

Surface Water and Ocean Topography Project

Algorithm Theoretical Basis Document

**Long Name: Level 2 KaRIn High Rate River Single Pass
Science Algorithm Software**

Short Name: SAS_L2_HR_RiverSP

Initial Release

Document Custodians:

Cassie Stuurman
JPL Algorithm Engineer

Date

Claire Pottier
CNES Algorithm Engineer

Date

Approved by:

Curtis Chen
JPL Algorithm System
Engineer

Date

Roger Fjørtoft
CNES Algorithm System
Engineer

Date

Concurred by:

Tamlin Pavelsky
JPL Hydrology Lead

Date

Jean Francois Cretaux
CNES Hydrology Lead

Date

Paper copies of this document may not be current and should not be relied on for official purposes. The current version is in the JPL Product Data Management System (EPDM: <https://epdm.jpl.nasa.gov>) and the CNES Product Data Management System

See Sect 1.4 for how to cite this document

July 13, 2023
JPL D-105505



National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology



Contributing Authors

Name	Affiliation
Cassie Stuurman	Jet Propulsion Laboratory
Francis J. Turk	Jet Propulsion Laboratory
Alexander Fore	Jet Propulsion Laboratory
Mike Durand	Ohio State University
Tamlin Pavelsky	University of North Carolina
Renato P. M. Frasson	Jet Propulsion Laboratory
Brent A. Williams	Jet Propulsion Laboratory
Rui Wei	Jet Propulsion Laboratory

Science Team Reviewers

Name	Affiliation
Larry Smith	Brown University
P. A. Garambois	INRAE, Aix Marseille Univ., RECOVER
S. Ricci	CERFACS, France

CHANGE LOG

VERSION	DATE	SECTIONS CHANGED	REASON FOR CHANGE
Initial Release	2023-07-13	ALL	Initial Release Approved for public release (URS317511/CL#23-3542)

Table of Contents

Contributing Authors	2
Science Team Reviewers	2
CHANGE LOG	3
Table of Contents.....	4
Table of Figures	6
Table of Tables.....	6
List of TBC Items	7
List of TBD Items	7
1 Introduction	8
1.1 Purpose	8
1.2 Scope	8
1.3 Document Organization.....	8
1.4 Citing This Document.....	9
2 Overview	9
2.1 Background and Context	9
2.2 Preliminaries	10
2.2.1 Prior River Database (PRD).....	10
2.2.2 Pixel Cloud Information (PIXC)	12
2.3 Functional Flow	13
3 Algorithm Descriptions	16
3.1 AssignPixels.....	16
3.1.1 Water Labelling	16
3.1.2 Pixel Assignment	20
3.2 Aggregate Nodes	26
3.2.1 Purpose	26
3.2.2 Input Data	26
3.2.3 Output Data	26
3.2.4 Mathematical Statement	26
3.2.5 Accuracy	31
3.3 Height-Constrained Geolocation.....	33
3.3.1 Purpose	33
3.3.2 Input Data	34
3.3.3 Output Data	34
3.3.4 Mathematical Statement	34
3.3.5 Accuracy	34
3.4 Aggregate Reaches	35
3.4.1 Purpose	35
3.4.2 Input Data	35

3.4.3	Output Data	35
3.4.4	Mathematical Statement	35
4	Accuracy of L2_HR_River Algorithms	45
5	References	50
Appendix A.	Acronyms	51
Appendix B.	Simulations.....	52

Table of Figures

FIGURE 1: REACHES AND NODES IN THE PRIOR RIVER DATABASE (PRD).	10
FIGURE 2: HIGH-LEVEL BLOCK DIAGRAM OF THE LEVEL 2 RIVER PROCESSING STEPS (FUNCTIONS) USED TO GENERATE THE L2_HR_RIVERSP PRODUCT.	15
FIGURE 3: ASSIGNPIXELS BLOCK DIAGRAM.	16
FIGURE 4: PRINCIPLES OF THE SEGMENTATION ALGORITHM (HIGHLIGHTED IN THE BLUE BOX) AND SUBSEQUENT PIXEL SUB-SELECTION (WITH THE LATTER DESCRIBED IN SECTION 3.1.2) FOR THE RIVERSP PIXEL ASSIGNMENT.	18
FIGURE 5: EXAMPLE OUTPUT OF THE SEGMENTATION ALGORITHM.	19
FIGURE 6: NODE COORDINATES USED IN THRESHOLDING.	21
FIGURE 7: THE DEFAULT CASE FOR THE SEARCH DISTANCE AND EXTREME DISTANCE THRESHOLDS.	23
FIGURE 8: STEPS TAKEN TO THRESHOLD THE PIXEL ASSIGNMENT USING THE EXTREME DISTANCE COEFFICIENT FROM THE PRD WHEN A LAKE IS EXPECTED TO BE NEAR A RIVER.	24
FIGURE 9: QUALITY FLAG HANDLING IN THE L2_HR_RIVERTILE AGGREGATENODES ALGORITHM.	27
FIGURE 10: AN EXAMPLE OF A BLOCKED WIDTH NODE.	32
FIGURE 11: PIXC AND PIXCVecRIVER HEIGHTS BEFORE AND AFTER HEIGHT-CONSTRAINED GEOLOCATION.	34
FIGURE 12: EXAMPLE ILLUSTRATING THE PIECEWISE LINEAR OUTLIER REJECTION ALGORITHM.	37
FIGURE 13. MAIN STEPS IN THE COMPUTATION OF THE ENHANCED SLOPE.	43
FIGURE 14: SIMULATED PERFORMANCE HISTOGRAMS FOR WSE, SLOPE, AREA, AND WIDTH FROM THE REPRESENTATIVE DATASET.	46
FIGURE 15: SIMULATED PERFORMANCE CDFs FOR THE REACH-LEVEL ABSOLUTE ERROR IN WSE, TOTAL AREA, AND SLOPE FROM THE REPRESENTATIVE DATASET.	47
FIGURE 16: REACH-LEVEL AREA (UPPER) AND WSE ERROR (LOWER) VERSUS REACH-LEVEL AREA AND POSITION IN THE SWATH FOR NOMINAL RIVERTILES IN THE REPRESENTATIVE DATASET.	48
FIGURE 17: RIVER SLOPE PERFORMANCE VERSUS RIVER WIDTH AND POSITION IN THE SWATH FOR NOMINAL RIVERTILES IN THE REPRESENTATIVE DATASET.	49
FIGURE 18: STATISTICS OF THE REPRESENTATIVE DATASET BEFORE AND AFTER FILTERING FOR REACH GEOMETRY AND POSITION IN THE SWATH.	54

Table of Tables

TABLE 1. CONTINENT CODES FOR THE REACH_ID ATTRIBUTE AND CORRESPONDING CONTINENT IDs FOR THE FILENAME.	11
TABLE 2: WATER BODY TYPE CODES FOR THE REACH_ID ATTRIBUTE.	12
TABLE 3: DETECTED WATER CLASSES FROM PIXC.	12
TABLE 4. HIGH-LEVEL DESCRIPTION OF THE FUNCTIONS USED TO GENERATE THE L2_HR_RIVERSP PRODUCT.	14
TABLE 5: PIXEL ASSIGNMENT THRESHOLDS FOR DOMINANT AND NON-DOMINANT LABELS IN THE ALONG-REACH AND CROSS-REACH DIRECTIONS.	22
TABLE 6: EXTREME DISTANCE CALCULATION FOR REACHES WITH CHANNEL WIDTHS GREATER THAN AND LESS THAN THE NODE SPACING. ...	22
TABLE 7: MAPPING FROM PIXC QUALITY FLAGS TO INTERNAL STATE VARIABLES IN THE RIVER QUALITY MODULE.	28
TABLE 8: SUMMARY STATISTICS FOR THE L2_HR_RIVERSP SIMULATED NOMINAL REACH-LEVEL PERFORMANCE USING L2_HR_PIXC SIMULATED DATA FROM THE REPRESENTATIVE DATASET.	45
TABLE 9: REPRESENTATIVE DATASET FILTERING CRITERIA FOR SWOT RIVER SCIENTIFIC REQUIREMENTS.	53

List of TBC Items

Page	Section

List of TBD Items

Page	Section

1 Introduction

1.1 Purpose

The purpose of this Algorithm Theoretical Basis Document (ATBD) is to describe the physical and mathematical basis for the science data processing algorithms that are used to generate the **SWOT Level 2 KaRIn High Rate River Single Pass Vector (L2_HR_RiverSP)** science data product. It describes the necessary input data (SWOT and auxiliary data), the methods used, the outputs derived, and the accuracy obtained by the L2_HR_RiverSP processor. The aim is to have a comprehensive reference for how SWOT pixel cloud (L2_HR_PIXC) information is processed into the river width, area, height, slope, and discharge information that forms the L2_HR_RiverSP product. For detailed information on the output product filetypes, attribute definitions, and metadata fields the reader is directed to the production description document [1].

Because the L2_HR_RiverSP product is simply the aggregate of L2_HR_RiverTile products to the continent scale [1], the algorithms used in L2_HR_RiverTile are, by definition, also described by this document. All subsequent discussions of L2_HR_RiverSP processor algorithms also apply to the L2_HR_RiverTile product.

1.2 Scope

The scope of this document is to:

1. Delineate the high-level processing steps within the SAS_L2_HR_RiverSP and their flow.
 - a. Each step is further described by the relevant software functions used in its processing (where “relevant” means it pertains to data processing and algorithms).
2. For each processing step, describe:
 - a. Its intended purpose.
 - b. The input data required.
 - c. The output information derived.
 - d. The mathematical basis of the algorithms used.
 - e. The expected accuracy and/or limitations of the algorithms.
3. Provide the relevant references for the algorithms described in this document.
4. Describe the simulation data and processing chain used in the design and testing of the algorithms.

1.3 Document Organization

Section 1 provides the purpose and scope of this document.

Section 2 provides the background and context of the algorithms described in this document as well as the functional flow of the primary functions.

Section 3 provides the algorithm description for each of the functions shown in the block diagram (Figure 2), including input data, output data, mathematical basis, and expected accuracy.

Section 4 provides a summary of the accuracy of the L2_HR_RiverSP processor described in this document.

Appendix A provides a listing of the acronyms used in this document.

1.4 Citing This Document

Please cite this document as follows:

JPL D-105505, “SWOT Algorithm Theoretical Basis Document: Level 2 KaRIn High Rate River Single Pass (L2_HR_RiverSP) Science Algorithm Software,” Jet Propulsion Laboratory Internal Document, 2023.

2 Overview

2.1 Background and Context

The Surface Water and Ocean Topography (SWOT) mission is a partnership between two communities -- physical oceanography and hydrology -- to share high vertical accuracy topography data produced by the payload, whose principal instrument is the Ka-band Radar Interferometer (KaRIn). The details of SWOT mission objectives and requirements can be found in the SWOT Science Requirements Document [2].

The input to the L2_HR_RiverSP product is the L2_HR_PIXC product, which provides geolocated heights from each continent-pass of the high-rate (HR) data stream of the SWOT KaRIn instrument. These data are generally produced for inland and coastal hydrology surfaces, as controlled by the reloadable KaRIn HR mask.

The L2_HR_RiverSP product specifically provides data for river reaches identified in the prior river database (PRD). Each reach is divided into a number of nodes in the PRD. The content and structure of the PRD are described in [3] (some of the content of the PRD is replicated in the L2_HR_RiverSP product for convenience). As discussed further in Section 2, each L2_HR_RiverSP product granule consists of data for both reaches and nodes.

Only rivers in the PRD are included in the L2_HR_RiverSP product. Information on lakes and unidentified water features is given in the L2_HR_LakeSP science data product [4].

This document describes the algorithms that are used to generate the L2_HR_RiverSP product. First, some high-level background is provided on the L2_HR_PIXC product, the prior river database (PRD), and a description of the functional flow of the algorithms. Section 3 contains the details of the algorithms that are used to generate the L2_HR_RiverSP product from the L2_HR_PIXC product and various auxiliary data inputs. Section 4 presents the overall accuracy assessment of the L2_HR_RiverSP algorithms.

2.2 Preliminaries

This section presents an overview of general concepts and terms used throughout the document and which are common among multiple subprocesses in the L2_HR_River algorithms. Since the L2_HR_PIXC product provides the input data to the L2_HR_River processing, a high-level description of the L2_HR_PIXC product is presented including a description of its coverage, resolution, sampling and accuracy.

2.2.1 Prior River Database (PRD)

The information contained in the prior river database (PRD) is central to the processing of the L2_HR_RiverSP products. The PRD used in SWOT processing is called the SWOT River Database (SWORD) and is described in more detail in [5]. RiverSP processing only reports data for rivers that are in the PRD. The PRD is fixed for each instance of the processing cycle (meaning for instance, that the positions of the center lines do not vary over time). In the PRD, rivers are classified initially by the continent in which they lie, with further refinements to identify the containing basin. Each PRD reach is defined by a high-resolution centerline with an approximate 30 m spacing. Reaches are further divided into nodes along the reach centerline with a node-to-node spacing of approximately 200 meters. Reach lengths of 10 km are typical, but they can vary from 5 to 20 km. Figure 1 shows the components of a typical reach.

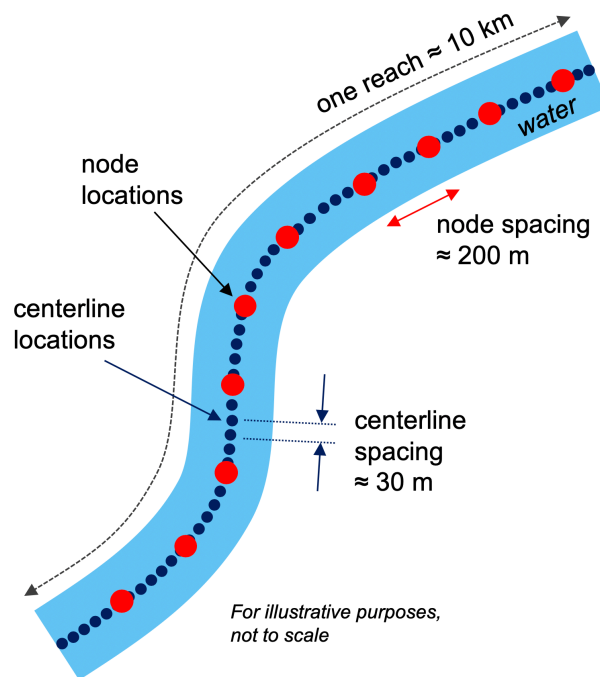


Figure 1: Reaches and Nodes in the Prior River Database (PRD). Illustration of the centerline locations (dark blue dots) that define the polyline shapes of the reach shapefile and the nodes (red dots) that define the point shapes of the node shapefile for a partial segment of one reach, per the PRD.

Most reaches are defined to have a minimum water surface area of 1 km² to allow for sufficient spatial averaging of the SWOT observations such that the estimated water surface elevation (WSE) meets the required precision. Other considerations for setting reach boundaries include hydrological and morphological features such as tributaries, dams, large islands or multiple channels, and edges of the KaRIn measurement swath. Generally, reaches are demarcated where a significant change in WSE is expected (e.g. dams) in order to preserve the integrity of the slope, WSE, and discharge calculations for a given reach.

Each reach record is associated with a unique identifier *reach_id* from the PRD. The format of the identifier is a 11-character string of the form *CBBBBRRRRRT*, where *C* = continent, *B* = basin, *R* = reach, and *T* = type. The *reach_id* provides the link between the observed reach location and its corresponding entry in the PRD.

- *reach_id*: Unique reach identifier from the prior river database.

The *reach_id* is based on the Pfafstetter coding system [6] that assigns identifications based on the topology of the river network. The code allows digits 0-9 at each hierarchy level. SWOT *reach_id* values always include six Pfafstetter levels of basins. Continent (*C*) and water body type (*T*) codes are provided in Table 1 and Table 2, respectively. Note that lake water bodies that are connected to the river topology of the PRD (*T* = 3) are processed as river reaches and will appear in the L2_HR_RiverSP product, although their slope, width, and area variables are filled with null values because the algorithms for computing such quantities may not be applicable to connected lakes. These connected lakes will typically also appear in the L2_HR_LakeSP product. Reaches with unreliable topology (*T* = 5) are treated the same as connected lake (*T* = 3) topologies if and only if the PRD *lakeflag* is set to 1. If the reach is flagged as a lake in the PRD, a bit indicating this is set in *node_q_b* and in *reach_q_b* (see Appendix C). Short reaches defined where dams occur (*T* = 4) have null-filled WSE, slope, width, and area. Ghost reaches (*T*=6) occur only at the headwaters and outlets of rivers. They are designed to buffer PRD reaches from other waterbodies such as oceans or non-PRD rivers. Ghost reaches are used only in processing and do not appear in the product.

Table 1. Continent codes for the *reach_id* attribute and corresponding continent IDs for the filename.

Continent Code	Continent	Continent ID
1	Africa	AF
2	Europe and Middle East	EU
3	Siberia	SI
4	Central and Southeast Asia	AS
5	Australia and Oceania	AU
6	South America	SA
7	North America and Caribbean	NA
8	North American Arctic	AR
9	Greenland	GR

Table 2: Water body type codes for the *reach_id* attribute.

Type Code (<i>T</i>)	Water Body Type
1	River
3	Connected Lake
4	Dam
5	Unreliable Topology
6	Ghost node or reach (not output to product)

2.2.2 Pixel Cloud Information (PIXC)

2.2.2.1 Detected Water Class

All pixels passed from the L2_HR_PIXC processor have a classification value describing the nature of the observed pixel (Table 3). The water class of a pixel may change how it is used in the river processor, and their usage is specified in the auxiliary parameter configuration file as a user input. See [7] for a description of how the detected water classes are determined in L2_HR_PIXC processing.

As this document discusses the algorithms that handle PIXC water classes, a table summarizing them is provided here for reference. The algorithmic handling of each water class in the river processor is discussed below in sections 3.1.1.4 and 3.2.4.

Table 3: Detected Water Classes from PIXC

Value	Water class name	Description	L2_HR_River Usage
-1	invalid	No valid water class	Do not use
1	land	Pixel is a land pixel that is not adjacent to water	Do not use
2	land_near_water	Pixel is a land pixel that is adjacent to water	Do not aggregate height Aggregate area with water fraction
3	water_near_land	Pixel is a water pixel that is adjacent to land	Aggregate height Aggregate area with water fraction
4	open_water	Pixel is a water pixel that is not adjacent to land	Aggregate height Aggregate area
5	dark_water	Pixel was determined to be dark water.	Do not aggregate height Aggregate area
6	low_coh_water_near_land	Pixel is low-coherence water near land	Aggregate height Aggregate area
7	low_coh_water	Pixel is low-coherence interior water	Aggregate height Aggregate area

2.2.2.2 Pixel Quality Data

All pixels passed from the L2_HR_PIXC processor contain detailed quality data that are used by the river processor to determine whether or not a pixel should be assigned and its information aggregated to node. The important PIXC quality outputs for the river processor are *classification_qual*, *geolocation_qual*, and *sig0_qual*. For a detailed description of the L2_HR_PIXC quality flags see [8]. For a description of how these flags are used in the river processor, see Section 3.2.4.1 of this document.

2.3 Functional Flow

Table 4 provides a high-level description of each of the processing functions that are used to generate the L2_HR_RiverSP product. Figure 2 illustrates the flow of these processing steps.

The L2_HR_RiverSP processing begins with the L2_HR_PIXC products [8]. The L2_HR_PIXC is a pixel cloud product that provides InSAR-derived pixel *heights* (with respect to the reference ellipsoid), pixel *geolocations*, and pixel *classifications* of the land and water regions observed by SWOT. It also provides information such as quality flags and measurement uncertainties that are necessary for tracking measurement quality. This product is subsequently referred to as the “pixel cloud” and its contents as “pixels”.

The L2_HR_RiverSP processor maps the pixel cloud pixels to the PRD in order to estimate river widths, areas, heights, and slope. At a high level, there are four main algorithmic steps, each of which is described in more detail later in this document:

1. Assign Pixels

- a. Wherein water pixels from the pixel cloud are segmented and spatially thresholded in order to separate river pixels (i.e. river water surfaces) from non-river (e.g., lake) pixels.
- b. Pixels identified as belonging to the water surface of a river are assigned to their nearest PRD node.

2. Aggregate Nodes

- a. Node-level quantities such as height, width, and area are computed using the aggregated and uncertainty-weighted values from the pixel cloud.

3. Aggregate Reaches

- a. Wherein PRD reach-level quantities such as height, width, slope, and discharge are computed using the node-level estimations from Aggregate Nodes.

4. Height-Constrained Geolocation

- a. Wherein the geolocations of assigned river water pixels are adjusted using the estimated reach height and slope to refine their horizontal positioning.

Section 3.1 describes Assign Pixels, where all observed water pixels are segmented and labelled to represent each unique water body in the scene. The segmentation label corresponding to the river channel is determined, and river pixels are assigned to their respective PRD-defined

node locations. After this step, each PRD node will typically have many pixels assigned. Corrections are then applied to convert the pixel ellipsoid-relative height to the water surface elevation (WSE) given in the L2_HR_RiverSP product, which is provided with respect to the geoid model used. Corrections for the solid-earth, pole, and load tides are also applied.

Section 3.2 describes Aggregate Nodes, where the quantities associated with each pixel are then aggregated to the node level for each PRD node, and geophysical quantities such as WSE, area, and width are calculated.

Section 3.3 describes the Aggregate Reaches algorithm. Similar to the pixel-to-node aggregation process, this step aggregates each of the node-level estimates to their parent reach location from the PRD. Nearly all of the node-level outputs have corresponding reach-level outputs, such as WSE, area, and width. Some additional reach variables are also calculated from the nodes, including the reach slope. The cross-sectional area and the discharge from multiple flow laws are then computed for each reach using information from the PRD; the main algorithms for these steps are in offline computation and are therefore outside the scope of this document (see [1] for more information and additional references).

Section 3.4 describes the Height-Constrained Geolocation algorithm. Once the reach slope estimates have been derived, they can be used to adjust the geolocation for only those pixels used from the L2_HR_PIXC product. This is referred to as height-constrained geolocation, as the reach fit heights are used to further constrain the geolocation (latitude, longitude, and height) for each pixel. It is performed by further smoothing the WSE in comparison with those of the input L2_HR_PIXC product and then translating the geolocated pixels along the iso-range/Doppler contour to their new, smoothed-WSE positions.

Table 4. High-level description of the functions used to generate the L2_HR_RiverSP product.

Function Name	Description
AssignPixels	Assigns pixels from L2_HR_PIXC product to node locations and differentiates river from non-river pixels.
AggregateNodes	Aggregates pixel-level values to the node-level values.
AggregateReaches	Aggregates node-level products into reach-level values, and computes reach-level slope, cross-sectional area and discharge.
HeightConstrained Geolocation	Compute the height-constrained geolocation for all pixels associated with a node location, using the node-level heights.

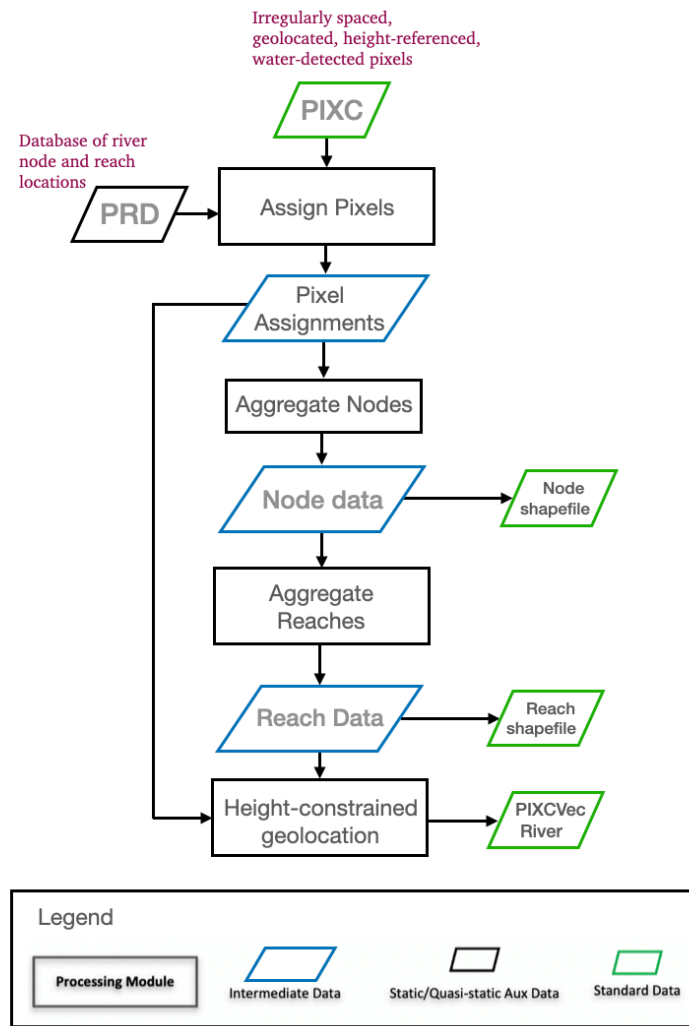


Figure 2: High-level block diagram of the Level 2 River processing steps (functions) used to generate the L2_HR_RiverSP product.

3 Algorithm Descriptions

3.1 AssignPixels

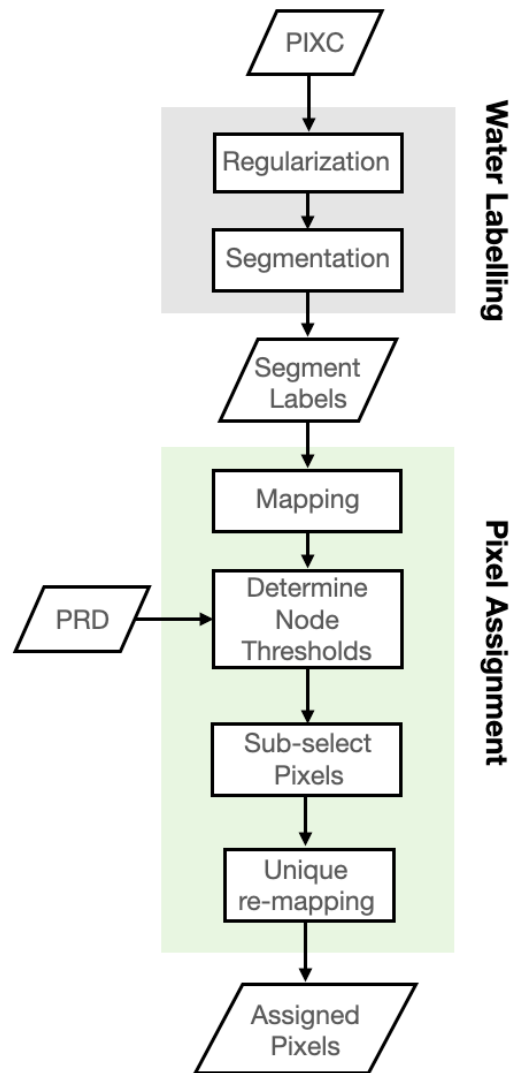


Figure 3: AssignPixels Block Diagram. At a high level, AssignPixels is made of two main steps: Water Labelling and Pixel Assignment. The Water Labelling algorithm separates distinct bodies of water pixels observed by SWOT. The Pixel Assignment algorithm identifies the segment corresponding to the river channel and maps pixels to river nodes as defined in the PRD.

3.1.1 Water Labelling

3.1.1.1 Purpose

The input LR_HR_PIXC product contains detected water pixels from both rivers and lakes, but only river pixels are needed by the RiverSP processor. The first step towards differentiating river from non-river pixels involves pixel cloud segmentation. Using input pixel cloud classifications and pixel indices, this function partitions the set of all water pixels into labelled

sets or “segments” of contiguous (i.e. connected) water features, including dark water. The water segment labels are passed to the Pixel Assignment algorithm.

3.1.1.2 *Input Data*

Description	Source
Pixel classifications and indices	L2_HR_PIXC data product

3.1.1.3 *Output Data*

Description
Integer-value segmentation labels to be used in pixel assignment.

3.1.1.4 *Mathematical Statement*

First, the 1-D array of pixel indices and water classification values from the L2_HR_PIXC product is arranged into a 2-D slant-plane array or matrix with dimensions corresponding to the maximum range and azimuth values. Non-water classes have a value of 0, and all water classes (including dark water) are assigned a value of 1. The slant-plane image is then segmented into disjoint, uniquely labelled water class features using a simple matrix connectivity algorithm. That is, elements that are horizontally or vertically connected are considered to be within the same feature and labelled with a unique integer that identifies the segment.

Finally, land edge segments are assigned to the nearest water feature using greyscale dilation. Greyscale dilation is similar to binary dilation but is used on non-boolean matrices. Segments are expanded such that land pixels are re-assigned the value of their nearest segment label. The end result is that each water segment label is expanded to include its land-edge pixels. Segmentation is illustrated in the left-hand portion of Figure 4.

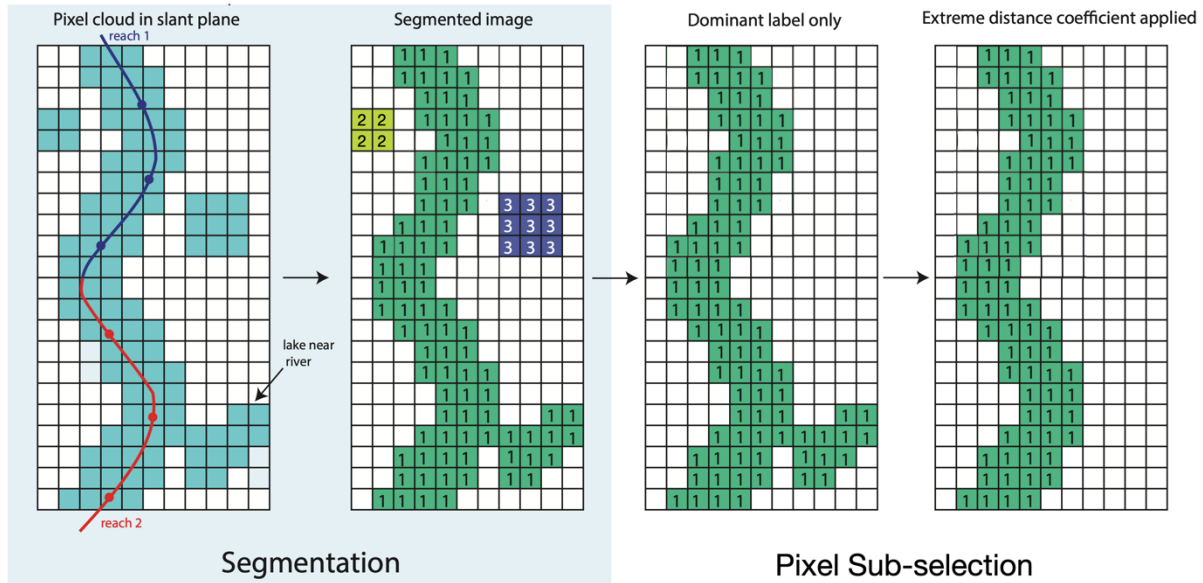


Figure 4: Principles of the segmentation algorithm (highlighted in the blue box) and subsequent pixel sub-selection (with the latter described in Section 3.1.2) for the RiverSP pixel assignment. Note this is for illustration purposes only and resolution is much coarser than a true pixel cloud image. For each reach, all contiguous water features as detected by SWOT in the slant plane are labelled (Section 3.1.1). The numbers in each water pixel in the figure represent the label values. The labels and pixel cloud information are then passed through pixel assignment, which prunes pixels that are not expected to be part of the river channel. The dominant label (Section 3.1.2.4) is assigned to the largest contiguous feature in each reach and is considered to be the river channel segment. Then, using the extreme distance coefficient, pixels that are contiguous with the dominant label but at a far distance from the reach centerline (farther than the extreme distance estimated as the prior width multiplied by the extreme distance coefficient, both present in the PRD) are removed, thus removing lakes that are near the river channel in the slant plane.

3.1.1.5 Segmentation Example

In Figure 5, an example pixel cloud in the slant-plane shows all pixels identified as water (including dark water) above an example output from the segmentation algorithm. This pixel cloud was generated by running a simulated SLC (single-look complex) image created using the SWOT Hydrology SLC Simulator [7] through the pixel cloud processor. Note that it includes both river pixels and pixels associated with non-river waterbodies such as nearby lakes. Unique segments identified as water in the pixel cloud are each assigned a different integer label value, represented in Figure 5 by different color values.

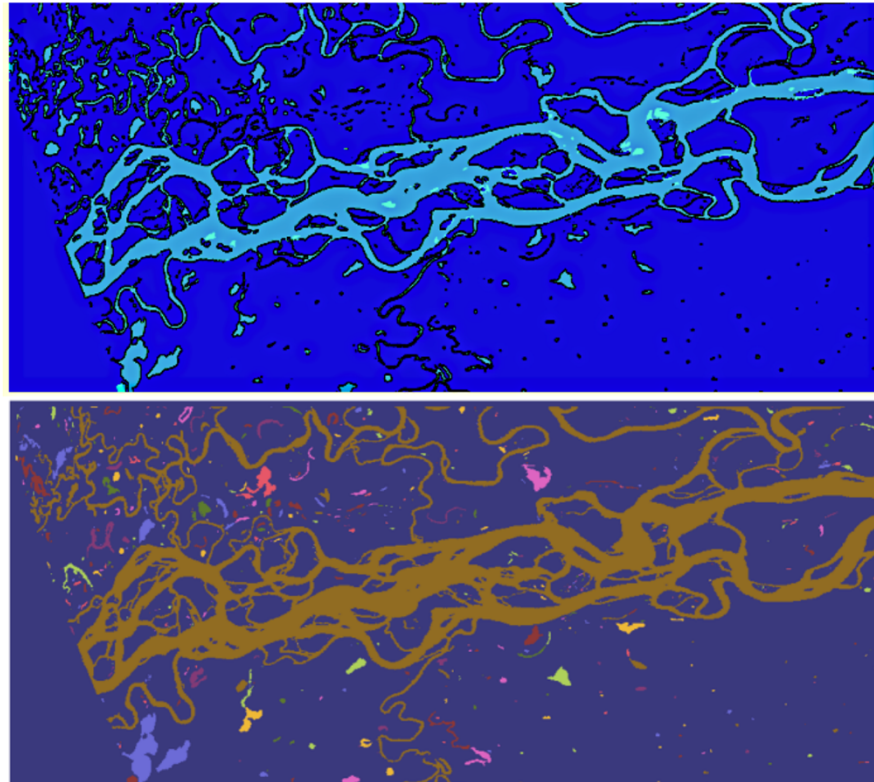


Figure 5: Example output of the segmentation algorithm. Upper image: The input pixel cloud colored by water class. Light blue and very light blue correspond to interior water classes, where the lighter color is dark water. The black outline marks the water-near-land pixels. The dark blue background is made up of land pixels. Lower image: The output segmentation labels from the Water Labelling algorithm. Each contiguous waterbody is given its own integer value, represented here by multicolored water segments. The dominant label (corresponding to the river channel) is determined by taking the segment with the most pixels (i.e. the mode of the integer labels) that is also within the in-channel bounds as defined by the PRD widths. In this case, the large multichannel river segment shown in brown is the dominant label.

3.1.1.6 Accuracy

During greyscale dilation, the Segmentation algorithm may attach land edge segments to the incorrect label if a land edge touches multiple water segments in the slant plane image. In this instance, the land edge pixel would be assigned to the water segment with the higher integer label (which is arbitrarily assigned). This means that if a land edge segment touches both a river channel and a lake segment with high label value it will be incorrectly assigned to the lake and excluded from the river label. This configuration is uncommon and theoretically would only affect a small number of water edge pixels. Thus, this issue should have minimal impact on algorithmic performance.

Depending on the parameters used, image dilation can sometimes over-regularize and result in an over-estimation of river width and area. This is most likely to be impactful for width and area estimates of multi-channel rivers with a large number of small interior land segments.

3.1.2 Pixel Assignment

3.1.2.1 Purpose

In order to compute the node-level water surface elevation and area, the pixels must first be mapped to nodes. The pixel assignment algorithm takes the input L2_HR_PIXC (pixel cloud) data as well as label information from the segmentation algorithm and assigns each pixel classified as water within the river channel to its appropriate node from the PRD. Thus, the pixel assignment algorithm must differentiate river pixels from non-river pixels and map these pixels to their appropriate PRD nodes.

3.1.2.2 Input Data

Description	Source
Pixel 3-D geolocation, area, and geophysical parameters	L2_HR_PIXC data product
Reach centerline, node locations, width threshold parameters	Prior River Database
Integer-value segmentation labels corresponding to river channel	Segmentation algorithm

3.1.2.3 Output Data

Description
Assigned pixels, L2_HR_PIXC pixels mapped to PRD nodes.

3.1.2.4 Mathematical Statement

At a high level, the pixel assignment algorithm differentiates river from non-river pixels using prior knowledge of river width combined with prior knowledge of the proximity of nearby lakes. River widths and node spacings from the prior river database (PRD) are translated into subselection *thresholds* that limit the maximum possible distance a pixel may be from a given river node. Pixels beyond these thresholds are excluded from river pixel assignment for that node. Note that the L2_HR_RiverSP product is not intended to be used for significant flooding events; the L2_HR_Raster product should be considered for such cases.

The *thresholds* for pixel assignment are defined in the local coordinate system of the node, where s is the along-reach distance away from the node, and n is the normal or ‘cross-reach’ distance away from the node (Figure 6), as defined by the PRD centerline. These coordinates are defined in the ground plane. Both the along-reach and cross-reach directions have inner and outer thresholds, where the inner threshold is referred to as the *search distance* and the outer threshold is called the *extreme distance*.

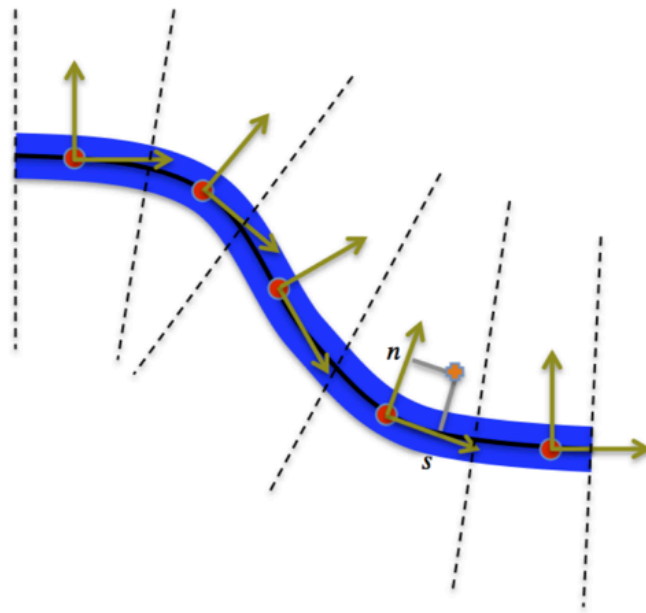


Figure 6: Node coordinates used in thresholding. Pixels are assigned to each node using thresholds defined in the n (cross-reach) and s (along-reach) directions.

The algorithm begins by extracting all nodes (and reaches) from the PRD that intersect the bounding box of the input L2_HR_PIXC. Pixels will only be assigned to these nodes. Each pixel is then mapped to its nearest node *without* thresholding in order to translate pixel range and azimuth image coordinates into the local coordinates of each node (s, n) by the *Centerline* function. For the purposes of this document, we refer to this initial function as pixel *mapping* and the final assignment following sub-selection as pixel *assignment*. The *Centerline* function returns the nearest-node index, Euclidean distance-to-node (d), and local coordinates of each pixel computed using the PRD centerlines for each node. Pixel mapping via the *Centerline* function is necessary before pixel assignment as each node may have different sub-selection thresholds. The mapping of pixels to nodes is completed using a nearest-neighbor algorithm for clustered point sets (described in [9]) that fundamentally minimizes the Euclidean horizontal distance between pixel and node.

Following pixel mapping and transformation to local node coordinates, pixel sub-selection is performed. An important concept to understand when considering pixel sub-selection is the *dominant label*. The dominant label is selected from the water segment labels output from the **Segmentation** algorithm (see Figure 5) and is needed to define the sub-selection thresholds. It is defined as the largest segment label within the search distance of each reach, found by taking the mode of the pixel labels within the search distance. This segment corresponds to the segment with the largest number of pixels within the search distance and is treated as the main river channel segment for the reach. Mathematical definitions of the search distances for s and n are given in Table 5. The dominant label affects which thresholds are applied to which pixels in the tile. In pseudocode, the dominant label is extracted as:

algorithm **get_dominant_label**(PIXC_pixels, PRD):
 inputs: PIXC_pixels from pixel cloud; PRD reach/node locations from database
 outputs: dominant_label a value corresponding to the river segment label

```

SET river_pixels = [ ]
FOR each reach in tile:
    FOR each node in reach:
        SET search_distance_n = (PRD_channel_width)/2
        SET search_distance_s = 3*(PRD_node_spacing)

        river_pixels.APPEND(
            pixels where n < search_distance_n AND
                s < search_distance_s
        )

    dominant_label = MODE(river_pixels.segmentation_labels)
RESET river_pixels
    
```

	Maximum distance for dominant label pixels	Maximum distance for non-dominant label pixels
Along-reach direction, <i>s</i>	Extreme Distance	Search Distance =3*node_spacing
Cross-reach direction, <i>n</i>	Extreme Distance	Search Distance =(PRD width)/2

Table 5: Pixel assignment thresholds for dominant and non-dominant labels in the along-reach and cross-reach directions. Pixels within these thresholds will be considered part of the river channel for the pixel assignment.

	PRD width/2 > node spacing	PRD width/2 < node spacing
Extreme distance	<i>ext_dist_coefficient</i> * (PRD width)/2	<i>ext_dist_coefficient</i> * node_spacing

Table 6: Extreme distance calculation for reaches with channel widths greater than and less than the node spacing. For narrow channels, the extreme distance is calculated using the node spacing to ensure all pixels in the along-reach direction are incorporated into the pixel assignment.

Two factors determine the pixel assignment thresholds used for each node (Table 5). One factor is whether the pixel is within the dominant label or not. The second considers whether the threshold is in the along-reach direction, *s*, or the cross-reach (i.e. normal) direction *n*. In both *s* and *n*, pixels that are within the dominant label are given larger thresholds equal to the *extreme distance*, meaning they are more generously included in the pixel assignment. This larger threshold is denoted the extreme distance as it uses the *extreme distance coefficient* in its computation. This coefficient was implemented to handle instances where there are known non-river waterbodies (such as lakes or artificial reservoirs) near the river channel. The extreme distance coefficient is pre-defined in the PRD and is described in more detail in [5].

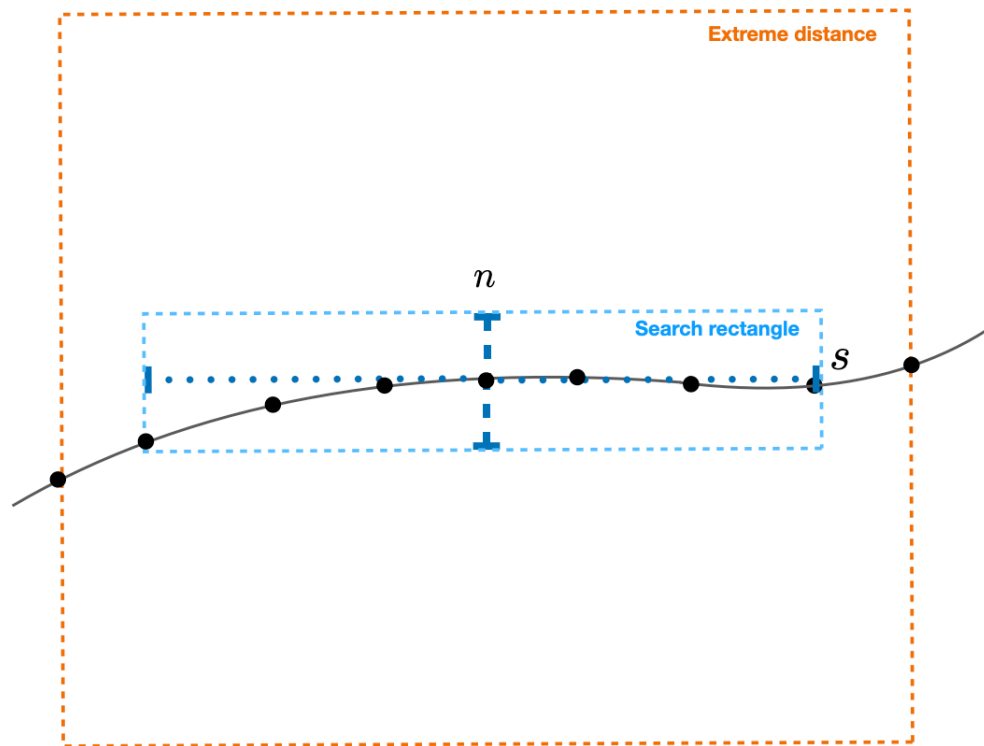


Figure 7: The default case for the search distance and extreme distance thresholds. In cases where no lakes near the river channel are expected, the extreme distance is far from the channel centerline (10 times the river reach width). The search rectangle extends half the river reach width in the cross-reach direction and three node spacings in the along-reach direction from the node (not to scale). A large extreme distance is useful for wide river channels with sharp bends to ensure that all connected pixels in the river channel are included in the pixel assignment.

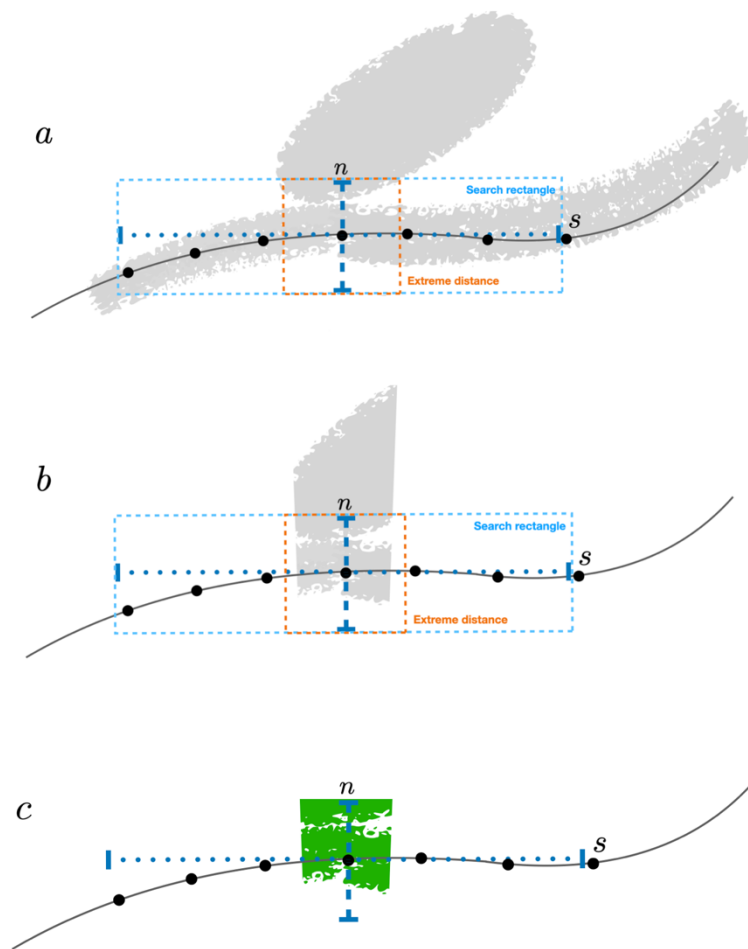


Figure 8: Steps taken to threshold the pixel assignment using the extreme distance coefficient from the PRD when a lake is expected to be near a river. The grey shaded area represents the set of all pixels detected by SWOT as water (cartoon representation, not-to-scale). (a) The search rectangle and extreme distance bounds are shown for a node where the extreme distance coefficient is 1 (b) All nearest-neighbor pixels mapped to the center node. (c) Pixels assigned to the node following thresholding by the extreme distance coefficient.

Figure 7 illustrates the subselection thresholds for a typical node that does not have a nearby lake. For each node, pixels that are *not* part of the dominant label and are less than 3 times the node spacing in along-reach distance and less than half the maximum river width (from the PRD) in the cross-reach direction will be included. Pixels that are a part of the dominant label segment will be included up to the *extreme distance* (10 times the PRD width for reaches without nearby lakes).

For nodes that are expected to have nearby lakes, the *extreme distance* coefficient is tuned offline in the PRD to prevent lake pixels from being incorporated into the pixel assignment (Figure 8). In cases where there are known Prior Lake Database (PLD, see [10]) lakes within 500 m of the river edge, the extreme distance coefficient is scaled depending on the type of river and

the river width. For lake-adjacent nodes within a single-channel river, the extreme distance coefficient is set to 1. For lake-adjacent nodes within a multi-channel river, the extreme distance coefficient is set to the ratio of the maximum river width and the channel width, with typical resulting values of 1-3.

The pixel subselection process is repeated for every reach in the subsetted PRD. The result is a set of pixel observations that are determined to be a part of the river channel mapped to their nearest-neighbor node. Following iterative reach-level assignment, some PIXC pixels may be assigned to multiple nodes near the beginning and end of each reach due to overlapping *search distance* boundaries. Each pixel is then checked for multiple assignments and only the nearest node assignment is preserved. This assures that each of the subsetted PIXC pixels are only assigned to the (single) node to which it is closest within a given tile.

The end result is a unique assignment of all pixels that are determined to be within the river channel to their closest PRD nodes by geographical distance.

3.1.2.5 Accuracy

Pixel assignment errors can be grouped into several broad categories:

1. River pixel misassignment; where a river pixel is assigned to the wrong node or to the wrong reach as defined by contiguity of the water belonging to a reach.
2. Non-PRD waterbody pixel misassignment; where a water pixel that corresponds to a water feature that is not represented in the PRD (such as a tributary that is not in the database or a disconnected lake) is assigned to a river node.
3. Water class misassignment; where a land pixel is incorrectly assigned to a river or where a river pixel is incorrectly identified as land.
4. Flood event misassignment.

The first class of pixel assignment error listed above is almost always of minimal impact. It occurs when node search distances overlap causing all pixels in the overlapping region to be assigned to whichever node is nearest. The most common instances of river pixel misassignment are around tight river bends where pixels may be assigned across the bend to the wrong node. Another common example is at river confluences where multiple reaches come together and nodes may lie within each other's search distances. River pixel locational misassignment typically affects a small number of pixels with heights that are not very different than the node they should be assigned to and thus does not have a significant effect on the node and reach level measurements.

The second category describes cases where detected water pixels that originate from features that are not PRD rivers are incorrectly assigned to river nodes. This generally results in a large number of contaminating pixels and can affect the accuracy of the WSE, width, and area measurements. This may only occur in cases where the search distance or extreme distance is incorrectly tuned and does not successfully prune nearby lakes from rivers prior to pixel assignment; the PRD centerline is inaccurate and intersects waterbodies that are not the river (e.g. due to river migration); or in cases where PRD reaches are missing over tributaries that are wide enough to be observed by SWOT. The solution for this class of error is to adjust the PRD so the algorithm will correctly reject or reassign the spurious water pixels.

The third class of pixel assignment error relates to incorrectly classified water pixels. This depends largely on the accuracy of the input L2_HR_PIXC data. See [7] for a description of water class detection accuracy in the pixel cloud processor.

Lastly, there are flood-event pixel misassignment errors. The L2_HR_River vector product was not designed to provide accurate width, area, WSE, and slope information in the case of extreme flood events. One result of this is that in the case of flood events for river reaches that are also adjacent to lakes, it is possible that the extreme distance coefficient could filter pixels far from the centerline that may in fact be part of the flooded river channel. See Section 4 for further discussion of L2_HR_River accuracy in the case of extreme flood events. See L2_HR_Raster [11] for a product appropriate for analysis of such events.

In our representative dataset simulations (see Appendix C Simulations) we estimate the error due to pixel assignment to be less than 1% for the total area estimates at the reach level and less than 2 cm for the reach-level WSE estimates.

3.2 Aggregate Nodes

3.2.1 Purpose

Following pixel assignment, information from individual pixels is combined and summarized at the node-level. The AggregateNodes algorithm aggregates L2_HR_PIXC pixel information to nodes using the pixel assignment output from the AssignPixels algorithm and input quantities from the pixel cloud processor. It computes node-level data including WSE, area, width, and their associated uncertainties and quality indicators. The results from AggregateNodes are written to the node-level shapefile of the L2_HR_RiverSP product.

3.2.2 Input Data

Description	Source
Pixel assignments to nodes	AssignPixels
Pixel 3-D geolocation, area, quality, and geophysical parameters	L2_HR_PIXC data product
Node separation distances and reach boundary definitions	Prior River Database

3.2.3 Output Data

Description
Node aggregated data

3.2.4 Mathematical Statement

Each node-assigned pixel from the AssignPixels algorithm has associated attributes derived from the pixel cloud (L2_HR_PIXC) processor, including pixel latitude and longitude, water class, height (with respect to the ellipsoid), area, observation time, atmospheric correction values, radar backscatter, uncertainty values, and quality indicators (see [8] for a detailed description of L2_HR_PIXC product). This section describes how these L2_HR_PIXC data are aggregated to the PRD node level in order to output river processing results to the L2_HR_RiverSP node shapefile.

3.2.4.1 PIXC quality handling

Input PIXC information is checked for quality before it is aggregated to river nodes. This is done using a quality handling module within the river processor. This module maps PIXC quality information to river data quality “states” that summarize their impact on river algorithms.

There are three internal state variables for each pixel that affect its aggregation to the node level. They are:

- WSE status: Possible values are good, suspect, degraded, and bad
- Area status: Possible values are good, suspect, degraded, and bad
- Sig0 status: Possible values are good, suspect, and bad

These internal states correspond to PIXC quality output parameters *classification_qual*, *geolocation_qual*, and *sig0_qual*, each of which is a bit flag. The internal states are defined using configurable bitwise mapping “words” (i.e. masks) for each quality parameter. Each mask specifies which bits from the corresponding input quality flag map to the associated internal state. Figure 9 shows how the quality state of a pixel affects its aggregation to the node level in the river processor. A summary of the algorithmic result for each state is also provided in Table 7, and a description of which PIXC quality bits are suspect, degraded, or bad is provided in [8].

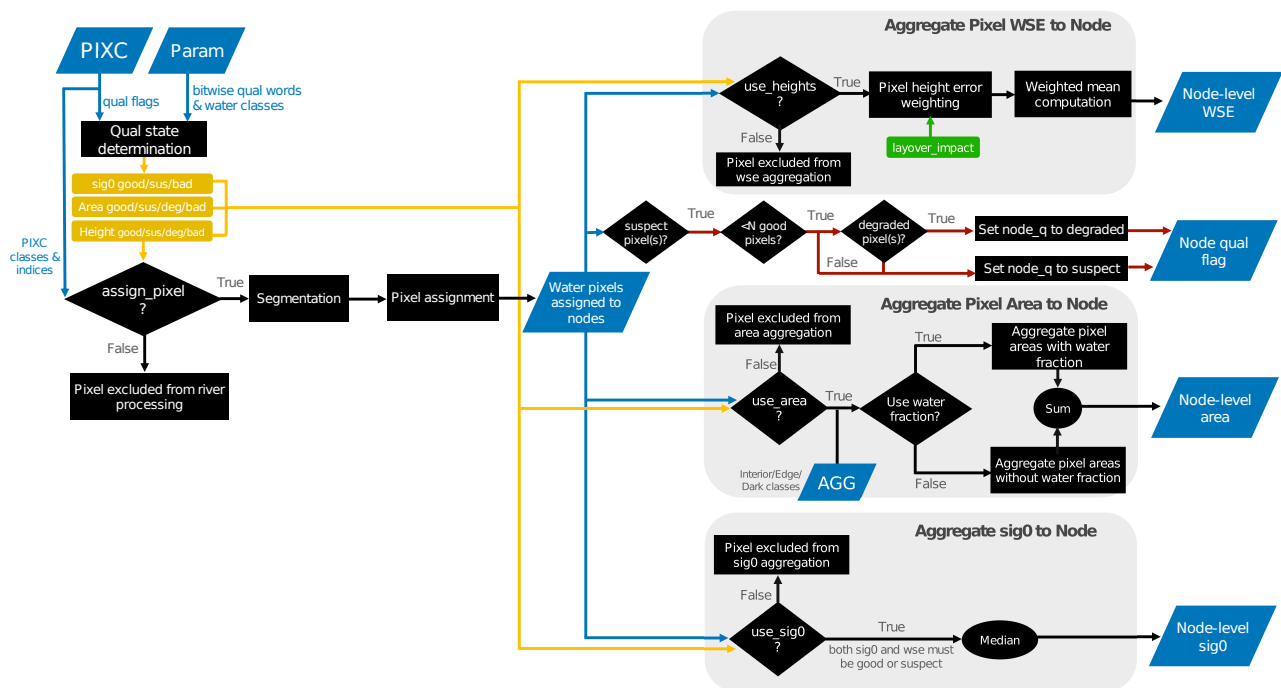


Figure 9: Quality flag handling in the L2_HR_RiverTile AggregateNodes algorithm. The L2_HR_PIXC product contains quality information in the form of the green indicators in the bottom left of the figure. These quality indicators are mapped to internal river quality “states” (yellow) that describe the expected quality of the area, height, and radar backscatter (denoted “sig0” in the diagram) measurements for each pixel. Generally, pixels

with bad quality are not used; pixels with suspect quality are aggregated normally (this results in a suspect output quality flag in the river products); pixels with degraded quality are not generally used unless there are no good or suspect pixels available (this results in a degraded output quality flag).

If *any* PIXC bit signifies bad area or WSE quality, then the pixel will be marked “bad” and is not used in pixel assignment nor aggregation. Otherwise, if *any* PIXC bit signifies degraded area or WSE quality, then the pixel will not be used to compute node-level area or WSE *unless* there are fewer than N good and suspect pixels available. Otherwise, if *any* PIXC bit signifies suspect quality, then the pixel will be treated as suspect in river processing. Suspect pixels are generally used in WSE and area aggregation but result in the node being flagged as suspect in the output (see Section 3.2.4.5 for a description of how node-level quality flags are set in the river processor). This threshold N is a configuration parameter whose nominal value is 1. Table 7 details the algorithmic result for each river quality state.

Table 7: Mapping from PIXC quality flags to internal state variables in the river quality module

River Config Param	PIXC qual reference	River Internal Qual State	Algorithm Result
geo_qual_wse_suspect	geolocation_qual	wse_suspect	Use in wse aggregation; mark geolocation_qual node_q_b bit as suspect; flag node as suspect
geo_qual_wse_degraded	geolocation_qual	wse_degraded	Use only if there are fewer than N good and suspect pixels; flag node as degraded if used.
geo_qual_wse_bad	geolocation_qual	wse_bad	Do not use pixel in wse aggregation
class_qual_area_suspect	classification_qual	area_suspect	Use in area aggregation; mark classification_qual node_q_b bit as suspect; flag node as suspect
class_qual_area_degraded	classification_qual	area_degraded	Use only if there are fewer than N good and suspect pixels; flag node as degraded if used.
class_qual_area_bad	classification_qual	area_bad	Do not use pixel in pixel assignment
sig0_suspect	sig0_qual	sig0_suspect	Use in sig0 aggregation; mark sig0_qual node_q_b bit as suspect; flag node as suspect
sig0_bad	sig0_qual	sig0_bad	Do not use pixel in sig0 aggregation

3.2.4.2 Area Aggregation

For the node-level area calculation, the AggregateNodes algorithm simply sums the estimated water area of each pixel assigned to the node in order to calculate total node area. The pixel areas are provided by the L2_HR_PIXC product. For pixels that are identified as water edge pixels, the estimated water area is equal to the pixel area multiplied by the estimated water

fraction for that pixel (also provided by the L2_HR_PIXC product). For pixels that are identified as interior (open) water, the water fraction is assumed by the river processor to always be 1 (even if that is not the value provided by the L2_HR_PIXC product). The total area is thus:

$$\hat{A}_f = \sum_x A(x) \hat{\alpha}(x) I_{f,\alpha}(x) \quad (1)$$

Where $A(x)$ is the area of each pixel x , $\hat{\alpha}$ is the estimated water fraction for each pixel, $I_{f,\alpha}$ is the pixel assignment indicator for pixels that are used in the area computation, and \hat{A}_f is the aggregated estimate of the water area for a given node.

The joint aggregation approach can be expressed as

$$A'_f = \sum_x A(x) (I_{dw,in}(x) + \hat{\alpha}(x) I_{de}(x)) I_f(x) \quad (2)$$

Where $I_{dw,in}(x)$ indicates detected water pixels that are interior to the water feature (i.e. not on the edge) and $I_{de}(x)$ indicates the detected edge pixels (both water and land edges).

Pixels passed from the L2_HR_PIXC processor each have a classification value describing the nature of the water pixel, including “*open_water*”, “*land_near_water*”, “*water_near_land*”, “*dark_water*”, etc. (see [7] for a more detailed description of all water classes). The node-level detected area (*area_detct*) calculation uses the *open_water* and *water_near_land* classes. This is the area that was directly detected by the SWOT observation, and is denoted by the *area_detct* attribute in the L2_HR_RiverSP node shapefile. For *open_water* pixels, the node-level area computation simply sums all pixel areas. For *water_near_land_pixels*, the water area is computed using the total pixel area scaled by the fractional inundation, which is a quantity from L2_HR_PIXC describing the amount of water relative to land for a given water edge pixel. *area_detct* does not include the area of water that was not observed directly by SWOT owing to a low radar echo level, which can occur over very smooth water surfaces or by significant attenuation of the radar signal due to propagation through rain, denoted as “dark water”. In the L2_HR_PIXC processing, areas of dark water are classified separately through the use of a prior water probability map. The sum of the *area_detct* and the area of the pixels classified as *dark_water* is given by *area_total*. This output contains the total area of the *area_detct* classes as well as *dark_water*, *low_coh_water_near_land*, and *low_coh_water*.

For a detailed mathematical description of the node-level area uncertainty computation, see [12].

3.2.4.3 Width Aggregation

The river width at the node is then computed by dividing the aggregated estimate of water area for a given node by the node length (*p_length*) provided in the PRD.

$$\hat{w}_f = \frac{\hat{A}_f}{ds_f} \quad (3)$$

where \hat{w}_f is the estimated node width, \hat{A}_f is the estimated water area, and ds_f is the node length from the PRD.

3.2.4.4 Height Aggregation

Pixels are aggregated to produce the node-level WSE using the heights of each pixel assigned to the node weighted by their uncertainties. All pixels that are both in the list of water classes usable for heights and also have good quality (see Sect. 3.2.4.1) will be used in height aggregation to node. Water classes included in the height aggregation are *open_water*, *low_coh_water_near_land*, and *low_coh_water*.

First, the geoid and tide corrections are applied to the ellipsoid-relative pixel-cloud heights to obtain the pixel WSE for all pixels:

$$h(x) = H(x) - \text{geoid_hght}(x) - \text{solid_tide}(x) - \text{load_tide}(x) - \text{pole_tide}(x)$$

where $h(x)$ is the WSE for each pixel-cloud pixel x , H is the ellipsoid-relative pixel height, *geoid_hght* is the geoid height relative to the ellipsoid, *solid_tide* is the solid-Earth (body) tide, *load_tide* is the load tide, and *pole_tide* is the pole tide. All quantities on the right-hand side of the equation above are reported for each pixel in the pixel-cloud product.

Next, the pixels are aggregated into node-level WSE values using a metric to estimate the mean of the height distribution over the feature. The optimal linear aggregator is the weighted average where the weights are chosen to minimize the variance (or uncertainty) of the estimate (which corresponds to inverse variance weighting of the pixels).

The general mathematical expression for the linear optimal aggregator is

$$\hat{h}_f = \sum_x w(x)h(x)I_{f,h}(x) \quad (4)$$

Where \hat{h}_f is the estimated node-level WSE, $h(x)$ is the WSE for pixel-cloud pixel x , $w(x)$ is the weight, and $I_{f,h}(x)$ is an indicator function (either 0 or 1) representing the pixels that are used for WSE aggregation. Assuming no correlation between pixels, the optimal weights are

$$\begin{aligned} w(x) &= \frac{\sigma_h^{-2}(x)}{\sum_x \sigma_h^{-2}(x)I_{f,h}(x)} \\ &= \frac{d_h^{-2}(x)\sigma_\phi^{-2}(x)}{\sum_x d_h^{-2}(x)\sigma_\phi^{-2}(x)I_{f,h}(x)} \end{aligned} \quad (5)$$

where

$$\sigma_h(x) = d_h(x)\sigma_\phi(x) \tag{6}$$

is the height noise standard deviation, $\sigma_\phi(x)$ is the phase noise standard deviation, and

$$d_h(x) = \frac{\delta h(x)}{\delta \phi} \tag{7}$$

is the height sensitivity to phase. $\sigma_\phi(x)$ and $d_h(x)$ are obtained from the pixel cloud product along with the height estimate.

Several other node-level quantities are computed using the weighted mean of all pixels used to compute the node-level WSE, where the weights used are identical to those used for the WSE aggregation. The node-level media-delay corrections for dry troposphere, wet troposphere, and ionosphere are aggregated in this way. The computation is identical for the various tide displacement models (see [1]). The node-level geoid height is also the weighted mean of the pixel-level geoid heights.

For a detailed mathematical description of the node-level height uncertainty computation, see [12].

3.2.4.5 Node Characteristics and Quality Indicators

The node shapefile product contains several outputs that describe the nature of the pixel-to-node aggregation, including *lat*, *lon*, *node_dist*, *area_wse*, *n_good_pix*, *partial_f*, *node_q*, and *node_q_b*.

The summary quality indicator checks numerous node and pixel-level parameters and propagates them to the *node_q* output. If any quality bit that is associated with bad data in *node_q_b* is 1, then *node_q* is set to bad. Otherwise, if any quality bit that is associated with degraded data in *node_q_b* is 1, then *node_q* is set to degraded. Otherwise, if any quality bit that is associated with suspect data in *node_q_b* is 1, then *node_q* is set to suspect. A detailed description of all node-level quality indicators and bit flags is provided in [1].

Nodes with “bad” quality are never used when aggregating their information to the reach level. Nodes with “degraded” quality will only be used if there are no “good” or “suspect” nodes available.

3.2.5 Accuracy

Node aggregation errors can be grouped into the following broad categories:

1. Pixel misassignment errors

2. PIXC quantity inaccuracies
3. Reach phenomenology biases

Pixel misassignment errors occur for reasons described in Section 3.1.2.5. They can cause errors at the node aggregation level if there are a large number of pixels misassigned to a given node or if the misassigned pixels have large errors. A typical example is when a tributary is observed by SWOT but is not in the PRD. Because the tributary pixels are connected, many pixels from the tributary might be erroneously assigned to the node at which the tributary joins the main stem. These can result in WSE errors up to several meters for a single node. Another common case is when a disconnected lake is misassigned to one or more nodes, causing width, area, and WSE errors following their aggregation.

The second category of error describes inaccurate quantities provided by the L2_HR_PIXC product. See [7] for a description of area and height estimation accuracy in the pixel cloud processor. The most common example of this type of error relates to inaccurate water fraction estimations for water edge pixels. If the water fraction estimates from L2_HR_PIXC are inaccurate then the aggregated water area will be wrong for the output L2_HR_RiverSP nodes.

The third class of error relates to the reach geometry itself. Because the pixel assignment algorithm uses the Euclidean distance to map pixels to nodes, the geometry of the centerline can sometimes result in pixel assignment “wedges” that do not span the full width of the river (Figure 10). This results in inaccurate node-level width estimations for affected nodes. This phenomenon does not tend to affect the reach-level width estimates because the errors across nodes tend to average out at the reach level. Nodes that are identified as “blocked” or having “blocked width” due to their geometry are flagged in the bitwise quality outputs.

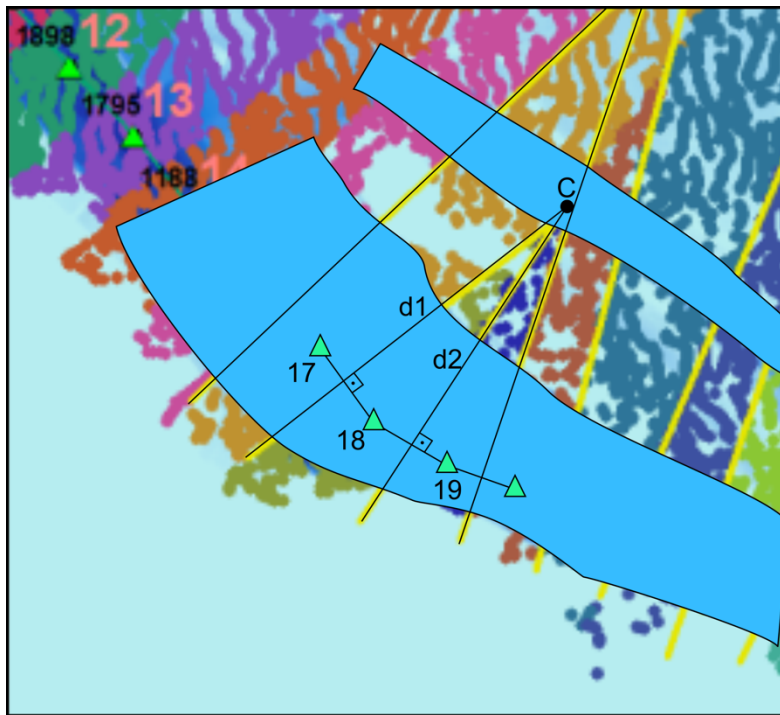


Figure 10: An example of a blocked width node. The blue polygons show the typical locations of the channels. The green triangles represent node locations. The marker “C” denotes the center of curvature of the main channel computed using node coordinates 17, 18, and 19.

3.3 Height-Constrained Geolocation

3.3.1 Purpose

The SWOT L2_HR_PIXC product provides an array of geolocated points identified as water pixels including latitude, longitude, and height information. Because of the SWOT viewing geometry, however, small height errors in the pixel 3-D locations can couple to large cross-track (horizontal) errors that may make pixel-level information difficult to use. To reduce geolocation error, L2_HR_River processing includes an algorithm to adjust the noisy L2_HR_PIXC geolocation of each river-assigned pixel using the aggregated node-level WSE information. This processing step is called “height-constrained geolocation”. The results of the new geolocation are used to produce pass-level height-constrained geolocations and smoothed heights in L2_HR_LakeSP processing.

The main idea underlying the improved geolocation algorithm is to replace the phase in the interferometric height reconstruction system (which uses the Doppler shift, slant-range, and absolute phase to geolocate a given pixel in the radar scene) with the processed and smoothed WSE fit to the reach. Thus, the noisy interferometric phase is replaced by a processed height computed using a fit to the node-level WSE values. By translating the pixels along the iso-range/Doppler contour from the position associated with the original, noisy, interferometric height to that associated with the smoothed, processed height, the horizontal geolocation of each river-assigned pixel is greatly improved at the expense of horizontal resolution in the estimated height. See the LakeSP ATBD for a more detailed algorithmic description of the height-constrained geolocation [10].

Because of the steep SWOT imaging geometry, a small error in height can introduce a large error in cross-track geolocation. The cross-track error is largest at low incidence angles at the near-range side of the swath, where it can be hundreds of meters. After a great deal of averaging, the height error can be reduced to centimeter scales and the cross-track geolocation error reduced to meter scales. Only pixels that were assigned to the river are averaged, though, so the slant-plane arrangement of these pixels preserves the horizontal resolution of the resulting feature shape. However, any small-scale variations in the WSE over the horizontal extent of the river are lost in the height-constrained geolocation result.

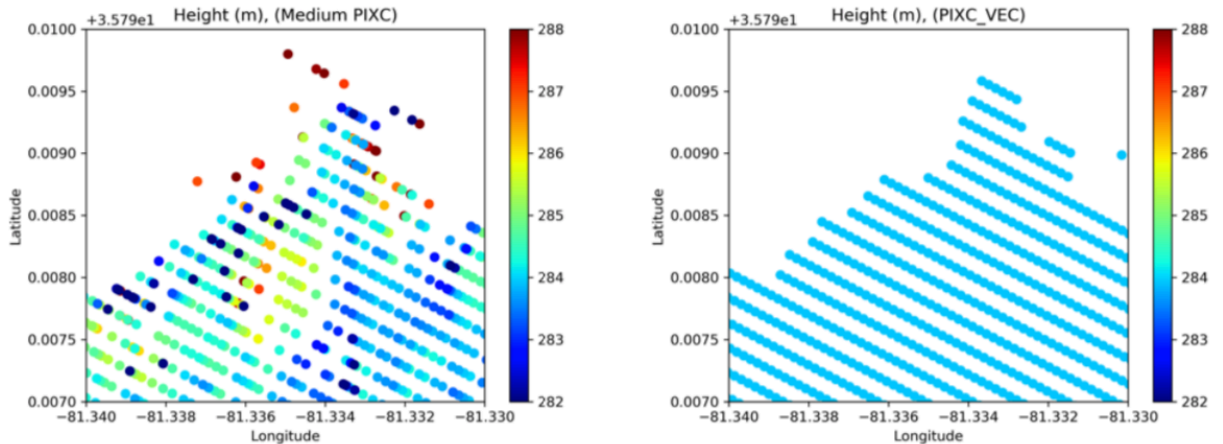


Figure 11: PIXC and PIXCVecRiver heights before and after height-constrained geolocation. (Left) Native pixel cloud heights with geolocation error. (Right) Pixel cloud after height-constrained geolocation.

3.3.2 Input Data

Description	Source
Mapping of L2_HR_PIXC pixels to nodes in the PRD.	Assign Pixels
PIXC geolocations and look geometry data	L2_HR_PIXC
Node aggregated WSE data	Aggregate Nodes

3.3.3 Output Data

Description
Updated, height-constrained geolocation (latitude, longitude, and height) of each pixel in the L2_HR_PIXC input files that was assigned to a node from the PRD.

3.3.4 Mathematical Statement

See the LakeSP ATBD [10] and CNES technical note [13] for a mathematical description of the height-constrained geolocation.

3.3.5 Accuracy

See LakeSP ATBD [10] for description of the height-constrained geolocation accuracy.

Note that the height-constrained geolocations are computed by assuming smooth WSE throughout the PRD reach. Therefore, the height-constrained geolocation values do not contain information about WSE variations at fine spatial scales even though they are finely sampled spatially.

3.4 Aggregate Reaches

3.4.1 Purpose

AggregateReaches combines node-level data to form reach-level area, cross-sectional area change, WSE, slope, and discharge values using the node information from AggregateNodes. This information is written to the reach shapefile comprising the L2_HR_RiverSP product.

3.4.2 Input Data

Description	Source
Node aggregated data	Aggregate Nodes
Node distances; reach boundary information; adjacent reach IDs; dam locations	Prior River Database

3.4.3 Output Data

Description
Reach aggregated data

3.4.4 Mathematical Statement

The reach-level area, WSE, cross-sectional area change, slope, and discharge values are computed along with their associated uncertainties. This section outlines each attribute computation sequentially in the same order that the processor computes them.

3.4.4.1 Reach-level area and width computation

First, the reach-level total area (*area_total*) is found by simply taking the sum of all node-level areas. Similarly, the reach-level detected area (*area_detct*) is computed by taking the sum of all node-level detected areas. We are not currently doing any outlier detection for the reach-level area computation. We also do not mask for the near or far range edges of each SWOT half swath.

$$\hat{A}_r = \sum_N \hat{A}_f(N) I_{r,A}(N) \tag{8}$$

Where $\hat{A}_f(N)$ is the area of each node N and $I_{r,A}$ is an indicator for nodes that were observed in the area computation, and \hat{A}_f is the aggregated estimate of the water area for a given node.

Next, the reach-level width \hat{w}_r (*width*) is calculated by dividing the reach-level total area by the reach-level length as provided by the PRD. The reach length ds_r is computed by summing the prior node lengths for all observed nodes in $I_{r,A}$.

$$\hat{w}_r = \frac{\hat{A}_r}{ds_r} \tag{9}$$

3.4.4.2 **Outlier detection and masking**

Node-level WSE outliers are detected and masked prior to reach-level WSE and slope estimation. Node-level height errors, if left undetected, could propagate to reach-level errors when deriving reach height profiles, thus affecting reach-level WSE estimation. Outlier detection and masking is designed to prevent this from occurring.

Outliers at the node-level may be caused by error generated during pixel assignment, water-water layover, or any instance where a population of pixels with anomalous heights are assigned to a single node (see Section 3.1.1.6 for a discussion of pixel assignment accuracy and [7] for a discussion of pixel-level WSE accuracy).

The node-level WSE anomalies are identified at each reach using an iterative piecewise linear regression approach. The piecewise algorithm finds breakpoints where the WSE gradient changes, which better handles non-linear height profiles and avoids false positives in the outlier rejection. More information on the piecewise regression package used for outlier detection can be found at [14]. River reach profiles are often non-linear in height (see Figure 12). As a result, a piecewise linear regression is used to obtain accurate outlier detection over reaches with non-linear height profiles. That is, each reach is split into several segments, each of which can be adequately approximated by a linear profile, and a linear regression is computed for each segment. After the regression and the fitting of the WSE as a function of the distance to the river outlet, the vertical distance from each node WSE to its fitted line is computed as a metric for outlier determination. The error metric is defined as:

$$residual(N) = abs(h_{fit}(N) - \hat{h}_f(N))$$

Where $residual(N)$ is the outlier error metric for node N , $h_{fit}(N)$ is the node-level fitted WSE, and $\hat{h}_f(N)$ is the estimated node-level WSE.

If the error metric from the piecewise fit of a given node is greater than a threshold, this node will be identified as an outlier. The threshold is set to be the greater of a configurable absolute threshold (nominally 1.5 m) and the 80th percentile of the error metric over all nodes in the reach. Therefore, no more than 20% of the nodes are ever discarded as outliers. This logic can be summarized by the following pseudocode:

```
algorithm flag_outliers(node_WSEs, node_ss, absolute_threshold):
inputs: node-level WSEs; flow distance to outlet for each node from database
outputs: outlier mask array, set to false if identified as outliers
```

```

FOR each reach in tile:
    CALL fitting tools
    SET metrics = abs(fitted_node_WSEs - node_WSEs)
    IF absolute_threshold < metric_80th_percentile THEN
        SET outlier_mask = metrics < metric_80th_percentile
    ELSE
        SET outlier_mask = metrics < absolute_threshold
    RETURN outlier_mask
    
```

Figure 12 below illustrates the outlier rejection algorithm behavior using both linear regression and piecewise linear regression.

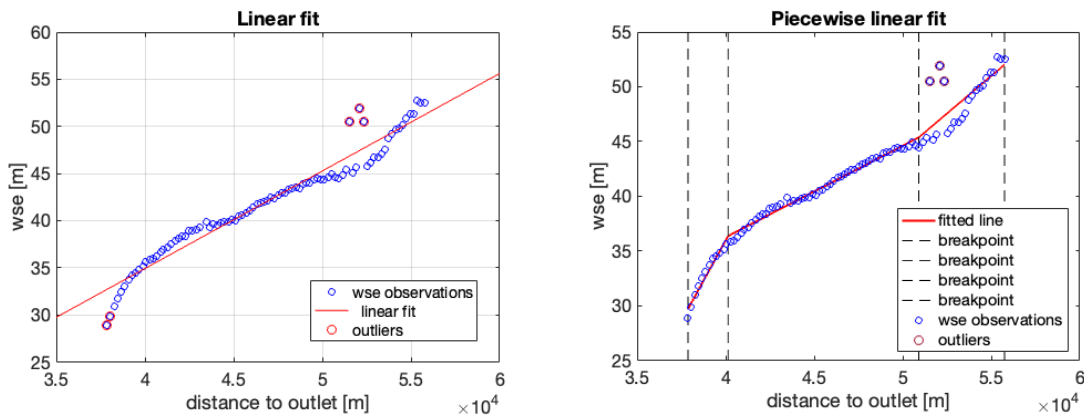


Figure 12: Example illustrating the piecewise linear outlier rejection algorithm. (Left) Linear fit to an example reach height profile. Due to the non-linear height profile of the reach, the two leftmost nodes marked in red are falsely rejected as outliers when the linear fit outlier algorithm is used. (Right) Piecewise linear fit to the same reach. The piecewise linear algorithm successfully rejects the outlier nodes while maintaining deviations from the fit due to curvature in the WSE profile.

3.4.4.3 Reach-level slope and WSE computation

Because the estimation of slope from noisy height data is generally very sensitive to the noise, the reach-level WSE and slope are computed by reconstructing the node-level heights into a smoothed river profile and then taking the mean WSE and slope of the smoothed heights. This algorithm is referred to as the Bayes Reconstruction algorithm.

The Bayes Reconstruction method for river reach WSE is a minimum-covariance smoothing algorithm that can also incorporate a prior-based bias correction to adjust the mean height over the entire reach (or, when available, multiple consecutive reaches). The prior can be constructed using actual prior data (which, in the case of SWOT, will not exist until it launches) or by estimating what the mean of the data should be using the observed data itself (e.g., take a linear fit to the observations and use that as the prior). The prior can be weighted or de-weighted depending on our confidence in it, thus tuning up or down its influence on the reconstruction.

The noisy measurement of node WSEs over a reach or reach series is modelled as a random process as

$$\vec{x} = H\vec{y} + \vec{v}$$

where \mathbf{H} is a sampling operator that maps the true node heights \vec{y} to the measured node heights subspace (spanned by \vec{x}) and \vec{v} is a zero-mean random process representing the measurement noise $\vec{v} = [v_1, v_2, \dots, v_N]^T$. Because $\bar{v} = E\{\vec{v}\} = 0$, the noise covariance is described by

$$\begin{aligned} cov(\vec{v}) &= E\{(\vec{v}\vec{v}^T)\} \\ &= R_v \end{aligned}$$

Given M noisy measurements \vec{x} of the N nodes over a reach or reach series, we can obtain an estimate of the true node heights \hat{y} using a Bayes framework that estimates the conditional expected loss given a distribution of the underlying random process and the measurement process. A common loss function is the mean-squared error (i.e. minimum covariance) which results in the estimate that is the conditional expected value of the true WSE given the measurements (i.e. the mean of the posterior distribution).

With this model, the estimate of all node heights \hat{y}_b can be expressed as

$$\begin{aligned} \hat{y}_b &= \underset{\hat{y}}{\operatorname{argmin}} E_{\vec{y}|\vec{x}} (\hat{y} - \vec{y})^T (\hat{y} - \vec{y}) \\ &= E\{\vec{y}|\vec{x}\} \\ &= \int \vec{y} f(\vec{y}|\vec{x}) d\vec{y} \end{aligned} \tag{10}$$

and it can be shown (see [15]) that the solution for the best estimate is

$$\hat{y}_b = K\hat{y} + \bar{K}\vec{x} \tag{11}$$

where K is the minimum covariance Bayes filter,

$$K = (R_y^{-1} + H^T R_v^{-1} H)^{-1} H^T R_v^{-1} \tag{12}$$

for the noisy measurements and

$$\bar{K} = (R_y^{-1} + H^T R_v^{-1} H)^{-1} R_y^{-1} \tag{13}$$

is the filter for the mean of the prior. Using these models combined with existing knowledge of what river profiles should look like, the posterior distribution of the node WSEs can be obtained.

3.4.4.3.1 Obtaining the posterior distribution

The Bayes Reconstruction method relies on knowledge of what the posterior distribution of the true river WSE profile should be. In our model, we explicitly assume knowledge of R_y , R_v , \bar{y} and \bar{v} and assume a multivariate Gaussian distribution, which is the maximum entropy (i.e. least information imposing) multivariate distribution with unconstrained extent.

To obtain the mean \bar{y} of the prior (over the distribution assumed by the Bayesian estimate), we look to the measurements themselves. In many contexts, the mean of the prior is assumed to be zero. For river reach estimates where several nodes in series may be missing, however, using a zero-mean prior generally produces unrealistic height profiles with sections of the river dropping towards zero height for the missing nodes, while also imposing smoothing across those discontinuities that make even the good measurements next to missing nodes drop to zero. Another simple approach for the prior mean \bar{y} would be to use the node WSEs from the prior river database (PRD), but the current prior height profile is expected to be a poor estimate of the absolute height (i.e. it is generally biased or offset in height). As we cannot currently trust the shape of the river profiles from the PRD, we have decided to initially use a weighted linear fit to the measured data itself as the prior. This results in a profile that is somewhere between a linear fit and just using the unfiltered measurements themselves as the estimate. For this case, the proximity of the resulting reconstruction to the linear fit versus the measurements can be controlled with various scaling and tuning parameters that are discussed below. Future implementations of the Bayes Reconstruction may use the PRD prior heights for \bar{y} as they improve and become more accurate with the availability of SWOT data.

The linear-fit prior is computed using the node-level WSE values (weighted by their respective WSE uncertainties) and the along-reach flow distance s as defined by *node_dist* in the PRD. Nodes masked due to bad or degraded quality are not used (see Section 3.2.4.5). The weights for the linear fit are defined as

$$w = \frac{1}{\sigma_{h,f}^2} \tag{14}$$

where $\sigma_{h,f}$ is the random uncertainty in the wse (*wse_r_u*) for a given node. See section 3.2.4.4 for a description of the node-level WSE uncertainties.

To obtain the autocovariance of the noise R_v we generate the autocovariance matrix of the random uncertainty of the node-level heights for all observed nodes (*wse_r_u*). To obtain the covariance of the expected true height profile R_y , we have assumed a k^{-2} exponential covariance model for river height profiles. This implies a red spectrum where higher wavenumbers occur less often, a structure that is common for many geophysical signals. Thus, we generate R_y using the exponential covariance function

$$c[n] = a \exp \frac{-|n|}{\tau}$$

(15)

which has a power spectrum that falls off as k^{-2} for wavenumbers larger than $1/\tau$. The full covariance matrix can be obtained from the exponential covariance by shifting the function to produce the Toeplitz (convolution) matrix.

In this model, a and τ can be estimated from the true height profiles or, alternatively, used as tuning parameters to control noise/resolution trade-off, where τ is the characteristic length of the imposition of structure and a is a relative scaling parameter to control (in conjunction with τ) how much the prior mean is weighted versus the measurements. To scale the covariance as a control for the noise/resolution and measurement/prior trade-offs, we simply normalize the covariance and scale by an imposed prior uncertainty. So, if we have a covariance matrix candidate $R_{y,0}$ and a desired uncertainty scalar u , we can obtain the covariance we wish to impose as

$$R_y = R_{y,0} \frac{u^2}{\max\{R_{y,0}\}} \quad (16)$$

Once we have \bar{y} , R_y and R_v we can compute the reconstructed heights $\hat{y}_b = E\{\vec{y}|\vec{x}\}$ as

$$\hat{y}_b = \bar{K}\bar{y} + K\vec{x} \quad (17)$$

The reach-level slope \hat{m}_b is then computed as the first-to-last slope over the reconstructed heights \hat{y}_b within the reach of interest:

$$\hat{m}_b = \frac{\hat{y}_{brN} - \hat{y}_{br1}}{ds} \quad (18)$$

where ds is the total flow distance spanned by the reach as defined by *node_dist* in the PRD. Note that edge node heights \hat{y}_{br1} and \hat{y}_{brN} for the current reach may be the reconstructed WSE for nodes that were masked and then filled in due to being unobserved, containing bad heights, or being rejected as outliers. The first-to-last definition of slope is chosen because it is equivalent to an unweighted mean of the slope over the reach.

The reach-level WSE is then estimated by taking the unweighted mean of \hat{y}_b over all reconstructed node heights within the reach of interest. The final output reach-level WSE is then

$$wse = \frac{1}{N} \sum \hat{y}_{br} \quad (19)$$

where, \hat{y}_{br} represents the reconstructed node height for nodes within the current reach of interest (not upstream or downstream reaches).

3.4.4.4 **Discharge calculation**

The reach and node-level width, height, and slope data generated by the river processor are used to compute various discharge outputs for the reach shapefile. For a description of each discharge output in the product, see [1]. For a detailed description of the different discharge algorithms and their performance assessments, see [16].

3.4.4.5 **Enhanced slope**

The reach shapefile of the L2_HR_RiverSP product also contains a second slope output commonly referred to as enhanced slope (*slope_2*). This algorithm smooths the node-level WSE across reach boundaries and then computes the slope using the finite difference between the first and last nodes of the reach of interest. See [17] for a detailed mathematical description of the enhanced slope algorithm.

The enhanced slope relies on the assumption that in the absence of natural or man-made discontinuities, e.g. dams or waterfalls, the water surface profile should be continuous and smooth across reach boundaries [17]. Thus, when a set of contiguous river reaches with no PRD flags for water discontinuities are simultaneously observed, the application of a smoothing operator across reach boundaries enables the determination of the water surface elevation at boundary nodes with a greater precision. The enhanced slope calculation comprises 5 steps: (1) assembling of the extended river height profile; (2) height profile flattening; (3) height smoothing; (4) profile reconstitution; (5) slope calculation, which are illustrated below in Figure 13.

The assembly of the extended river height profile is done by querying the PRD for the existence of adjacent river reaches. If the reach of interest has up- and/or downstream reaches that are not flagged as containing a dam or a waterfall, and if their nodes have valid height observations, these nodes are appended to the reach of interest. Nodes are sorted by node flow distance, yielding an extended profile containing up to three reaches, that is, the reach of interest (central), an upstream reach, and a downstream reach (Figure 13A). Next, the height profile is flattened by subtracting the least-squares slope from the extended profile (Figure 13B shown in green). Note that flattening described here should not be confused with the flattening of interferometric phase in other processing. The flattening of the profile allows for the mitigation of potential errors induced by the lack of a valid adjacent reach. The flattening process is followed by the application of a smoothing operator, which produces the smooth profile shown as triangles in Figure 13B.

The smoothing operator consists of a weighted moving average that accounts for all valid nodes located within a distance of 5 km to the central node measured along the river centerline in the up or downstream directions. The smoothed height at the central node (\bar{h}) is computed as:

$$\bar{h} = \frac{1}{\sum_{i=1}^N w_i} \sum_{i=1}^N w_i h_i$$

where N is the number of nodes inside the averaging window, w_i is the weight applied to a node i contained within the averaging window and h_i is the unsmoothed, flattened height of the node i . The weights used for the node height smoothing used in the enhanced slope calculation are computed using:

$$w_i = \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{s_i-s}{\sigma}\right)^2}$$

where s stands for the flow distance of the central node, s_i represents the flow distance of the node i contained within the averaging window, σ governs how quickly the weight decays with the node distance to the central point. The nominal value of σ is 2 km.

After the smoothing, the least squares slope removed in the flattening step is added back, resulting in the smooth profile in Figure 13C. This profile is used to compute the enhanced slope by dividing the difference in elevation between the last and first nodes ($\Delta\bar{h}$ in Figure 13C) in the central reach by the flow distance between them (Δs). Here, “last” denotes the node closest to the outlet, and “first” represents the node farthest from the outlet.

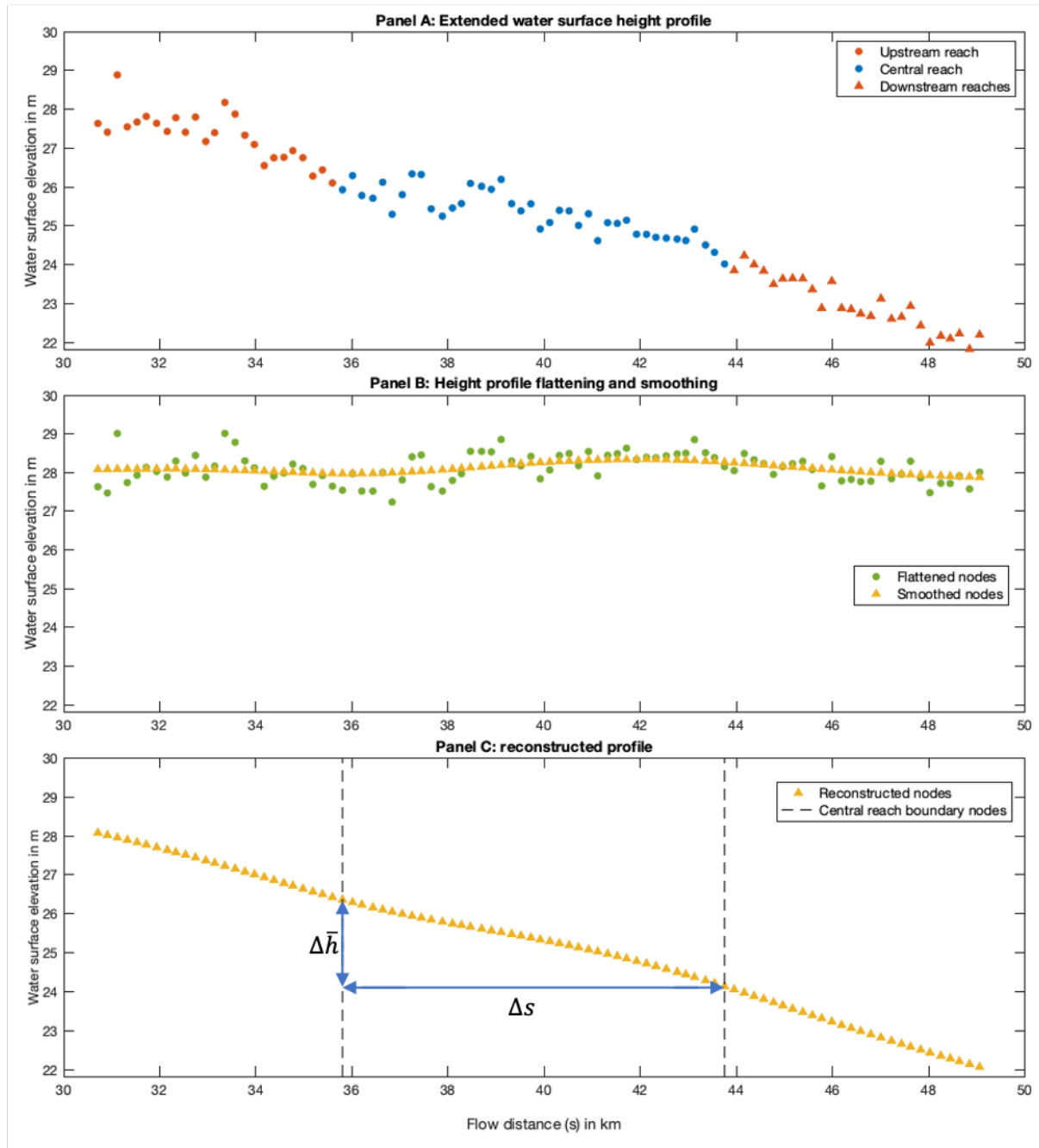


Figure 13. Main steps in the computation of the enhanced slope. A: Extension of the water surface profile by adding valid up and downstream reaches. B: height profile flattening and effect of the smoothing operator. C: reconstructed profile and identification of the change in elevation and flow distance used for the computation of the enhanced slope.

3.4.4.6 Reach Characteristics and Quality Indicators

The primary quality indicators for all reach-level information are the summary flag *reach_q* and its bitwise counterpart *reach_q_b*. Reach quality may have a value of good, suspect,

degraded, or bad. A detailed description of all reach-level quality indicators and bit flags is provided in [1].

4 Accuracy of L2_HR_River Algorithms

This section summarizes the overall accuracy of the L2_HR_River algorithms. Table 8 describes the simulated performance statistics at the reach level for all reaches in the representative dataset to which the scientific requirements are applicable (see Appendix B for a description of the representative dataset used in simulations of L2_HR_RiverTile performance). These statistics do not account for reach or node quality information (see Section 3.2.4.5 and Section 3.4.4.6) as many quality aspects were not meaningfully captured in the simulated dataset. Figure 14 shows the performance histograms for reach-level area, width, and slope and reach- and node-level WSE and is a graphical representation of the information in Table 8. Figure 15 contains the performance CDFs for each metric.

Metric	68%ile	50%ile	Mean	Reach Count
Area_total (%)	14.605	3.308	0.6277	341
Area_detct (%)	15.766	1.840	-1.727	341
WSE (cm)	7.696	0.826	-5.3606	341
Slope (cm/km)	1.046	-0.047	-1.181	341
Enhanced Slope (cm/km)	0.809	-0.008	-0.577	341

Table 8: Summary statistics for the L2_HR_RiverSP simulated nominal reach-level performance using L2_HR_PIXC simulated data from the representative dataset.

Figure 16 and Figure 17 show the simulated effect of reach area, reach width, and position in the swath on area, slope, and WSE performance. Relative area performance at the reach level generally improves towards the center of the swath and as the reach total area increases. Area performance degrades as reach area decreases and there is a positive bias in area error for small reaches in the near-swath. These experimental results are consistent with the analytic L2_HR_PIXC predictions of detected water performance described in [7] and is also seen in the area performance histogram in Figure 14. WSE performance is less sensitive to position in the swath and reach area for reaches above the required minimum area.

The total and detected area estimates have a positive bias of 3.3% and 1.8% in our simulated dataset respectively. Width estimates are also positively biased with a median error of +8 m. There are several potential sources of positive bias in the area and width estimation algorithms: non-PRD water pixel misassignment error; land pixel misassignment error; dark water over-detection; and L2_HR_PIXC water fraction error. Due to the design of the pixel assignment thresholds, misassignment error often maps extra pixels to node and almost never prunes river pixels incorrectly (see Section 3.1.2.5). This results in a slight over-estimate of area over the simulated dataset. The increase in relative area estimation bias from detected to total area suggests that dark water over-prediction may also contribute to the bias. See [7] for a description of the accuracy of L2_HR_PIXC quantities such as dark water and water fraction values. It is worth noting that the bias observed in the simulated data will likely differ from the performance of real SWOT observations as the simulated water maps do not model water bodies with perfect fidelity (and may over-represent tributaries, for example).

There is also a slight positive bias in reach- and node-level WSE estimations of 0.8 cm and 1.1 cm respectively. Because this is found at both the node and reach level it is likely that pixel assignment error is the main contributor. Pixel assignment errors often map non-PRD water pixels (which are often higher in elevation, such as tributaries) to the main channel and cause a positive bias in the WSE estimates. Land pixel misassignment also results in an over-estimation

of WSE. Despite the bias in WSE, the slope estimates are unbiased with a median signed error of 0.0 cm/km.

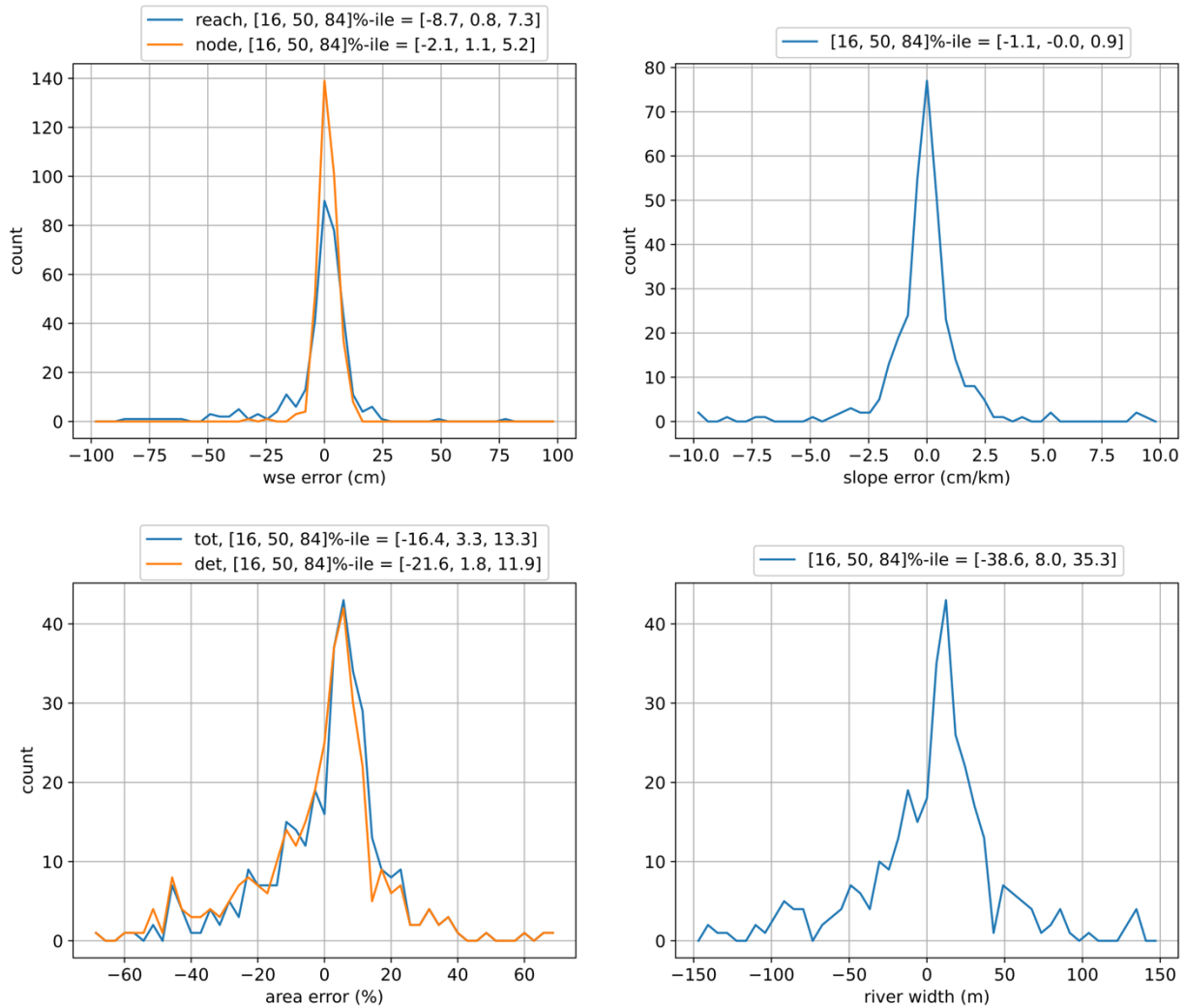


Figure 14: Simulated performance histograms for WSE, slope, area, and width from the representative dataset. The WSE histogram (upper left) contains both reach- and node- level data while all others contain reach-level data only.

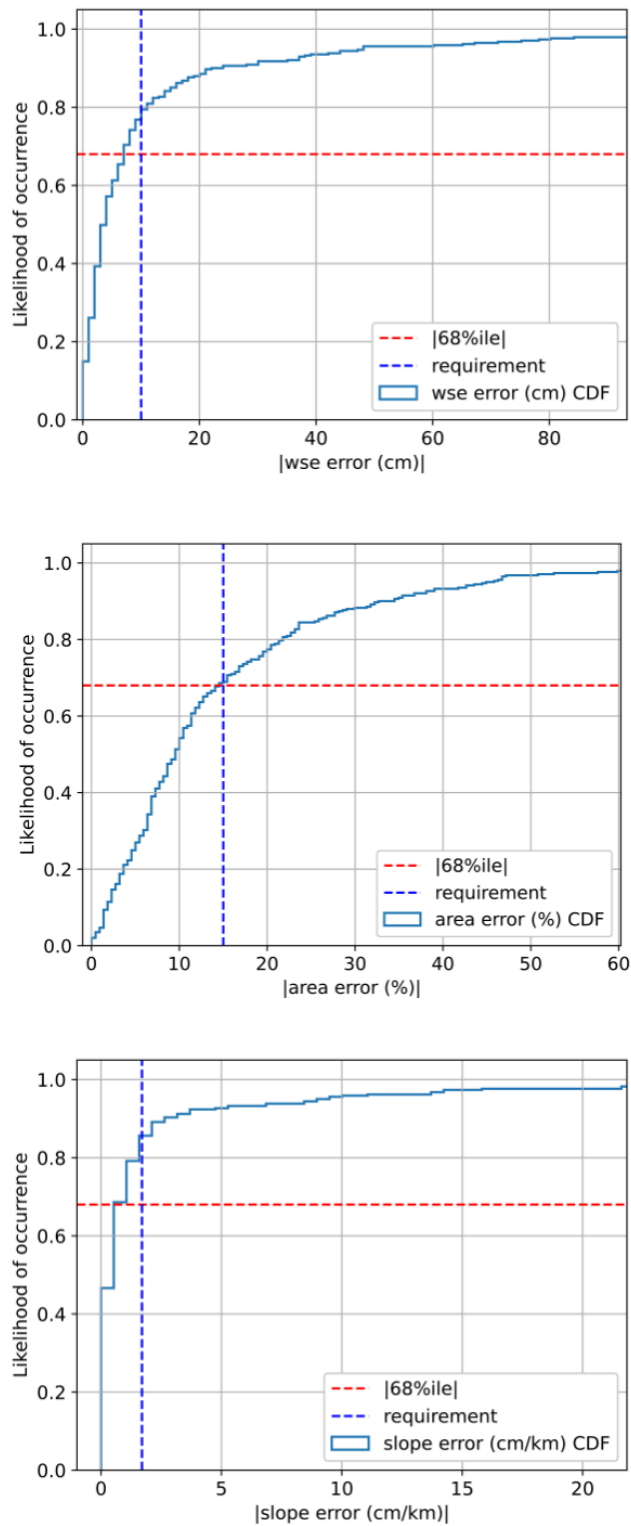


Figure 15: Simulated performance CDFs for the reach-level absolute error in WSE, total area, and slope from the representative dataset. Performance requirements and the 68th percentile error are also shown for each plot.

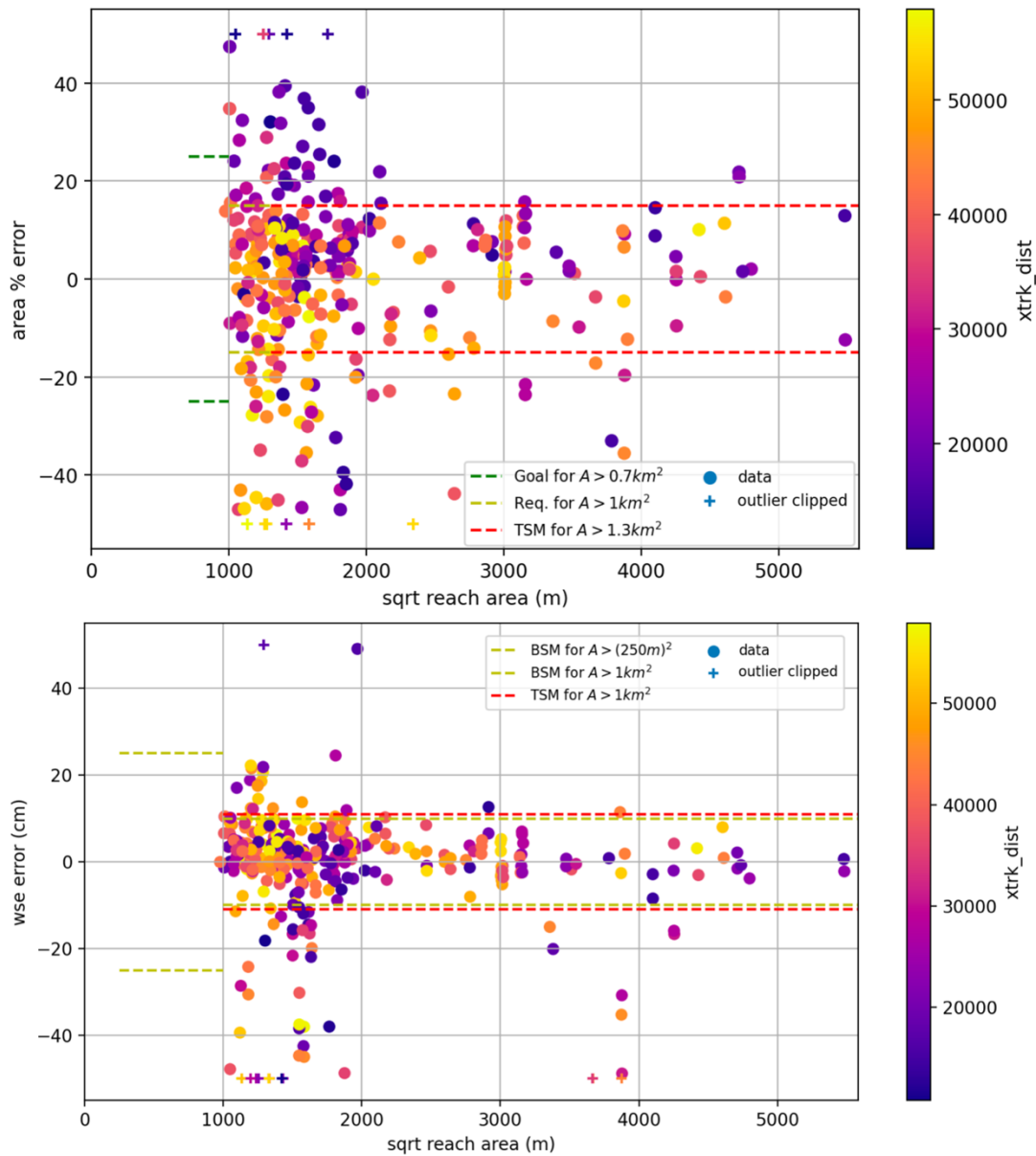


Figure 16: Reach-level area (upper) and WSE error (lower) versus reach-level area and position in the swath for nominal rivertiles in the representative dataset. Dotted lines show the TSM (threshold science mission), BSM (baseline science mission), and goal performance for each metric.

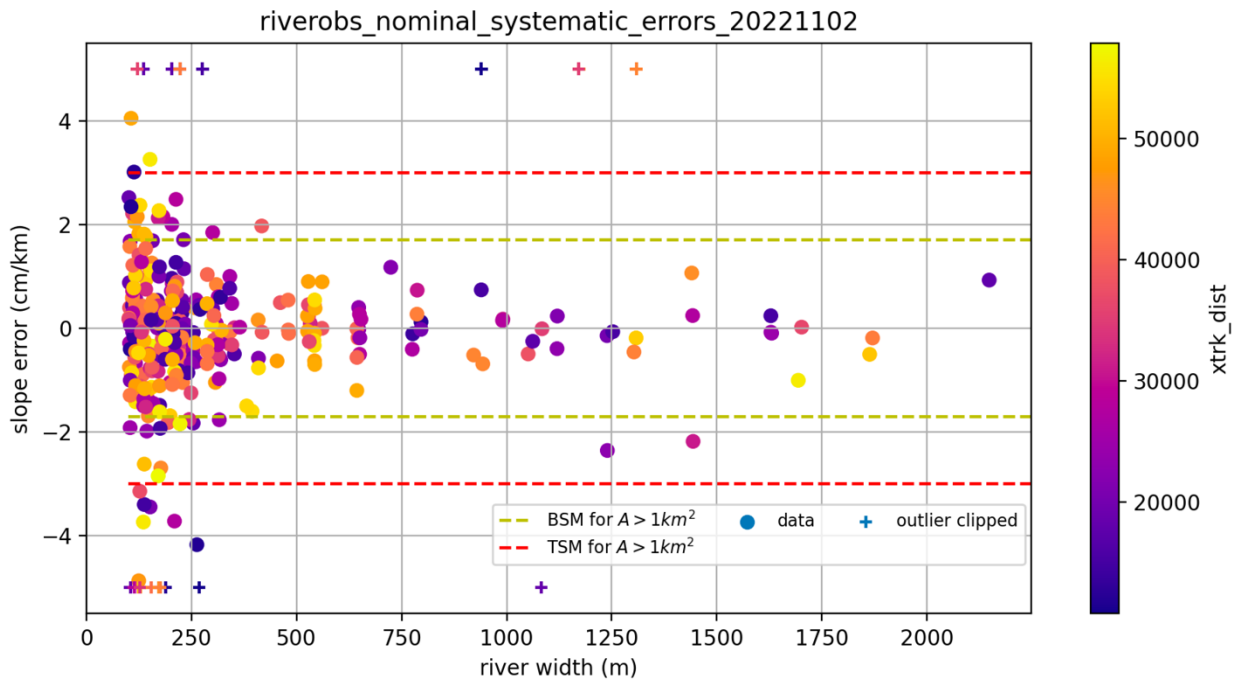


Figure 17: River slope performance versus river width and position in the swath for nominal rivertiles in the representative dataset. Dotted lines show the TSM (threshold science mission) and BSM (baseline science mission) performance requirements.

5 References

- [1] JPL D-56413, "SWOT Level 2 KaRIn High Rate River Single Pass Vector Product," Jet Propulsion Laboratory Internal Document, 2023.
- [2] JPL D-61923, "SWOT Science Requirements Document," Jet Propulsion Laboratory Internal Document, 2018.
- [3] JPL D-105518, "SWOT Prior River Database Auxiliary Data Description," Jet Propulsion Laboratory Internal Document, 2022.
- [4] C. Pottier and C. Stuurman, "SWOT Level 2 KaRIn High Rate Lake Single Pass Vector Product," SWOT-TN-CDM-0673-CNES, CNES, Toulouse, 2022.
- [5] E. Altenau, T. Pavelsky, M. Durand and R. Frasson, "SWORD Product Description Document," U. North Carolina, 2020.
- [6] K. L. Verdin and J. P. Verdin, "A topological system for delineation and codification of the Earth's river basins," *Journal of Hydrology*, vol. 218, no. 1-2, pp. 1-12, 1999, [https://doi.org/10.1016/S0022-1694\(99\)00011-6](https://doi.org/10.1016/S0022-1694(99)00011-6).
- [7] JPL D-105504, "SWOT Level 2 KaRIn High Rate Pixel Cloud Algorithm Theoretical Basis Document," Jet Propulsion Laboratory Internal Document, 2023.
- [8] JPL D-56411, "SWOT Level 2 KaRIn High Rate Water Mask Pixel Cloud Product," Jet Propulsion Laboratory Internal Document, 2023.
- [9] S. Maneewongvatana and D. M. Mount, Analysis of approximate nearest neighbor searching with clustered point sets, arXiv preprint cs/9901013, 1999.
- [10] C. Pottier and C. Stuurman, "SWOT Level 2 High Rate Lake Single Pass Algorithm Theoretical Basis Document," SWOT-NT-CDM-1753-CNES, CNES, Toulouse, 2023.
- [11] JPL D-56416, "SWOT Level 2 High Rate Raster Product," Jet Propulsion Laboratory Internal Document, 2023.
- [12] B. A. Williams, "SWOT Hydrology Height and Area Uncertainty Estimation," Jet Propulsion Laboratory Internal Document, Unpublished, 2018.
- [13] D. Desroches, "Height-constrained geolocation," *DTN/TPI/TR-2022/00500*, CNES, no. Technical Note, 2022.
- [14] C. Pilgrim, "Piecewise Regression (aka segmented regression) in Python," *The Journal of Open Source Software*, vol. 6, no. 68, p. 3859, 2021.
- [15] B. A. Williams, "Optimal River Height Profile Reconstruction from SWOT Measurements," Jet Propulsion Laboratory Internal Document, Unpublished, 2021.
- [16] R. P. Frasson, M. T. Durand, K. Larnier, C. Gleason, K. M. Andreadis and M. Hagemann, "Exploring the factors controlling the error characteristics of the Surface Water and Ocean Topography mission discharge estimates," *Water Resources Research*, vol. 57, no. e2020WR028519, 2021.
- [17] R. Prata de Moraes Frasson, R. Wei, M. J. Durand and J. Minear, "Automated River Reach Definition Strategies: Applications for the Surface Water and Ocean Topography Mission," *Water Resources Research*, vol. 53, no. 10, pp. 8164-8186, 2017.

Appendix A. **Acronyms**

AD	Applicable Document
AMR	Advanced Microwave Radiometer
API	Application Interface
ATBD	Algorithm Theoretical Basis Document
CNES	Centre National d'Études Spatiales
JPL	Jet Propulsion Laboratory
NASA	National Aeronautics and Space Administration
OBP	On-Board Processor
PGE	Product Generation Executable
PIXC	L2_HR_Pixel Cloud product
PRD	Prior River Database
RD	Reference Document
SAS	Science Algorithm Software
SDS	Science Data System
SWOT	Surface Water Ocean Topography
TBC	To Be Confirmed
TBD	To Be Determined
WSE	Water Surface Elevation

Appendix B. Simulations

All simulated L2_HR_RiverSP outputs were created using ideal and nominal simulated L2_HR_PIXC products. See [7] for a description of the representative dataset simulation process up to the beginning of L2_HR_RiverTile processing.

The representative dataset contains a total of 47 unique simulated scenes covering 118 unique tiles observed by 74 SWOT passes, yielding 359 valid (i.e., containing river data) scene-pass-tile combinations in total. When intersecting these simulated data with the PRD [5], the representative dataset contains 717 unique river reaches and 1857 reach-passes in total. Of these, there are 181 unique river reaches and 341 reach-passes that meet filtering criteria based on the applicability of the science requirement for SWOT [2]. Simulated reaches must meet a minimum width, length, and area to be included in the river performance statistics. They must also be located between 10 km and 60 km cross track. That is, if any portion of the reach is outside this extent, the reach is not considered in the performance statistics. Note that useful data may still be reported for such reaches in the SWOT products, but the reaches are ignored when compiling the performance statistics reported in this document. A summary of the filtering criteria and the number of reaches that are excluded based on these criteria is provided in Table 9. Statistics of the representative dataset characteristics before and after filtering are provided in Figure 18.

Note that some of the criteria are less restrictive than the applicability of the science requirements in order to preserve as many reaches as possible for evaluation. For example, the minimum reach area and length criteria have thresholds of 0.8 km² and 8 km, whereas the science requirements are applicable to reaches that have areas and lengths of at least 1 km² and 10 km. Performance generally improves with increased reach area and length, so the lower thresholds used here are conservative with respect to performance analyses.

Simulated SWOT performance estimates require both truth and nominal processed data. River “truth” data were generated by evenly distributing water observation pixels over truth water masks and assigning WSEs to each pixel from the truth heights (based on airborne lidar data) used as inputs to the simulation in order to form an artificial L2_HR_PIXC product. Directly mapping truth heights to pixel heights eliminates sources of error due to LR_HR_PIXC or L1_HR_SLC processing. These artificial L2_HR_PIXC products are then processed through L2_HR_RiverTile processing to create “truth” L2_HR_RiverTile products. All truth river profiles in the representative dataset were manually reviewed by experienced hydrologists and non-physical or incomplete profiles were removed. Truth profiles may be spurious due to unrealistic height profiles resulting from artifacts in the height truth, inaccurate “truth” water masks, or discrepancies between the water elevation and extent. Moreover, unrealistic discrepancies between the truth data and the reference data due to temporal changes (e.g., river migration) are possible.

It is important to note that the “truth” rivertiles are processed with different configuration parameters than the “nominal” tiles. Because smoothing on the truth node heights is unnecessary, the slope algorithm used for truth processed reaches is a simple “first-to-last” algorithm that computes the slope using the first and last node heights of the truth processed reach. Outlier rejection is also not performed on the truth processed reaches.

Filter type	Filtering criterion	Number of reach-passes excluded from the representative dataset due to each criterion
Area	Minimum 0.8 km ²	738
Length	Minimum 8 km	508
Width	Minimum 100 m	752
Cross-track position	Between 10 km and 60 km	1072
Spurious truth data	Unrealistic or incomplete water height profiles	519

Table 9: Representative dataset filtering criteria for SWOT river scientific requirements. Rivers must meet a minimum width, length, and area in order to be included in the performance assessment statistics. They must also be located entirely within the science requirement bounds for cross-track position. The truth simulations must also be free of artifacts or non-physical height profiles. This table summarizes the number of reaches removed from the representative dataset for each criterion individually. The sum of the values in the rightmost column is greater than the total number of reaches excluded because some reaches fail multiple criteria simultaneously.

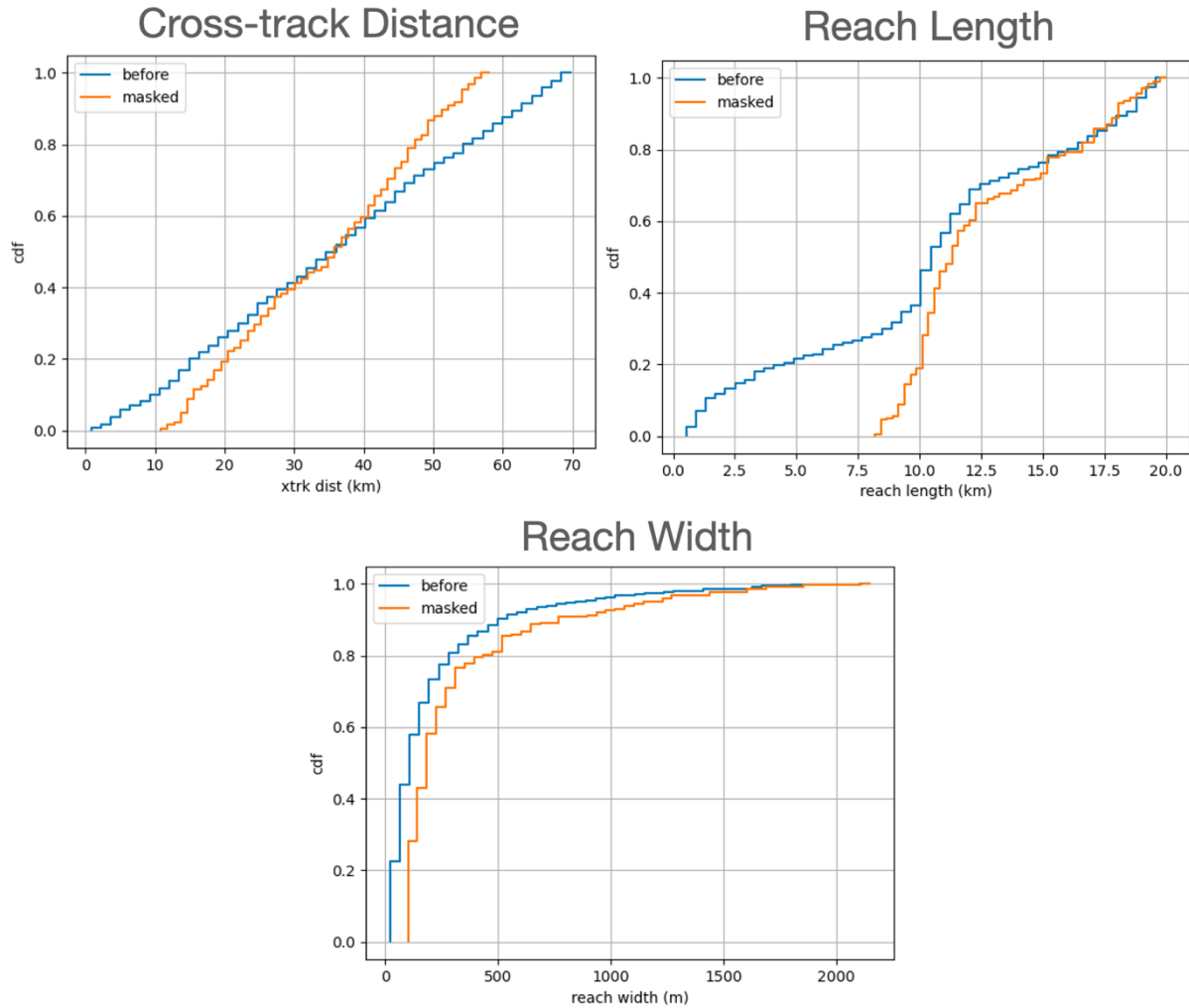


Figure 18: Statistics of the representative dataset before and after filtering for reach geometry and position in the swath. Each plot shows the before/after masking CDF for a given feature, where “before” corresponds to all 1857 reach-passes in the representative dataset and “masked” corresponds to the filtered dataset with 359 reach-passes.