Surface Water and Ocean Topography (SWOT) Project

SWOT Auxiliary Data Description Document
Long Name: Prior Lake Database
Short Name: LakeDatabase

Revision C

Document Custodian:

Pottier Claire
Signature numérique de Pottier Claire
Date : 2025.04.03 14:45:38 +02'00'

Claire Pottier
CNES Algorithm Engineer

Date

Approved by:

Fjortoft Roger
Signature numérique de Fjortoft Roger
Date : 2025.04.02 09:39:38 +02'00'

Roger Fjørtoft
CNES Algorithm System Engineer

Date

Concurred by:

Brunet Pierre-Marie
Signature numérique de Brunet Pierre-Marie
Date : 2025.04.02 09:53:28 +02'00'

Lionel Zawadzki
CNES Hydrology Distribution Center Manager

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 Engineering Product Data Management System (EPDM: https://epdm.jpl.nasa.gov).

See Sect 1.3 for how to cite this document

March 31, 2025
SWOT-IS-CDM-1944-CNES

cnes

CHANGE LOG

VERSION	DATE	SECTIONS CHANGED	REASON FOR CHANGE
Initial Release	2022-03-16	All	Initial Release
Revision A	2023-09-22	All	Revision A, addition of some attributes and one table
Revision B	2023-12-08	5.1 Appendix B Appendix C	Addition of coefficients to compute storage change for 26 lakes A more detailed description of the parent lake database is provided. Addition of description of specific lake database files distributed to end users
Revision C	2025-03-31	1.1 1.3 2.3 3.1.1 3.2, 3.3 4.1 5.1 5.2 Appendix B	Additional and updated data sources How to cite this document Size of files Coordinates removed from “basin” table for pass-granule LakeDatabase Figure 3 updated, Figure 4 added New data sources, updated *ref_wse*, *date_t0*, and *ds_t0* definitions Data history table updated Updated graphs and figures Addition of flags to estimate the stability of the lake from Pekel occurrence mask

Table of Contents

CHANGE LOG **2**

Table of Contents **3**

Table of Tables **5**

Table of Figures **5**

List of TBC Items **6**

List of TBD Items **6**

1 Introduction **7**
1.1 Purpose **7**
1.2 Document Organization **8**
1.3 Citing This Document **8**

2 General Description of Auxiliary Files **9**
2.1 Auxiliary File Format **9**
2.2 File Naming Convention **9**
2.3 Size of files **10**

3 LakeDatabase Files **11**
3.1 Overview **11**
3.1.1 Pass-granule LakeDatabase 11
3.1.2 Basin-granule LakeDatabase 12
3.2 The “lake” table **12**
3.3 The “lake_influence” table (pass-granule only) **14**
3.4 The “lake_catchment” table (pass-granule only) **14**
3.5 The “basin” table **15**
3.6 The “pass_tile” table (pass-granule only) **15**
3.7 The “hypso_curve” table **15**

4 Population of the LakeDatabase **16**
4.1 The “lake” table **16**
4.2 The “lake_influence” table **19**
4.3 The “lake_catchment” table **20**
4.4 The “basin” table **20**
4.5 The “pass_tile” table **20**
4.6 The “hypso_curve” table **21**

5 Notes **22**
5.1 Data history **22**
5.2 Statistics **23**
5.2.1 Distribution of lakes per continent 23
5.2.2 Lake number and size 23
5.2.3 Cycle flag distribution 24
5.2.4 Coverage of large lakes per pass 25
5.2.5 Minimum distance to neighbor lake 26

6 References **28**

Appendix A. **Acronyms 30**

Appendix B. **Parent LakeDatabase 31**

Appendix C. **Science LakeDatabase 34**

Table of Tables

TABLE 1. DATA VOLUME OF EACH GRANULE TYPE OF THE LAKEDATABASE AUXILIARY PRODUCT. 10
TABLE 2. WATER BODY TYPE CODES FOR THE *LAKE_ID* ATTRIBUTE. 17
TABLE 3. CONTINENT CODES FOR THE *LAKE_ID* ATTRIBUTE, AND CONTINENT IDS FOR THE BASIN-GRANULE LAKEDATABASE FILENAMES. 17
TABLE 4. CONTENT OF LAKEDATABASE VERSIONS. 22
TABLE 5. DISTRIBUTION OF LAKES PER CONTINENT. 23
TABLE 6. STATISTICS ON MINIMUM DISTANCE TO NEIGHBOR LAKES PER CONTINENT. 26
TABLE 7. DATA VOLUME OF THE SCIENCE LAKEDATABASE. 35

Table of Figures

FIGURE 1. MODEL OF THE PASS-GRANULE LAKEDATABASE. 11
FIGURE 2. MODEL OF THE BASIN-GRANULE LAKEDATABASE. 12
FIGURE 3. PLD LAKES (BLUE) AND THEIR INFLUENCE AREA (ORANGE). 14
FIGURE 4. PLD LAKES (BLUE) AND THEIR CATCHMENT AREA (ORANGE). 14
FIGURE 5. HYPSOMETRIC CURVE FOR THE ORIENT LAKE. 16
FIGURE 6. GEOGRAPHICAL DELINEATION OF CONTINENTS FROM HYDROBASINS [4]. 17
FIGURE 7. EXAMPLE OF THE PFAFSTETTER BASIN CODING SYSTEM OVER NORTH AMERICA, LIMITED TO THE 3 FIRST LEVELS USED IN THE IDENTIFIERS OF THE LAKE PRODUCTS. 18
FIGURE 8. ILLUSTRATION OF THE COMPUTATION OF THE INFLUENCE AREAS OF TWO LAKES. 20
FIGURE 9. NUMBER OF LAKES AND MAXIMUM AND MEAN LAKE SIZE FOR EACH CONTINENT. 24
FIGURE 10. GLOBAL DISTRIBUTION OF CYCLE FLAG. 24
FIGURE 11. CYCLE FLAG DISTRIBUTION BY CONTINENT. 25
FIGURE 12. LAKES COVERED BY THE GREATEST NUMBER OF TILES FOR A SINGLE PASS. 26
FIGURE 13. CUMULATIVE DISTRIBUTION FUNCTION OF DISTANCE TO CLOSEST LAKE FOR DIFFERENT CATEGORIES OF SQRT(LAKE AREA). 27
FIGURE 14. MODEL OF THE PARENT LAKEDATABASE. 31
FIGURE 15. MODEL OF THE SCIENCE LAKEDATABASE. 34

List of TBC Items

These items are to be completed when document is ready to enter configuration control.

Page	Section

List of TBD Items

These items are to be completed when document is ready to enter configuration control.

Page	Section
13	§3.2: format of *ice_dyn_flag* of table “lake”

1 Introduction

1.1 Purpose

This document describes auxiliary data files that represent the Prior Lake Database (PLD). This dataset is also referenced by the short name LakeDatabase. The PLD provides prior data on known lakes, making it possible to link SWOT lake observations over time and to compute storage change.

The current version of the LakeDatabase is mainly based on the

- CIRCA-2015 lake water extent map [1],
- Georeferenced global Dams And Reservoirs dataset (GeoDAR) v1.1 [2],
- Post2k reservoir dataset (Fan et al., 2023) [3],
- And also specific lakes provided by scientists [4],

but also relies on the following datasets:

- The Global Reservoir and Dam database (GRanD) [5] [6]
- The HydroBASINS dataset [7]
- The SWOT Prior River Database (RiverDatabase) SWORD [8] [9] [10]
- The Lake-TopoCat dataset [11]
- SWOT Orbits [12]:
    - 0-80km width swath was considered to build the pass-granule PLD (cf. Section 3.6)
    - 10-60km width swath was considered to compute the fields related to the lake coverage by SWOT in the “lake” table (see Section 4.1)
- Several databases for lake names:
    - The IGN Carthage database (covering France) [13]
    - Open Street Map (OSM) [14]
    - Global Lakes and Wetlands Database (GLWD) [15]
    - Natural Earth Data [16]
    - Vector Map Level 0 (VMap0) [17]
    - HydroLAKES [18]

The LakeDatabase files are organized in two different ways:

- by pass number, both for the calibration orbit and the nominal science orbit,
- by basin, as defined by the HydroBASINS Level-2 dataset [7].

The three Science Algorithm Software (SAS) processors dedicated to lake products use this auxiliary dataset:

- Both Level 2 KaRIn high rate lake tile science algorithm software

(SAS_L2_HR_LakeTile) and Level 2 KaRIn high rate lake single pass science algorithm software (SAS_L2_HR_LakeSP) [19] use the pass-granule LakeDatabase,

- The Level 2 KaRIn high rate lake average science algorithm software (SAS_L2_HR_LakeAvg) uses the basin-granule LakeDatabase.

1.2 Document Organization

Section 2 provides a general description of the LakeDatabase and its application.

Section 3 describes the LakeDatabase files.

Section 4 details how the LakeDatabase is populated.

Section 5 summarizes the data history and presents some statistics.

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

Appendix B describes the parent files used to generate the operational LakeDatabase.

Appendix C describes the LakeDatabase files distributed to the science community.

1.3 Citing This Document

This document should be cited as a corporate document, without the names of the document custodians, e.g.:

> SWOT-IS-CDM-1944-CNES, Revision C, "SWOT Auxiliary Data Description Document: Prior Lake Database (LakeDatabase)," Centre National d'Etudes Spatiales, 2025.

or

> Centre National d'Etudes Spatiales, "SWOT Auxiliary Data Description Document: Prior Lake Database (LakeDatabase)," SWOT-IS-CDM-1944-CNES, Revision C, 2025.

2 General Description of Auxiliary Files

2.1 Auxiliary File Format

The LakeDatabase adopts the SQLite format [20]. SQLite is an embedded SQL database engine. Unlike most other SQL databases, SQLite does not have a separate server process. SQLite reads and writes directly to ordinary disk files. A complete SQL database with multiple tables, indices, triggers, and views, is contained in a single disk file. SQLite files are directly readable by most GIS software.

In addition to the LakeDatabase files described in the next sections, used for operational processing of SWOT data, Appendix C describes the LakeDatabase files that are made available to the science community and other end users, and that contain some additional attributes.

2.2 File Naming Convention

Each LakeDatabase file adopts the following file naming convention:

SWOT_LakeDatabase_<Granule>_<ValidityBeginningDateTime>_<ValidityEndingDateTime>_<CreationDateTime>_<Version>.sqlite

- <*Granule*> takes the form:
  - <*Orbit*>_<*PassID*> for pass-granule LakeDatabase files, as defined below:
    - <*Orbit*> is either “Cal” or “Nom” for the calibration and nominal orbit phases, respectively. Note that pass is unique within a given orbit phase.
    - <*PassID*> is the zero-padded, three-digit, numeric pass number of the granule. Calibration orbit contains 28 passes from ‘001’ to ‘028’ and nominal orbit contains 584 passes from ‘001’ to ‘584’.
  - <*Continent*>_<*BasinID*> for basin-granule LakeDatabase files, as defined below:
    - <*Continent*> is a two-letter continent identifier, as defined in Table 3.
    - <*BasinID*> corresponds to the basin number in the continent (i.e. the two first digits of the *lake_id*; see section 4.1).
- <*ValidityBeginningDateTime*> gives the UTC date and time of SWOT observations in the form *YYYYMMDD*T*hhmmss* after which the information in the file becomes valid. For static auxiliary data, this field may be populated by a date and time before launch such as 20000101T000000.
- <*ValidityEndingDateTime*> gives the UTC date and time of SWOT observations in the form *YYYYMMDD*T*hhmmss* after which the information in the file is no longer valid. For static auxiliary data, this field may be populated by a date and time far beyond the expected life of the mission such as 20991231T235959.
- <*CreationDateTime*> gives the UTC file creation date and time in the form YYYYMMDDThhmmss.
- <*Version*> is a string that takes the form “vMmm” where “v” is a fixed character that

denotes the version field, and Mmm are three numeric digits. The first (major) digit (M) describes the version of the file format, and the latter two (minor) digits (mm) describe the version of the file contents (e.g. “v203”).

2.3 Size of files

There are 28 pass-granule files for the calibration orbit, 584 pass-granule files for the nominal orbit and 61 basin-granule files.

Table 1 provides the expected size of the LakeDatabase files and the overall data volume (uncompressed).

Table 1. Data volume of each granule type of the LakeDatabase auxiliary product.

Granule	Statistic	Number of lakes	Volume/Granule (GB)
Cal	Minimum (pass id)	**4** (005)	**0.0057** (005)
	Median	**43592**	**0.21**
	Maximum (pass id)	**148948** (024)	**0.71** (024)
		For the granule type	**7.05**
Nom	Minimum (pass id)	**1** (365)	**0.0057** (059;337)
	Median	**34909**	**0.19**
	Maximum (pass id)	**288438** (160)	**1.2** (466)
		For the granule type	**154.14**
Basin	Minimum (basin CB)	**22** (54)	**0.0054** (54)
	Median	**29538**	**0.054**
	Maximum (basin CB)	**832732** (72)	**1.40** (72)
		For the granule type	**9.23**
		Total	**170.4**

3 LakeDatabase Files

3.1 Overview

The operational LakeDatabase files described in Sections 3.1.1 and 3.1.2 are built from a parent database detailed in Appendix B.

3.1.1 Pass-granule LakeDatabase

The pass-granule LakeDatabase files follow the model given in Figure 1 below. Its tables and attributes are derived from the parent database detailed in Appendix B and keep only the elements needed for LakeTile and LakeSP processes.

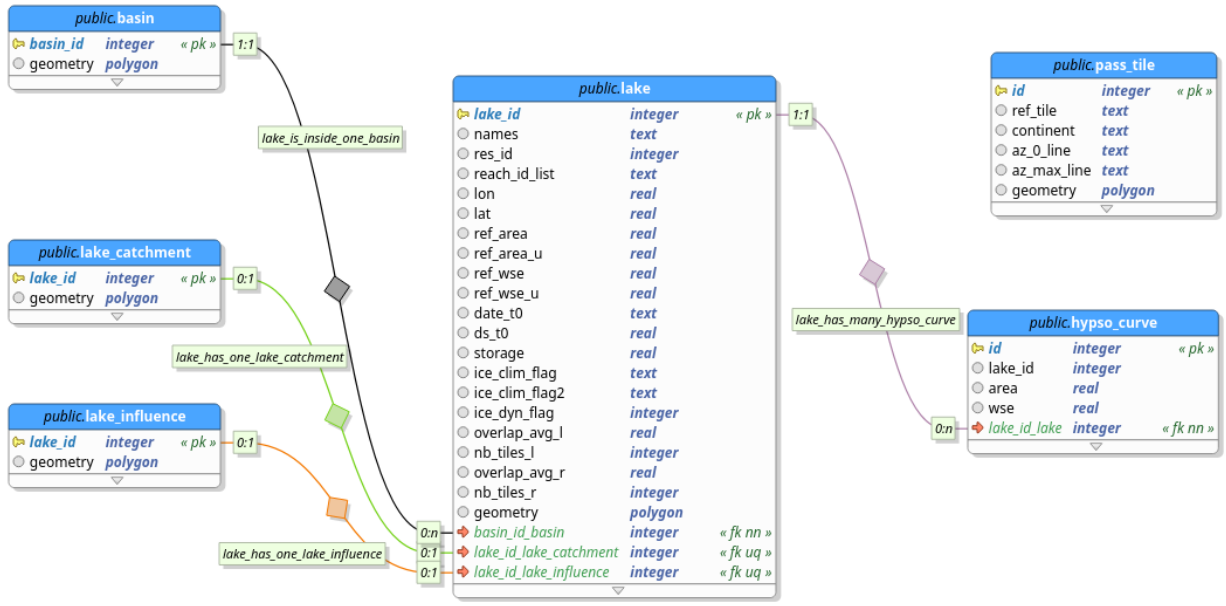


Figure 1. Model of the pass-granule LakeDatabase.

The role of these tables is as follows:

- The “lake” table contains a polygon and a set of attributes for each prior lake located on the selected orbit and pass. It is used to link observed polygons to prior lakes, by intersecting their geometries. It also contains basic information to populate the L2_HR_LakeTile [21] and L2_HR_LakeSP [22] products and to compute storage change. This table also contains an attribute to link the LakeDatabase and the Prior River Database (RiverDatabase) [8] [9] [10].
- The “lake_influence” geometries delineate the area of influence of each of the prior lakes stored in the “lake” table. These polygons are used to accelerate the assignment of pixel cloud pixels to LakeDatabase lakes.
- The “lake_catchment” geometries provide a spatial partition, considering hydrological constraints and topography while “lake_influence” geometries only take in account distances between lakes.

- The “basin” table delineates Level-3 HydroBASINS basins. These are used to label observed water features and separate them into continents.
- The “pass_tile” table delineates theoretical polygons of the left and right tiles for the selected orbit and pass.
- The “hypso_curve” table stores the information of lake hypsometry for computing water storage changes.

Each table is further described in the following sub-sections.

3.1.2 Basin-granule LakeDatabase

The basin-granule LakeDatabase files follow the model given in Figure 2 below. Its tables and attributes are derived from the parent database detailed in Appendix B and keep only the elements needed for LakeAvg processing.

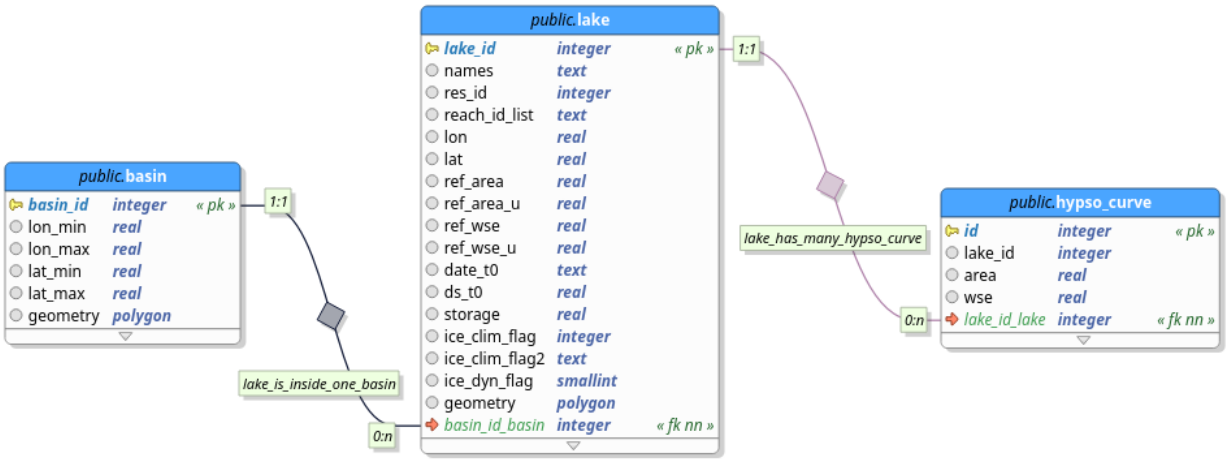


Figure 2. Model of the basin-granule LakeDatabase.

3.2 The “lake” table

The “lake” table contains the geometry and associated attributes of each prior lake corresponding to the selected granule.

Each prior lake feature is represented as an OGR.MultiPolygon, delineating its boundary; any island inside the lake is represented as an inner polygon.

Each prior lake feature has the following attributes:

- *lake_id*: identifier of the lake. Primary key of the “lake” table.
- *basin_id*: foreign key linking a prior lake to its watershed. The association of a lake with a basin is done by simple intersection between the polygon of the lake and the “basin” layer. A lake belongs to only one basin. In very particular cases where the intersection returns two or more basins (due to the imprecision of the basin boundaries), the basin chosen is the one that contains the centroid of the lake. If a lake is not located in a watershed (due to the

imprecision of the basin boundaries), the lake is affected to the closest watershed.

- *names*: known name(s) of the prior lake. This field is populated from different databases. The correspondence between the prior lake and the lakes in these databases is computed by direct intersection. The names are in capital letters to avoid accents or other problems. If there are several names, they are separated by a semicolon ";".
- *res_id*: reservoir identifier from the Global Reservoir and Dam database (GRanD) [5] [6]. The identifier is assigned by intersection of the prior lake polygon with reservoirs from GRanD. If the prior lake intersects more than one reservoir, only the identifier of the GRanD water body containing the prior lake's centroid is kept. Note that if the prior lake corresponds to a reservoir, its corresponding *lake_id* identifier ends with digit 3.
- *reach_id_list*: for connected lakes, list of the identifiers of the reaches from RiverDatabase [8] [9] [10] that intersect the prior lake. If there are more than one reach, they are separated by a semicolon ";".
- *lon*: longitude of the centroid of the prior lake.
- *lat*: longitude of the centroid of the prior lake.
- *ref_area* / *ref_area_u*: maximum area and associated uncertainty of the prior lake. This area is computed by projecting the polygon in UTM Zone. It corresponds to the same state of the lake as *ref_wse*.
- *ref_wse* / *ref_wse_u*: median water surface elevation and associated uncertainty of the prior lake, with respect to the reference geoid.
- *date_t0*: reference date from which the storage change is computed. The reference date will be populated at the first update of the LakeDatabase with the date of the first observation of the lake.
- *ds_t0:* storage change between the lake state at reference water surface elevation (*ref_wse*) and the lake state at the reference date (*ds_t0*).
- *storage:* absolute volume of the prior lake computed from SWOT data.
- *ice_clim_flag*: Climatological ice cover flag indicating whether the lake is ice-covered for the days from January to June, based on climatological ice coverage. It is formatted as a string of 182 digits, each one corresponding to a day in the chronological order: the first digit corresponds to January 1st, the second to January 2nd, …, and the last to June 30th. Note that a leap year is considered, so the 60th digit corresponds to February 29th and the 61st digit corresponds to March 1st. The value of each digit is 0, 1, and 2 indicate that the lake is likely not ice-covered, may or may not be partially or fully ice-covered, and likely fully ice-covered, respectively.
- *ice_clim_flag2:* same as *ice_clim_flag*, for days from July to December.
- *ice_dyn_flag*: dynamic ice cover flag. Format TBD.

The following attributes are present only in the "lake" table of the pass-granule LakeDatabase:

- *overlap_avg_[l/r]*: average overlap of lake and tile of current pass and swath L or R. If the

lake covers the whole tile, overlap = 1, if the lake overs 50% of the tile, overlap_avg = 0.5.

- *nb_tiles_[l/r]*: number of tiles covered by a lake in swath L or R.

3.3 The “lake_influence” table (pass-granule only)

The “lake_influence” table contains the influence area of each prior lake referenced in the “lake” table. An example is shown in Figure 3. Each lake influence feature is represented as an OGR.MultiPolygon, delineating the influence area of the prior lake.

Each feature of this table is referenced by:

- *lake_id:* lake identifier to link the “lake_influence” geometry to the “lake” table. Primary key of the “lake_influence” table.

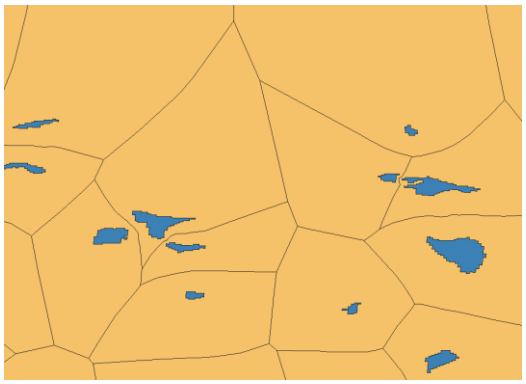


Figure 3. PLD lakes (blue) and their influence area (orange).

3.4 The “lake_catchment” table (pass-granule only)

The “lake_catchment” table contains the catchment area of each prior lake referenced in the “lake” table, taking hydrological constraints and topography into account. Each lake catchment feature is represented as an OGR.MultiPolygon. An example is shown in Figure 4

Each feature of this table is referenced by:

- *lake_id:* lake identifier to link the “lake_catchment” geometry to the “lake” table. Primary key of the “lake_catchment” table.

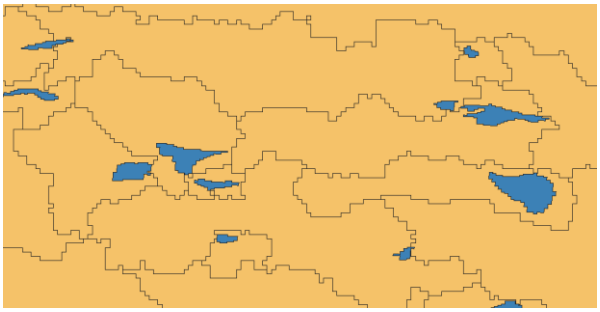


Figure 4. PLD lakes (blue) and their catchment area (orange).

3.5 The “basin” table

The “basin” table contains the watershed geometries corresponding to the selected granule.

Each feature of this table has the following attributes:

- *basin_id*: identifier of the basin. Primary key of the “basin” table.
- *lon_min*: minimum longitude of the basin.
- *lon_max*: maximum longitude of the basin.
- *lat_min*: minimum latitude of the basin.
- *lat_max*: maximum latitude of the basin.

3.6 The “pass_tile” table (pass-granule only)

The “pass_tile” table contains the identifier of each tile of the selected orbit and pass. It also has a geometry field with the boundaries of each tile.

Each feature of this table is referenced by:

- *id (ogc_fid):* automatically generated unique identifier. Primary key of the “pass_tile” table.
- *ref_tile:* identifier of the tiles of the selected orbit and pass. The format of *ref_tile* is a 4-character string of the form **PPP[L/R]** where **PPP**=zero-padded, three-digit, numeric tile number, and **[L/R]**=single character that identifies whether the granule if on the left (“L”) or right (“R”) of the nadir track.
- *continent:* two-letter continent identifier of the continent intersecting current tile, as defined in Table 3 . If the tile intersects several continents, the identifiers are separated by a “;”.
- *az_0_line*: azimuth 0 line of current tile in Well-known text (WKT) LineString format.
- *az_max_line*: azimuth max line of current tile in WKT LineString format.

3.7 The “hypso_curve” table

The “hypso_curve” table contains several area/wse couples corresponding to different states of the lake water storage. Figure 5 shows an example for the Orient lake.

Each feature of this table is referenced by:

- *id (ogc_fid):* automatically generated unique identifier. Primary key of the “hypso_curve” table.
- *lake_id:* lake identifier to link the “hypso_curve” data to the “lake” table.
- *area:* lake surface corresponding to the wse.
- *wse:* lake water surface elevation corresponding to the area.

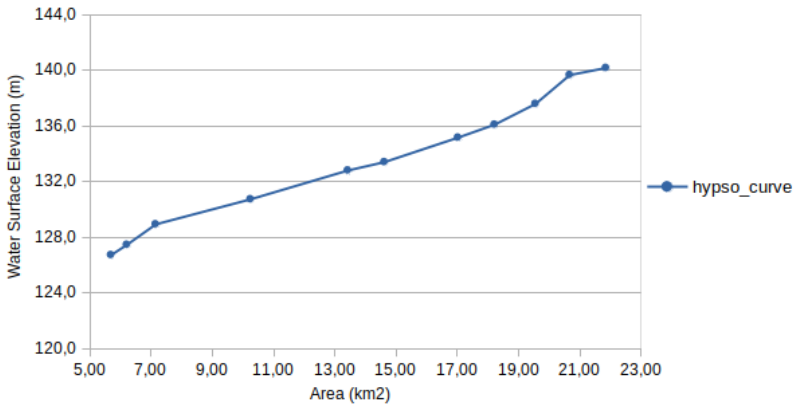


Figure 5. Hypsometric curve for the Orient lake.

4 Population of the LakeDatabase

The LakeDatabase has been populated based on various information sources as described in the following sections. It will be updated based on SWOT data and possibly other data sources during the mission.

4.1 The “lake” table

- The geometries are directly retrieved from:
    - the CIRCA-2015 lake water extent map [1]
    - the Georeferenced global Dams And Reservoirs dataset (GeoDAR) v1.1 [2],
    - the Post2k reservoir dataset (Fan et al., 2023) [3],
    - and also from scientists [4].
- The format of the *lake_id* identifier is a 10-character string of the form **CBBNNNNNNT** where **C**=continent code, **BB**=basin, **NNNNNN**=lake counter in the basin, and **T**=water body type.

lake_id is based on the Pfafstetter coding system [23] that is constructed using the topology of river networks. The code allows digits 0-9 at each hierarchy level. Continent code (**C**) and water body type code (**T**) are provided in Table 3 and Table 2, respectively. The geographical delineation of the continents is shown in Figure 6.

As indicated in Table 2, lakes are separated into two types, connected lakes (T=3) and disconnected lakes (T=2). Note that lake water bodies that are connected to the river topology (T=3) are included both in the river and lake products. Connectivity is defined specifically in relation to the PRD, so lakes that are termed disconnected (T=2) here may, in fact, be connected to a river network, but via channels too small to be included in the PRD.

Table 2. Water body type codes for the *lake_id* attribute.

Type Code (T)	Water Body Type
1	River *(not used in this product)*
2	Disconnected lake
3	Connected lake
4	Dam *(not used in this product)*
5	No topology *(not used in this product)*

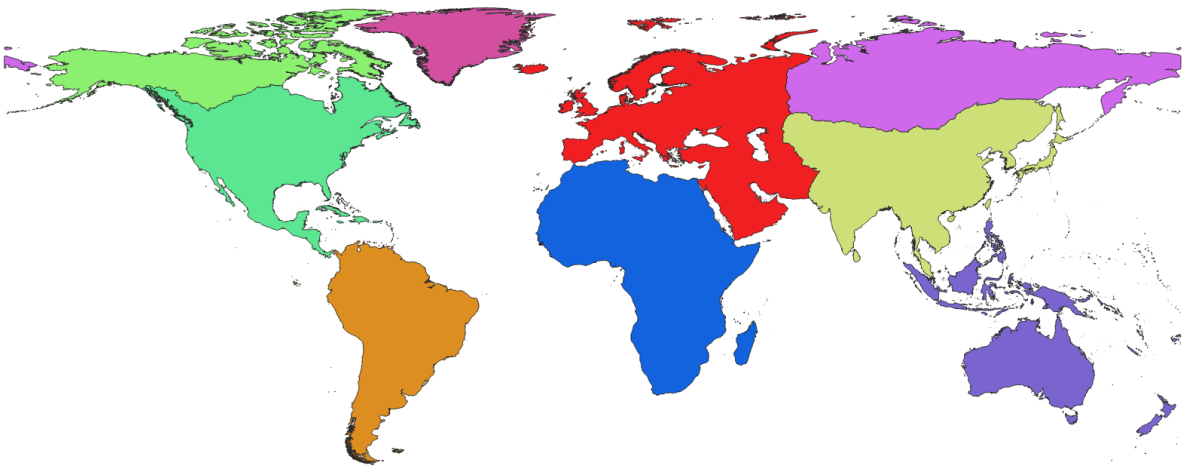


Figure 6. Geographical delineation of continents from HydroBASINS [4].

Table 3. Continent codes for the *lake_id* attribute, and continent IDs for the basin-granule LakeDatabase filenames.

Continent Code (C)	Continent	Continent ID
1	Africa	AF
2	Europe and Middle East	EU
3	Siberia	SI
4	Central and South-East 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

The continent code (**C**) is Level 1 in the Pfafstetter code. As indicated in the template above, *lake_id* value is based upon Pfafstetter Level 3, leading to 3 digits (**CBB**). Note that the continent codes in Table 3 are consistent with the continent coding used in the HydroBASINS dataset [7].

Figure 7 shows an example of what the coding may look like at Level 1 ($C = 7$) for the basin encompassing Mississippi (first level $B = 4$). Within each basin level, the lake is numbered with 000001 to a maximum of 999999 (i.e., a zero-padded six-digit number, represented as **NNNNNN**). Continent and basin polygons are available from [7].

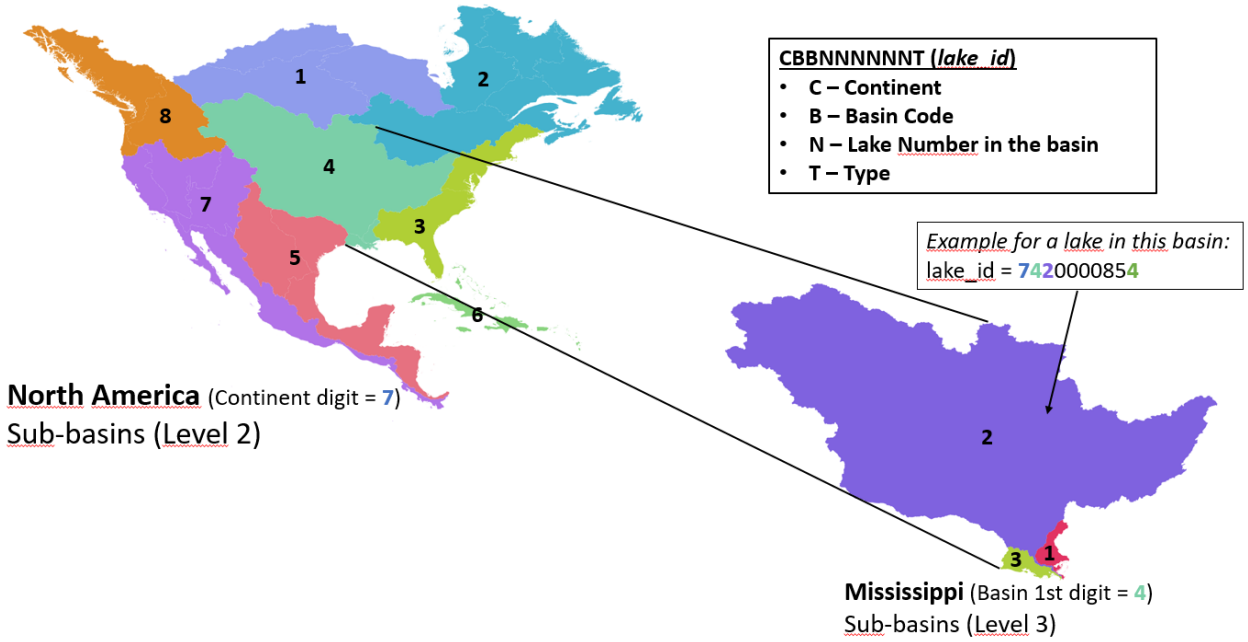


Figure 7. Example of the Pfafstetter basin coding system over North America, limited to the 3 first levels used in the identifiers of the lake products.

- The *basin_id* attribute is populated by intersecting the polygon of the prior lake with basin geometries from HydroBASINS [7], as described in section 3.2.
- The *lake_num* and *lake_type* (see Appendix B) corresponds respectively to the **NNNNNN** and **T** characters of the *lake_id* described above.
- The *names* attribute is populated from the following databases:
  - IGN Carthage database (covering France) [13]
  - Open Street Map (OSM) [14]
  - Global Lakes and Wetlands Database (GLWD) [15]
  - Natural Earth Data [16]
  - Vector Map Level 0 (VMap0) [17]
  - HydroLAKES [18]
- The *res_id* attribute is populated by intersecting the polygon of the lake with reservoir geometries from GRanD [5] [6].
- The *reach_id_list* attribute is populated using the list of reaches that intersect the prior lake, obtained by intersecting polygons of the "lake" table with reaches from SWORD [10].
- The *lon* and *lat* attributes correspond to the coordinates of the centroid of the polygon geometry of the feature.
- The *ref_area* attribute is computed by projecting the lake polygon in the corresponding UTM zone.
- The *ref_wse* is the median value of the LakeSP water surface elevations such as *ice_clim_f*=0, and outliers filtered when outside the mean +/- twice the standard-deviation.
- The *date_t0* corresponds to the date of the first SWOT observation obtained after the same

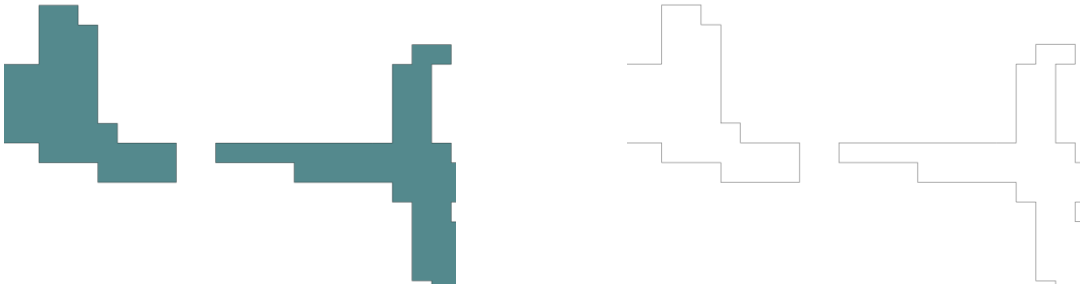
filtering as described for *ref_wse*.

- *The ds_t0* is the storage change computed between the lake state at reference water surface elevation (*ref_wse*) and the lake state at *date_t0*.
- The *ice_clim_flag*, *ice_clim_flag2* and *nb_ice_clim* attributes are populated from [24].
- The *pass_full_[cal/nom]*, *pass_part_[cal/nom]*, *nb_pass_full_[cal/nom*], *nb_pass_part_[cal/nom]* and *cycle_flag_[cal/nom]* attributes are populated using 10-60km swath width computed from [12] for both Cal/Val and nominal orbits.
- The *req_size* compares *ref_area* and SWOT requirements.
- The *min_dist_lake* and *min_dist_lake_id* are determined by computing lakes distance to each other's. The *min_dist* is by default equal to the *min_dist_lake*, until *min_dist_river* is computed.
- The *ref_area_u*, *ref_wse_u*, *storage*, *ice_dyn_flag, dist_flag, obs_flag, min_dist_river* and *min_dist_river_id* attributes are populated with "no_data".
- The *sources* attribute is populated with the source of the data (Circa2015, Circa2000, digitize, post2kGSW, GeoDARv11, HydroLAKESv10, GLAKES, or scientist name)

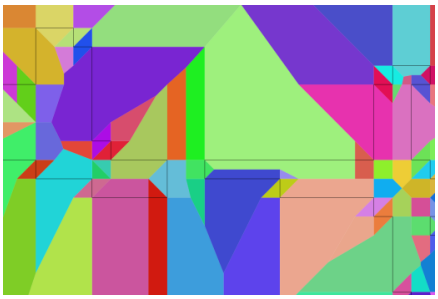
4.2 The "lake_influence" table

The "lake_influence" table is populated with polygons defining the influence area of each lake of the "lake" table, based on Voronoi diagrams as illustrated in Figure 8.

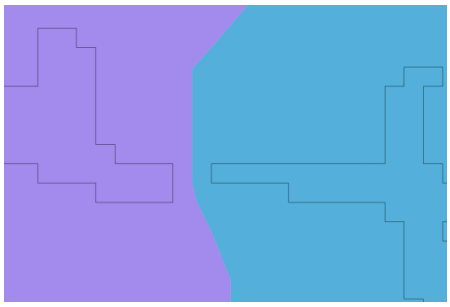
- The *lake_id* corresponds to the one of "lake" table.
- The *neighbour_basin* attribute is retrieved by intersecting the lake_influence geometry and the geometries from HydroBASINS basins corresponding to Pfafstetter Level-3 [7].



Two lake polygons Corresponding lake boundaries



Voronoi diagram of lake boundaries



Influence area map for the two lakes

Figure 8. Illustration of the computation of the influence areas of two lakes.

4.3 The "lake_catchment" table

The "lake_catchment" table is populated with lake catchment polygons from Lake-TopoCat **[11]** for each lake of the "lake" table.

- The *lake_id* corresponds to that of the "lake" table.

4.4 The "basin" table

The geometries directly come from HydroBASINS basins corresponding to Pfafstetter Level-3 [7]. The database contains 291 geometries delineating watershed boundaries. Basin geometry can be either an OGR.Polygon or an OGR.MultiPolygon: islands located in the same geographical area are considered to belong to the same basin.

- The *basin_id* attribute contains the level-3 Pfafstetter code from HydroBASINS [7].
- The *name* attribute is not yet populated.
- The *lon_min*, *lon_max*, *lat_min* and *lat_max* attributes contain the basin's coordinates.
- The *swot_obs* attribute is populated with SWOT science orbit number intersecting the basin.

4.5 The "pass_tile" table

- The *ref_tile* attribute is computed from [12], considering an extended 0-80 km swath width for both Cal/Val and nominal orbits.
- The *pass_num* and *pass_type* attributes are populated using a 10-60 km swath width computed from [12] for both Cal/Val and nominal orbits.
- The *continent* attribute is given by the level-1 Pfafstetter code from HydroBASINS [7] intersecting the tile.
- The *az_0_line* and *az_max_line* attributes are computed from the pass_tile geometry.

4.6 The “hypso_curve” table

The “hypso_curve” table is populated with *in situ* data available for 27 lakes located in different basins (19 in basin 725, 3 in basin 642, 3 in basin 724, 1 in basin 822 and 1 in basin 232).

5 Notes

5.1 Data history

This section provides notes on the history of the auxiliary data for internal record-keeping purposes. Information in this section is not necessary to use the data; it is kept here only for context.

Table 4 relates versions and creation time stamps of the LakeDatabase auxiliary data, and lists the content of the updates between the different versions.

Table 4. Content of LakeDatabase versions.

Version number	LakeDatabase Creation Timestamp	List of updates from previous version
v001	20211103T133032	First version, from CIRCA-2015
v002	20220601T000000	Adds *overlap_[r|l]* and *nb_tiles_[r|l]* to point out PLD lakes too huge to be handled by the LakeSP processor and that need to be split and processed by the LakeTile processor before being gathered at the LakeSP level
v100	20221121T110000	Adds missing lakes from CIRCA-2015 due to computation bug [Launch version]
v101	20230209T000000	Populates *ice_clim_flag* and additional *ice_clim_flag2* attributes
v102	20230614T000000	Updates *reach_id_list* attribute wrt SWORD v15
v103	20230921T000000	Adds lake_catchment table and new lakes from Jida Wang, Marielle Gosset and Manuella Grippa
v104	20230927T000000	Adds the coefficients needed to compute storage change for 26 lakes: • NA: Achigan; Argent; Becancour; Bois-Verts; Bouard; Brome; Brompton; Camatose; Canimina; Carre; Elgin; Kiamika; Lovering; Massawippi; Montagne-Noire; Nord-Est; Stukely; Theodore; Victoria; Waterloo • SA: Bariri; Caconde; Segredo • EU: Estanyol; Fontargente; Orient
v105	20231017T000000	Fixes duplicated tiles

v106	20240913T140000	• Adds new lakes:
		  ○ having a LOCSS station (https://www.locss.org/)
		  ○ from scientists: Jida Wang (University of Illinois, USA), Jean-François Crétaux (CNES-LEGOS, France), and Cécile Gomez (IRD)
		• Removes Caspian Sea
		• Updates *reach_id_list* attribute by using SWORD v16
		• Updates *min_dist_river* and *min_dist_river_id* attributes by using SWORD_v16
		• Updates *lake_catchment* table

v200	20241129T140000	• Updates *reach_id_list* attribute w.r.t. SWORD v17 • Adds *ref_wse, date_t0, ds_t0* from SWOT LakeSP products (5 370 219 lakes updated)
V201	20250109T163600	• Removal of spurious lakes in the ocean and their influence areas • Correction of *lon* attribute for 15 lakes at +/-180° longitude • Adds *sources* for 56 lakes where they were missing • Correction of the *name* of some large lakes • Updates some *date_t0* and *ds_t0* set to 0 instead of NULL • Updates *cycle_flag* and **pass** attributes
V202	20250331T170000	• Updates *reach_id_list* .w.r.t. new connected lakes in SWORD v17

5.2 Statistics

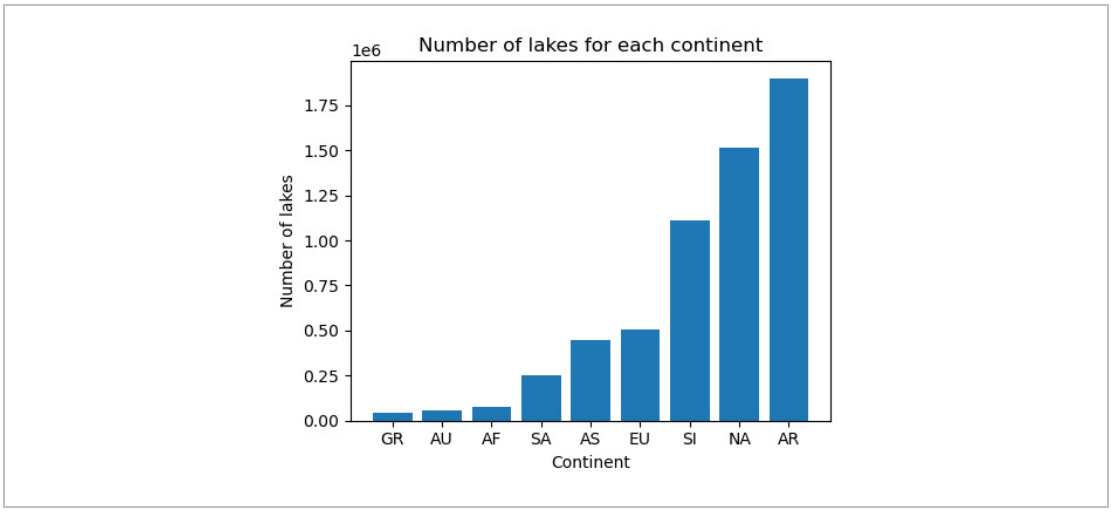
5.2.1 Distribution of lakes per continent

Table 5. Distribution of lakes per continent.

Continent ID	Continent code	Continent name	Number of lakes
AF	1	Africa	73 781
EU	2	Europe and Middle East	504 435
SI	3	Siberia	1 113 265
AS	4	Central and South-East Asia	446 773
AU	5	Australia and Oceania	57 338
SA	6	South America	250 215
NA	7	North America and Caribbean	1 512 137
AR	8	North American Arctic	1 899 238
GR	9	Greenland	40 759
		Total	***5 897 941***

5.2.2 Lake number and size

The number of lakes varies from 40 759 for Greenland (GR) to 1 899 238 for the North American Arctic (AR), as shown in Table 5 and Figure 9. The North American Arctic (AR) features 32% of the world's lakes, North America (NA) 26%, and Siberia (SI) 19%. The largest lake is Lake Superior in North America (NA), with a size of 8 221 820 ha. Africa (AF) has the highest mean lake mean, with an average of 330 ha.



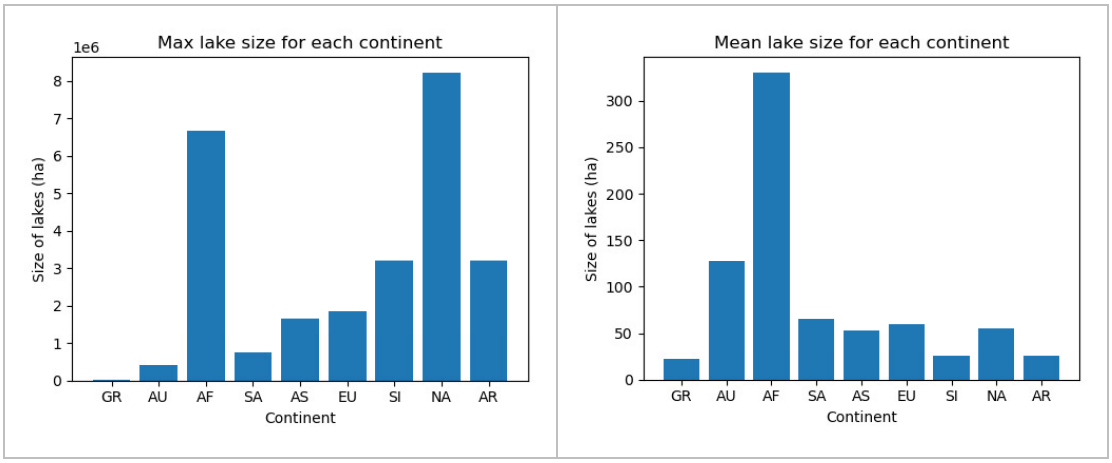


Figure 9. Number of lakes and maximum and mean lake size for each continent.

5.2.3 Cycle flag distribution

The cycle flag qualifies the theoretical observation rate of a lake by SWOT during a cycle:

- 0 = The lake is never observed within the nominal swath (10-60 km)
- 1 = The lake is only partially observed, and not fully covered combining multiple passes
- 2 = The lake is only partially observed, but fully covered combining multiple passes
- 3 = The lake is fully observed by at least one pass

Figure 10 shows that 97.41% of the LakeDatabase lakes (i.e. 5 745 323 lakes) are fully observed during a cycle, and that 2.30% (i.e. 135 508 lakes) are never observed. ~16 000 lakes are not fully covered using all the passes within a cycle. Less than 1 000 lakes are fully covered using multiple passes within a cycle.

Per continent (Figure **11**), the proportion of cycle flags equal to 1 is almost always lower than 1%. The proportion of cycle flags equal to 2 is always lower than 0,07%.

Siberia (SI) and the North American Arctic (AR) feature the highest proportion of fully observed lakes in a single pass.

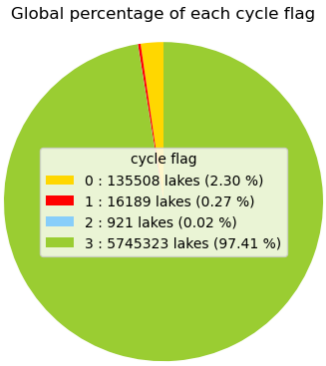


Figure 10. Global distribution of cycle flag.



Figure 11. **Cycle flag distribution by continent.**

5.2.4 Coverage of large lakes per pass

Lake Michigan (*lake_id* = 7250080513) in North America (NA) and the Malawi lake (*lake_id* = 1220056903) in Africa (AF) have the highest number of tiles covered by a single pass (Figure 12), with 17 tiles on pass 509 and 19 tiles on pass number 068, respectively.



(a) Michigan Lake – pass 509 (b) Malawi Lake – pass 068

Figure 12. Lakes covered by the greatest number of tiles for a single pass.

5.2.5 Minimum distance to neighbor lake

Overall, the mean distance to the closest neighbor lake is 627 m, and 50% of the lakes have a distance lower than 240 m (median value). In the North American Arctic (AR), the median value is as low as 183 m (Table 6).

Table 6. Statistics on minimum distance to neighbor lakes per continent.

Continent ID	Median dist_min	Mean dist_min	Min dist_min	Max dist_min
AF	496 m	2 504 m	5 m	1 062 668 m
AR	183 m	307 m	0 m	728 761 m
AS	684 m	1 326 m	0 m	142 897 m
AU	402 m	2 101 m	0 m	1 054 510 m
EU	360 m	950 m	0 m	231 394 m
GR	300 m	601 m	14 m	72 690 m
NA	256 m	544 m	0 m	1 080 605 m
SA	543 m	1 455 m	1 m	99 482 m
SI	210 m	470 m	0 m	112 397 m
Total	**240 m**	**627 m**	**0 m**	**1080605 m**

Figure 13 shows the CDF of the distance to the closest lake as a function of the lake area. Large lakes tend to be closer to another lake than small lakes. In Siberia (SI), North America (NA), the North American Arctic (AR) and Greenland (GR), more than 80% of the lakes have a minimum distance to a neighbor lake below 1 km.

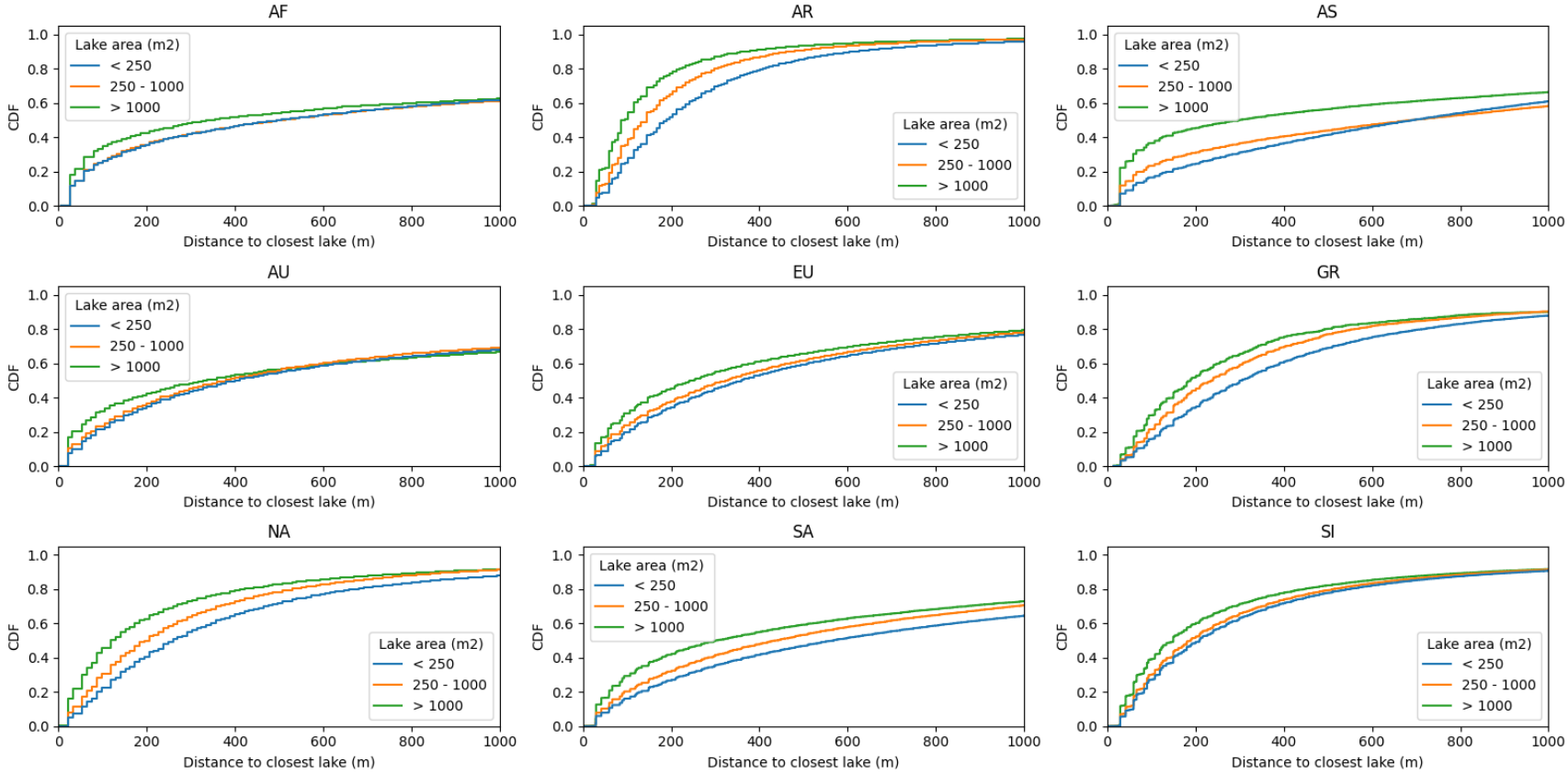


Figure 13. Cumulative distribution function of distance to closest lake for different categories of sqrt(lake area).

6 References

[1] Y. Sheng, C. Song, J. Wang, E. Lyons, B. Knox, J. Cox and F. Gao, "Representative lake water extent mapping at continental scales using multi-temporal Landsat-8 imagery," *Remote Sensing of Environment,* vol. 185, pp. 129-141, 2016.

[2] J. Wang, B. A. Walter, F. Yao, C. Song, M. Ding, A. S. Maroof, J. Zhu, C. Fan, J. M. McAlister, S. Sikder, Y. Sheng, G. H. Allen, J.-F. Crétaux and Y. Wada, "GeoDAR: georeferenced global dams and reservoirs dataset for bridging attributes and geolocations," *Earth System Science Data,* vol. 14, no. 4, pp. 1869-1899, 2022.

[3] C. Fan and C. Song, "An underrated abundance of young reservoirs revealed by new data," *Science Bulletin,* 2023 (In review).

[4] J. Wang, C. Pottier, C. Cazals, M. Battude, Y. Sheng, C. Song, M. S. Sikder, X. Yang, L. Ke, M. Gosset, R. Reis, A. Oliveira, M. Grippa, F. Girard, G. H. Allen, S. Biancamaria, L. C. Smith, J.-F. Créteaux and T. M. Pavelsky, "The Surface Water and Ocean Topography Mission (SWOT) Prior Lake Database (PLD): Lake mask and operational auxiliaries," *Water Resources Research,* 2024 (submitted).

[5] B. Lehner, C. R. Liermann, C. Revenga, C. Vörösmarty, B. Fekete, P. Crouzet, P. Döll, M. Endejan, K. Frenken, J. Magome, C. Nilsson, J. C. Robertson, R. Rödel, N. Sindorf and D. Wisser, "High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management," *Frontiers in Ecology and the Environment,* vol. 9, pp. 494-502, 2011.

[6] B. Lehner, C. R. Liermann, C. Revenga, C. Vörösmarty, B. Fekete, P. Crouzet and C. Nilsson, "Global reservoir and dam (GRanD) database," Technical Documentation, Version1, 1-14, 2011.

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

[8] JPL D-105518, Revision A (Draft), "SWOT Auxiliary Data Description Document: Prior River Database (RiverDatabase)," Jet Propulsion Laboratory Internal Document, 2025.

[9] E. H. Altenau, T. M. Pavelsky, M. T. Durand, X. Yang, R. P. d. M. Frasson and L. Bendezu, "The Surface Water and Ocean Topography (SWOT) mission River Database (SWORD): A globl river network for satellite data products," *Water Resources Research,* vol. WRCS25408, 2021, https://doi.org/10.1029/2021WR030054..

[10] E. H. Altenau, T. M. Pavelsky, M. T. Durand and E. Collins, "SWOT River Database (SWORD) Product Description Document," Release v17, 2024.

[11] S. Sikder, J. Wang, G. H. Allen, Y. Sheng, D. Yamazaki, C. Song, M. Ding, J.-F. Crétaux and T. M. Pavelsky, "Lake-TopoCat: A global lake drainage topology and catchment," Earth System Science Data Discussion, 2023. [Online]. Available: https://essd.copernicus.org/preprints/essd-2022-433/essd-2022-433.pdf.

[12] JPL D-105500, "SWOT Auxiliary Data Description Document: Reference Orbit Track Files (RefOrbitTrack)," Jet Propulsion Laboratory Internal Document, 2022.

[13] IGN, "Carthage database," [Online]. Available: http://www.geoinformations.developpement-durable.gouv.fr/bd-carthage-r363.html.

[14] "Open Street Map (OSM)," [Online]. Available: https://www.openstreetmap.org.

[15] B. Lehner and P. Döll, "Development and validation of a global database of lakes, reservoirs and wetlands," *Journal of Hydrology, 296(1-4),* pp. 1-22, 2004.

[16] "Natural Earth Data," [Online]. Available: https://www.naturalearthdata.com.

[17] "Vector Map Level 0 (VMap0)," [Online]. Available: https://mdl.library.utoronto.ca/collections/geospatial-data/vector-map-level-0-vmap0.

[18] M. L. Messager, B. Lehner, G. Grill, I. Nedeva and O. Schmitt, "Estimating the volume and age of water stored in global lakes using a geo-statistical approach," *Nature Communications,* vol. 13603,

no. doi: 10.1038/ncomms13603, 2016.

[19] SWOT-NT-CDM-1753-CNES, Revision A, "SWOT Algorithm Theoretical Basis Document: Level 2 KaRIn High Rate Lake Single-Pass Science Algorithm Software (L2_HR_LakeSP)," Centre National d'Etudes Spatiales, 2025.

[20] "Sqlite," [Online]. Available: https://www.sqlite.org/index.html.

[21] SWOT-TN-CDM-0672-CNES, Revision C, "SWOT Product Description Document: Level 2 KaRIn High Rate Lake Tile Vector Product (L2_HR_LakeTile)," Centre National d'Etudes Spatiales, 2025.

[22] SWOT-TN-CDM-0673-CNES, Revision C, "SWOT Product Description Document: Level 2 KaRIn High Rate Lake Single Pass Vector Product (L2_HR_LakeSP)," Centre National d'Etudes Spatiales, 2025.

[23] J. Furnans and F. Olivera, "Water Shed Topology - The Pfafstetter System," Esri.

[24] X. Yang, T. M. Pavelsky and G. H. Allen, "The past and future of global river ice," *Nature,* vol. 577, pp. 69-73, 2020, https://doi.org/10.1038/s41586-019-1848-1.

Appendix A. **Acronyms**

CNES	Centre National d'Études Spatiales
GIS	Geographic Information Systems
JPL	Jet Propulsion Laboratory
NASA	National Aeronautics and Space Administration
PLD	Prior Lake Database
PRD	Prior River Database
SAS	Science Algorithm Software
SWOT	Surface Water Ocean Topography
TBC	To Be Confirmed
TBD	To Be Determined

Appendix B. Parent LakeDatabase

The operational LakeDatabase files described in Section 3 are built from a parent LakeDatabase, which is not used operationally. It is divided in 291 level-3 basins defined in the HydroBASINS dataset [7]. Its model is shown in Figure 14.

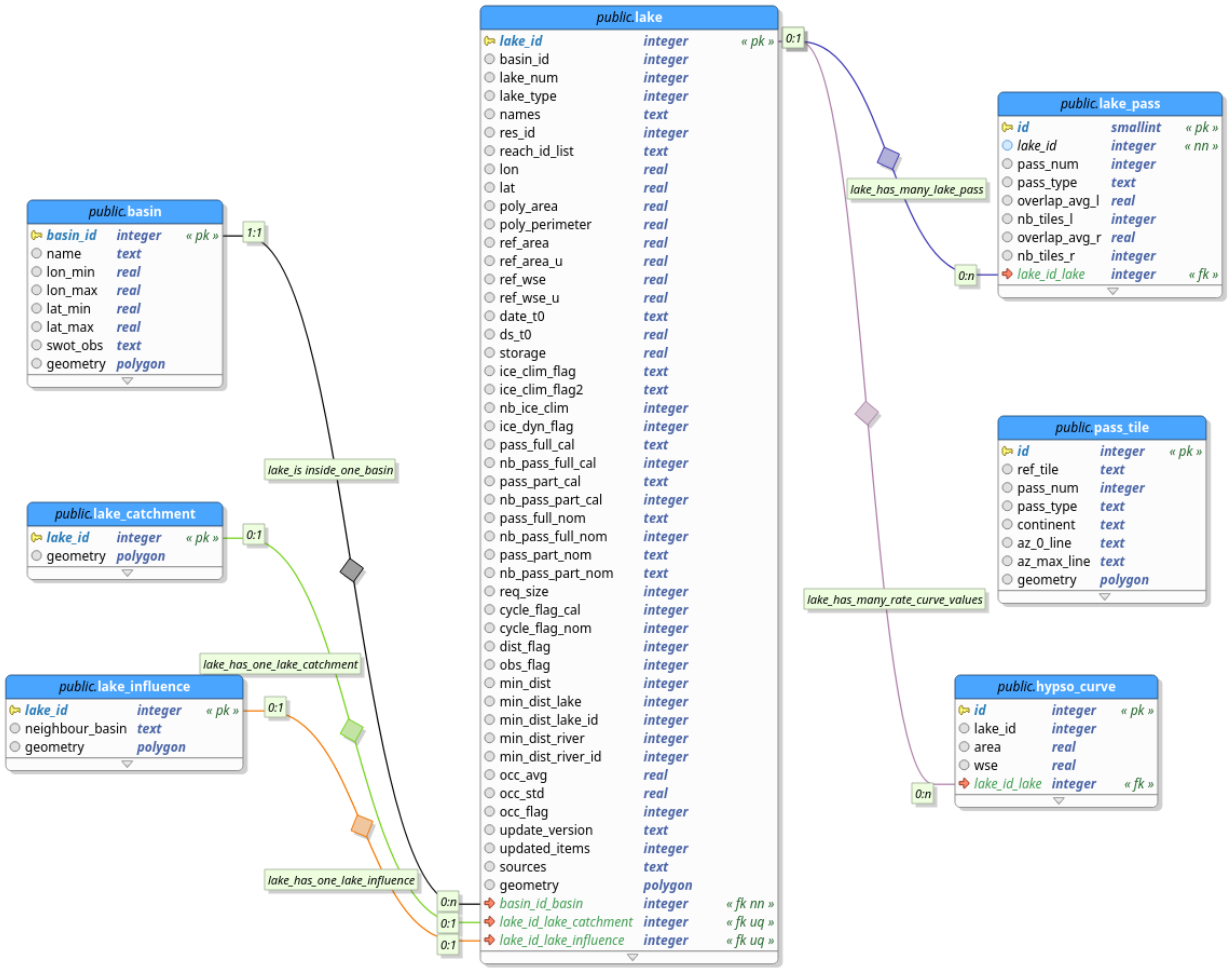


Figure 14. Model of the parent LakeDatabase.

Its tables are similar to those used for operational files (described in Section 3.1.1). The differences are described below.

The “lake” table contains the same attributes as the “lake” table of the basin-granule LakeDatabase described in sections 3.1.2 and 3.2, as well as the following additional attributes:

- *lake_num*: lake counter in the basin
- *lake_type*: water body type (see Table 2)
- *poly_area / poly_perimeter*: polygon area and polygon perimeter
- *nb_ice_clim:* number of days the lake is ice-covered based on climatological data.

- *pass_[full/part]_[cal/nom]*: list of the identifiers of the Cal/Val or nominal passes that may entirely or partially cover the prior lake; this is computed by intersecting the prior lake polygon and the 10-60 km swaths associated with the SWOT orbit plus a 5 km buffer area. The buffer area is used to take SWOT orbit variations into account. If there are more than one pass number, they are separated by a semicolon ";".
- *nb_pass_[full/part]_[cal/nom]:* number of Cal/Val or nominal passes that may entirely or partially cover the prior lake. This is related to the *pass_full* and *pass_part* lists described above
- *req_size*: indicates whether the lake as a ref_area higher than the requirement size (6 ha)
- *cycle_flag_[cal/nom]*: flag characterizing lake observation by SWOT; this flag synthetizes the *pass_full* and *pass_part* lists described above:
    - 0: the lake is expected to never be observed by SWOT; in this case, *pass_full* and *pass_part* are empty fields.
    - 1: the lake is expected to be only partially observed by SWOT, even after aggregation over a cycle; in this case, *pass_full* is empty and *pass_part* is populated.
    - 2: the lake is expected to be entirely observed by SWOT, but only after aggregation over a cycle, i.e. never in a single pass; in this case, *pass_full* is empty and *pass_part* is populated.
    - 3: the lake is expected to be entirely observed by SWOT in a single pass at least once per cycle; in this case, at least pass_full is populated.
- *dist_flag*: binary flag that indicates whether the lake is close to another lake (minimum distance < 300 m)
- *obs_flag*: flag characterizing the complexity of the observations (based on statistics computed on SWOT lake products over a relatively long time period):
    - 0: the PLD lake has never been observed
    - 1: 1 PLD lake matches 1 observation
    - 2: 1 PLD lake matches N observations
    - 3: N PLD lakes match 1 observation
    - 4: N PLD lakes match M observations
- *min_dist*: smallest distance out of *min_dist_lake* and *min_dist_river* described below.
- *min_dist_lake* and *min_dist_lake_id*: distance to the closest lake and its lake_id
- *min_dist_river* and *min_dist_river_id*: distance to the closest river and its reach_id
- *occ_avg*: average occurrence measured from Pekel pixels that intersect the PLD polygon for the *lake_id*
- *occ_std*: standard deviation of the occurrence values of selected Pekel pixels intersecting the PLD polygon for the *lake_id*
- *occ_flag:* 0 for stable lakes (occ_avg > 90 and occ_std < 20), 1 for stable lakes of lower

reliability (occ_avg > 80 and occ_std < 35), 0 for non-stable lakes

- *sources*: indicates the origin of the lake geometry.

The “lake_influence” table contains one additional attribute:

- *neighbour_basin:* list of *basin_id* that intersect the *lake_influence* area (except the one with the highest overlap, used to compute the *lake_id*)

The “basin” table contains the following additional attributes:

- *name*: name of the basin.
- *swot_obs*: list of SWOT passes that entirely or partially cover the basin. These pass numbers are separated by a “;”.

The “pass_tile” table contains the following additional attributes:

- *pass_num:* number of the pass
- *pass_type:* type of the pass (Cal or Nom)

There is also an additional “lake_pass” table, which provides the number of tiles covered by the lakes in the L and R swaths, and the average overlap. Each feature of this table is referenced by:

- *id (ogc_fid):* automatically generated unique identifier. Primary key of the “lake_pass” table.
- *lake_id:* lake identifier to link the “lake_pass” feature to the “lake” table.
- *pass_num:* number of the pass.
- *pass_type:* type of the pass (Cal or Nom orbits).
- *overlap_avg_[l/r]* and *nb_tiles_[l/r]* which are in the “lake” table for the pass-granule LakeDatabase (cf. Section 3.2).

The “lake_pass” table has been populated based on various information sources:

- The *lake_id* corresponds to the identifier of the “lake” table.
- The *pass_num* and *pass_type* attributes are populated using 10-60km swath width computed from [12] for both Cal and Nom orbits.
- The *overlap_avg_[l/r]* and *nb_tiles_[l/r]* attributes are populated using the lake geometry and tile geometry from the ”pass_tile” table (see Section 4.5).

Appendix C. Science LakeDatabase

The parent LakeDatabase also allows to generate the files that are the distributed to the scientific community and other end users. These files contain all the attributes needed to run the operational lake algorithms and generate the SWOT standard lake products, as well as some additional attributes that may be of interest.

This dataset, fully described in [4], is divided into 61 level-2 basins as defined in the HydroBASINS dataset [7]. Its model is shown in Figure 15. Its tables are basically the same as those used for operational files described in Section 3.1.1.

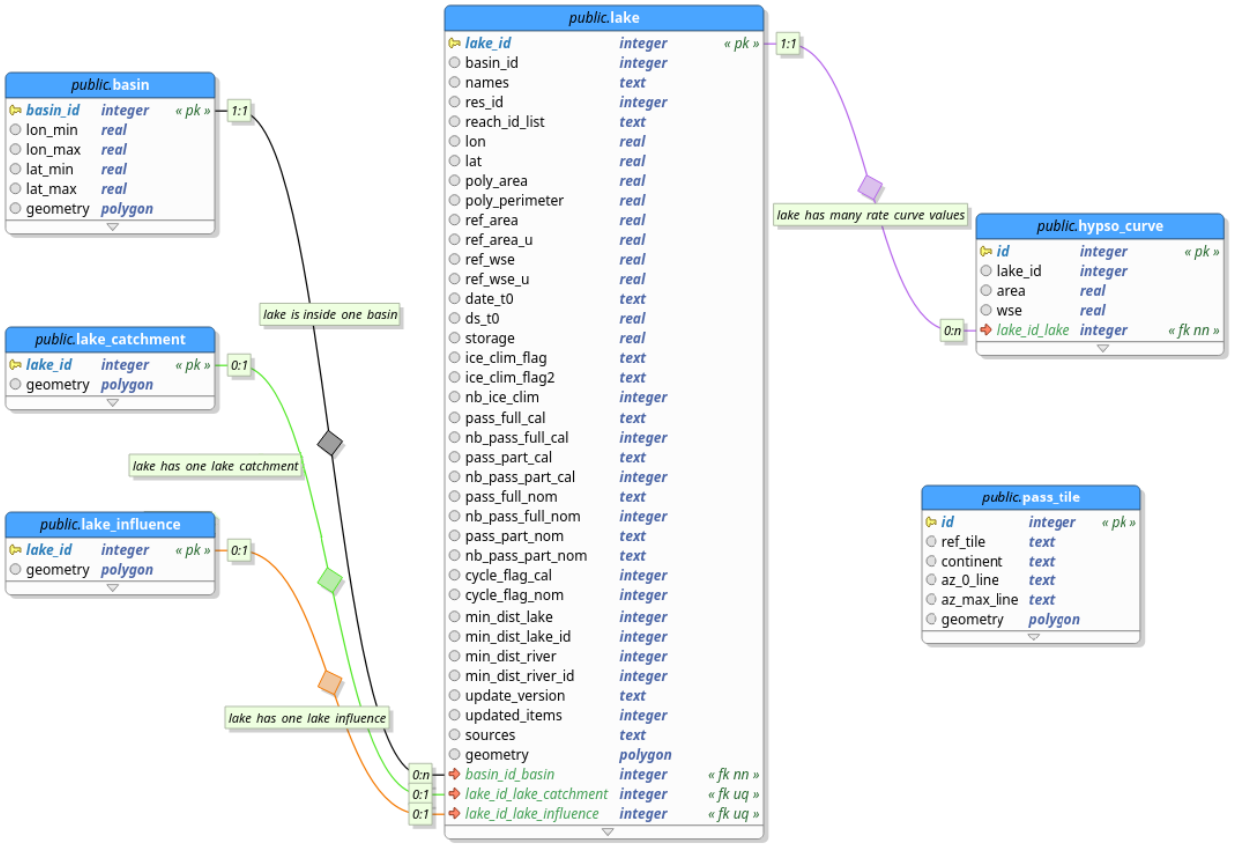


Figure 15. Model of the science LakeDatabase.

This “lake” table contains the same attributes as the “lake” table of the parent LakeDatabase described in Appendix B, with a few exceptions (no *lake_num*, *lake_type*, *req_size*, *dist_flag*, *obs_flag, occ_avg, occ_std,* nor *occ_flag* attributes are included).

Each file of the science LakeDatabase adopts the following file naming convention (refer to

Section 2.2 for further details):

*SWOT_LakeDatabase**Sci**_<Continent>_<BasinID>_<ValidityBeginningDateTime>_<ValidityEndingDateTime>_<CreationDateTime>_<Version>.sqlite*

Table 7 provides the expected size and data volume (uncompressed) of this dataset.

Table 7. Data volume of the science LakeDatabase.

Statistic	Number of lakes	Volume/Granule (GB)
Minimum (basin id)	**22** (54)	**0.017** (54)
Median (basin id)	**29 538** (42)	**0.21** (12)
Maximum (basin id)	**832 732** (72)	**3.69** (72)