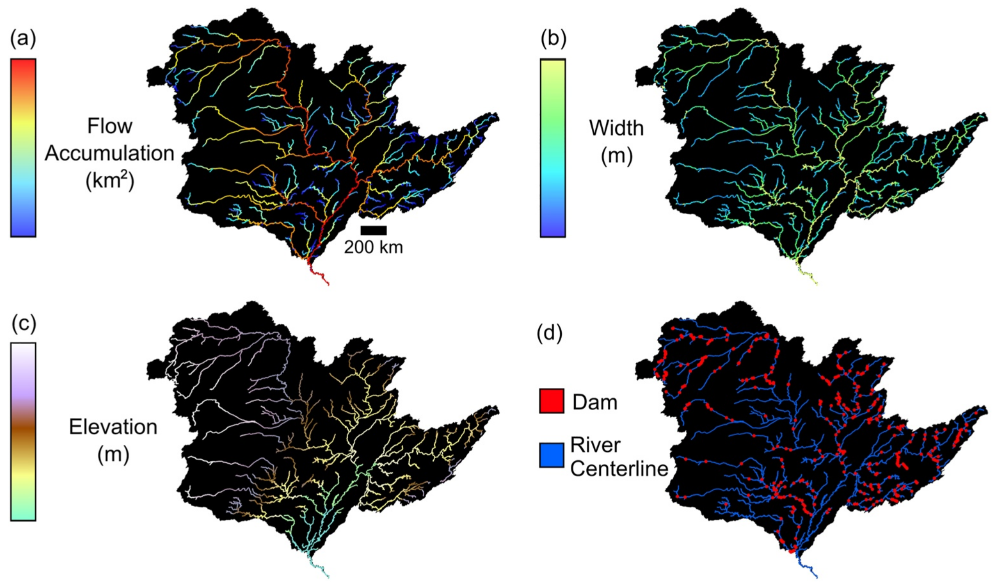

Figure 3. MERIT Hydro flow accumulation (a) and elevation (c), GRWL widths (b), and GROD obstruction locations (d) along the GRWL centerlines in the Mississippi Basin after the database merge. GROD features represent four types of obstructions: dams, channel dams, locks, and low permeable dams.

make it difficult to determine which river tributaries are associated with the GRWL rivers. We identify the river centerline locations from the MERIT Hydro rasters by setting a threshold for flow accumulation and extracting the remaining pixel locations as the river centerlines. We use a flow accumulation threshold of 10 km^2 because the river network extent covers the GRWL river network while excluding much smaller streams that would introduce noise during the merging process (Figure S2). After we isolate the MERIT Hydro centerlines, we use a spatial k-d tree (Maneewongvatana & Mount, 1999) to find the 10 closest MERIT Hydro pixels to each GRWL centerline point. Then, we take the median value for elevation and flow accumulation across the 10 closest MERIT Hydro pixels. Given MERIT Hydro was trained on the GRWL centerlines (Yamazaki et al., 2019), we deem a pixel-based mapping scheme to be sufficient for mapping MERIT Hydro elevation and flow accumulation values to GRWL in the majority of cases. Cases where merging the values between the two databases produce more errors are primarily in large, anabranching river environments where the centerlines do not match up as well (Figure S3). To mitigate flow accumulation artifacts in these areas, we filter the flow accumulation values along each GRWL river segment to remove outliers (defined as ± one standard deviation away from the median flow accumulation value along a GRWL river segment) and to ensure flow accumulation increases in the downstream direction (Figure S2). We primarily use the flow accumulations to help define SWORD's global topology scheme (Section 2.2.4) and therefore prioritize maintaining the general magnitude and direction of flow accumulation over preserving the exact values. Since we apply a filter to the flow accumulation values, users are advised to treat flow accumulation as an estimate, and general topology tool, not as a precise value.

Once we attach the MERIT Hydro attributes to the GRWL network, we use a nearest neighbor search with a distance threshold of 500 m to assign the GROD and HydroFALLS obstruction locations to the GRWL centerlines. A simple nearest neighbor search is sufficient to match the GROD features to the appropriate GRWL centerline points because GROD was built based on the GRWL network. Additionally, upon visual inspection, the relevant HydroFALLS data tends to fall within 500 m of the GRWL centerlines. Finally, we use geopanda's spatial join feature (https://geopandas.org/docs/user_guide/mergingdata.html#spatial-joins) to merge the GRWL centerline points with the HydroBASINS, global deltas, and SWOT track polygons. Figure 3 shows example attributes along the GRWL centerlines for the Mississippi Basin after the full database merge process is complete.

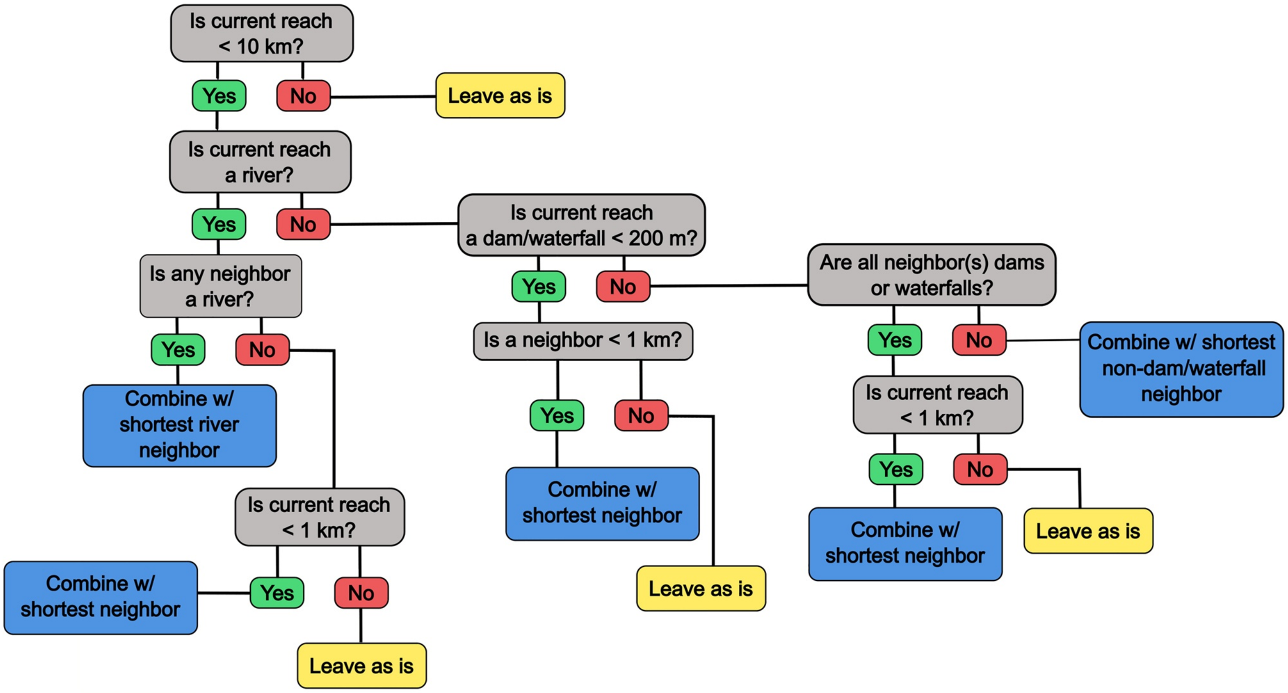

Figure 4. Flow chart demonstrating the general criteria used to aggregate short reaches.

2.2.3. Reach and Node Definition

The goal for reach definition in SWORD is to maximize SWOT observation coverage along a reach, while respecting natural and anthropogenic hydrologic boundaries and barriers. A ~10 km reach length is required by the SWOT mission to obtain the desired elevation, slope, and discharge accuracies (Biancamaria et al., 2016; Rodriguez et al., 2020). Given that natural and anthropogenic barriers impact hydraulics, it is impossible to ensure an exact reach length of 10 km everywhere. Therefore, we aim to incorporate these hydraulic boundaries while preserving an average reach length as close to 10 km as possible. In some cases, this process may result in slightly longer or shorter reaches where elevation and slope accuracies may be slightly higher (for a longer reach) or lower (for a shorter one).

To define reaches, we first divide the GRWL centerlines at level 6 Pfafstetter basin boundaries, tributary junctions, dams, waterfalls, lakes, and SWOT swath boundaries. After the initial divisions, we proceed through several steps to obtain an average reach length of ~10 km. One exception to the desired 10 km reach length is dam/waterfall reaches. Dam and waterfall reaches are primarily needed to prevent surrounding reaches from including large elevation variations that would affect both elevation and slope accuracies. SWOT's pixel cloud product, which gets mapped onto SWORD vectors, includes all observed elevations in 10–60 m pixels. In order to avoid mixing pixels across a waterfall or dam into a river reach, we create dam reaches for the obstruction pixels to be averaged along. We determine a ~400-m reach length, or approximately two nodes, one above and one below the obstruction, to be sufficient for dam and waterfall reaches in order to absorb the elevation differences around the obstruction, without significantly impacting surrounding reach lengths.

Once the centerlines are cut into initial reaches, we divide the remaining reaches >20 km into equal intervals with lengths of ~10 km. To do this, we divide the reach length by 10 km and round the resulting number to get the number of divisions for the reach. For example, if a reach is 20 km in length, we divide the reach in half to produce two 10 km reaches. If a reach is 25 km, we divide the reach into three 8.3 km reaches and so on. Next, we use a set of criteria to aggregate small reaches <10 km with neighboring reaches (Figure 4). The criteria are applied iteratively until all short reaches that can be aggregated are at least 10 km long.

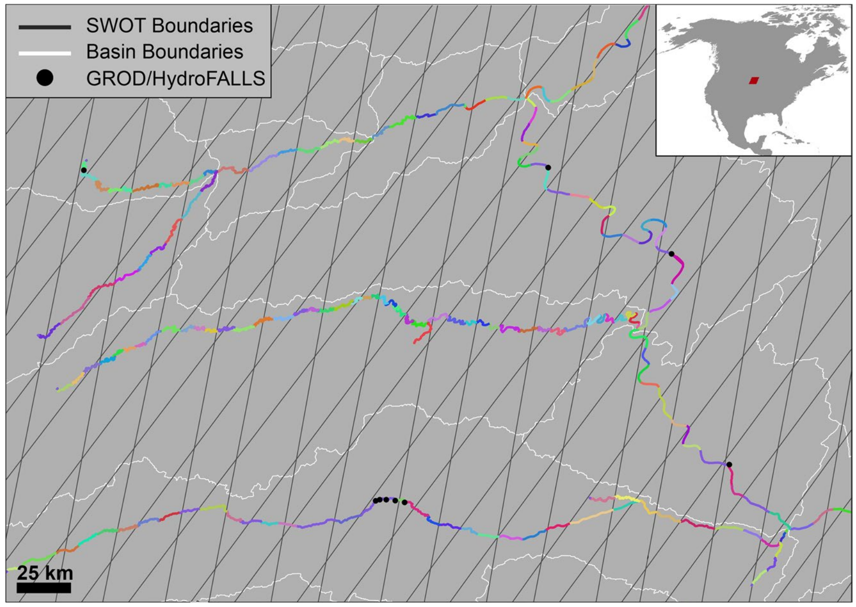

Figure 5. Example reaches in the upper Mississippi Basin. Upper right inset shows the selected reaches' location within North America. Colored lines depict different reaches, black dots show GROD and HydroFALLS features, black lines represent SWOT orbit boundaries, and white lines indicate Pfafstetter level 6 basin boundaries.

Boundaries that we never aggregate across include basin boundaries and tributary junctions. Furthermore, we only combine reaches identified as rivers that fall between two non-river reaches if they are <1 km in length. Additionally, we only aggregate reaches identified as dams or waterfalls if they are <200 m in length. While dam/waterfall reaches are designed to be 400 m, there are rare cases during the initial segmentation of reaches that result in dam/waterfall reaches being too short to represent the structure at the node resolution (200 m). Therefore, these cases need to be aggregated. After aggregating the reaches, we again divide any resulting reaches that are >20 km into equal intervals of ~10 km.

Finally, we define ~200 m "ghost" reaches at the headwaters and outlets of every river network. Ghost reaches are needed for SWOT processing algorithms to ensure that SWOT observations beyond the beginning and end of river systems (i.e., in the ocean or large lake into which a river flows) are not included in calculating SWOT river data products. These reaches will not be processed by SWOT algorithms and will not appear in the final SWOT vector products effectively making them "dummy" reaches. The ghost reaches are ~200 m long (i.e., one node length) in order to preserve the reach lengths of the remaining SWORD reaches. Figure 5 shows an example of the reach definitions for a section of the upper Mississippi Basin.

When the final reach boundaries are complete, we divide each reach into nodes (i.e., points) spaced ~200 m apart. Nodes are included as part of the SWOT vector products to provide a finer-resolution product in addition to the reaches. The SWOT science team determined that a node length of 200 m was a good balance between preserving SWOT elevation and geolocation accuracies while maintaining a finer spatial resolution for representing river processes. To define nodes, we calculate flow distance along the high resolution GRWL centerlines for each reach. Next, we divide the maximum flow distance (i.e., reach length) by the desired node length (200 m) to get the number of nodes for a reach. Once the number of nodes is determined, we use this number to find break points along the high-resolution centerlines based on the flow distance and node length. Finally, we use the median location value of the high-resolution centerlines between each 200 m break point to define the node locations. This results in the nodes spaced ~200 m apart within a reach while starting and ending ~100 m from the reach boundaries.

2.2.4. Topology Structure (Reach and Node IDs)

The SWORD reach and node ids make up the foundation of the database topology. Therefore, the id structure needs to be globally consistent and have an intuitive method for ordering and identifying rivers. We built the foundation of the reach and node ids based on the HydroBASINS nested-basin system because it is

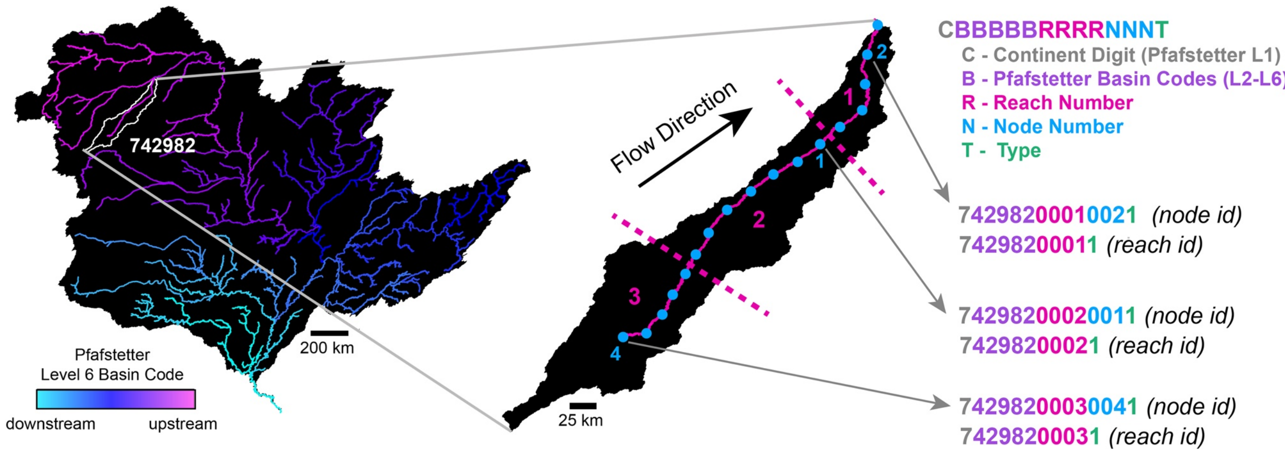

Figure 6. Schematic of reach and node ordering inside a Pfafstetter level 6 basin. Note: reaches and nodes are not shown to scale within the level 6 basin, in order to better illustrate the identification scheme.

global and widely known throughout the hydrologic community (Section 2.1.3). After the basin codes, we attach reach and node numbers and an identifier which indicates the type of reach or node (i.e., river vs. lake/reservoir). The general id structure is an 11-digit code of the format CBBBBBRRRRT for the reach ids and a 14-digit code CBBBBBRRRRNNNT for the node ids. C stands for continent, B for basin, R for reach, N for node, and T for type. The first 6 digits of the id (CBBBBB) are the HydroBASINS Pfafstetter level 6 basin code, where the first digit represents one of nine continental regions (1 = Africa, 2 = Europe, 3 = Siberia, 4 = Asia, 5 = Oceania, 6 = South America, 7 = North America, 8 = Arctic, 9 = Greenland), and the remaining digits are nested basin levels 2–6 (Figure S1). Users are encouraged to refer to the HydroBASINS documentation for further details on the Pfafstetter basin system (Lehner & Grill, 2013).

Following the basin code is the reach number of four digits (RRRR) and the node number of three digits (NNN). The reach id does not contain the node numbers. We determine the order of the reach and node numbers (RRRR/NNN) within a level 6 basin by the flow accumulation and elevation values. Within each level 6 basin, $R = 0001$ at the downstream end of the basin and increases in the upstream direction. Similarly, within each reach, $N = 001$ at the downstream end of the reach and increases in the upstream direction (Figure 6).

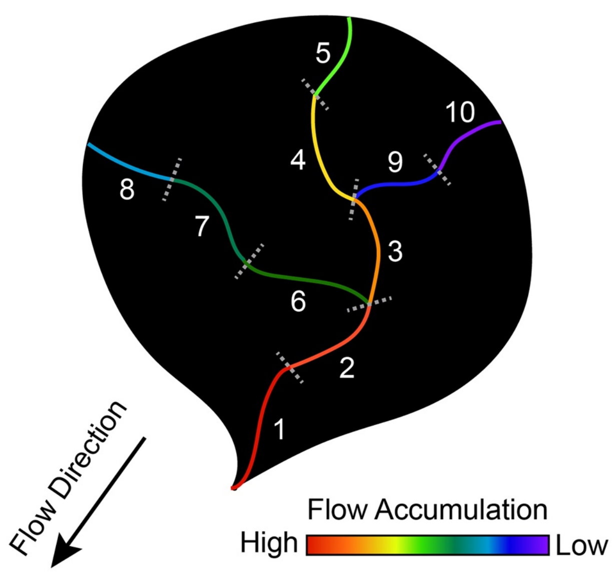

Figure 7. Schematic of reach numbering when there are multiple tributaries inside a level 6 basin. The black area represents a Pfafstetter level 6 drainage basin while the colored lines display different reaches along the river centerlines.

If there are multiple tributaries within a level 6 basin, the reach numbers start at the downstream end of the largest channel (determined by flow accumulation) and increase until the upstream end of that channel. Once all reaches for that channel are determined, the reach numbers continue at the downstream end of the next largest tributary and so forth (Figure 7). Given there are occasional artifacts in the flow accumulation and elevation attributes, reach numbers may not always increase sequentially upstream along a tributary in basins with multiple tributaries. This is more common in complex river environments such as large anabranching rivers or coastal areas. In these cases, users can determine the local topology using the neighboring reach attributes ("rch_id_up"; "rch_id_dn") (Table S2). The "rch_id_up" and "rch_id_dn" attributes catalog the upstream and downstream neighbors for each reach in the SWORD database. After the topology is defined, we use the information to calculate the distance from outlet ("dist_out") for every reach and node. The "dist_out" attribute is the upstream distance (in meters) from the river system outlets for every reach and node. We determine the river outlets using the flow accumulation values then calculate "dist_out" based on topology and neighboring reach and node lengths.

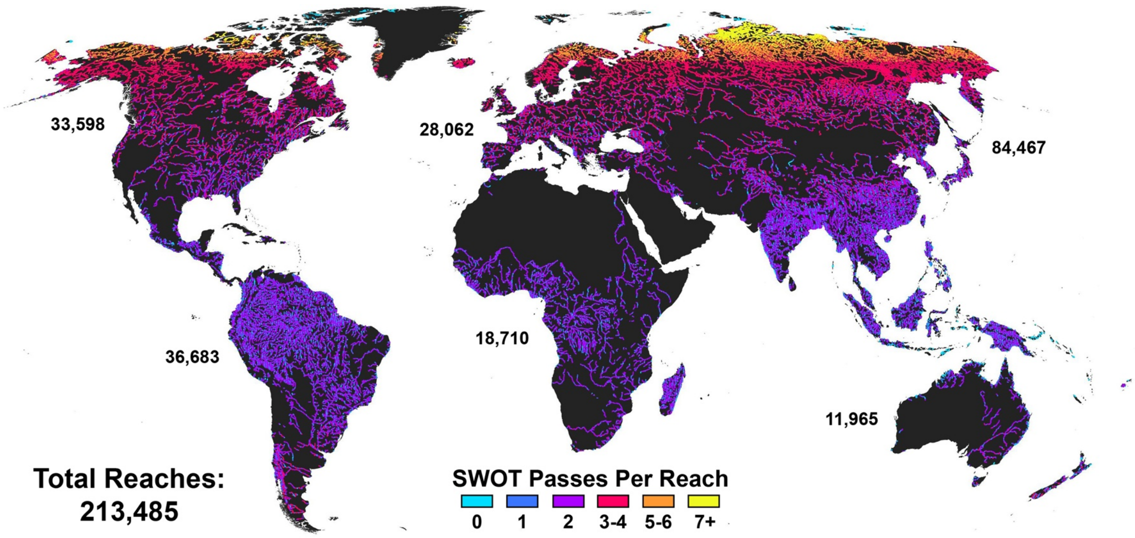

Figure 8. Number of SWORD reaches per continent (not including ghost reaches). Colors display the number of SWOT passes per reach during the 21 day orbit cycle.

Finally, the last digit in the id is a type identifier (T) where 1 = river, 2 = lake off river, 3 = lake on river (such as a reservoir), 4 = dam/waterfall, 5 = unreliable topology, and 6 = ghost reach/node. The type 2 category will be used in a SWOT prior lake database analogous to SWORD; there are no type 2 reaches in the SWORD database because, by definition, it does not include lakes disconnected from the river network it represents. In the future, this type identifier will be important when integrating the SWOT lake and river databases. "Unreliable topology" reaches (type = 5) are located in complex, coastal areas where it is very difficult to determine upstream from downstream flow directions. We identify the "unreliable topology" reaches by using the tidal water body information from the GRWL database and the global delta maps extent. After we define all reaches and nodes and assign them a topological id, we calculate several attributes for each reach and node, which are detailed in Section 3.

3. Results

SWORD contains a total of 213,485 reaches (Figure 8) and 10.7 million nodes, excluding ghost reach and node types. Ghost reaches and nodes will not be included in the official SWOT vector products, and users are encouraged to exclude these reach and node types from their analyses. As such, we do not include ghost reaches and nodes in any of our reported statistics (see Section 2.2.3 for ghost node/reach description).

Overall, SWORD reach types are primarily rivers (73.6% globally, Figure 9). Dam/waterfall reach types are the second most prevalent reach type at 10.7% globally, though they account for just 0.5% of total river length. This is due to the short length of the dam and waterfall reaches at ~400 m on average. Median reach lengths per continent range from 9.7 (Europe) to 10.7 km (Africa). If we isolate the reaches to river reach types, the range in median reach lengths increases slightly from 10.7 (Europe) to 11.1 km (Asia). Table 2 presents reach length statistics per continent for all reach types (excluding ghost reaches), as well as river reach types only. Globally, 64.1% of all reach lengths and 77.3% of river reach lengths are between 10 and 20 km, with a median reach length of 10.5 km.

In addition to defining the nodes and reaches that will establish the foundation of the SWOT river vector products, SWORD includes other important hydrologic and morphological attributes for each reach and node including elevation, width, slope, and number of channels (Figure 10). We generate these attributes based on the merged data set values along the high-resolution river centerlines (Section 2.2.2). We calculate SWORD slopes using a linear regression along the MERIT Hydro elevations for each reach. Globally, 91% of reach slopes are <500 cm/km with a median slope value of 31 cm/km. SWORD's median slope value is higher than that of Frasson, Pavelsky, et al. (2019) who report a median global slope of 19 cm/km for

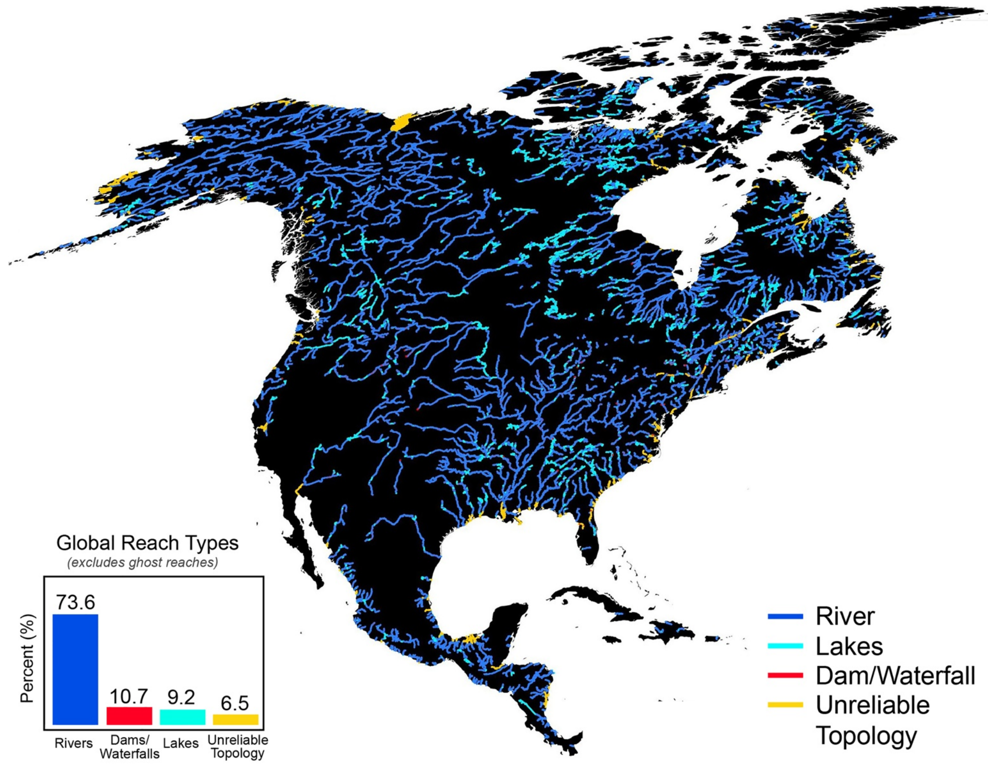

Figure 9. North America reach types. Inset displays global reach type percentages, excluding ghost reaches. River obstruction reaches (dams or waterfalls) are depicted in red but difficult to see given their short reach lengths of ∼400 m.

reaches of ∼10 km in length. The discrepancy in median slope values likely stems from the difference in the data included when calculating slope statistics. Frasson, Pavelsky, et al. (2019) exclude reaches with slopes >300 cm/km, widths <90 m, and their data set only includes rivers below 60°N, while here we include all of SWORD's reach data. We calculate the other attributes (elevation, width, etc.) by taking the median, min, max, or mode of the merged data set values for each reach and node location. More description on SWORD's data distribution formats (Text S1 and Table S1) and attributes (Table S2) are detailed in the supplementary information.

An important science objective for the SWOT mission is to provide a global discharge vector product that will be attached to the SWORD river reaches. In order to estimate discharge, a reach needs to have at least 50% SWOT observation spatial coverage for any given pass. Therefore, we aim to preserve as much SWOT

Table 2
Reach Lengths Per Continent (Excluding Ghost Reaches)

Reach length (L)	NA	SA	AS	EU	AF	OC	Global
$L < 5$ km	24.8% (13.6%)	11.5% (5.6%)	23.7% (9.8%)	37.9% (18.6%)	11.3% (5.0%)	11.4% (4.8%)	21.9% (9.8%)
5 km $\le L <$ 10 km	15.5% (12.2%)	14.1% (11.9%)	13.28% (13.6%)	13.6% (14.9%)	12.3% (10.6%)	18.6% (13.3%)	14.0% (12.9%)
10 km $\le L \le$ 20 km	59.8% (74.1%)	74.3% (82.4%)	63.0% (76.7%)	48.4% (66.6%)	76.4% (84.4%)	69.9% (81.9%)	64.1% (77.3%)
$L >$ 20 km	0.01% (0.02%)	0.02% (0.01%)	0% (0%)	0.04% (0.06%)	0.02% (0.01%)	0.01% (0.01%)	0.01% (0.01%)
Mean	9.7 km (11.4 km)	11.3 km (12.2 km)	9.9 km (11.7 km)	8.1 km (10.7 km)	11.5 km (12.4 km)	11.1 km (12.2 km)	10.1 km (11.7 km)
Median	10.4 km (11.0 km)	10.6 km (10.9 km)	10.5 km (11.1 km)	9.7 km (10.7 km)	10.7 km (10.9 km)	10.5 km (10.9 km)	10.5 km (10.9 km)

Notes. Percentages are calculated based on total reach numbers. Numbers in parentheses are calculated for reaches with river type identifiers only.

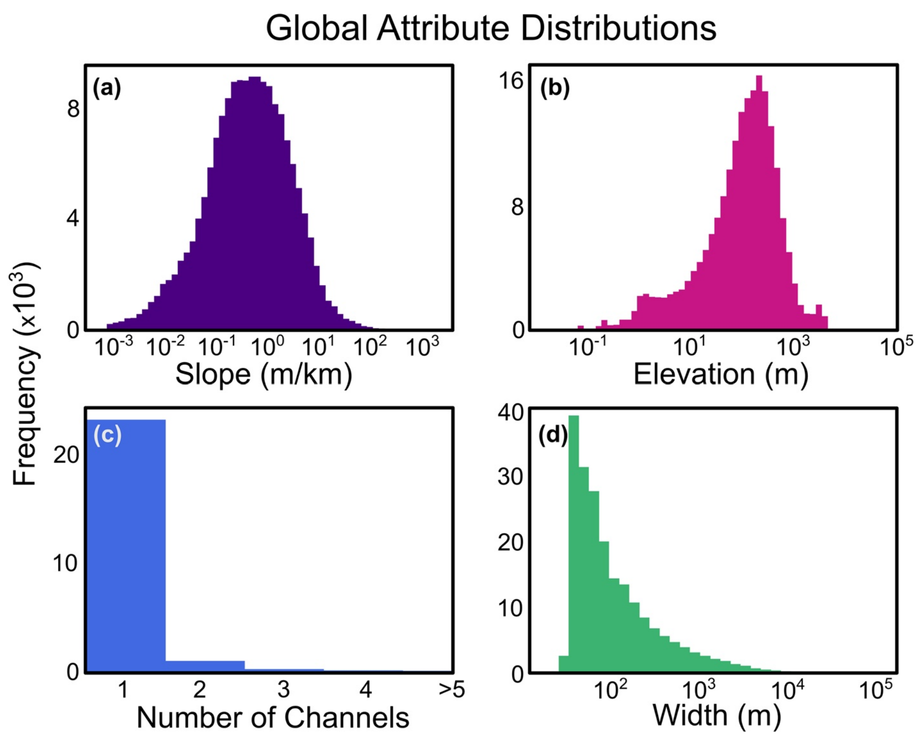

Figure 10. Global distributions of (a) reach slope, (b) reach elevation, (c) number of channels per reach, and (d) reach width. The *x*-axis for slope, elevation, and width are shown in log scale.

coverage per reach as possible while balancing other hydrologic considerations in the reach definition procedure (Section 2.2.3). To assess SWOT coverage in SWORD, we calculate the spatial coverage extent for every SWOT pass along each reach. For river type reaches ≥10 km in length, 95% have at least one SWOT pass that covers ≥50% of the reach during the 21 days of orbit cycle. Therefore, under a zero SWOT error scenario, 95% of SWORD reaches ≥10 km have sufficient SWOT coverage to provide discharge estimates at least once per orbit cycle. Furthermore, many reaches will likely have sufficient SWOT coverage to provide discharge estimates multiple times per orbit cycle given the median number of SWOT passes per reach is 2 globally with many high latitude reaches receiving >4 SWOT passes per cycle (Figure 9).

Currently, there are minimal manual adjustments made to SWORD, which results in the database containing some topological errors due to artifacts in the flow accumulation and elevation values that occur during the merging process. As mentioned in Section 2.2.2, we apply a filter to automatically reduce these errors (Figure S2) as much as possible; however, the remaining artifacts can translate into incorrect topology definitions (Figure S3). To estimate how frequently these errors occur, we calculate the inconsistencies present in the upstream and downstream reach attributes. More specifically, we look at the upstream and downstream neighbors for each reach and see if the current reach is correctly identified by its neighbors. Based on these calculations, we estimate that <2% of reaches have topological inconsistencies. While this calculation does not necessarily capture all types of topological inaccuracies, it gives us a good indication of how prevalent and where these errors occur. Additionally, the "dist_out" attribute (Section 2.2.4), which defines the distance from the river outlet for every reach and node, is highly sensitive to topology. Therefore, large discrepancies in the "dist_out" variable can result from minor errors in the topological structure. We applied a filter to reduce these errors in the "dist_out" attribute, but users are advised to use "dist_out" as an estimate and not a precise value. In future SWORD versions released closer to the launch of SWOT in 2022, we plan to include manual adjustments where automatic methods fail, which will address the remaining topological inconsistencies. Additionally, we plan to use future SWOT elevations to improve errors in SWORD after launch.

4. Discussion

As described above, SWORD is unique relative to other global river databases because it provides consistent network topology and combines a unique set of attributes that allows sensible global reach definitions. The GRWL data set (Allen & Pavelsky, 2018), which provides the basis for the river network presented here, provides important information on river location, width, and number of observed channels based on observations from Landsat imagery. However, it also requires input from the primarily DEM-based MERIT Hydro (Yamazaki et al., 2019), which provides information on channel elevation and flow accumulation. This combination represents a considerable upgrade on the conceptually similar data set analyzed by Frasson, Pavelsky, et al. (2019), which, for example, primarily included rivers south of 60 N. In addition, SWORD includes, for the first time in such a data set, comprehensive databases of anthropogenic river obstructions (Whittemore et al., 2020) and waterfalls (in the form of HydroFALLS), which are natural breaking points for reaches. Finally, SWORD is adapted for SWOT specifically by including the locations of swath boundaries. With these tools, we have created a global river reach (and node) database that reflects both hydrologic priorities (i.e., reaches that respect tributary junctions, dams, etc.) and satellite-specific priorities (i.e., optimizing for reaches that will be completely observed in one SWOT pass).

As such, SWORD is essential for development of SWOT vector data products. Only in these data products, is it possible to calculate river surface slopes, estimate river discharges, and conduct efficient multitemporal analyses using SWOT data. We anticipate that these products will be essential for future river science at large regional to global scales that would be difficult to pursue using other SWOT data products. These scientific opportunities include assimilation of SWOT data into hydrologic and hydrodynamic models (Biancamaria et al., 2011; Häfliger et al., 2019; Li et al., 2020; Wongchuig-Correa et al., 2020), studies of flood events (Frasson, Schumann, et al., 2019), improved understanding of river morphology (Langhorst et al., 2019; Larnier et al., 2019; Tourian et al., 2017; Yoon et al., 2012), and facilitate applications such as reservoir management (Munier et al., 2015).

Along with these opportunities, the future applicability of SWORD requires ongoing work. Because global river networks are complex, there are some remaining cases in which topology is ambiguous or poorly defined. Some of these cases may require manual editing to correct, and we will conduct this work closer to the launch of SWOT. Moreover, after SWOT launches we plan to update SWORD to reflect features detected in actual SWOT observations, including detection of currently unobserved discontinuities (e.g., dams, waterfalls, rapids), improved understanding of optimal reach definition, etc. (Frasson et al., 2017; Samine Montazem et al., 2019). Finally, the state of SWOT-observed rivers depends on inputs from smaller tributaries, and improved representation of these tributaries could allow for easier incorporation of knowledge about lateral inflows important to SWOT discharge algorithms (Nickles et al., 2020) and upstream reservoir management. In a system as complex as the global river network, ongoing updates will be important, and we plan to offer continued improvements at least through the lifetime of the SWOT mission.

Beyond SWOT, SWORD also provides a potentially very useful framework for integrating a broad array of satellite observations of rivers. Because it is based on GRWL, it already contains information on mean river widths, but it could easily serve as a framework for storing and using multitemporal width data sets such as those recently analyzed by Allen et al. (2020) and Ishitsuka et al. (2020). In addition, recent advances in nadir altimetry processing methods and sensor technology are making the broad-scale observation of river WSE a reality (Biancamaria et al., 2018). Integration of these data into the SWORD framework would also make sense. Finally, there is increasing interest in studying river water quality at large scales from space (Gardner et al., 2020; Ross et al., 2019), and the SWORD network is a logical candidate for performing such analyses globally. Ultimately, the real potential of a network like SWORD is to serve as a means of providing easy access to satellite observations and derived products from many sensors, allowing global-scale analysis that leverages their combined strengths. Within the SWORD framework, for example, SWOT discharge estimates can be used to create rating curves with width from optical and radar imagers as well as WSE from nadir altimeters. The resulting data products could dramatically increase the global availability of discharge estimates even in areas where in situ gauge data are unavailable to the global scientific community.

Data Availability Statement

The SWORD data set is freely available for download from Zenodo (Altenau et al., 2021; https://zenodo.org/record/3898569).

Acknowledgments
This research was funded by contracts from the SWOT Project Office at the Jet Propulsion Laboratory, California Institute of Technology to the University of North Carolina and The Ohio State University. The authors would like to thank Jon Schwenk and an anonymous reviewer who provided comments that helped improve the quality of the manuscript.

References

Allen, G. H., & Pavelsky, T. M. (2018). Global extent of rivers and streams. *Science*, *361*(6402), 585–588. https://doi.org/10.1126/science.aat0636

Allen, G. H., Yang, X., Gardner, J., Holliman, J., David, C. H., & Ross, M. (2020). Timing of Landsat overpasses effectively captures flow conditions of large rivers. *Remote Sensing*, *12*(9), 1510.

Altenau, E. H., Pavelsky, T. M., Durand, M. T., Yang, X., Frasson, R. P. D. M., & Bendezu, L. (2021). *SWOT River Database (SWORD) (Version v1) [Data set]*. Zenodo. https://doi.org/10.5281/zenodo.3898569

Biancamaria, S., Durand, M., Andreadis, K. M., Bates, P. D., Boone, A., Mognard, N. M., et al. (2011). Assimilation of virtual wide swath altimetry to improve Arctic river modeling. *Remote Sensing of Environment*, *115*(2), 373–381. https://doi.org/10.1016/j.rse.2010.09.008

Biancamaria, S., Lettenmaier, D. P., & Pavelsky, T. M. (2016). The SWOT mission and its capabilities for land hydrology. In *Remote sensing and water resources* (pp. 117–147). Springer. https://doi.org/10.1007/978-3-319-32449-4_6

Biancamaria, S., Schaedele, T., Blumstein, D., Frappart, F., Boy, F., Desjonquères, J. D., et al. (2018). Validation of Jason-3 tracking modes over French rivers. *Remote Sensing of Environment*, *209*, 77–89. https://doi.org/10.1016/j.rse.2018.02.037

Domeneghetti, A., Schumann, G. J. P., Frasson, R. P. D. M., Wei, R., Pavelsky, T. M., Castellarin, A., et al. (2018). Characterizing water surface elevation under different flow conditions for the upcoming SWOT mission. *Journal of Hydrology*, *561*, 848–861. https://doi.org/10.1016/j.jhydrol.2018.04.046

Durand, M., Fu, L. L., Lettenmaier, D. P., Alsdorf, D. E., Rodriguez, E., & Esteban-Fernandez, D. (2010). The surface water and ocean topography mission: Observing terrestrial surface water and oceanic submesoscale eddies. *Proceedings of the IEEE*, *98*(5), 766–779. https://doi.org/10.1109/jproc.2010.2043031

Durand, M., Gleason, C. J., Garambois, P. A., Bjerklie, D., Smith, L. C., Roux, H., et al. (2016). An intercomparison of remote sensing river discharge estimation algorithms from measurements of river height, width, and slope. *Water Resources Research*, *52*(6), 4527–4549. https://doi.org/10.1002/2015wr018434

Frasson, R. P. D. M., Pavelsky, T. M., Fonstad, M. A., Durand, M., Allen, G. H., Schumann, G., et al. (2019). Global relationships between river width, slope, catchment area, meander wavelength, sinuosity, and discharge. *Geophysical Research Letters*, *46*(6), 3252–3262. https://doi.org/10.1029/2019GL082027

Frasson, R. P. D. M., Schumann, G. J. P., Kettner, A. J., Brakenridge, G. R., & Krajewski, W. F. (2019). Will the Surface Water and Ocean Topography (SWOT) satellite mission observe floods? *Geophysical Research Letters*, *46*(17–18), 10435–10445. https://doi.org/10.1029/2019gl084686

Frasson, R. P. D. M., Wei, R., Durand, M., Minear, J. T., Domeneghetti, A., Schumann, G., et al. (2017). Automated river reach definition strategies: Applications for the surface water and ocean topography mission. *Water Resources Research*, *53*(10), 8164–8186. https://doi.org/10.1002/2017wr020887

Gardner, J. R., Yang, X., Topp, S. N., Ross, M. R., Altenau, E. H., & Pavelsky, T. M. (2020). The color of rivers. *Geophysical Research Letters*, *48*, e2020GL088946. https://doi.org/10.1029/2020GL088946

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

Häfliger, V., Martin, E., Boone, A., Ricci, S., & Biancamaria, S. (2019). Assimilation of synthetic SWOT river depths in a regional hydrometeorological model. *Water*, *11*(1), 78. https://doi.org/10.3390/w11010078

HydroFALLS. (n.). Retrieved from http://wp.geog.mcgill.ca/hydrolab/hydrofalls/

Ishitsuka, Y., Gleason, C. J., Hagemann, M. W., Beighley, E., Allen, G. H., Feng, D., et al. (2020). Combining optical remote sensing, McFLI discharge estimation, global hydrologic modelling, and data assimilation to improve daily discharge estimates across an entire large watershed. *Water Resources Research*, *56*, e2020WR027794. https://doi.org/10.1029/2020WR027794

JPL Internal Document. (2020a). *Surface Water and Ocean Topography Mission Level 2 KaRIn high rate water mask pixel cloud product, JPL D-56411*. Retrieved from https://podaac-tools.jpl.nasa.gov/drive/files/misc/web/misc/swot_mission_docs/pdd/D-56411_SWOT_Product_Description_L2_HR_PIXC_20200810.pdf

JPL Internal Document. (2020b). *Surface Water and Ocean Topography Mission Level 2 KaRIn high rate river single pass vector product, JPL D-56413*. Retrieved from https://podaac-tools.jpl.nasa.gov/drive/files/misc/web/misc/swot_mission_docs/pdd/D-56413_SWOT_Product_Description_L2_HR_RiverSP_20200825a.pdf

Langhorst, T., Pavelsky, T. M., Frasson, R. P. D. M., Wei, R., Domeneghetti, A., Altenau, E. H., et al. (2019). Anticipated improvements to river surface elevation profiles from the surface water and ocean topography mission. *Frontiers of Earth Science*, *7*, 102. https://doi.org/10.3389/feart.2019.00102

Larnier, K., Monnier, J., & Garambois, P. A. (2019). Discharge and bathymetry estimations of rivers from SWOT like data. *Geophysical Research Abstracts* (Vol. 21).

Lehner, B., & Grill, G., (2013). Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171–2186. https://doi.org/10.1002/hyp.9740

Lehner, B., Verdin, K., & Jarvis, A. (2008). New global hydrography derived from spaceborne elevation data. *Eos, Transactions American Geophysical Union*, *89*(10), 93–94. https://doi.org/10.1029/2008eo100001

Li, D., Andreadis, K. M., Margulis, S. A., & Lettenmaier, D. P. (2020). A data assimilation framework for generating space-time continuous daily SWOT river discharge data products. *Water Resources Research*, *56*(6), e2019WR026999. https://doi.org/10.1029/2019wr026999

Maneewongvatana, S., & Mount, D. M. (1999). It's okay to be skinny, if your friends are fat. In *Center for Geometric Computing 4th Annual Workshop on Computational Geometry* (Vol. 2, pp. 1–8).

Munier, S., Polebistki, A., Brown, C., Belaud, G., & Lettenmaier, D. P. (2015). SWOT data assimilation for operational reservoir management on the upper Niger River Basin. *Water Resources Research*, *51*(1), 554–575. https://doi.org/10.1002/2014wr016157

Nickles, C., Beighley, E., Durand, M., & Prata de Moraes Frasson, R. (2020). Integrating lateral inflows into a SWOT mission river discharge algorithm. *Water Resources Research*, *56*(10), e2019WR026589. https://doi.org/10.1029/2019wr026589

10.1029/2021WR030054

Nickles, C., Beighley, E., Zhao, Y., Durand, M., David, C., & Lee, H. (2019). How does the unique space-time sampling of the SWOT mission influence river discharge series characteristics? *Geophysical Research Letters*, *46*(14), 8154–8161. https://doi.org/10.1029/2019gl083886

Rodriguez, E., Durand, M., & Frasson, R. P. D. M. (2020). Observing rivers with varying spatial scales. *Water Resources Research*, *56*(9), e2019WR026476. https://doi.org/10.1029/2019wr026476

Rodriguez, E., Esteban Fernandez, D., Peral, E., Chen, C. W., De Bleser, J. W., & Williams, B. (2018). Wide-swath altimetry: A review. In *Satellite altimetry over oceans and land surfaces* (p. 2). CRC Press.

Ross, M. R., Topp, S. N., Appling, A. P., Yang, X., Kuhn, C., Butman, D., et al. (2019). AquaSat: A data set to enable remote sensing of water quality for inland waters. *Water Resources Research*, *55*(11), 10012–10025. https://doi.org/10.1029/2019wr024883

Samine Montazem, A., Garambois, P. A., Calmant, S., Finaud-Guyot, P., Monnier, J., Medeiros Moreira, D., et al. (2019). Wavelet-based river segmentation using hydraulic control-preserving water surface elevation profile properties. *Geophysical Research Letters*, *46*(12), 6534–6543. https://doi.org/10.1029/2019gl082986

SWOT Orbits. (n.d). Retrieved from https://www.aviso.altimetry.fr/en/missions/future-missions/swot/orbit.html

Tessler, Z. D., Vörösmarty, C. J., Grossberg, M., Gladkova, I., Aizenman, H., Syvitski, J. P. M., & Foufoula-Georgiou, E. (2015). Profiling risk and sustainability in coastal deltas of the world. *Science*, *349*(6248), 638–643. https://doi.org/10.1126/science.aab3574

Tourian, M. J., Elmi, O., Mohammadnejad, A., & Sneeuw, N. (2017). Estimating river depth from SWOT-Type observables obtained by satellite altimetry and imagery. *Water*, *9*(10), 753. https://doi.org/10.3390/w9100753

Verdin, K. L., & Verdin, J. P. (1999). A topological system for delineation and codification of the Earth's river basins. *Journal of Hydrology*, *218*(1–2), 1–12. https://doi.org/10.1016/s0022-1694(99)00011-6

Whittemore, A., Ross, M. R., Dolan, W., Langhorst, T., Yang, X., Pawar, S., et al. (2020). A participatory science approach to expanding instream infrastructure inventories. *Earth's Future*, *8*(11), e2020EF001558. https://doi.org/10.1029/2020ef001558

Wongchuig-Correa, S., de Paiva, R. C. D., Biancamaria, S., & Collischonn, W. (2020). Assimilation of future SWOT-based river elevations, surface extent observations and discharge estimations into uncertain global hydrological models. *Journal of Hydrology*, *590*, 125473. https://doi.org/10.1016/j.jhydrol.2020.125473

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

Yoon, Y., Durand, M., Merry, C. J., Clark, E. A., Andreadis, K. M., & Alsdorf, D. E. (2012). Estimating river bathymetry from data assimilation of synthetic SWOT measurements. *Journal of Hydrology*, *464*, 363–375. https://doi.org/10.1016/j.jhydrol.2012.07.028