The Surface Water and Ocean Topography Mission (SWOT) Prior Lake Database (PLD): Lake mask and operational auxiliaries

Jida Wang, Claire Pottier, Cécile Cazals, Marjorie Battude, Yongwei Sheng, Chunqiao Song, Md Safat Sikder, Xiao Yang, Linghong Ke, Marielle Gosset, Rafael Reis Alencar Oliveira, Manuela Grippa, Félix Girard, George H. Allen, Sylvain Biancamaria, Laurence C. Smith, Jean-Francois Crétaux, Tamlin Pavelsky

Author affiliations:

Jida Wang: Department of Geography and Geographic Information Science, University of Illinois Urbana-Champaign, Urbana, IL, USA. jidaw@illinois.edu

Claire Pottier: Centre National d'Études Spatiales (CNES), Toulouse, France. Claire.Pottier@cnes.fr

Cécile Cazals: CS Group, Toulouse, France. cecile.cazals@csgroup.eu

Marjorie Battude: CS Group, Toulouse, France. marjorie.battude@csgroup.eu

Yongwei Sheng: Department of Geography, University of California, Los Angeles, CA, USA. ysheng@geog.ucla.edu

Chunqiao Song: Key Laboratory of Watershed Geographic Sciences, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China; University of Chinese Academy of Sciences, Nanjing, Nanjing, China. cqsong@niglas.ac.cn

Md Safat Sikder: Department of Geography and Geospatial Sciences, Kansas State University, Manhattan, KS, USA. msikder@ksu.edu

Xiao Yang: Department of Earth Sciences, Southern Methodist University, Dallas, TX, USA. xnayang@smu.edu

Linghong Ke: College of Hydrology and Water Resources, Hohai University, Nanjing, 210098, China; State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, 210098, China. kelinghong@hhu.edu.cn

Marielle Gosset: Géosciences Environnement Toulouse (Université de Toulouse, CNRS, IRD, CNES). marielle.gosset@ird.fr

Rafael Reis Alencar Oliveira: Ceará State's Foundation for Meteorology and Water Resources (FUNCEME, UFC, UT, GET), Fortaleza, Brazil. rafael.reis@funceme.br

Manuela Grippa: Géosciences Environnement Toulouse (Université de Toulouse, CNRS, IRD, CNES). manuela.grippa@get.omp.eu

Félix Girard: Géosciences Environnement Toulouse (Université de Toulouse, CNRS, IRD, CNES). felix.girard@get.omp.eu

George H. Allen: Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. geoallen@vt.edu

Sylvain Biancamaria: Laboratoire d'Études en Géophysique et Océanographie Spatiales (LEGOS), Centre National d'Études Spatiales (CNES), Toulouse, France. sylvain.biancamaria@univ-tlse3.fr

Laurence C. Smith: Department of Earth, Environmental and Planetary Sciences, Brown University, Providence, RI, USA. Laurence_Smith@Brown.edu

Jean-Francois Crétaux: Laboratoire d'Études en Géophysique et Océanographie Spatiales (LEGOS), Centre National d'Études Spatiales (CNES), Toulouse, France. jean-francois.cretaux@cnes.fr

Tamlin Pavelsky: Department of Earth, Marine and Environmental Sciences, University of North Carolina, Chapel Hill, NC, USA. pavelsky@email.unc.edu

Key Points:

- SWOT Prior Lake Database (PLD) provides the foundation for generating SWOT vector lake products including area, height, and storage change.
- PLD inventories 6 million lakes with a 1-ha minimum area, 76% of which are smaller than 10 ha and 96% are fully observed per orbit cycle.
- PLD contains multiple operational auxiliaries to ease lake assignment, storage change computation, and vector lake product distribution.

Abstract

Lakes are the most prevalent and predominant water repositories on land surface. A primary objective of the Surface Water and Ocean Topography (SWOT) satellite mission is to monitor the surface water elevation, area, and storage change in Earth's lakes. To meet this objective, prior information of global lakes, such as locations and benchmark extents, is required to organize SWOT's KaRIn observations over time for computing lake storage variation. Here, we present the SWOT mission Prior Lake Database (PLD) to fulfill this requirement. This paper emphasizes the development of the "operational PLD", which consists of (1) a high-resolution mask of ~6 million lakes and reservoirs with a minimum area of 1 ha, and (2) multiple operational auxiliaries to assist the lake mask in generating SWOT's standard vector lake products. We built the prior lake mask by harmonizing the UCLA Circa-2015 Global Lake Dataset and several state-of-the-art reservoir databases. Operational auxiliaries were produced from multi-theme geospatial data to provide information necessary to embody the PLD function, including lake catchments and influence areas, ice phenology, relationship with SWOT-visible rivers, and spatiotemporal coverage by SWOT overpasses. Globally, over three quarters of the prior lakes are smaller than 10 ha. Nearly 96% of the lakes, constituting over half of the global lake area, are fully observed at least once per orbit cycle. The PLD will be recursively improved during the mission period and serves as a critical framework for organizing, processing, and interpreting SWOT observations over lacustrine environments with fundamental significance to lake system science.

1 Introduction

Natural lakes and manmade reservoirs, hereafter "lakes", are among the most predominant components of land hydrology (Messager et al., 2016; Verpoorter et al., 2014). They collectively store nearly 90% of the liquid freshwater on the Earth's surface, providing the most readily accessible water resource for societal use (Abbott et al., 2019; Oki & Kanae, 2006). Lakes also represent diverse and complex aquatic ecosystems, offering unique aesthetic appeals in the landscape and indispensable sources of biodiversity, food, and recreation outlets (Herdendorf, 1984). Although considered as lentic systems, lakes are often dynamic, with water storage and quality reflective of basin-scale hydrology and/or anthropogenic activities (Fergus et al., 2017; Wurtsbaugh et al., 2017; Yang, O'Reilly, et al., 2022). Lakes also sequester a large amount of carbon from the watersheds and modulate terrestrial carbon cycling through water storage variation and lacustrine-fluvial interactions (Mendonca et al., 2017; Tranvik et al., 2009). For these reasons, lakes serve as both "sentinels" and "regulators" of climate change (Adrian et al., 2009; Schindler, 2009) and are recognized as an "Essential Climate Variable" by the Global Climate Observing System (GCOS) of the World Meteorological Organization (WMO, 2022). Monitoring the dynamics in global lakes, including water extent and level that are essential to deriving storage variability, has important ramifications to hydrology, ecology, the carbon cycle, and water sustainability (Yao et al., 2023).

Our capability to monitor global lake dynamics has been rapidly advancing with the expanding Earth-observing system (Cretaux et al., 2016). But, until recently, individual satellite missions for surface hydrology measured either water extent, such as through spectral radiometers and Synthetic Aperture Radar (SAR) imagers, or water surface elevation (WSE), such as through nadir-looking radar and lidar altimeters. This dilemma challenged the monitoring of water storage variation, which requires a synchronous acquisition or coordination of both variables. In addition, conventional radar altimeters usually have coarse footprint sizes (~10 km^2 or greater) and large inter-track distances (~50–100 km or wider), limiting adequate measurements to a few thousand largest lakes (Busker et al., 2019; Cretaux et al., 2016; Cretaux et al., 2011; Schwatke et al., 2015; Yao et al., 2023). With improvements of the waveform processing methods, SAR-mode altimeters such as those onboard Sentinel-3A and Sentinel-3B showed potential for measuring WSEs of lakes as small as a few hectares (Boy et al., 2022). Smaller footprints (~11–70 m) were also enabled by laser altimeters such as the Ice, Cloud, and land Elevation Satellite (ICESat) and its successor ICESat-2. However, their multi-month repeat cycles, along with discrete nadir footprints, limit the temporal density of WSE measurements for medium-sized and small lakes (Cooley et al., 2021; Luo et al., 2022). Fortunately, these technical challenges have been largely overcome by the Surface Water and Ocean Topography (SWOT) satellite mission (Biancamaria et al., 2016), recently launched on December 16, 2022.

The main payload of SWOT is a Ka-band (8.6 mm wavelength) radar interferometer (KaRIn). As the first of its kind, KaRIn provides synchronous, wide-swath, and orbital surveys of both surface water extent and elevation, allowing for the derivations of river discharge and lake storage change (Biancamaria et al., 2016; Durand et al., 2010). SWOT's lake observation requirement includes all enclosed water bodies larger than 250×250 m^2 (i.e., 6.25 ha) between 77°N and 77°S covering 90% of the continental surface, and the observation goal is lakes as small as 100×100 m^2 (i.e., 1 ha) (Biancamaria et al., 2016). Owing to the wide-swath (2×50 km) configuration, more than 90% of the global lakes larger than 1 ha are expected to be observed by SWOT at least once within each 21-day cycle of the three-year science or nominal orbit period

(JPL internal document, 2018). While these spatiotemporal coverages will reveal unprecedented details of global lake storage variability, a prerequisite for facilitating SWOT lake data production is the preparation of a Prior Lake Database (PLD).

The fundamental purpose of the SWOT PLD is to provide prior data on known lake locations, hereafter "prior lakes", making it possible to link KaRIn observations over time and to compute lake storage variation. KaRIn observes terrestrial water features (e.g., lakes and rivers) at a high-rate (HR) mode with fine spatial resolution (~5 m ×10–70 m) (Biancamaria et al., 2016). To accommodate user needs, the HR raw data are processed by the Science Algorithm Software (SAS) to different levels of products, which range from Level 1 single-look complex SAR images (L1B_HR_SLC) (JPL internal document, 2022c) intended only for highly specialized applications, to the standard Level 2 vector products delivering readily usable variables specific to rivers and lakes. The initial HR product suitable for general hydrological purposes is "pixel cloud" (L2_HR_PIXC) (JPL internal document, 2022d), which consists of geolocated pixel points with measured water heights but is not organized to distinct water features. With the help of the SWOT River Database (SWORD) (Altenau et al., 2021), pixels associated with prior rivers are first extracted to process the standard vector river products (JPL internal document, 2022a, 2022b, 2023). Such river pixels, except those also on SWORD-connected lakes, are eliminated from further lake processing. The remaining pixels are then segmented to individual water regions based on statistical clusters of the pixel heights. PIXC geolocations, however, contain noise from the interferogram (Desroches et al., 2016). By smoothing pixel heights across individual water regions, PIXC geolocations are corrected to a less noisy pixel cloud (L2_HR_PIXCVec) (CNES internal document, 2022c) for vectorization. With assistance of the PLD, the corresponding water features are processed to the standard vector lake products, which deliver the dynamics and uncertainties of WSE, area, and storage change (when applicable) for each prior lake per orbit pass (L2_HR_LakeSP) or cycle (L2_HR_LakeAvg) (CNES internal document, 2022a, 2022b).

Two primary components are required to fulfill the purpose of the PLD (Fig. 1). As lakes are often dynamic over time, their water surface may split and coalesce, and new lakes may emerge whereas others disappear. Without defining lakes a priori, it would be difficult to sort out how water features observed in different periods are spatially related to each other, which would then pose a challenge for effectively comparing lake changes. So, the first component of the PLD is a comprehensive prior mask that inventories global lakes larger than SWOT's observation goal (1 ha). Albeit temporally static, this lake mask offers a standardized spatial reference, based on which observed water features can be assigned, aggregated, or partitioned to the corresponding prior lakes. This ensures water dynamics, especially storage change, to be characterized and delivered consistently at the scale of each known lake. On the other hand, the lake mask also identifies observed water features that cannot be assigned to any prior lake. The unassigned features will be used to recursively improve the prior lake mask as SWOT data accumulate and to investigate the changes in wetlands, newly emerged lakes, and other relevant phenomenology. To make the prior lake mask functional, we need the second component of the PLD, namely "operational auxiliaries", which supplement the prior lakes with other necessary attributes, geometries, and logical information. The additional prior information works synergistically to ease the linkage of SWOT observations to the prior lakes, the calculation of lake storage change, and the population of the vector lake products.

An accurate and up-to-date prior lake mask is essential to the function of the PLD. We consider a lake mask qualified for the PLD to be "exhaustive" (including all lakes ≥ 1 ha), "exclusive" (excluding non-lake features), and "representative" (with lake polygons representing intermediate rather than extreme inundation conditions). At the current stage of the SWOT mission, we prefer intermediate water extent because it presents how a lake normally appears when being observed by SWOT, which eases the spatial linkage of SWOT observations to the prior lake. Despite the recent proliferation of global lake datasets, none of them alone can meet all three criteria. Two fine-resolution and publicly accessible global lake masks are HydroLAKES (v1.0) (Messager et al., 2016), which inventories 1.4 million lakes larger than 10 ha, and GLAKES, which comprises 3.4 million polygons depicting the maximum lake water extents large than 3 ha (Pi et al., 2022). The primary data source of HydroLAKES for the landmass below 60° N is the Shuttle Radar Topography Mission Water Body Dataset (SWBD) (Farr et al., 2007), where lake extents were based on water occurrence during February 2000. This timing concurred with the dry winter season across a large proportion of the northern hemisphere, meaning the sizes of many lakes in HydroLAKES are likely skewed towards their seasonal minimums. In addition, SWBD was acquired over twenty years ago, predating the recent prominent lake changes such as the shrinkage of many saline lakes (Wang et al., 2018; Wurtsbaugh et al., 2017), the expansion of glacial lakes (Nie et al., 2017; Shugar et al., 2020; Song et al., 2017), and the boom of new reservoir construction (Wang et al., 2022; Wu et al., 2023; Yao et al., 2023). Therefore, HydroLAKES may no longer accurately reflect the latest boundaries of many lakes in the world. In comparison, GLAKES used Landsat-derived Global Surface Water Occurrence (GSWO) dataset (Pekel et al., 2016) to extract all-time water area maximum during 1984 to 2019, where non-lake features (e.g., rivers, estuaries, and floodwaters) were removed by a deep-learning algorithm (Pi et al., 2022). While GLAKES is more up to date, the lake polygons correspond to the maximum extreme. Based on our visual inspection, these maximum extents occasionally include inundated riparian zones, floodplains, and paddy fields. Critically, neither dataset reaches a minimum lake size of 1 ha, meaning that lakes potentially visible to SWOT are not exhaustively inventoried.

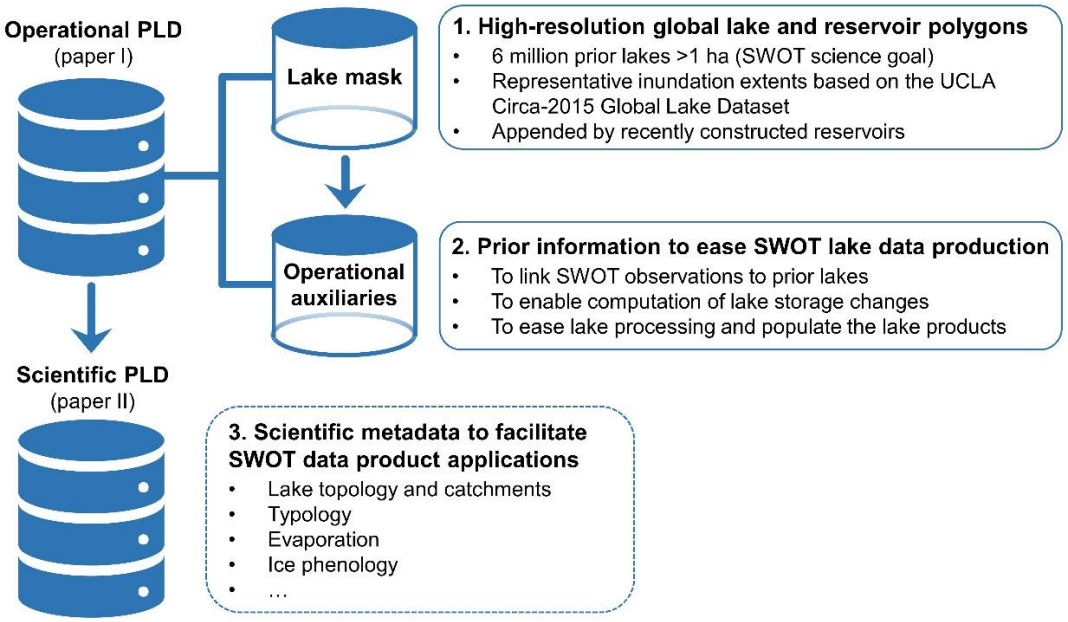


Figure 1. Conceptual structure of the SWOT Prior Lake Database (PLD).

Tailored to the SWOT mission needs, we describe the development of the SWOT PLD in a pair of companion papers (Fig. 1). The first paper (this article) emphasizes the prior lake mask and the operational metadata, which constitute the "operational PLD", addressing the above-described fundamental purpose for assisting SWOT lake data production. The second paper (in preparation) will focus on the development of a "scientific PLD", which consists of multi-theme scientific metadata to facilitate a wide range of limnological applications of the SWOT lake data products. Following the introduction, we describe the input data sources (section 2) and the methods (section 3) to construct the operational PLD. This is followed by the results (section 4) that present the prior lake mask comparatively with other lake datasets, the theoretical coverage for global lakes during a nominal orbit cycle, and the functionality of the operational metadata. With a primary focus on data development instead of algorithms (SASs), this paper does not elaborate how lake storage change is computed. However, we do describe the purpose of each prior attribute including those for computing lake storage change (section 3) and illustrate how the PLD works to ease SWOT lake data production (section 4). We then conclude the paper by discussing the plans of future PLD improvements and versioning (section 5).

2 Input data sources

We leveraged multiple data sources to compose the operational PLD. These input datasets and their contributions are summarized in Table 1. The primary data source is the UCLA Circa-2015 Global Lake Dataset (Sheng et al., 2016), which provides most of the polygons in the high-resolution prior lake mask. A collection of other datasets, covering the themes of lake name, reservoir identity, prior river locations, hydrography, and SWOT orbits, were used to populate the prior attribute information. Details of each input dataset are described below.

Table 1. Data sources used to develop the operational PLD

Data source	Contribution
UCLA Circa-2015 Global Lake Dataset (Sheng et al., 2016)	Provides the main source of prior lake mask
Georeferenced global Dams And Reservoirs (GeoDAR) dataset v1.1 (Wang et al., 2022)	Supplements the UCLA Circa-2015 Global Lake Dataset with additional recently constructed reservoirs
Post2k reservoir dataset (Fan et al., 2023)	
Other regional reservoirs (see Section 3.2)	
Global Reservoir and Dam database (GRanD) v1.3 (Lehner et al., 2011)	Provides the identities of large reservoirs
HydroBASINS (Lehner & Grill, 2013)	Populates basin IDs
SWORD (Altenau et al., 2021)	Identifies lakes on prior rivers, which are included for both lake and river data products
SWOT orbits (https://www.aviso.altimetry.fr)	Populates attributes related to the lake coverage by SWOT in the "lake" table
Global Lakes and Wetlands Database (GLWD) (Lehner & Doll, 2004)	Populates lake names
HydroLAKES v1.0 (Messager et al., 2016)	
Natural Earth Data (scale 1:30,000,000) (https://www.naturalearthdata.com)	
OpenStreetMap (OSM;	

https://www.openstreetmap.org)	
The IGN Carthage database (BD CARTHAGE®) (https://services.sandre.eaufrance.fr/telechargement/geo/ETH/BDCarthage/FXX/2017)	
Vector Map Level 0 (VMap0) (https://mdl.library.utoronto.ca/collections/geospatial-data/vector-map-level-0-vmap0)	

2.1 UCLA Circa-2015 Global Dataset

The foundation of the PLD, i.e., the prior lake mask, mainly comes from the UCLA Circa-2015 Global Lake Dataset (Sheng et al., 2016), or hereafter "Circa-2015" lake dataset. The Circa-2015 lake dataset inventories representative inundation extents of 9.0 million open-water lakes and reservoirs larger than 0.4 ha (i.e., four 30-m-resolution Landsat pixels) in the world. These lakes were mapped from a selection of high-quality Landsat-8 images acquired during the initial 2.5 years of the mission operation (May 2013 to August 2015). The complete mapping procedure, including image selection, water extraction algorithm, quality assurance and quality control (QA/QC), and multi-scene composition, was articulated for the case of Oceania by (Sheng et al., 2016), and the rest of the world was subsequently mapped using the same methods.

Compared to other global lake data, a unique merit of the Circa-2015 dataset is the emphasis on representative lake extents, echoing one of the three criteria expected for the SWOT prior lake mask (section 1). Specifically, the images selected for mapping were acquired during the "lake stable season" to minimize the misrepresentation of lake size due to intra-annual inundation extremes. The lake stable season was defined as the period after the rainy season, when inflows equal outflows and the lake thus reaches a stable condition within the annual cycle. To implement this idea, an image selection tool "LakeTime" (Lyons & Sheng, 2018) was developed using long-term climate data to determine the lake stable season independently for each Landsat tile. Cloud-free images were then collected tile by tile during the ideal period for lake mapping. This image selection process rendered a total of ~60,000 Landsat-8 scenes across the continents, with an average of about 6 scenes per tile.

For each selected scene, open water was segmented from land using a hierarchical and self-adaptive algorithm to ensure lakes across different landscapes can be mapped as accurately and thoroughly as possible (see (Sheng et al., 2016) for details). Together with a minimum mapping unit of 0.4 ha, the result satisfies the second criterion "exhaustive" for the prior lake mask. Since lakes are diverse aquatic systems, multiple factors such as water turbidity, mineral and chlorophyll contents, ice and snow, and mountain shadows can all complicate their spectral characteristics. To tackle this challenge, the adaptive mapping algorithm was automated to simulate how a human operator segments lakes from the background landscapes (Li & Sheng, 2012). In brief, each Landsat scene was first transformed to a normalized difference water index (NDWI) image (McFeeters, 1996) to enhance water appearance and suppress others. Then, the algorithm performs a two-step "global-to-local" segmentation. In the global segmentation, a loose preliminary NDWI threshold was used to flag potential lake extents across the entire scene. In the local segmentation, each flagged lake was re-exampled as an object, and the boundary was fine-tuned by an updated NDWI threshold, determined only using the spectral histogram based on the vicinity of the lake. The local segmentation was implemented iteratively until the result

converged to a stable water extent. Through this design, the final threshold and lake extent were tailored optimally to the unique spectral condition for each lake.

Following the automated mapping, a rigorous QA/QC process aided by a semi-automated editing tool (Wang et al., 2014) was performed to remove free-flowing river segments and to correct the remaining omission and commission errors. The resultant mapping contained only water bodies deemed as lakes and thus satisfies the third criterion "exclusive" for the SWOT prior lake mask. The quality-controlled lake extents from multi-temporal scenes were then composited across the Landsat tiles. With assistance of a previously produced circa-2000 reference lake map (Sheng et al., 2016), the median water extent during the lake stable season was selected as the final representative extent for each lake. To comply with SWOT's observation goal, the subset of the Circa-2015 lake dataset with lake size equal to or larger than 1 ha was used as the PLD prior lake mask.

2.2 Additional reservoir polygons

To ensure that the prior lake mask presents major reservoirs as thoroughly as possible, we supplemented the Circa-2015 lake dataset by another two global reservoir inventories. They are the Georeferenced global Dams And Reservoirs dataset (GeoDAR) v1.1 (Wang et al., 2022) and the Post2k reservoir dataset (Fan et al., 2023). GeoDAR v1.1 consists of 24,783 dam points and their associated reservoir polygons when detectable. The dam points harmonized the Global Reservoir and Dam database (GRanD) v1.3 (Lehner et al., 2011) (see section 2.4) and a georeferenced subset of the World Register of Dams from the International Commission on Large Dams (ICOLD; https://www.icold-cigb.org). Reservoir polygons were retrieved for each of the dam points by jointly using the water masks of HydroLAKES v1.0, GRanD v1.3, and the Circa-2015 lake dataset. This led to 21,515 reservoir polygons with a total area of 496,313.8 km^2, representing a cumulative storage capacity of 7216.1 km^3.

Post2k reservoir dataset contains 6,760 global reservoirs constructed after the year 2000. These post-2000 reservoirs were detected by comparing composite water occurrence probabilities before and after 2000, using the multi-decadal remote sensing products Global Surface Water (GSW) database (Pekel et al., 2016) and the Global Land Analysis and Discovery (GLAD) database (Pickens et al., 2020). Polygons of the verified post-2000 reservoirs were then retrieved using the maximum water occurrence maps of GSW and GLAD, such that each polygon represents the maximum inundation area of the reservoir from the construction to about 2020 and has a minimum size threshold of 0.5 km^2. These post-2000 reservoir polygons have a total area of 53,183.9 km^2, corresponding to a cumulative storage capacity of 1,287.7 km^3.

2.3 SWORD

SWORD is the official a priori river database for SWOT (Altenau et al., 2021). It defines the global networks of mainstems and tributaries potentially visible to SWOT (i.e., wider than 50 m according to SWOT's observation goal) (Biancamaria et al., 2016) and serves as the framework for the SWOT vector river products. Because its primary data source is the Global River Widths from Landsat (GRWL) database (Allen & Pavelsky, 2018), SWORD also contains river reaches with mean annual flow widths as narrow as 30 m. In total, SWORD consists of 213,485 river reaches (centerlines) with a median length of 10.5 km, comprising 10.7 million nodes with ~200 m spacing. The SWOT river vector products, which contain WSE, width, slope, and discharge, will be disseminated at the scales of both river reach and node. In addition,

SWORD also used multiple auxiliary datasets to provide a wide range of hydrological and morphological attributes such as reach sinuosity, average width, slope, natural and human-created obstructions, and the topology structure among the reaches and nodes. These attributes facilitate the processing of SWOT river products as well as their scientific applications. Here we used SWORD version 15 to identify the PLD lakes that are directly connected to the river networks visible to SWOT, and the intersecting water bodies will be considered in both lake and river products.

2.4 GRanD

GRanD is one of the most comprehensive spatial repositories of large dams and reservoirs in the world (Lehner et al., 2011). GRanD was constructed by harmonizing a collection of open-access dam and reservoir data, including the United Nations Food and Agricultural Organization (FAO) AQUASTAT (https://www.fao.org/aquastat/en/databases/dams) and multiple regional inventories and registers, to form a single, congruent global database. The latest version v1.3 contains 7320 georeferenced dams and their associated reservoir polygons when possible. Each reservoir feature is also provided with over 50 attributes such as reservoir name, storage capacity, and purpose. While the primary goal is to inventory all reservoirs with a storage capacity greater than 0.1 km^3, GRanD v1.3 includes 3992 smaller reservoirs, leading to a total storage capacity of 6881 km^3 in the entire database. The reservoirs also include 119 regulated natural lakes such as Lake Victoria and Lake Ontario. While a more exhaustive inclusion of smaller and/or newer reservoirs is important, we used GRanD v1.3 to flag some of the largest manmade reservoirs and regulated lakes as an a priori attribute for the operational PLD. Polygons in GRanD were not used to construct the geometry of the PLD lake mask.

2.5 HydroBASINS

HydroBASINS (Lehner & Grill, 2013) offers a global tessellation of hierarchically nested basins and subbasins at various scales, derived primarily from the HydroSHEDS hydrography dataset at a grid resolution of 15 arc seconds (~500 m at the equator) (Lehner et al., 2008). Following the Pfafstetter coding system (Verdin & Verdin, 1999), the basin hierarchy in HydroBASINS is broken down to 12 nesting levels. They range from level 1 containing 9 continental or subcontinental boundaries, to level 12 encompassing about 1.0 million subbasins at a scale of only tens of square kilometers. In other words, the basins of a lower level consecutively comprise the subbasins of a higher level. For clarity, the subbasins corresponding to each level are organized as a different data layer. We used the data layers at Pfafstetter levels 3 in HydroBASINS v1.c, which contains 291 basin polygons together with their associated Pfafstetter codes at level 2 (corresponding to 62 larger basins) and level 1 (corresponding to 9 continental and subcontinental divisions). These level-3 basin boundaries and their Pfafstetter codes were used to help structure the prior lake identifier (*lake_id* attribute) and partition the PLD into level-2 basin granules (section 3.1).

2.6 SWOT orbit files

The SWOT mission is split in two phases related to two different orbits (JPL internal document, 2022e). The initial Calibration/Validation (Cal/Val) phase, up to July 11st, 2023, was related to a 1-day orbit at an 857 km of altitude: by frequent revisits of specific sites, this phase enabled the calibration of radar system parameters in the shortest time; it also allowed the study

of rapidly changing phenomena. The science orbit, on which SWOT has been placed since July 21st, 2023, is a non-sun-synchronous 21-day orbit, at an 890.6 km of altitude. Combined with the swath of the satellite, this orbit allows a quasi-global coverage until 77° of latitude north and south and, with its 10-day sub-cycle, is a good compromise for the temporal sampling as a region may be observed between once a cycle (at the Equator) to more than ten times a cycle (at the highest latitudes).

Different from nadir-pointing altimetry missions, which provide measurements just below the satellite, SWOT KaRIn makes observations with a 120 km wide swath, from approximately 10 to 60 km of its nadir, on both "right" and "left" sides. The terms "left" and "right" are defined as if one stands on the Earth surface at the spacecraft nadir point facing in the direction of the spacecraft velocity vector. The 10 to 60 km width swath orbit file (for the CalVal orbit: https://www.aviso.altimetry.fr/fileadmin/documents/missions/Swot/sph_calval_swath.zip; for the Science orbit: https://www.aviso.altimetry.fr/fileadmin/documents/missions/Swot/swot_science_orbit_sept2015-v2_10s_swath.zip) was considered to compute the observability of each lake by SWOT. It provides the full swath per pass. A "pass" is a half revolution of the Earth by the satellite from pole to pole (south to north latitudes for ascending passes, and north to south latitudes for descending passes). There are 28 passes for the 1-day orbit, and 584 passes for the 21-day orbit.

2.7 Databases for lake names

Multiple databases or open-source online repositories were jointly used to populate lake names for the PLD polygons as thoroughly as possible. These sources include the IGN Carthage database (BD CARTHAGE®) to cover lakes in France (https://services.sandre.eaufrance.fr/telechargement/geo/ETH/BDCarthage/FXX/2017), the OpenStreetMap (OSM; https://www.openstreetmap.org), the Global Lakes and Wetlands Database (GLWD) (Lehner & Doll, 2004), the Natural Earth Data (scale 1:30,000,000) (https://www.naturalearthdata.com), the Vector Map Level 0 (VMap0) (https://mdl.library.utoronto.ca/collections/geospatial-data/vector-map-level-0-vmap0), and HydroLAKES (v1.0) (Messager et al., 2016).

3 Database development

3.1 Overview of the operational PLD

Conceptually, the operational PLD is comprised of two primary components (Fig. 1): (1) the prior lake mask, which inventories the polygon geometries of global lakes potentially visible to SWOT (i.e., ≥1 ha); and (2) the operational auxiliaries, which facilitate the linkage of SWOT observations to the prior lakes and assemble prior information necessary to compute lake storage change and populate the lake products. Analogously, the prior lake mask sets up the data infrastructure, while the operational auxiliaries assist the SAS in embodying the functionality of the data infrastructure.

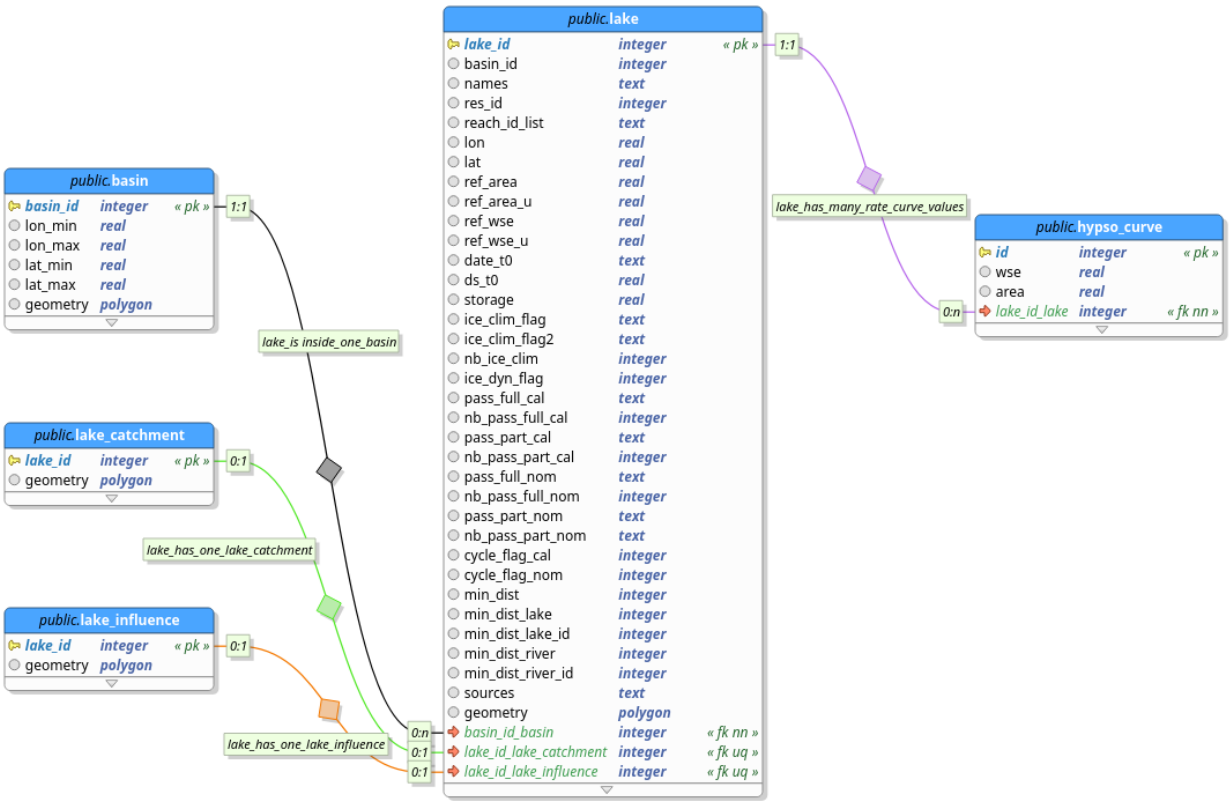


Figure 2. Structural model of the operational PLD.

Structurally, the operational PLD is a relational database (Fig. 2) which ties the central "lake" table to four auxiliary tables: "lake_catchment", "lake_influence", "basin", and "hypso_curve". Following the terminology of data science, here "table" refers to an arrangement of records that may contain fields of both geometry (e.g., raw polygons) and non-geometry (e.g., other numeric and text prior attributes). The central "lake" table consists of the polygons of the prior lake mask and a set of attributes for each prior lake. The prior lake mask is used to link SWOT-observed water features to the prior lakes by intersecting their geometries. The other attributes store prior information to calculate lake water storage change and to help populate the vector products at two granule levels, including the standard lake single pass vector product (L2_HR_LakeSP) in continental-pass granules (CNES internal document, 2022b) and the standard lake average cycle product (L2_HR_LakeAvg) in Pfafstetter level-2 basin granules (CNES internal document, 2022a). The "lake" table also contains an attribute to link the prior lakes and prior rivers (SWORD), such that pixels of lakes connected to the prior river networks, the so-called "connected lakes", are also included in lake data processing. For clarity, we refer to the non-geometric attributes of the "lake" table and the entirety of the other ancillary tables ("lake_catchment", "lake_influence", "basins", and "hypso_curve") as operational auxiliaries, which collectively supplement the prior lake mask to enable the expected functions of the operational PLD.

The "lake_catchment" and "lake_influence" tables contain ancillary geometries to accelerate the assignment of SWOT observations to the prior lakes. The issue is that the spatial linkage between SWOT-observed water features and the prior lakes does not always follow a one-to-one relationship. Particularly, complexities arise when an observed water feature is

intersected by multiple prior lakes, leading to ambiguity regarding how the pixels in the observed water feature should be assigned to each of the prior lakes. To tackle the issue, these two assignment tables delineate a spatial partition for each prior lake stored in the "lake" table. Each lake assignment polygon defines the spatial domain within which the associated prior lake is allowed to "expand" before it infringes the domain of another prior lake. In other words, the lake assignment polygons disambiguate the vicinity of each prior lake so that lake assignment in complex spatial relationships can be eased (see examples in section 4.4). The "lake_catchment" geometries provide a spatial partition that takes into account hydrological constraints and topography while "lake_influence" geometries take into account only distances between lakes.

The "hypso_curve" table stores the information of lake hypsometry for computing water storage changes. This ancillary table will be added to the PLD approximately one year into the SWOT mission. More specifically, this table will contain discrete WSE and water area points on the hypsometric curve (i.e., WSE-area relationship) for each prior lake. The curve will be fitted using the pairs of SWOT WSE and water area measurements to be collected from the first valid observation of this lake throughout a certain mission period. Since the hypsometric points account for the variation in lake bathymetry, they will allow for lake storage changes to be estimated in an "incremental" approach (Cretaux et al., 2016) (CNES internal document, 2023a), which is theoretically more accurate than the "direct" approach assuming an invariant bathymetric shape.

The operational PLD is organized by HydroBASINS level-2 basins (section 2.5), which results in 61 valid basin-granule PLD files. The "lake" table in each basin-granule PLD includes only the prior lakes intersected by the associated level-2 basin. The "lake_catchment" and "lake_influence" tables include the catchment and influence polygons intersecting this level-2 basin, respectively. The "basin" table delineates the full boundaries of HydroBASINS level-3 basins nested within this level-2 basin (section 2.5), together with the associated basin Pfafstetter codes. This table is used to label the observed water features in different continents and basins. More details on the development of each table, except "hypso_curve", are given in the following subsections.

3.2 Prior lake mask

The Circa-2015 lake dataset (Sheng et al., 2016) (section 2.1) was used as the primary source of the prior lake mask. To improve the representation of reservoirs, particularly those constructed after 2015, a few state-of-the-art global and regional reservoir databases (section 2.2) were integrated to the Circa-2015 dataset to form the final prior lake mask.

The reservoirs in GeoDAR v1.1 (Wang et al., 2022) and Post2k (Fan et al., 2023) databases that are not intersected by any Circa-2015 polygon were first added successively to the prior lake mask. The remaining Post2k reservoirs were next investigated based on their spatial relationship with the updated prior lakes. High-resolution Esri and Google Earth imagery were also employed to assist in visual inspection. When a prior lake spatially conflicts with more than one Post2k reservoir, we examined whether this prior lake overshoots the dam location and mistakenly spans multiple reservoirs. If verified, this prior polygon was manually split to multiple reservoirs. When a prior lake intersects with only one Post2k reservoir, we examined whether the Post2k reservoir was substantially overrun by its intersecting prior lake. This possibility was identified when the Post2k reservoir is well included (>75%) by the prior polygon but the latter is less well covered (<75%) by the former. We then visually inspected if

this prior polygon mistakenly annexed the reservoir depicted by Post2k; if verified, this prior polygon was truncated, allowing the Post2k polygon to be added as a new reservoir without topological conflicts. For the rest of the cases (one Post2k reservoir intersected by one or multiple prior lakes), we classified them based on the spatial agreement between the two data sources. If their overlapping area covers at least 50% of the lake area in both sources, we considered the Post2k reservoir and its intersecting prior lake(s) in good agreement and thus excluded them from further investigation. Otherwise, the case was visually inspected, and when necessary, the prior polygon was split or replaced by the intersecting Post2k reservoir.

The improved prior lakes were next compared with the remaining GeoDAR reservoirs. The procedure was overall similar to the one for Post2k reservoirs, except that we employed a more qualitative approach in comparing GeoDAR and prior polygons, and that the comparison was focused on large reservoirs only. This was because many small and medium-sized reservoir polygons in GeoDAR are already sourced from the Circa-2015 lake dataset, and the other polygons sourced from HydroLAKES and GRanD usually exhibit coarser shorelines (Wang et al., 2022). Nevertheless, when a GeoDAR polygon shows a major superiority in representing the reservoir integrity (e.g., with improved shoreline connectivity and reduced surface water patchiness), the GeoDAR polygon was used to replace the intersecting prior lake(s). In occasional cases where a single source is not a sufficient solution, we performed manual digitization to modify and merge multiple sources. The data sources and harmonizing methods were reflected in the attribute *source* of the prior lake mask (Table 2).

Additional regional improvements were further made on the prior lake mask after the integration of global reservoir databases. In particular, we included nearly 7,000 reservoirs in the Crateús and Banabuiú basins of Brazil to refine the completeness and accuracy of reservoir mapping in this hotspot region. These Brazilian reservoirs were mapped from Landsat surface reflectance images using water index spectral thresholds as in (Fisher et al., 2016) to represent the interannual water area maximum during 2008 to 2019. We also improved the mapping of several critical reservoirs in semi-arid western Africa, which are typically covered by aquatic vegetation and difficult to delineate using global algorithms. These reservoirs were extracted following a supervised classification of Sentinel-2 images using the Active Learning for Cloud Detection (ALCD) algorithm (Papa et al., 2023) and/or spectral thresholding of the modified NDWI (MNDWI) with an ad hoc threshold for each lake (de Fleury et al., 2023). Finally, the updated prior lakes were post-processed such that polygons sharing a common vertex were concatenated by a narrow channel and polygons sharing a common border were separated by a small gap. This post-processing reduced the number of original prior polygons by a minor extent but improved the connectivity of lake surface and eliminated topological ambiguity.

Table 2. Key attributes in the operational PLD tables.

Attribute name	Description
"lake" table	
lake_id	Unique identifier (ID) of the prior lake.
lat / lon	Latitude and longitude (in decimal degree) of the centroid of the prior lake
names	Known name(s) of the prior lake. If one lake has several names, the names are separated by semicolons.
res_id	Reservoir ID from the Global Reservoir and Dam database (GRanD v1.3), if the prior lake intersects a GRanD reservoir

reach_id_list	List of the IDs of the SWORD reaches that intersect the prior lake. If there are more than one reach, the IDs are concatenated by semicolon.
ref_area / ref_area_u	Maximum area and the associated uncertainty (in m^2) of the prior lake. Values correspond to the same state as *ref_wse*.
ref_wse / ref_wse_u	Maximum water surface elevation (WSE) and the associated uncertainty (in m) of the prior lake, with respect to the reference geoid (EGM 2008).
date_t0	Reference date from which the storage change is computed. The reference date will be populated as the date of the first SWOT observation of the lake.
ds_t0	Storage change (in m^3) between the lake at the maximum WSE (*ref_wse*) and the lake state at the reference date (*date_t0*).
storage	Maximum water storage variation (in m^3) observed by SWOT, i.e., the storage difference between the maximum and minimum WSEs observed by SWOT
ice_clim_flag / ice_clim_flag2	Flag characterizing the presence of ice covering the lake during each day of a calendar year based climatological data: 0: never frozen 1: can be frozen 2: always frozen *ice_clim_flag is a text string containing the ice flags from January 1st to June 30th, and ice_clim_flag2 contains information from July 1st to December 31st.*
pass_[full/part]_[cal/nom]	List of the IDs of the SWOT passes fully or partially covering the prior lake during a calibration or nominal orbit cycle. The IDs are separated by semicolons.
cycle_flag_[cal/nom]	Flag characterizing the scenario of lake observation by SWOT during a calibration or nominal orbit cycle. 0: never observed 1: only partially observed 2: fully observed after aggregating partial observations in multiple passes 3: fully observed by a single pass at least once
min_dist_[lake/river][_id]	Geodesic distance (in m) to the closest lake or river and the corresponding *lake_id or reach_id*
sources	Data source of the prior lake polygon
"lake_catchment" table	
lake_id	ID of the prior lake that the lake_catchment polygon encompasses
"lake_influence" table	
lake_id	ID of the prior lake that the lake_influence polygon encompasses
"hypso_curve" table (expected to be generated approximately one year into the SWOT mission)	
id	ID of the WSE-area pair
lake_id	ID of the prior lake with which the WSE-area pair is associated
wse /area	WSE (in m) and water area (in m^2) values of a discrete point on the hypsometric curve of the prior lake. The hypsometric curve is fitted using available SWOT measurements of the lake WSE and water area since the first observation of this prior lake.
"basin" table	
basin_id	ID of the HydroBASINS Pfafstetter level-3 basin
lat_[min/max]	Minimum and maximum latitudes (in decimal degree) of the basin boundary
lon_[min/max]	Minimum and maximum longitudes (in decimal degree) of the basin boundary

3.3 Attributes in "lake" table

The attributes in the "lake" table (Table 2) provide multi-theme information for each prior lake polygon, which covers basic lake identities, relationship with drainage basins and prior

rivers, reference WSE and water area for deriving lake storage change, and SWOT overpass statistics to enable data processing and product distribution. Accompanying the attribute definitions in Table 2, we provide additional details that are necessary for understanding the attribute format, purpose, and populating method.

3.3.1 Lake identities

The primary key, *lake_id*, is a ten-character string in the format CBBNNNNNNT, where C is a one-digit continent code (Table 3), BB a two-digit basin code, NNNNNN a zero-padded, six-digit sequence representing the ordinal index of the lake within its associated basin, and T a one-digit code indicating the water body type (Table 4). Only integers 0 to 9 are allowed for each digit. The first three digits (CBB) are based on the Pfafstetter coding system used in HydroBASINS. The continental code (C) corresponds to level-1 divisions (Table 3), and BB concatenates the codes of level-2 and level-3 basins representing increasing drainage details. This hierarchy organized the global prior lakes to 291 subbasins at the scale of Pfafstetter level 3 (see section 3.5), and the coding assignment was based on geometric intersection with HydroBASINS boundaries. Following the Pfafstetter coding system, prior lakes within each level-3 basin are then indexed from 000001 to a maximum of 999999 based on a random order.

Table 3. PLD lake abundance in each Pfafstetter level-1 continental divisions.

Continent code (ID)	Continent name	Lake count	Lake area (ha)	Mean / median area (ha)	Lake density (%)
1 (AF)	Africa	73,781	24,377,083.2	330.4 / 3.2	0.8
2 (EU)	Europe and Middle East	504,426	28,023,908.5	55.6 / 3.7	1.6
3 (SI)	Siberia	1,113,275	28,945,547.3	26.0 / 4.0	2.2
4 (AS)	Central and Southeast Asia	446,736	23,450,116.9	52.5 / 2.5	1.1
5 (AU)	Australia and Oceania	57,338	7,332,072.6	127.9 / 3.8	0.7
6 (SA)	South America	250,212	16,435,120.5	65.7 / 3.6	0.9
7 (NA)	North America and Caribbean	1,512,139	82,537,606.4	54.6 / 3.9	5.2
8 (AR)	North American Arctic	1,899,665	47,750,812.6	25.1 / 3.5	7.6
9 (GR)	Greenland	40,759	894,185.3	21.9 / 3.3	0.4
Global		**5,898,331**	**259,746,453.3**	**44.0 / 3.6**	**1.9**

The last digit in *lake_id* classifies global prior lakes to two water body types based on their geometric connectivity with prior rivers (Table 4). The same water body type codes are also used for the primary key in SWORD (Altenau et al., 2021). As listed in Table 4, each prior lake was categorized to either a connected lake (T = 3) or a disconnected lake (T = 2). A connected lake is defined as any prior lake polygon intersected by one or more prior reaches and is included in both river and lake data processing. It is worth noting that this connectivity was determined specifically in relation to SWORD. This means a “disconnected” prior lake may also be

hydrologically connected to a river, but the river is too narrow to be observed by SWOT and is therefore not inventoried in SWORD.

Table 4. Water body type codes in the prior lake and river IDs

Type code (T)	Water body type
1	River (only applicable to SWORD)
2	Disconnected lake
3	Connected lake
4	Dam (only applicable to SWORD)
5	No topology (only applicable to SWORD)

Figure 3 illustrates an example to help interpret the hierarchy of *lake_id*. This example covers the Pfafstetter coding system in the North American continent (C = 7), which encompasses eight level-2 basins. One of them (first B = 4) contains the Mississippi River Basin (second B = 2) at level 3. There are 109,861 prior lakes in the Mississippi River Basin, which all share "742" as the first three digits in *lake_id*. Among them is an example lake "7420469602", indicating that this lake is indexed to be the 46960th in the basin (NNNNNN = 046960) and is disconnected from any prior rivers in SWORD (T = 2).

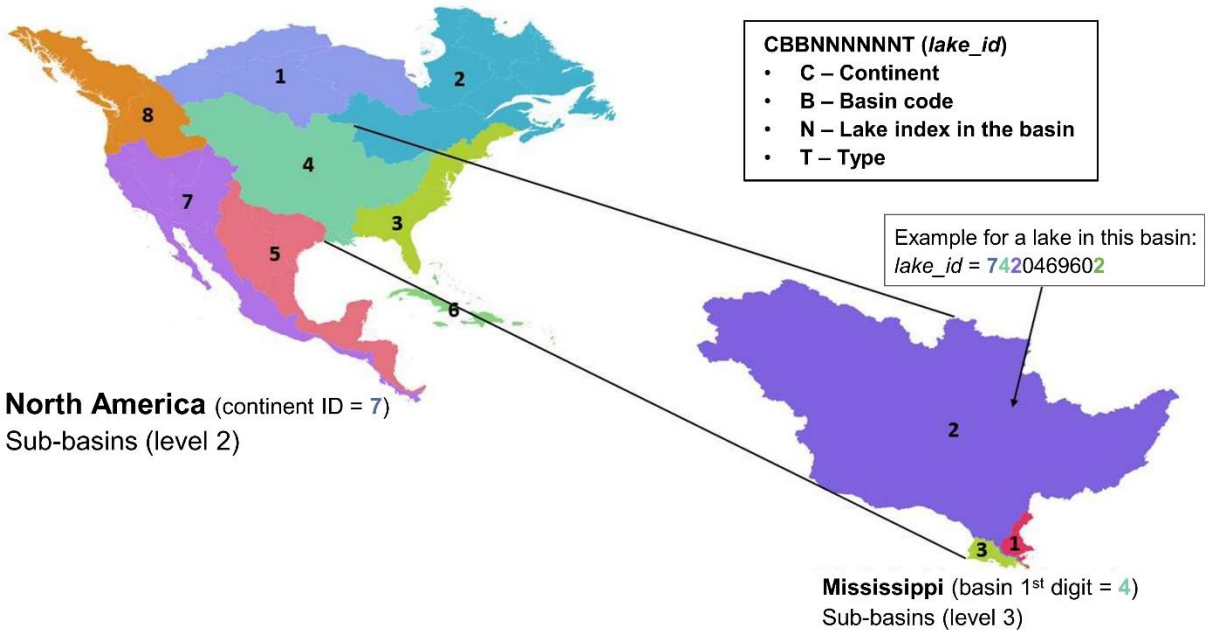


Figure 3. Hierarchical structure of the 10-character *lake_id* for prior lakes. The example is given to a disconnected prior lake (T = 2) in a Pfafstetter level-3 basin (BB = 42, the Mississippi River Basin) of the North American continent (C = 7).

The *names* attribute inventories the known names of global prior lakes as thoroughly as possible for the convenience of PLD and SWOT science data users. The lake names were populated through "spatial join" from multiple open-source atlases and databases, including the IGN Carthage database for France, OSM, GLWD, the Natural Earth Data, VMap0, and HydroLAKES (v1.0) (section 2.7; Table 1). All names are in capital letters to avoid accents and

other spelling discrepancies. The same name can be shared by several prior polygons if they are disconnected portions of the same lake due to either mapping issues or seasonal variation. From this aspect, the *names* attribute is potentially useful for dissolving patchy water bodies with known lakes to improve the integrity of their prior extents and the completeness of storage change estimates. A total of 152,260 lake names were assigned to 329,376 prior lake polygons, which account for 5.6% of the global lakes by count and 61.8% by area.

The prior lake polygons include both natural lakes and artificial reservoirs. While classifying lake typology is not the priority of the operational PLD, the "lake" table does provide a *res_id* attribute, which flags about 7300 large reservoirs using the IDs of GRanD v1.3. These IDs were populated by intersecting the prior lakes with GRanD reservoir polygons. If a prior lake intersects more than one reservoir, only the ID of the GRanD reservoir containing the prior lake centroid was used. Although GRanD focuses on the world's largest reservoirs (e.g., with storage capacity exceeding 0.1 km^3), this flag allows for a preliminary attribution of SWOT-measured water storage changes to either climate or human regulation. More comprehensive information about reservoirs and other lake types is available in the "scientific PLD".

3.3.2 Relations with SWOT-visible rivers

The *reach_id_list* attribute identifies each river-connected prior lake by the IDs of the intersecting SWORD reaches. For each identified prior reach, SWOT-detected water pixels that correspond to the lake portion are kept for both lake and river data processing whereas the other pixels on the reach are eliminated from further lake processing. The specific reach IDs will also facilitate a potential synergy of SWOT lake and river data products. One example is the LakeFlow algorithm (Riggs et al., 2023), which uses both products and the concept of lake-river mass conservation to improve the estimates of lake inflow and outflow. The *reach_id_list* attribute identified 16,499 prior lakes connected to 43,247 prior river reaches, and these connected lakes account for 38.4% of the global lake area. More advanced information on lake drainage topology and lake-river connectivity will be available in the "scientific PLD".

3.3.3 Prior information for computing lake storage change

An essential role of the operational PLD is to assist the SAS in turning lake area and WSE repeatedly measured by SWOT to lake water storage variation. For this purpose, the "lake" table reserves several attributes associated with the reference water state for each prior lake, based on which water storage change (i.e., the output variable *delta_s* in the lake products) can be computed. These attributes start with *date_t0*, which defines the date of the first valid SWOT observation of each prior lake. The WSE and water area on this initial date set up the reference state for computing *delta_s*. In other words, even though lake storage algorithms in the SAS vary in bathymetrical model (linear or quadratic) and integration approach (direct or incremental), the output *delta_s* conceptually always represents the storage change from the observed state (i.e., WSE and water area at a given time t_i) to the reference state defined by *date_t0* (see Fig. 4).

For practical reasons in bathymetric and hypsometric modeling, the calculation is first performed for the lake storage change ($\Delta V(t_i)$) between t_i and a high water level state defined by the *ref_wse* and *ref_area* attributes. Specifically, *ref_wse* quantifies the maximum WSE of the prior lake during a certain SWOT observation period, and *ref_area* stores the inundation area corresponding to *ref_wse*. Their associated uncertainties are given in *ref_area_u* and *ref_wse_u*, which are needed for propagating storage change errors. The storage difference between the two

states (i.e., the high level state and the reference state on *date_t0*) is provided in the *ds_t0* attribute. This way, *delta_s* can be derived by subtracting *ds_t0* from $\Delta V(t_i)$ (Fig. 4). Technical details on lake storage change algorithm and error propagation are beyond the scope of this paper but are available in the Algorithm Theoretical Basis Document (CNES internal document, 2023a).

It is important to note that the reference state on *date_t0* does not necessarily correspond to the minimum level of the prior lake. However, the "lake" table provides another attribute *storage*, which quantifies the storage change between the maximum and minimum WSEs for each lake during the same period for calculating *ref_wse* and *ref_area*. This attribute estimates the magnitude of possible storage variation per prior lake, which is needed for assessing the scales of intermediate storage variation relative to the maximum storage change magnitude. Lacking sufficient SWOT observations so far, *ref_area* is temporarily populated as the area of the prior lake polygon whereas the other attributes are filled with "no data" and will be populated at the first major update of the operational PLD (see section 5).

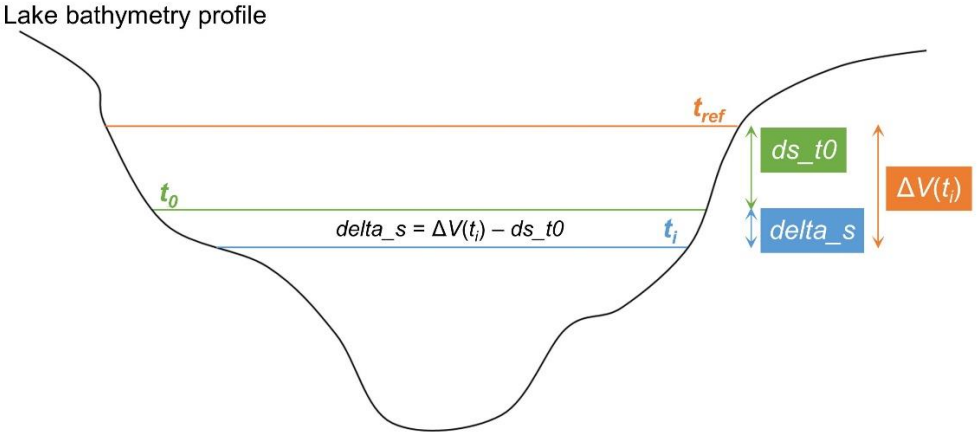


Figure 4. Illustration of different water states used in lake storage change calculations.

3.3.4 SWOT overpasses and lake coverage

Lastly, the "lake" table contains a few more attributes that describe SWOT's coverage of the prior lakes in relation to orbit passes. These attributes inform how well each prior lake can be observed under a single pass or after aggregating multiple passes during a calibration or nominal orbit cycle. The *pass_full* and *pass_part* attributes list the IDs of the passes covering each prior lake fully and partially, respectively. Their values were configured by intersecting the prior lakes and the orbit passes with swaths covering 10-60 km from nadir (section 2.6). The intersection applied a 5-km buffer to take into account SWOT orbit jitter. These two attributes can be used to quantify how many times each lake can be observed partially, completely, or both during an orbit cycle (see section 4.3). Using this information, the *cycle_flag* attribute summarizes SWOT's lake coverage into four scenarios. Scenario "0" flags the prior lakes that will never be observed by SWOT. This was determined by the lakes where both *pass_full* and *pass_part* values are empty. Scenario "1" indicates that the lake will only be partially observed even after aggregating all passes over a cycle, and scenario "2" indicates that the lake can be fully observed by SWOT, but only after pass aggregation over a cycle. In both scenarios, *pass_part* has valid pass IDs while *pass_full* is empty. Finally, scenario "3" flags all prior lakes that will be observed fully by at

least one single pass. This was determined by the prior lakes where *pass_full* has valid pass IDs regardless of *pass_part*.

3.3.5 Lake ice flag

The goal of *ice_clim_flag* (climatological ice flag) is to help the data user make decisions on removing potentially ice-affected SWOT lake products, and to allow the SAS to calculate the *ice_clim_f* attribute (i.e., a climatological flag indicating whether the lake is ice-covered on the day of the observation based on *ice_clim_flag*) in the vector lake product (CNES internal document, 2022b). Climatological ice flags are estimated ice conditions for a typical year, averaging ice conditions between January 1st 2010 and January 1st 2020. Here we briefly describe the two steps taken to develop the lake ice flag.

Development of a lake ice fraction empirical model. To develop a priori ice conditions for all prior lakes, we applied an empirical lake ice fraction model by matching same-day ice fractional data derived from Landsat 5, 7, and 8 images, whenever cloud-free conditions were observed, with daily surface air temperature from ERA5 climate reanalysis data (Copernicus Climate Change Service, 2017). The lake ice fraction was calculated based on the lake ice detection algorithm (SLIDE) (Yang, Pavelsky, et al., 2022) for each prior lake polygon. By modeling the lake ice fraction with daily-mean air temperature, we identified the following logistic regression:

$$\log\left(odds(P_{ice})\right) = -0.46 \cdot SAT_{30} - 0.02 \cdot SAT_{30} \cdot Period + 0.85 \qquad (1)$$

where P_{ice} denotes the lake ice area fraction; SAT_{30} denotes the prior 30-day mean surface air temperature; and $Period$, a categorical variable, denotes whether the calculation was carried out during the breakup months ($Period$ = 1 when Julian day is between [70, 227]; $Period$ = 0 otherwise). Adding the variable $Period$ allowed the model to accommodate the difference in ice dynamics during the breakup and freeze-up, a difference that has been previously identified in other types of freshwater bodies (Lacroix et al., 2005).

Estimating lake ice flag. For each point geometry representing the prior lake centroid, and for each day during the period between January 1st 2010 and January 1st 2020, we estimated the ten-year mean lake ice fraction by inputting daily mean surface air temperature from ERA5 reanalysis database (variable: mean_2m_air_temperature) to the empirical lake ice model above. Then, a climatological mean lake ice fraction was estimated by averaging lake ice fraction across the ten years for each Julian day. At last, the continuous ice fraction was converted to three discrete integer values to represent ice conditions for SWOT ice flag: mean ice fraction < 0.2: 0; 0.2 ≤ mean ice probability < 0.8: 1; and mean ice probability ≥ 0.8: 2.

This flag can suggest likely ice cover conditions at the given time of year for a given prior lake based on modeled historical ice conditions. However, factors such as interannual variability for ice phenology, multiple freeze-thaw events during cold seasons, and non-stationarity in climate mean that users are encouraged to seek ice conditions that are more recent and locally relevant whenever those sources are available. When no other sources are available, the climatological flag provides a reasonable representation of the average ice condition.

3.4 "Lake_catchment" and "lake_influence" tables

The "lake_catchment" and "lake_influence" tables store the assignment polygons for

each of the prior lakes referenced in the "lake" table. By definition, a lake assignment polygon should encompass the associated prior lake as well as its water fluctuation zone; meanwhile, it should not overlap those of any other prior lakes but collectively, the assignment polygons partition the entire continental surface. This way, when it is unclear how a SWOT-detected water region should be assigned to the prior lakes using the prior lake geometry alone, the assignment polygons can help determine the rule for executing lake assignment (see sections 3.1 and 4.4). In addition to the geometries, each assignment polygon is also indexed by the ID of the encompassed prior lake, *lake_id* (section 3.3.1), which links the "lake_catchment" and "lake_influence" tables to the "lake" table.

We considered two rationales for constructing lake assignment polygons. The first rationale follows the concept of lake hydrological catchment, which defines the sub-basin between the outlets of a prior lake and its immediate upstream prior lake(s). If a prior lake is in the headwater (meaning no lakes further upstream), the catchment is then the entire watershed upstream to the outlet of this lake. As water dynamics in a lake are confined by its own catchment boundary, this rationale complies with the ideal definition of lake assignment polygons described above. To implement this rationale, we applied the algorithm recently developed for the global Lake Topology and Catchment (Lake-TopoCat) database (Sikder et al., 2023) on the prior lake mask and the 90-m-resolution MERIT-Hydro hydrography data (Yamazaki et al., 2019). Results of the algorithm are fine-detailed catchments for each of the prior lakes, which compose the geometries of the "lake_catchment" table. A regional example is given for part of western Africa in Fig. 5a.

The second rationale relies on geometric vicinity. Specifically, we employed the Voronoi tessellation (Aurenhammer, 1991) to partition the continental surface into proximal regions based on the geodesic distance to the prior lakes, and the resultant regions, also known as Voronoi cells or Thiessen polygons, are the geometries of the "lake_influence" table (see the example of Fig. 5b). Mathematically, the Voronoi tessellation decomposes a plane with a finite number of objects, or the so-called "seeds", into the same number of Thiessen polygons. Each Thiessen polygon corresponds to one seed object, e.g., a prior lake in our case, and every virtual point within this polygon is closer to its seed prior lake than to any other prior lake. Because of this proximal characteristic, Thiessen polygons are often regarded as the "areas of influence" in computational geometry and have been widely applied in hydrology, meteorology, and geo-statistics (Evans & Jones, 1987). Although these influence features do not follow the exact lake catchment boundaries (Fig. 5), it is important to note that assignment polygons are not needed for every case of lake assignment. When they are indeed needed, the Thiessen polygons provide a computationally efficient alternative to ease the linkage of SWOT observations to the prior lakes. An example of when lake assignment polygons are required and how they function to ease lake linkage is given in section 4.4.

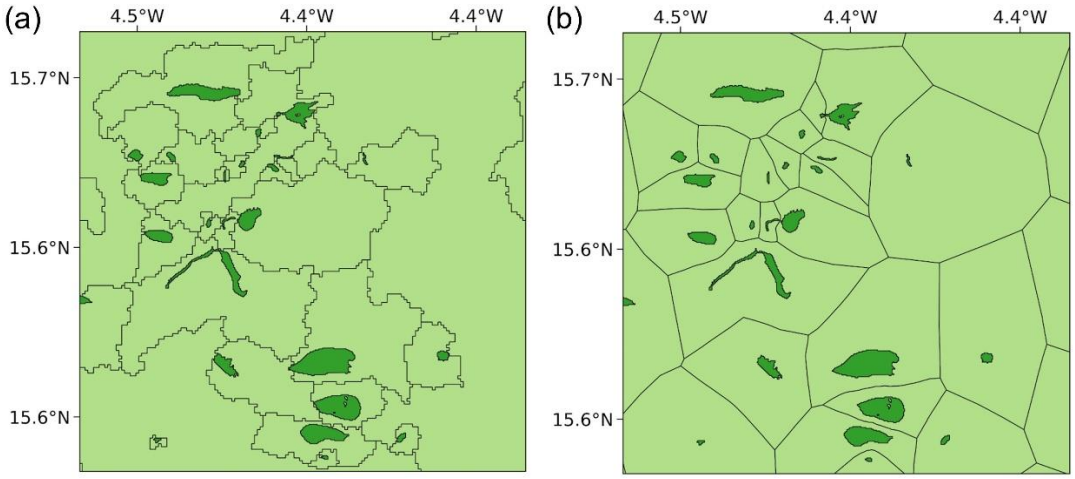


Figure 5. An example of SWOT prior lakes in part of western Africa (deep green) and their associated assignment domains (light green). (a) Lake hydrological catchments as in the "lake_catchment" table. (b) Lake influence features as in the "lake_influence" table.

3.5 "Basin" table

The "basin" table contains the geometries of Pfafstetter level-3 basins corresponding to a level-2 basin granule (see section 3.1 for PLD organization). The basin boundaries were retrieved from the HydroBASINS dataset, with a total number of 291 level-3 basins on the global continents except Antarctica. Each basin feature in this table is provided with five attributes as listed in Table 2. The *basin_id* attribute is the basin identifier, containing the level-3 Pfafstetter code from HydroBASINS. The value of this attribute is identical to the first three digits in *lake_id* (i.e., CBB) of the "lake" table, which links each prior lake to its associated basin. The basin geometries and *basin_id* values are used to separate the water features observed by SWOT, including those not intersected by any prior lakes, to different continents and basins, which is needed for populating the vector lake products at different granule scales.

4 Results and discussion

4.1 Prior lake abundance and distribution

As the primary component of the SWOT PLD, the prior lake mask contains 5,898,331 polygons larger than 1 ha (Fig. 6), mostly representing the intermediate water extents of global lakes during their stable seasons. These prior lakes have a total area of 2,597,464.5 km^2, covering about 2% of the global land surface excluding Antarctica. The Caspian Sea, including the Garabogazköl lagoon, is excluded from the PLD due to its large size and dual characteristics of both lake and ocean (Zimnitskaya & Geldern, 2011).

Table 3 summarizes the lake abundance in each of the nine Pfafstetter-1 continental divisions. The lake count ranges from less than 80,000 per division in Africa (AF), Australia and Oceania (AU), and Greenland (GR) to more than 1 million in Siberia (SI), North America and Caribbean (NA), and North American Arctic (AR). In general, the divisions with larger lake counts also tend to exhibit a greater total lake area and lake density. Despite a global average of 1.9%, lake density varies substantially from only 0.4–0.8% in GR, AU, and AF, to 5.2% in NA and as high as 7.6% in AR. On a continental scale, lake abundance appears to be negatively

correlated to aridity and positively correlated to the degree of glaciation or periglacial processes (except GR). The divisions with less lake abundance, however, tend to have a greater mean lake size (e.g., 127.9 ha in AU and 330.4 ha in AF), implying fewer but larger lakes are more likely to develop in arid regions. On the other hand, the lake-dense circum-Arctic regions (AR and SI) are dominated by smaller lakes with an average size of 25.1 ha, substantially below the global average 44.0 ha. In comparison, the median lake sizes are more consistent among the continents and range subtly between 2.5 ha to 4.0 ha.

With a minimal lake size of 1 ha, the prior mask reveals an unprecedented detail of global lake distribution. About 65% of the total lake count or 40% of the total lake area is clustered in the sparsely populated high-latitude regions above 55°N (Fig. 6b), where glacial activities prevailed in the last ice age. Lakes are particularly ubiquitous across the Canadian Shield and Scandinavia as a result of glacial erosions during the Pleistocene (Shilts et al., 1987) and the boreal permafrost lowlands (e.g., in Siberia and Alaska) associated with thermokarst (Kokelj & Jorgenson, 2013; Manasypov et al., 2014; Smith et al., 2005; Wik et al., 2016). While lake count gradually declines southward, lake area continues to plateau till 40°N, owing to the presence of some of the most gargantuan lakes in the world such as the Laurentian Great Lakes, Lake Balkhash, and Lake Baikal. As a result, more than 70% of the global lake area is concentrated above 40°N, a latitudinal belt accounting for only one-third of the global landmass (excluding Antarctica). In comparison, the temperate and tropical zones between 40°N and 40°S are home to about 85% of the global population (estimated based on the Gridded Population of the World (GPW v4) (CIESIN, 2018)) but only 16% of the global lake count or a quarter of the lake area, highlighting the unequal spatial distribution of lake water resources. Longitudinally, 64% of the global lakes (or 59% by area) are distributed in the land-lacking western hemisphere (Fig. 6c) due to disproportionate lake densities in Alaska, the Canadian Shield, the Amazon floodplain, and alpine Patagonia. A spike of lake area is also seen around 30°E, which is associated with Lake Victoria and a few elongated large lakes in the East African Rift System such as Lakes Tanganyika and Malawi. Another cluster of lake abundance occurs in the longitudinal belt of 60°E to 90°E, which is contributed by thousands of thermokarst lakes across the North Siberian Lowlands and the alpine and glacial lakes on the Tibetan Plateau.

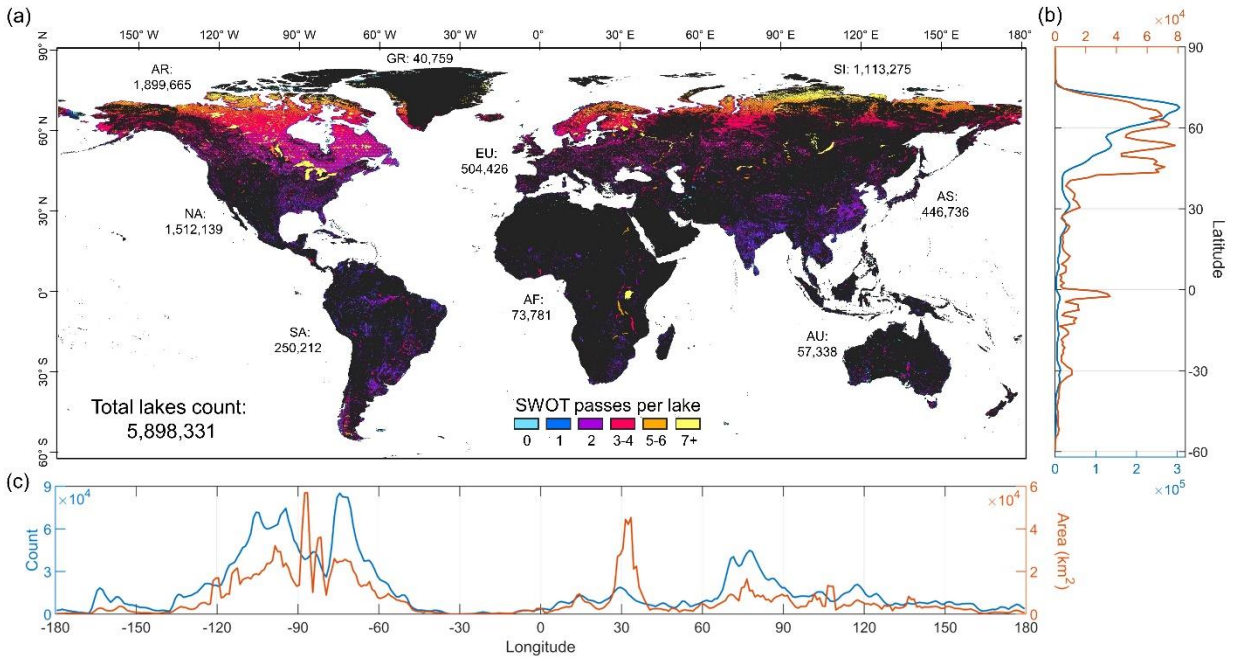


Figure 6. Global map and distribution of the SWOT prior lakes. (a) Global map of prior lakes, with numbers labeling the count of lake polygons per Pfafstetter level-1 division and colors displaying the number of SWOT overpasses per lake during each 21-day orbit cycle. (b) Count and total area of the PLD lakes per latitudinal degree. (c) Count and total area of the PLD lakes per longitudinal degree. The location of a lake polygon was determined by the latitude and longitude coordinates of the centroid of the lake polygon. Values in both latitudinal and longitudinal histograms (b and c) were smoothed by a 3-degree average window to enhance aesthetic appearance and take into account that lakes can span multiple 1-degree intervals.

4.2 Comparison with other global lake masks

We compare the PLD prior lake mask with HydroLAKES, GLAKES, and the entirety of the Circa-2015 lake dataset, to further understand the capability of the PLD in helping SWOT achieve its science objectives for global lake monitoring. The comparison emphasizes the characteristics of lake size distribution, shoreline fractality, and lake mask accuracy across different landscapes, in addition to summary statistics on global lake abundance. While the prior lake mask is, to a large extent, a subset of the Circa-2015 lake dataset (section 2.1), we include the latter for comparison in order to understand the abundance of small lakes that are inventoried but beyond SWOT's science goal (<1 ha).

As shown in Fig. 7a, all datasets concur that the distribution of the Earth's lake area is asymmetric and lake abundance increases as lake size decreases. When lakes are larger than a scale of ~100 ha (1 km^2), the size-abundance relationship conforms to a power-law or Pareto distribution, where the cumulative lake count increases linearly with the decrease of lake size in logarithmic space. Lakes smaller than this scale, however, gradually deviate from a power-law distribution. Since 100 ha well exceeds the minimum lake size in any of the four datasets, the power-law deviation is not attributable to incomplete mapping of small lakes, but instead suggests that lakes behave as self-similar fractals until a lower size limit is reached (Mandelbrot, 1982). Cael and Seekell (Cael & Seekell, 2016) explained that such a lower size limit exists

because topographic characteristics at sub-kilometer scales are less self-similar, and that the development of small lakes is more subject to external dynamics that are scale dependent. Pi et al. (Pi et al., 2022) also noted that lakes <100 ha, despite accounting for only ~15% of the global lake area, dominated the lake area variability over the past four decades, further highlighting the unique roles of small lakes in representing regional geomorphic processes and regulating surface water dynamics.

For the above reasons, the capability of characterizing small lake abundance is critical to the SWOT PLD. Through experimentation based on the PLD, we suggest the lower size limit of the power-law distribution to be 30 (±5) ha, where the fitting slope as a function of lake size threshold reaches the inflection point and remains stable. Using the subset of lakes ≥ 30 ha, we fitted a power-law function for each of the lake datasets (fitting for the PLD shown in Fig. 7a), which rendered a similar tail exponent of ~1.00, close to 1.05 predicted by percolation theory (Cael & Seekell, 2016). While this consistency suggests that the four datasets are comparable in representing the abundance of large lakes, the major difference is their capabilities of characterizing smaller lakes that deviate from a power-law distribution. As shown in Fig. 7a, the pattern of how this deviation develops becomes increasingly clear as the minimum lake size decreases from 10 ha in HydroLAKES, 3 ha in GLAKES, 1 ha in PLD, to 0.4 ha in the Circa-2015 lake dataset. Put in the context of SWOT, the deviation is to the extent that there are 55% fewer lakes meeting SWOT's science requirement (≥ 6.25 ha) than would be expected if the lakes conformed to power law across the entire size range, and the deviation was amplified to a factor of two (221% fewer lakes) for the lakes meeting SWOT's science goal (≥ 1 ha).

Besides size distribution, the perimeter-area scaling relations are plotted in Fig. 7b to compare lake shoreline convolutedness (complexity) among the datasets. As fractals are self-similar and scale-invariant, their perimeters and areas are related to each other by power law (Cheng, 1995). The exponent, equivalent to the slope of perimeter-area scaling in logarithmic space, defines the fractal dimension (d) measuring how irregular the fractal boundaries are relative to perfect circles ($d = 1$). As expected, the perimeters and areas of lakes ≥ 30 ha in all datasets conform to power-law relationships. The fitted d ranges from 1.25 for HydroLAKES, 1.30 for PLD (fitting shown in Fig. 7b) and Circa-2015, to 1.33 for GLAKES, which are overall consistent with 1.33 predicted by percolation theory (Cael & Seekell, 2016). The smaller d for HydroLAKES was likely because the scales of some of the source data (e.g., the MODerate resolution Imaging Spectro-radiometer (MODIS) MOD44W water mask (Carroll et al., 2009)) underrepresented real shoreline complexity, in combination with additional shoreline smoothing during data post-processing (Messager et al., 2016) (Fig. 8). As the area and fractality decrease among lakes < 30 ha, the lake masks with finer resolutions, particularly the PLD and the Circa-2015 dataset, reveal a subtle transition of d towards 1 (Fig. 7b), echoing the finding of (Cael & Seekell, 2016) based on high-resolution Swedish lakes that the shapes of small lakes are less convoluted. This comparison highlights the advantage of PLD in representing reliable shoreline morphology for both sizable and small lakes.

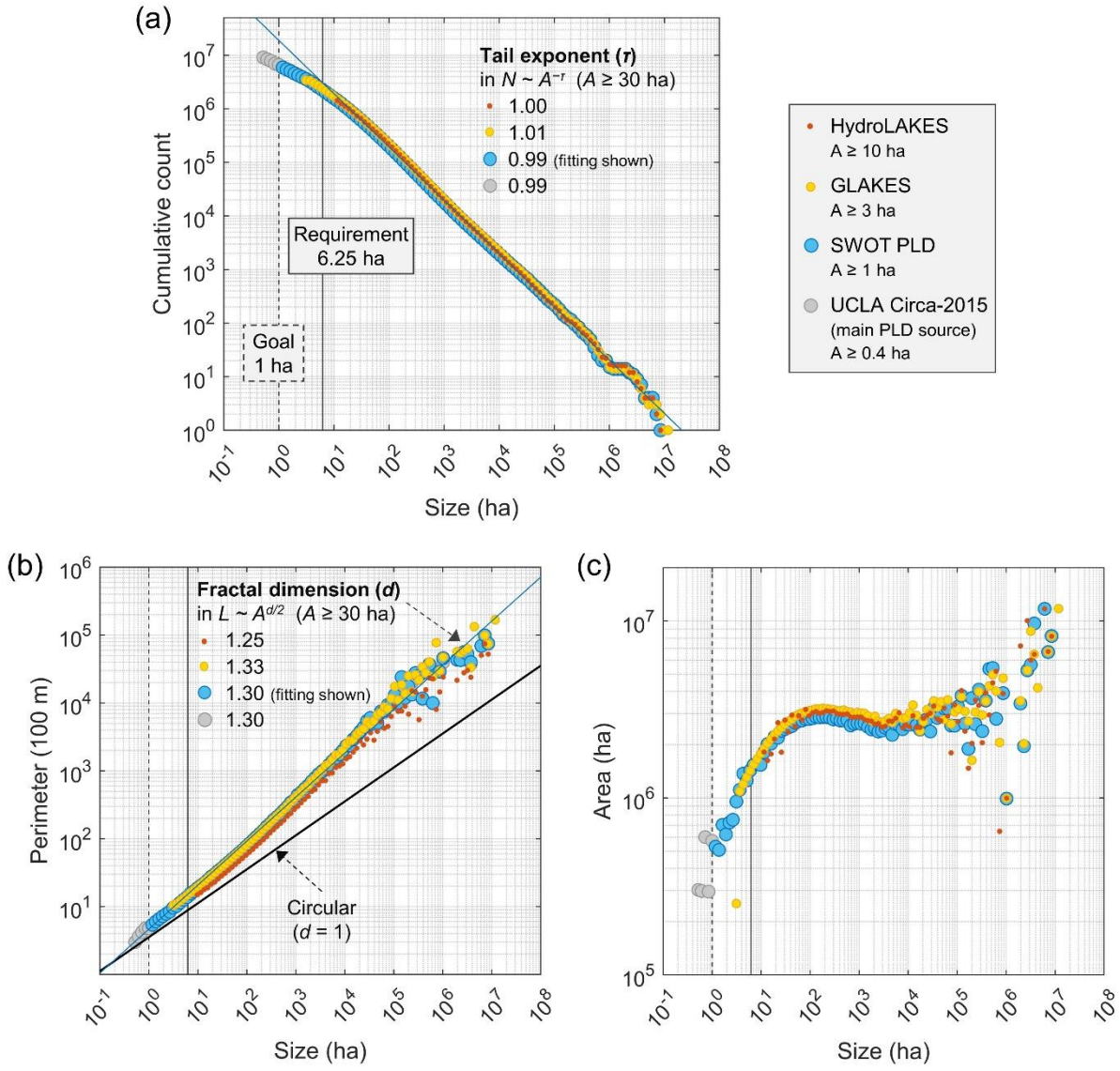


Figure 7. Comparison of lake abundance and distribution among different lake masks (SWOT PLD, HydroLAKES, GLAKES, and Circa-2015). (a) Cumulative abundance or count (N) of lakes as lake size (A) decreases. Lakes ≥ 30 ha are power-law distributed with tail exponents (τ) for all lake masks highly consistent with 1.05 predicted by percolation theory. For clarity, only the fitting line for PLD (blue line) is shown. (b) Lake perimeter (L) in relation to lake size. Perimeters are plotted as the logarithmic medians within the size bins. Similar to the abundance-size distribution, lakes ≥ 30 ha are power-law distributed with fractal dimensions (d) close to the theoretical prediction 1.33. Only the fitting for PLD (blue line) is shown. The solid black curve represents a hypothetical condition where lakes are perfect circles (d = 1). (c) Total lake area per lake size bin. In all plots, size bins represent 105 equal intervals between the minimum and maximum lake areas (~0.4 ha to 11,733,200 ha) in logarithmic space. Solid and dashed vertical lines mark the lake sizes for the SWOT observation requirement (6.25 ha) and goal (1 ha), respectively. The Caspian Sea, including the Garabogazköl lagoon, was excluded from statistics.

We further compare the lake masks using their summary statistics (Table 5) and discuss the implications of discrepancies among them. The total lake count in the PLD (~6.0 million ≥ 1 ha) is nearly double that in GLAKES (3.4 million ≥ 3 ha) and more than quadruple that in HydroLAKES (1.4 million ≥ 10 ha). These multi-fold differences reflect an unparalleled ability of the PLD to characterize the sheer number of small but SWOT-visible lakes. This improvement is exemplified by two high-latitude lake-rich regions: one in the Kanin Peninsula of Russia dotted with circularly shaped thermokarst lakes and bogs (Fig. 8a), and the other in the interior of the Canadian Shield, which is dominated by more convoluted, elongated lakes largely controlled by structural geology (Fig. 8b). In both examples, the PLD shows superiority in representing local lake density, geolocations, and shoreline morphology. The Circa-2015 lake dataset includes another 3.1 million lakes in the world beyond SWOT's observation goal (< 1 ha), although these tiny lakes and ponds add only 3% to the total lake area. Despite a significantly greater lake population in the PLD, its total lake area (2,597,464.5 ha) is 7% lower than that of GLAKES (2,787,115.5 ha) and exceeds that of HydroLAKES (2,537,863.1 ha) by only 2%.

For more detailed comparisons, we broke down the statistics into size classes determined by the minimum lake areas of each of the datasets as well as SWOT's observation requirement (Table 5). For lakes smaller than 6.25 ha and larger than 3 ha (the minimum size in GLAKES), the abundance in the PLD exceeds that in GLAKES by ~2% for both count and area. Based on our visual comparisons, we attribute this difference to an overall greater omission error for small lakes in GLAKES (e.g., Fig. 8a and Fig. 8b), probably related to a conservative nature of its non-parametric "expert system" for water detection (Pekel et al., 2016). For lakes ≥ 6.25 ha, however, the total abundance in the PLD becomes 6–9% less than that in GLAKES and 1–4% less than that in HydroLAKES. To investigate if such lower abundance is skewed to any individual lakes or size groups, we calculated how the total lake area and population (count) are distributed across detailed size bins (Fig. 7c). The patterns are highly consistent among the datasets: as lakes grow in size, their population decreases monotonically, but the total lake area exhibits a three-phase change. In phase one, the total lake area increases as the size grows towards ~100 ha, suggesting for smaller lakes, the area gain due to size growth outpaces the area loss due to population decline. In phase two, the total lake area decreases subtly and then stabilizes as the size grows to ~10,000 ha, suggesting for medium-sized lakes, the area gain due to size growth generally compensates for the area loss due to population decline. In phase three, the total area rapidly increases again when lake size exceeds 10,000 ha, indicating a dominant impact of large lakes on area statistics albeit a limited population. Regardless of this multi-phase pattern, the relative difference in lake abundance between the PLD and the other datasets remains overall uniform across the size bins: excluding lakes ≤ 6.25 ha, both area and count in the PLD are centered around ~10% less than those in GLAKES and 5% less than those in HydroLAKES.

Table 5. Statistical comparison among SWOT PLD, HydroLAKES, GLAKES, and Circa-2015 lake dataset. The Caspian Sea, including the Garabogazköl lagoon, was excluded from the statistics.

Statistics		**HydroLAKES**	**GLAKES**	**SWOT PLD**	**Circa-2015**
Minimum lake size (ha)		10	3	1	0.4
Count	All	1,427,686	3,426,387	5,898,331	9,092,158
	1–3 ha	---	---	2,590,538	2,659,773

	3–6.25 ha	---	1,302,135	1,327,205	1,345,113
	6.25–10 ha	---	616,915	574,923	573,042
	≥10 ha	1,427,686	1,507,337	1,405,665	1,414,396
Area (km^2)	All	2,537,863.1	2,787,115.5	2,597,464.5	2,596,665.3
	1–3 ha	---	---	46,076.6	47,276.0
	3–6.25 ha	---	56,509.8	57,583.0	58,326.4
	6.25–10 ha	---	48,665.5	45,276.0	45,020.6
	≥10 ha	2,537,863.1	2,681,940.1	2,448,529.0	2,425,359.2

The discrepancy in the lake abundance for lakes ≥ 6.25 ha reflects the differences in mapping standard, quality, timespan, and reference sources among the datasets. A higher lake abundance in GLAKES is expected because its polygons represent all-time water area maximum during 1984 to 2019 whereas most lakes in the PLD depict intermediate water extents during circa 2015. Although both datasets were derived from Landsat imagery, the differences in mapping period and standard, in theory, led to not only a larger lake area in GLAKES, but also a greater lake quantity given that not all intermittent lakes were inundated during circa 2015. While the Circa-2015 lake dataset was supplemented by recently constructed reservoirs (sections 2.2 and 3.2), natural lakes that disappeared before or emerged after circa 2015 are not included in the PLD. On the other hand, HydroLAKES was a compilation of eight independent lake sources with publication dates spanning a decade (Messager et al., 2016). Variation among these data sources may contribute intricately to a higher abundance (for lakes ≥ 10 ha) in HydroLAKES than the PLD.

For instance, the acquisition time of SWBD (February 2000), a major source of HydroLAKES for 56°S to 60°N, may explain the smaller areas in the reservoirs of northwestern India (Fig. 8g), where water levels were low during the drier monsoon season. In another relevant case, a number of important reservoirs in western Africa were built after February 2000. These include the Ziga Reservoir (completed in July 2000) in Burkina Faso (12.5°N, 1.1° W) that is absent from HydroLAKES 1.0. On the other hand, this acquisition time of SWBD coincided with the warmer season in the southern hemisphere. Meanwhile, SWBD as a radar-derived product (Slater et al., 2006) is less sensitive to surface spectral disturbance such as remnant lake ice and snow. Both factors might lead to a more complete inventory of glacier lakes in HydroLAKES across the southern Andes (Fig. 8h).

Another example in Fig. 8e highlights a portion of the Yukon River Valley in Alaska, where thermokarst lakes and their drained lake basins develop dynamically atop the permafrost (Grosse et al., 2013). While the PLD polygons appear highly consistent with the recent thermokarst lake extents, HydroLAKES depicts the much larger drained thaw lake basins. These outdated lake boundaries are sourced from the US National Hydrography Dataset (U.S.-Geological-Survey, 2013) and contribute partially to an overestimated area abundance in HydroLAKES.

In addition, part of the higher abundance in HydroLAKES and GLAKES may be ascribed to commission errors such as mountain shadows and forest patches, as shown in the examples of Fig. 8c and Fig. 8d. Such commission errors were largely eliminated from the PLD owing to a rigorous QA/QC procedure (Sheng et al., 2016) (section 2.1). Other factors such as lake definition and mapping objective could also lead to discrepancies in lake abundance. In Fig. 8f,

GLAKES and HydroLAKES include a large quantity of aquaculture ponds in coastal China, which were not considered as lakes in the PLD.

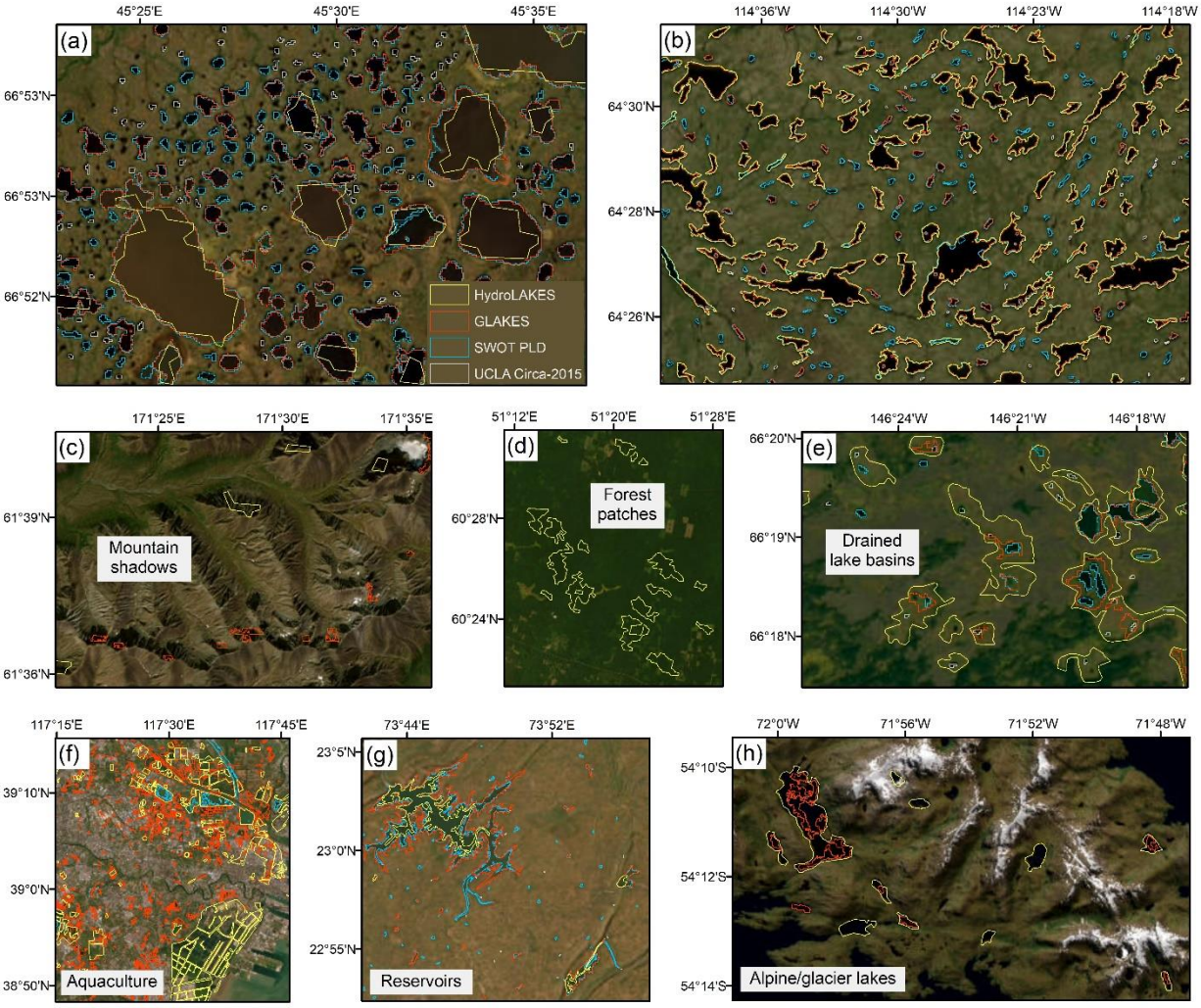


Figure 8. Regional comparisons among the PLD, HydroLAKES, GLAKES, and the Circa-2015 lake dataset. (a) Thermokarst lakes in the southern Kanin Peninsula, the Nenets Autonomous Okrug, Russia. (b) Structurally controlled lakes in the Canadian Shield, Northwest Territories, Canada. (c) Commission errors (mountain shadows misclassified as lakes in HydroLAKES and GLAKES) in Kamchatka Krai, Siberia, Russia. (d) Commission errors (forest patches misclassified as lakes in HydroLAKES) in southern Komi Republic, Russia. (e) Thermokarst lakes and drained thaw lake basins in the Yukon River Valley, eastern Alaska. (f) Aquaculture ponds near the Bohai coastline, Tianjin, China. (g) Reservoirs in eastern Gujarat, India. (h) Alpine and glacier lakes in the southern Andes.

4.3 Lake spatiotemporal coverage

SWOT coverage for the land surface is determined jointly by orbit characteristics, the KaRIn swath width (2 × 50 km), the nadir gap width (20 km) between the two swaths, and the orbit crossover density which is a function of latitude (Biancamaria et al., 2016). In addition, the spatiotemporal coverage for lakes also depends on the size and shape of each lake. With all these

factors considered, Fig. 6a shows the frequency of SWOT observations over each prior lake during every 21-day orbit cycle, which was calculated by summing the counts of unique overpasses in both *pass_full_nom* and *pass_part_nom* attributes (section 3.3.4). As summarized in Fig. 9, 96.5% of the global lakes, covering 98.2% of the total lake area, are observed by SWOT at least once per orbit cycle. More than 65% of the global lakes, covering nearly 80% of the lake area, are observed at least weekly on average (i.e., three times per cycle). About 3.5% of the global lakes, or 1.8% of the lake area, may never be observed due to a combination of nadir gaps and orbit intertrack gaps. This lake coverage complies with the SWOT science requirements, which states that "SWOT shall collect data over a minimum of 90% of all ocean and land area covered by the orbit inclination for 90% of the operation time" (JPL internal document, 2018).

Despite complexity in the global pattern (Fig. 6a), lake observation frequency tends to increase with higher latitudes and larger lake sizes. As latitude increases, the orbit crossover densifies and the overlap among adjacent swaths increases. This gradually leads to the closure of orbit intertrack gaps at 25°S/N and then the closure of nadir gaps at about 60.5°S/N. As lake size increases, the chance of one lake overlapped by multiple passes also increases. As a result, unobserved lakes between 10°S and 10°N account for about 10% of the local lakes (in terms of both count and area), but the proportion decreases to less than 1% (0.9% in lake count and 0.5% in lake area) over the latitudes above 60°S/N. Since lake abundance is skewed towards higher latitudes in both count and size (section 4.1), these factors also explain why SWOT's coverage gap for global lake area (1.8%; Fig. 9b) is significantly smaller than that for the entire land surface (3.6%) (Biancamaria et al., 2016).

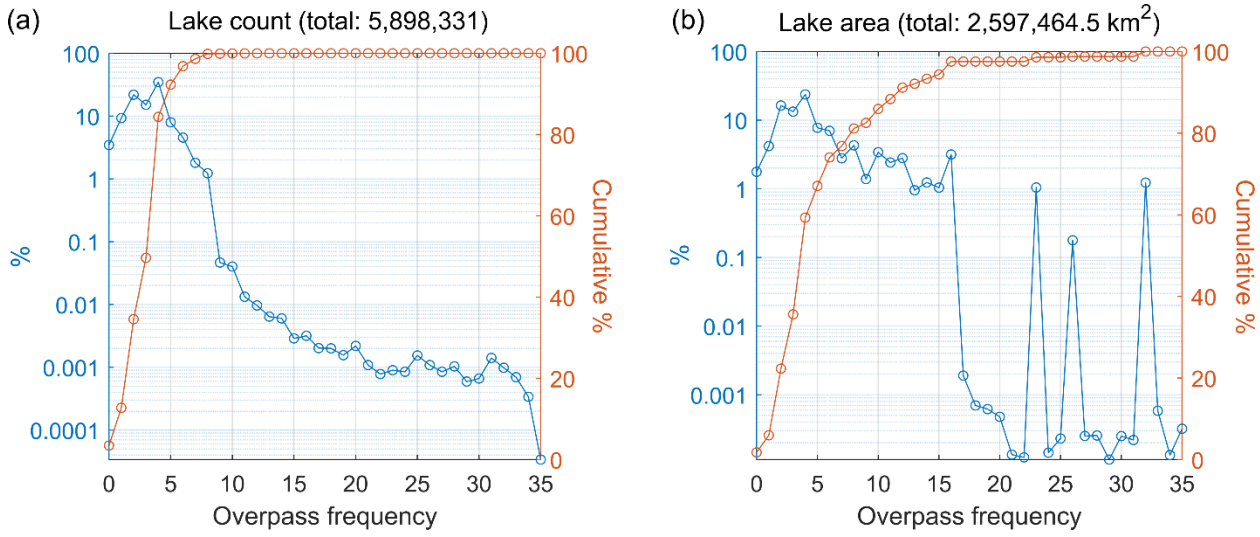


Figure 9. Distributions of lake overpass frequency within a SWOT nominal orbit cycle (21 days). (a) Density (left y-axis, blue) and cumulative distribution (right y-axis, orange) of overpass frequency in terms of lake count. Note the left y-axis is in logarithmic scale while the right y-axis is in linear scale. (b) The same as panel a but in terms of lake area.

Globally speaking, the median observation frequency is maintained at about twice per orbit cycle for lakes between 50°S and 50°N and smaller than 100 km^2. The median frequency increases to three times per cycle for larger lakes over this latitudinal band and for lakes between 50–60°N/S regardless of the lake size. The median frequency increases further to four times per

cycle above 60°N. On the other hand, some of the highest observation frequencies are found in the world's largest lakes regardless of latitudinal distribution. For example, nearly all lakes larger than 10,000 km^2, except Lake Malawi with an elongated shape parallel with SWOT passes (Fig. 6a), are observed from six times per cycle to more than twenty times per cycle (i.e., every day).

A higher overpass frequency does not always warrant a better spatial coverage. Nearly 6% of the prior lakes, constituting 67.3% of the global lake area, fall on the edge of at least one overpass. These lakes will appear incomplete in some of the granules of the single-pass product (L2_HR_LakeSP). However, with a higher overpass frequency, there is an increasing chance that the lake can be observed fully by at least one pass per cycle, or the aggregation of multiple passes can lead to a full extent to represent the average inundation condition during the cycle. The latter reflects the value of the cycle-average product (L2_HR_LakeAvg). To evaluate how lakes are spatially covered per cycle, we calculated the percentage of lakes for each of the *cycle_flag_nom* scenarios (section 3.3.4) and how the percentages vary in lake size. As shown in Fig. 10, smaller lakes, albeit overall less frequently observed, are easier to be seen with a full extent. About 90% of the lakes smaller than 10 km^2 are fully observed at least once per cycle (scenario 3). As lake size increases, the proportion of scenario 3 monotonically declines; meanwhile, lakes that are observed fully only after pass aggregation (scenario 2) and lakes that remain observed partially after pass aggregation (scenario 1) increase at similar paces. The three scenarios cross at ~300 km^2, beyond which lakes of scenario 3 are no longer the majority. The proportion of scenario-2 lakes peaks at nearly 40% around 500 km^2. Lakes larger than this size are gradually dominated by scenario 1 until 10,000 km^2, beyond which lakes can only be observed partially despite very high overpass frequencies. The proportion of lakes that can never be observed (scenario 0) remains less than 5% regardless of size, and more than 97% of them are smaller than 1 km^2.

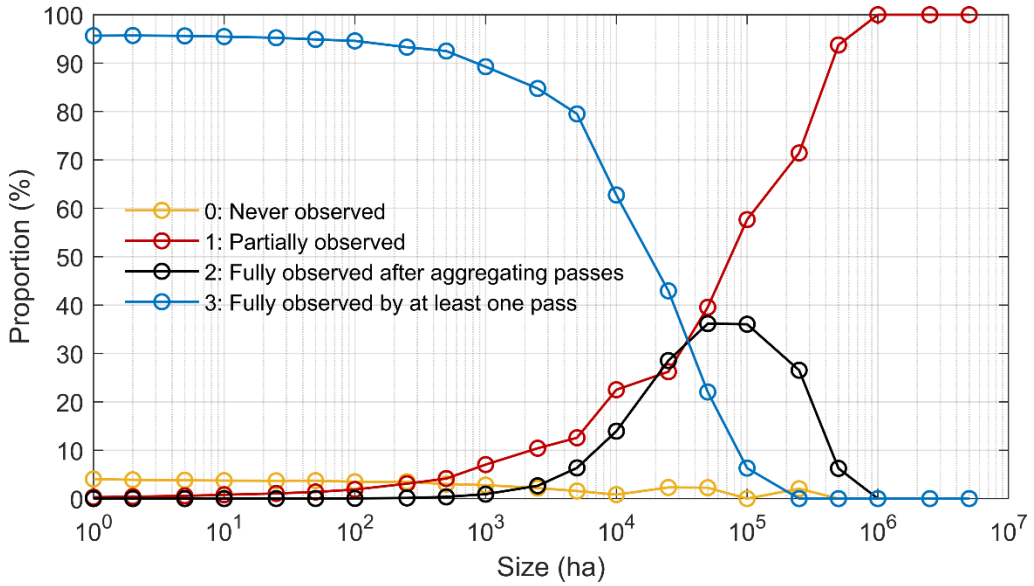


Figure 10. Lake spatial coverage (*cycle_flag_nom*) as a function of lake size during each SWOT nominal orbit cycle.

Synthesizing all lake sizes, Fig. 11 shows that 95.5% of the global prior lakes, constituting 50.4% of the total lake area, are fully observed during a nominal cycle, and 3.9% (or 1.9% by

966 area) are never seen. Less than 1,000 prior lakes, accounting for 8.7% of the global lake area, can
967 be fully covered after aggregating multiple passes per cycle, whereas the remaining 0.6%
968 (34,849) lakes, accounting for 38.9% of the global lake area, cannot be fully covered in a cycle.
969 For these partially observed lakes, complete water areas could be estimated with assistance of
970 other sensors and/or an auxiliary water probability or contour map (such as GSWO (Pekel et al.,
971 2016)). In each of the Pfafstetter-1 (sub)continents, the proportion of lakes that are partially
972 observed is lower than 1% except GR. In SI and AR, more than 96% of the prior lakes are fully
973 covered by a single pass, while in AU, this proportion is only 74.1%, and a quarter of the lakes,
974 most of which are small, are not observed by SWOT at all. It is also worth noting that lakes in AU
975 will be observed by the low-rate (LR) products but not by the HR products. Despite this regional
976 limit, the LR products can still be useful, especially for understanding the dynamics in larger lakes.

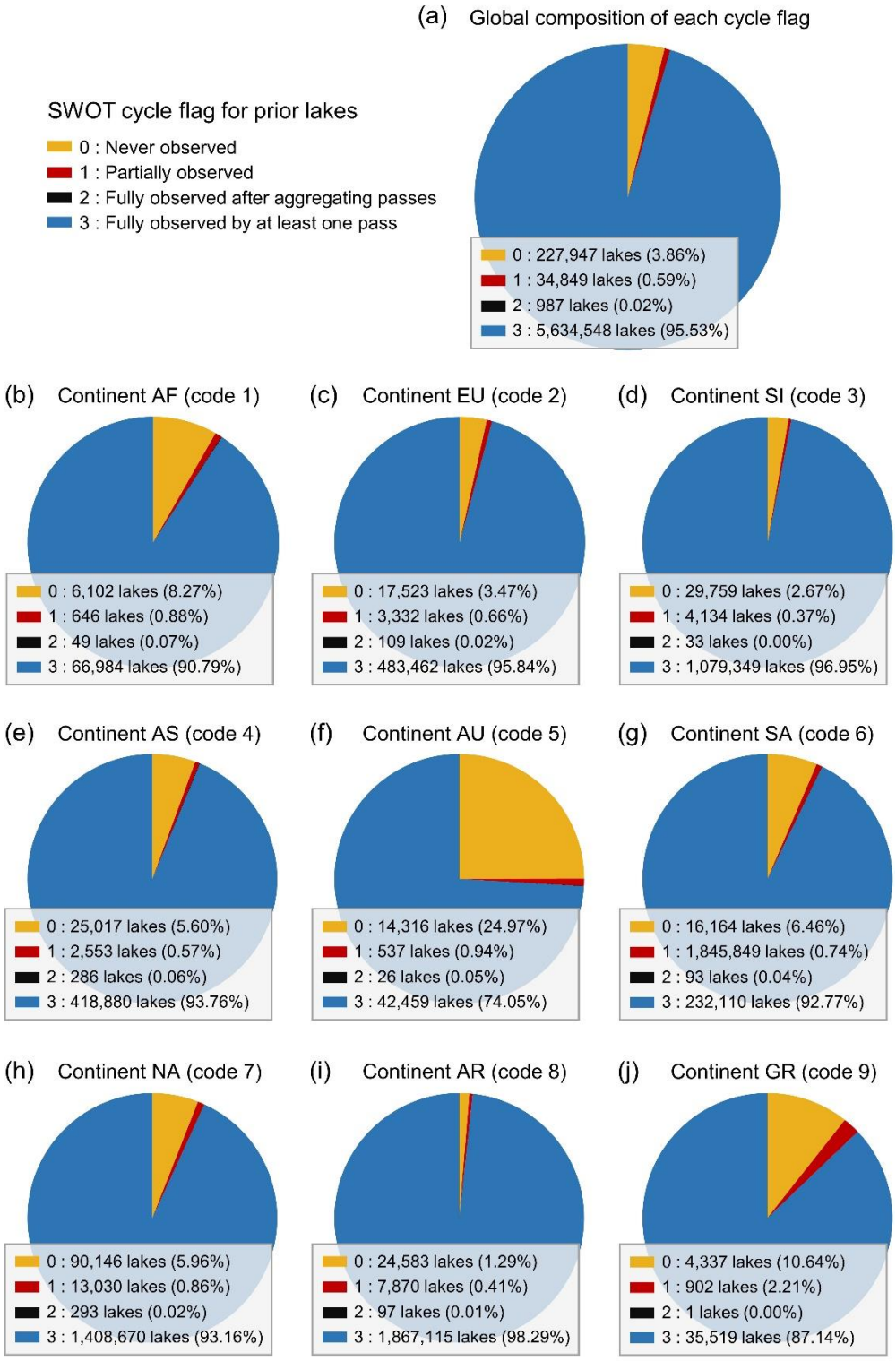


Figure 11. Distributions of lake coverage scenarios during a SWOT nominal orbit cycle for each continent or subcontinent.

4.4 Example of linking SWOT observations

Here we provide a conceptual example to demonstrate how the operational PLD assists the SAS in linking KaRIn observations to the prior lakes and generating the L2_HR_LakeSP vector product. More technical details are given in the Algorithm Theoretical Basis Document (CNES internal document, 2023a). As introduced in section 1, the lake processing pipeline starts from the subset of the pixel cloud (L2_HR_PIXC) after the removal of pixels associated with prior rivers. The remaining non-river pixels are segmented to distinct water regions based on height clusters, and the pixel geolocations are further regularized by the average height per region. The resulting pixels with height-constrained geolocations are used to vectorize water regions, and the attributes such as water area and average WSE are computed for each vectorized water feature. These processes are directly based on SWOT observations and are independent from the PLD.

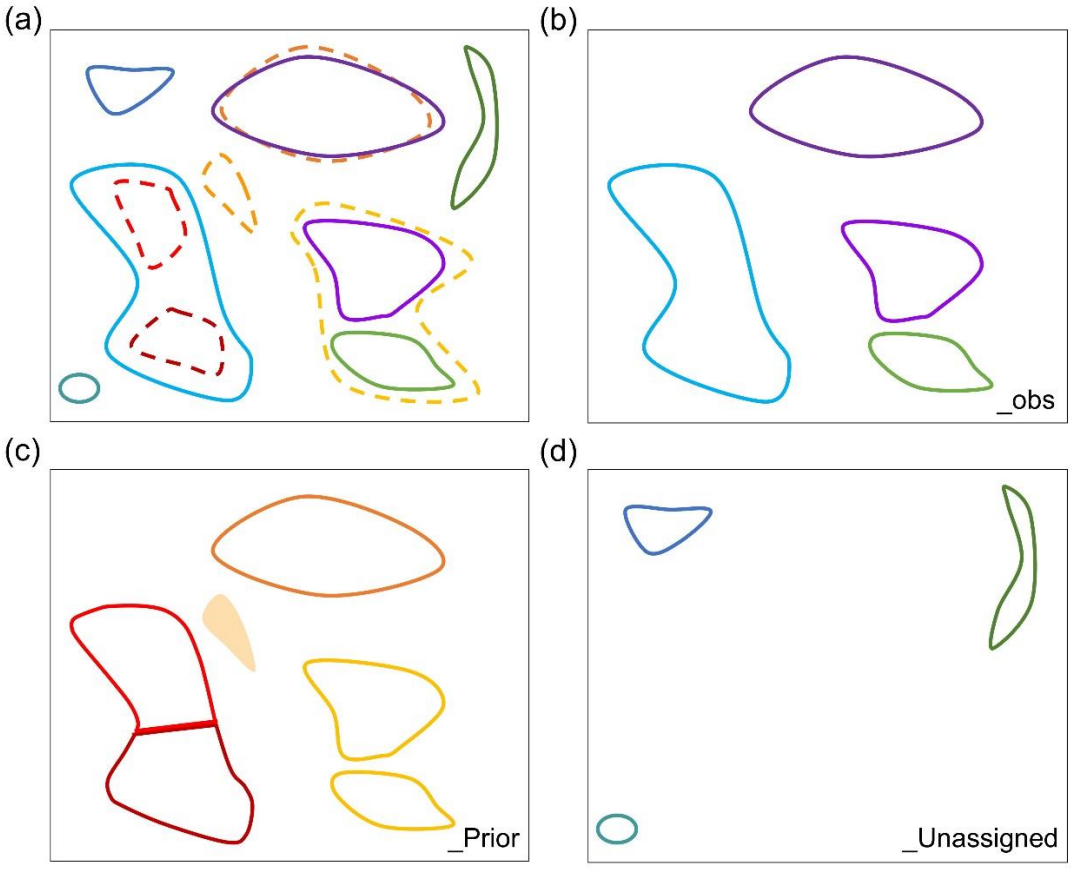


Figure 12. Illustration of how the PLD is used to organize SWOT-observed water features into the three vector files of the L2_HR_LakeSP product. (a) Observed water features (solid) and prior lakes (dashed) in a hypothetical region. Different colors represent different water features or prior lakes. (b) Result of the observation-oriented file (L2_HR_LakeSP_Obs). (c) Result of the PLD-oriented file (L2_HR_LakeSP_Prior). The unobserved prior lake is an empty geometry with only prior attributes, shown as a filled polygon. An observed feature intersecting two prior lakes is partitioned to two feature entities (red and dark red), whereas two observed features intersecting the same prior lake (yellow) are dissolved to a multipart entity. (d) Result of the observation-oriented unassigned file (L2_HR_LakeSP_Unassigned).

The observed water features are next compared with the prior lake polygons to establish spatial linkage between them. Depending on the relationship, the observed water features are organized into three product files (Fig. 12): L2_HR_LakeSP_Obs, L2_HR_LakeSP_Prior, and L2_HR_LakeSP_Unassigned. As illustrated in Fig. 12a, observed water features (solid) and prior lake polygons (dash) do not always exhibit a one-to-one relationship. A linkage is considered valid if an observed feature intersects at least one prior lake with sufficient overlap, typically defined as 2% or larger (CNES internal document, 2023b). In this case, the water feature is considered a lake and stored in L2_HR_LakeSP_Obs (Fig. 12b). Otherwise, the feature is gathered in L2_HR_LakeSP_Unassigned (Fig. 12d). Both product files are observation-oriented, meaning that the water features maintain the geometries as observed by SWOT, and the output attributes, such as area and WSE, are the same as those of the input observed features.

To enable storage change calculation, each observed water feature must be linked to a reference water state. However, reference states are only provided for prior lakes (section 3.3.3), which often exhibit complex topological relations with observed features. Such spatial inconsistency requires water features in L2_HR_LakeSP_Obs to be reorganized (grouped or split) according to the prior lakes, so that the resulting features and the prior lakes have a one-to-one relationship. The resulting features are gathered in L2_HR_LakeSP_Prior (Fig. 12c). This process is straightforward when the original feature intersects only one prior lake. In this case, the geometry of the water feature remains unchanged, and the intersected prior lake with its water reference state is assigned to this water feature. When a prior lake intersects more than one water feature, all intersected features are grouped to a multipart geometry (i.e., an entity composed of several distinct polygons that represent only one set of attributes), and this prior lake is assigned to the multipart feature.

A more complicated case is one observed water feature intersecting multiple prior lakes. When this occurs, the assignment polygons of the intersected prior lakes in either the "lake_catchment" table or the "lake_influence" table can be utilized to split the observed water feature. Figure 13 illustrates an example using the "lake_influence" table. In this example, an observed feature in the northeast overlaps two prior lakes (*lake_id* 232008092 and 232009412). To partition this feature, each of its PIXC pixels is assigned to the prior lake whose influence area contains the pixel (Fig. 13c). Since the influence areas are Thiessen polygons (section 3.4), this assignment essentially groups the water pixels based on the closest prior lake. The pixels are then re-vectorized based on their prior lake assignment to form separate water features, and the corresponding WSE and water areas are recalculated. Eventually, water storage change for each feature is computed using the reference water state of the prior lake assigned to the feature. Any prior lake that is not observed under an overpass, such as an intermittent lake during the dry season, is also added to L2_HR_LakeSP_Prior but as an empty geometry with only prior attributes. Water storage change is not calculated for L2_HR_LakeSP_Unassigned where features are not linked to any prior lake, thus lacking a reference water state to effectively derive storage change.

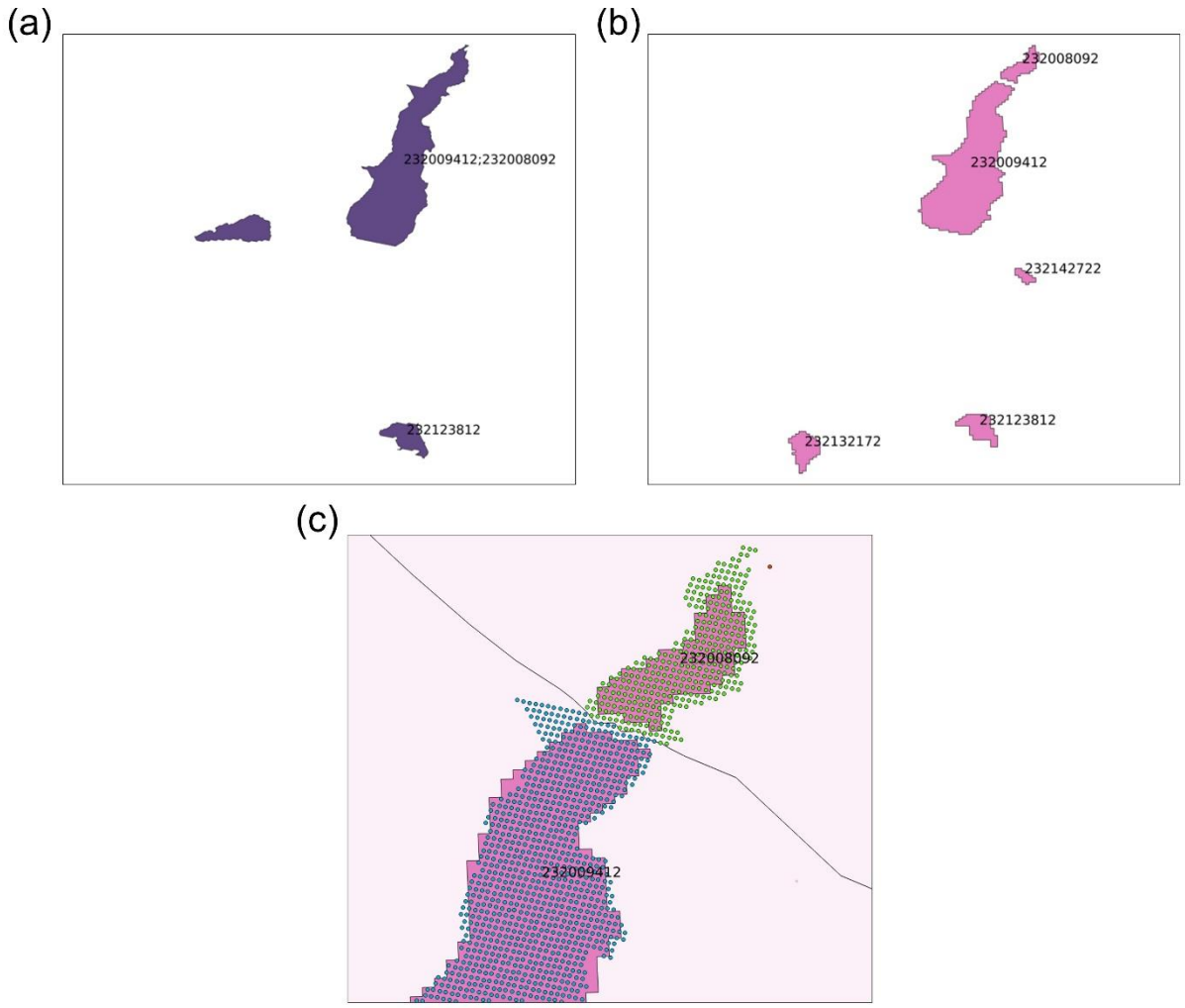


Figure 13. Example of lake assignment using the operational PLD. (a) SWOT-observed water features in a hypothetical region. (b) Associated prior lakes. Prior lakes 232142722 (*lake_id*) and 232132172 are not observed by this overpass and will be gathered by L2_HR_LakeSP_Prior as empty geometries with only prior attributes. The observed water feature in the central east is linked to no prior lake and will be gathered by L2_HR_LakeSP_Unassigned. The observed water feature in the northeast intersects both prior lakes 232008092 and 232009412. It will be a single feature in L2_HR_LakeSP_Obs but will be split into two separated features in L2_HR_LakeSP_Prior. The observed feature associated with prior lake 232123812 will be gathered by both L2_HR_LakeSP_Obs and L2_HR_LakeSP_Prior with identical geometry. (c) Zoom-in of the case where one observed feature intersects two prior lakes and how the pixels of this water feature are reorganized by the assignment polygons in the "lake_influence" table.

5 Versioning plan

The operational PLD introduced in this paper represents the initial version that is used to generate the official SWOT vector lake products. With the accumulation of SWOT observations throughout the mission period, the PLD will be recursively updated to improve the functionality and quality, according to the versioning plan currently configured below.

5.1 Five update levels

We envision five levels (Levels 0 to 4) of PLD update depending on the quality of the prior lake polygons and the attributes computed from the SWOT vector lake products. **Level 0** refers to manual inputs from data users. **Level 1** updates lake storage parameters, i.e., *ref_wse*, *ref_area*, *date_t0*, and *ds_t0*. As described in Section 3.3.3, *ref_area* in the initial PLD version is populated as the area of the prior lake polygon, and *ref_wse* are filled with "no data". With SWOT measurements being available, these two attributes will be updated using the values of *wse* and *area_total* attributes in the LakeSP product (CNES internal document, 2022b) corresponding to the 80th percentile of the time series for each prior lake during the update cycle (see timeline in Section 5.3). Accordingly, the storage change parameters at the reference state (*date_t0* and *ds_t0*) will be computed with reference to the date of the first valid SWOT observation of the prior lake. **Level 2** generates and updates the *hypso_curve* table. The *hypso_curve* table will be generated by fitting the (*wse*, *area*) pairs in the LakeSP product from the first valid observation for the prior lake (section 3.1). Each time the table is updated, the fitting will be redone using all (*wse*, *area*) pairs available from the first valid observation to the end of the update cycle. **Level 3** updates the geometry of each existing prior lake. This will be done by intersecting the polygons associated with the three highest *wse* values, of this prior lake in the LakeSP_Prior product. **Level 4** adds new prior lakes that are absent from the previous PLD version. New prior lakes will be obtained from the water features that are observed to be persistent in the LakeSP_Unassigned product.

5.2 Three priority categories

Along with the five update levels, we will classify the prior lakes into three categories (P1 to P3) based on how easy or complex the update can be, with consideration of the prior lake geometry, SWOT coverage, and relationship with SWOT-observed water features. These classes will be used to guide the update priority. **Class P1** represents the "easiest" prior lakes and is defined as any lake that satisfies the following criteria: (1) having a size compliant with the SWOT observation requirement (*area_total* in the vector lake product > 6.25ha); (2) being fully observed by SWOT at least once per cycle (*cycle_flag* = 3); (3) being fairly isolated from other lakes (*min_dist* > 300 m); and (4) exhibiting low complexity in relation to SWOT observation, i.e., one prior lake generally corresponds to one SWOT-observed lake feature. **Class P2** contains the prior lakes that have the same first three criteria as Class P1 but exhibit higher complexity in relation to SWOT observation, where one prior lake corresponds to many SWOT-observed lake features. **Class P3** refer to all the other prior lakes.

5.3 General timeline

The first major update of the PLD will be applied only on the prior lakes overflown during the Cal/Val phase (under the 1-day fast-track orbit) using the initial validated product. This update is expected to occur around 16 months after the launch of SWOT (i.e., April 2024, about one year into the mission after the Cal/Val phase), when the initial validated product is released. Prior lakes of Class P3 will go through a Level-0 update (manual inputs); lakes of Class P2 will experience a Level-1 update (populating *ref_wse*, *ref_area*, *date_t0*, and *ds_t0* attributes); and lakes of Class P1 will have a Level-2 update (generating the *hypso_curve* table). The second major update of the PLD will occur approximately one year after the first update. This update will involve prior lakes that are covered by the nominal orbit, using the same

methodology described for the first major update. Level-3 (geometry) and Level-4 (new lakes) updates will not be considered before the third major update of the PLD, which may occur approximately 3 years after the launch of SWOT (i.e., December 2025 or later). These expected PLD updates will reflect an improved understanding of global lake distribution and dynamics as SWOT observations accumulate, and in return, the updated PLD will also improve the processing of SWOT vector lake products. In addition to facilitating SWOT data production, the PLD, with its high-resolution lake mask and multiple operational tables, can be applied to benefiting a wide range of disciplines such as limnology, hydrological modeling, ecology, and climate science.

Acknowledgements

This work is in part funded by the NASA Surface Water and Ocean Topography (SWOT) Science Team grant (#80NSSC20K1143).

Data Availability Statement

The operational SWOT PLD is freely accessible for download from the Hydroweb-next website (https://hydroweb.next.theia-land.fr) under Etalab 2 license (https://www.etalab.gouv.fr/wp-content/uploads/2018/11/open-licence.pdf).

Open Research

All underlying data required to produce the operational PLD are openly accessible from the respective download sources specified in Table 1 and Section 2 of this paper. Codes and algorithms used in PLD production and analysis are available upon request.

References

Abbott, B. W., Bishop, K., Zarnetske, J. P., Minaudo, C., Chapin, F. S., Krause, S., . . . Pinay, G. (2019). Human domination of the global water cycle absent from depictions and perceptions. *Nature Geoscience*, *12*(7), 533-540. https://doi.org/10.1038/s41561-019-0374-y

Adrian, R., O'Reilly, C. M., Zagarese, H., Baines, S. B., Hessen, D. O., Keller, W., . . . Winder, M. (2009). Lakes as sentinels of climate change. *Limnology and Oceanography*, *54*(6), 2283-2297. https://doi.org/10.4319/lo.2009.54.6_part_2.2283

Allen, G. H., & Pavelsky, T. M. (2018). Global extent of rivers and streams. *Science*, *361*(6402), 585-587. https://doi.org/10.1126/science.aat0636

Altenau, E. H., Pavelsky, T. M., Durand, M. T., Yang, X., Frasson, R. P. D., & Bendezu, L. (2021). The Surface Water and Ocean Topography (SWOT) Mission River Database (SWORD): A global river network for satellite data products. *Water Resources Research*, *57*(7), e2021WR030054. https://doi.org/10.1029/2021WR030054

Aurenhammer, F. (1991). Voronoi diagrams - a survey of a fundamental geometric data structure. *Computing Surveys*, *23*(3), 345-405. https://doi.org/10.1145/116873.116880

Biancamaria, S., Lettenmaier, D. P., & Pavelsky, T. M. (2016). The SWOT Mission and Its Capabilities for Land Hydrology. *Surveys in Geophysics*, *37*(2), 307-337. https://doi.org/10.1007/s10712-015-9346-y

Boy, F., Cretaux, J. F., Boussaroque, M., & Tison, C. (2022). Improving Sentinel-3 SAR mode processing over lake using numerical simulations. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 5220518. https://doi.org/10.1109/Tgrs.2021.3137034

Busker, T., de Roo, A., Gelati, E., Schwatke, C., Adamovic, M., Bisselink, B., . . . Cottam, A. (2019). A global lake and reservoir volume analysis using a surface water dataset and satellite altimetry. *Hydrology and Earth System Sciences*, *23*(2), 669-690. https://doi.org/10.5194/hess-23-669-2019

Cael, B. B., & Seekell, D. A. (2016). The size-distribution of Earth's lakes. *Scientific Reports*, *6*, 29633. https://doi.org/10.1038/srep29633

Carroll, M. L., Townshend, J. R., DiMiceli, C. M., Noojipady, P., & Sohlberg, R. A. (2009). A new global raster water mask at 250 m resolution. *International Journal of Digital Earth*, *2*(4), 291-308. https://doi.org/10.1080/17538940902951401

Cheng, Q. M. (1995). The Perimeter-Area Fractal Model and Its Application to Geology. *Mathematical Geology*, *27*(1), 69-82. https://doi.org/Doi 10.1007/Bf02083568

CIESIN. (2018). *Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11* Version Revision 11) NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4JW8BX5

CNES internal document (2022a). Level 2 KaRIn high rate lake average vector product (short name: L2_HR_LakeAvg), Revision A. Surface Water and Ocean Topography (SWOT) Project Product Description Document. SWOT-TN-CDM-0676-CNES, September 30, 2022. Accessed from https://archive.podaac.earthdata.nasa.gov/podaac-ops-cumulus-docs/web-misc/swot_mission_docs/pdd/SWOT-TN-CDM-0676-CNES_Product_Description_L2_HR_LakeAvg_20220930_RevA.pdf on October 16, 2023.

CNES internal document (2022b). Level 2 KaRIn high rate lake single pass vector product (short name: L2_HR_LakeSP), Revision A. Surface Water and Ocean Topography (SWOT) Project Product Description Document. SWOT-TN-CDM-0673-CNES, September 30, 2022. Accessed from https://archive.podaac.earthdata.nasa.gov/podaac-ops-cumulus-docs/web-misc/swot_mission_docs/pdd/SWOT-TN-CDM-0673-CNES_Product_Description_L2_HR_LakeSP_20220930_RevA.pdf on October 16, 2023.

CNES internal document (2022c). Level 2 KaRIn high rate pixel cloud vector attribute product (short name: L2_HR_PIXCVec), Revision A. Surface Water and Ocean Topography (SWOT) Project Product Description Document. SWOT-TN-CDM-0677-CNES, September 30, 2022. Accessed from https://archive.podaac.earthdata.nasa.gov/podaac-ops-cumulus-docs/web-misc/swot_mission_docs/pdd/SWOT-TN-CDM-0677-CNES_Product_Description_L2_HR_PIXCVec_20220930_RevA.pdf on October 16, 2023.

CNES internal document (2023a). Level 2 KaRIn high rate lake single pass science algorithm software: Level 2 processing (short name: SAS_L2_HR_LakeSP: Level 2 processing), Initial Release. Surface Water and Ocean Topography (SWOT) Project Algorithm Theoretical Basis Document. SWOT-NT-CDM-1753-CNES, July 26, 2023. Accessed from https://archive.podaac.earthdata.nasa.gov/podaac-ops-cumulus-docs/web-misc/swot_mission_docs/atbd/SWOT-NT-CDM-1753-CNES_ATBD_LakeSP_20230726_Initial_w-sigs.pdf on October 16, 2023.

CNES internal document (2023b). Level 2 KaRIn high rate lake tile auxiliary parameter file (short name: Param_L2_HR_LakeTile). Surface Water and Ocean Topography (SWOT) Project Auxiliary Data Description Document. 2023, to be released.

Cooley, S. W., Ryan, J. C., & Smith, L. C. (2021). Human alteration of global surface water storage variability. *Nature*, *591*(7848), 78-81. https://doi.org/10.1038/s41586-021-03262-3.

Copernicus Climate Change Service (C3S) (2017). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate.

Cretaux, J. F., Abarca-del-Rio, R., Berge-Nguyen, M., Arsen, A., Drolon, V., Clos, G., & Maisongrande, P. (2016). Lake Volume Monitoring from Space. *Surveys in Geophysics*, *37*(2), 269-305. https://doi.org/10.1007/s10712-016-9362-6

Cretaux, J. F., Jelinski, W., Calmant, S., Kouraev, A., Vuglinski, V., Berge-Nguyen, M., . . . Maisongrande, P. (2011). SOLS: A lake database to monitor in the Near Real Time water level and storage variations from remote sensing data. *Advances in Space Research*, *47*(9), 1497-1507. https://doi.org/10.1016/j.asr.2011.01.004

de Fleury, M., Kergoat, L., & Grippa, M. (2023). Hydrological regime of Sahelian small waterbodies from combined Sentinel-2 MSI and Sentinel-3 Synthetic Aperture Radar Altimeter data. *Hydrology and Earth System Sciences*, *27*(11), 2189-2204. https://doi.org/10.5194/hess-27-2189-2023

Desroches, D., Fjørtoft, R., Gaudin, J.-M., Ruiz, C., & Blumstein, D. (2016, Jun. 2016). Precise geolocation of water bodies in SWOT HR InSar data. Proceedings of EUSAR 2016: 11th European Conference on Synthetic Aperture Radar, Hamburg, Germany.

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

Evans, D. G., & Jones, S. M. (1987). Detecting Voronoi (area-of-influence) polygons. *Mathematical Geology*, *19*(6), 523-537. https://doi.org/10.1007/Bf00896918

Fan, C., Song, C., & TBD. (2023). An underrated abundance of young reservoirs revealed by new data. *Science Bulletin*, *In review*.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., . . . Alsdorf, D. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, *45*(2), 2005RG000183. https://doi.org/10.1029/2005rg000183

Fergus, C. E., Lapierre, J. F., Oliver, S. K., Skaff, N. K., Cheruvelil, K. S., Webster, K., . . . Soranno, P. (2017). The freshwater landscape: lake, wetland, and stream abundance and connectivity at macroscales. *Ecosphere*, *8*(8), e01911. https://doi.org/10.1002/ecs2.1911

Fisher, A., Flood, N., & Danaher, T. (2016). Comparing Landsat water index methods for automated water classification in eastern Australia. *Remote Sensing of Environment*, *175*, 167-182. https://doi.org/10.1016/j.rse.2015.12.055

Grosse, G., Jones, B., & Arp, C. (2013). Thermokarst lakes, drainage, and drained basins. In J. F. Shroder, R. Giardino, & J. Harbor (Eds.), *Treatise on Geomorphology* (Vol. Vol. 8, Glacial and Periglacial Geomorphology, pp. 325–353). Academic Press.

Herdendorf, C. E. (1984). *Inventory of the morphopmetric and limnologic characteristics of the large lakes of the world*, Technical Bulletin OHSU-TB-17, The Ohio State University Sea Grant Program, March 1984.

JPL internal document (2018). Surface Water and Ocean Topography Mission (SWOT) Project Science Requirements Document, Rev B. JPL D-61923, January 24, 2018. Accessed from https://swot.jpl.nasa.gov/system/documents/files/2176_2176_D-61923_SRD_Rev_B_20181113.pdf on October 16, 2023.

JPL internal document (2022a). Level 2 KaRIn high rate river average vector product (short name: L2_HR_RiverAvg), Revision A, Surface Water and Ocean Topography (SWOT) Project Product Description Document. JPL D-56414, September 27, 2022. Accessed from https://archive.podaac.earthdata.nasa.gov/podaac-ops-cumulus-docs/web-misc/swot_mission_docs/pdd/D-56414_SWOT_Product_Description_L2_HR_RiverAvg_20220927a_RevA.pdf on October 16, 2023.

JPL internal document (2022b). Level 2 KaRIn high rate river single pass vector product (short name: L2_HR_RiverSP), Revision A, Surface Water and Ocean Topography (SWOT) Project Product Description Document. JPL D-56413, September 16, 2022. Accessed from https://archive.podaac.earthdata.nasa.gov/podaac-ops-cumulus-docs/web-misc/swot_mission_docs/pdd/D-56413_SWOT_Product_Description_L2_HR_RiverSP_20220916a_RevA.pdf on October 16, 2023.

JPL internal document (2022c). Level 1B KaRIn high rate single look complex data product (short name: L1B_HR_SLC), Revision A, Surface Water and Ocean Topography Mission (SWOT) Project Product Description Document. JPL D-56410, July 27, 2022. Accessed from https://archive.podaac.earthdata.nasa.gov/podaac-ops-cumulus-docs/web-misc/swot_mission_docs/pdd/D-56410_SWOT_Product_Description_L1B_HR_SLC_20220727_RevA.pdf on October 16, 2023.

JPL internal document (2022d). Level 2 KaRIn high rate water mask pixel cloud product (short name: L2_HR_PIXC), Revision A, Surface Water and Ocean Topography (SWOT) Project Product Description Document. JPL D-56411, July 27, 2022. Accessed from https://archive.podaac.earthdata.nasa.gov/podaac-ops-cumulus-docs/web-misc/swot_mission_docs/pdd/D-56411_SWOT_Product_Description_L2_HR_PIXC_20220727b_RevA.pdf on October 16, 2023.

JPL internal document (2022e). Reference Orbit Track (RefOrbitTrack). Surface Water and Ocean Topography (SWOT) Project Auxiliary Data Description. JPL D-105500, September 13, 2022.

JPL internal document (2023). Level 2 KaRIn high rate river single pass science algorithm software (short name: SAS_L2_HR_RiverSP), Initial Release. Surface Water and Ocean Topography Project Algorithm Theoretical Basis Document. JPL D-105505, July 13, 2023. Accessed from https://archive.podaac.earthdata.nasa.gov/podaac-ops-cumulus-docs/web-misc/swot_mission_docs/atbd/D-105505_SWOT_ATBD_L2_HR_RiverSP_20230713__w-sigs.pdf on October 16, 2023.

Kokelj, S. V., & Jorgenson, M. T. (2013). Advances in Thermokarst Research. *Permafrost and Periglacial Processes*, *24*(2), 108-119. https://doi.org/10.1002/ppp.1779

Lacroix, M. P., Prowse, T. D., Bonsal, B. R., Duguay, C. R., & Menard, P. (2005, Sept. 15-16, 2005). *River ice trends in Canada* 13th Workshop on Ice Covered Rivers, Hanover, NH.

Lehner, B., & Doll, P. (2004). Development and validation of a global database of lakes, reservoirs and wetlands. *Journal of Hydrology*, *296*(1-4), 1-22. https://doi.org/10.1016/j.jhydrol.2004.03.028

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

Lehner, B., Liermann, C. R., Revenga, C., Vorosmarty, C., Fekete, B., Crouzet, P., . . . Wisser, D. (2011). High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, *9*(9), 494-502. https://doi.org/10.1890/100125

Lehner, B., Verdin, K., & Jarvis, A. (2008). New global hydrography derived from spaceborne elevation data. *Eos, Transactions American Geophysical Union*, *89*(10), 93-104. https://doi.org/10.1029/2008EO100001

Li, J., & Sheng, Y. (2012). An automated scheme for glacial lake dynamics mapping using Landsat imagery and digital elevation models: a case study in the Himalayas. *International Journal of Remote Sensing*, *33*(16), 5194-5213. https://doi.org/10.1080/01431161.2012.657370

Luo, S. X., Song, C. Q., Ke, L. H., Zhan, P. F., Fan, C. Y., Liu, K., . . . Zhu, J. Y. (2022). Satellite laser altimetry reveals a net water mass gain in global lakes with spatial heterogeneity in the early 21st century. *Geophysical Research Letters*, *49*(3), e2021GL096676. https://doi.org/10.1029/2021GL096676

Lyons, E. A., & Sheng, Y. W. (2018). LakeTime: Automated seasonal scene selection for global lake mapping using Landsat ETM plus and OLI. *Remote Sensing*, *10*(1), 54. https://doi.org/10.3390/rs10010054

Manasypov, R. M., Pokrovsky, O. S., Kirpotin, S. N., & Shirokova, L. S. (2014). Thermokarst lake waters across the permafrost zones of western Siberia. *Cryosphere*, *8*(4), 1177-1193. https://doi.org/10.5194/tc-8-1177-2014

Mandelbrot, B. B. (1982). *The Fractal Geometry of Nature*. WH Freeman.

McFeeters, S. K. (1996). The use of the normalized difference water index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, *17*(7), 1425-1432. https://doi.org/10.1080/01431169608948714

Mendonca, R., Muller, R. A., Clow, D., Verpoorter, C., Raymond, P., Tranvik, L. J., & Sobek, S. (2017). Organic carbon burial in global lakes and reservoirs. *Nature Communications*, *8*, 1694. https://doi.org/10.1038/s41467-017-01789-6

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*, 13603. https://doi.org/10.1038/ncomms13603

Nie, Y., Sheng, Y. W., Liu, Q., Liu, L. S., Liu, S. Y., Zhang, Y. L., & Song, C. Q. (2017). A regional-scale assessment of Himalayan glacial lake changes using satellite observations from 1990 to 2015. *Remote Sensing of Environment*, *189*, 1-13. https://doi.org/10.1016/j.rse.2016.11.008

Oki, T., & Kanae, S. (2006). Global hydrological cycles and world water resources. *Science*, *313*(5790), 1068-1072. https://doi.org/10.1126/science.1128845

Papa, F., Crétaux, J. F., Grippa, M., Robert, E., Trigg, M., Tshimanga, R. M., . . . Calmant, S. (2023). Water resources in Africa under global change: Monitoring surface waters from space. *Surveys in Geophysics*, *44*(1), 43-93. https://doi.org/10.1007/s10712-022-09700-9

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes [Letter]. *Nature*, *540*(7633), 418-422. https://doi.org/10.1038/nature20584

Pi, X. H., Luo, Q. Q., Feng, L., Xu, Y., Tang, J., Liang, X. Y., . . . Bryan, B. A. (2022). Mapping global lake dynamics reveals the emerging roles of small lakes. *Nature Communications*, *13*(1), 5777. https://doi.org/10.1038/s41467-022-34140-9

Pickens, A. H., Hansen, M. C., Hancher, M., Stehman, S. V., Tyukavina, A., Potapov, P., . . . Sherani, Z. (2020). Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series. *Remote Sensing of Environment*, *243*(15), 111792. https://doi.org/10.1016/j.rse.2020.111792

Riggs, R. M., Allen, G. H., Brinkerhoff, C. B., Sikder, M. S., & Wang, J. (2023). Turning lakes into river gauges using the LakeFlow algorithm. *Geophysical Research Letters*, *50*(10), e2023GL103924. https://doi.org/10.1029/2023GL103924

Schindler, D. W. (2009). Lakes as sentinels and integrators for the effects of climate change on watersheds, airsheds, and landscapes. *Limnology and Oceanography*, *54*(6), 2349-2358. https://doi.org/10.4319/lo.2009.54.6_part_2.2349

Schwatke, C., Dettmering, D., Bosch, W., & Seitz, F. (2015). DAHITI - an innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry. *Hydrology and Earth System Sciences*, *19*(10), 4345-4364. https://doi.org/10.5194/hess-19-4345-2015

Sheng, Y. W., Song, C. Q., Wang, J. D., Lyons, E. A., Knox, B. R., Cox, J. S., & Gao, F. (2016). Representative lake water extent mapping at continental scales using multi-temporal Landsat-8 imagery. *Remote Sensing of Environment*, *185*, 129-141. https://doi.org/10.1016/j.rse.2015.12.041

Shilts, W. W., Aylsworth, J. M., Kaszycki, C. A., & Klassen, R. A. (1987). Canadian Shield. In W. L. Graf (Ed.), *Geomorphic systems of North America* (Vol. Centennial Special Volume 2, pp. 119-161). Geological Society of America.

Shugar, D. H., Burr, A., Haritashya, U. K., Kargel, J. S., Watson, C. S., Kennedy, M. C., . . . Strattman, K. (2020). Rapid worldwide growth of glacial lakes since 1990. *Nature Climate Change*, *10*(10), 939-945. https://doi.org/10.1038/s41558-020-0855-4

Sikder, M. S., Wang, J., Allen, G. H., Sheng, Y., Yamazaki, D., Song, C., . . . Pavelsky, T. M. (2023). Lake-TopoCat: A global lake drainage topology and catchment database. *Earth System Science Data*, *15*, 3483–3511. https://doi.org/10.5194/essd-15-3483-2023

Slater, J. A., Garvey, G., Johnston, C., Haase, J., Heady, B., Kroenung, G., & Little, J. (2006). The SRTM data "finishing" process and products. *Photogrammetric Engineering and Remote Sensing*, *72*(3), 237-247. https://doi.org/10.14358/Pers.72.3.237

Smith, L. C., Sheng, Y., MacDonald, G. M., & Hinzman, L. D. (2005). Disappearing Arctic lakes. *Science*, *308*(5727), 1429-1429. https://doi.org/10.1126/science.1108142

Song, C., Sheng, Y., Wang, J., Ke, L. H., Madson, A., & Nie, Y. (2017). Heterogeneous glacial lake changes and links of lake expansions to the rapid thinning of adjacent glacier termini in the Himalayas. *Geomorphology*, *280*(1), 30-38. https://doi.org/10.1016/j.geomorph.2016.12.002

Tranvik, L. J., Downing, J. A., Cotner, J. B., Loiselle, S. A., Striegl, R. G., Ballatore, T. J., . . . Weyhenmeyer, G. A. (2009). Lakes and reservoirs as regulators of carbon cycling and climate. *Limnology and Oceanography*, *54*(6), 2298-2314. https://doi.org/10.4319/lo.2009.54.6_part_2.2298

U.S. Geological Survey. (2013). *National Hydrography Geodatabase: Alaska*. http://nhd.usgs.gov/

Verdin, K. L., & Verdin, J. P. (1999). A topological system for delineation and codification of the Earth's river basins. *Journal of Hydrology*, *218*(1-2), 1-12. https://doi.org/10.1016/S0022-1694(99)00011-6

Verpoorter, C., Kutser, T., Seekell, D. A., & Tranvik, L. J. (2014). A global inventory of lakes based on high-resolution satellite imagery. *Geophysical Research Letters*, *41*(18), 6396-6402. https://doi.org/10.1002/2014gl060641

Wang, J. D., Sheng, Y. W., & Tong, T. S. D. (2014). Monitoring decadal lake dynamics across the Yangtze Basin downstream of Three Gorges Dam. *Remote Sensing of Environment*, *152*, 251-269. https://doi.org/10.1016/j.rse.2014.06.004

Wang, J. D., Song, C. Q., Reager, J. T., Yao, F. F., Famiglietti, J. S., Sheng, Y. W., . . . Wada, Y. (2018). Recent global decline in endorheic basin water storages. *Nature Geoscience*, *11*(12), 926-932. https://doi.org/10.1038/s41561-018-0265-7

Wang, J. D., Walter, B. A., Yao, F. F., Song, C. Q., Ding, M., Maroof, A., . . . Wada, Y. (2022). GeoDAR: georeferenced global dams and reservoirs dataset for bridging attributes and geolocations. *Earth System Science Data*, *14*(4), 1869-1899. https://doi.org/10.5194/essd-14-1869-2022

Wik, M., Varner, R. K., Anthony, K. W., MacIntyre, S., & Bastviken, D. (2016). Climate-sensitive northern lakes and ponds are critical components of methane release. *Nature Geoscience*, *9*(2), 99-105. https://doi.org/10.1038/ngeo2578

WMO. (2022). *The 2022 GCOS Implementation Plan. GCOS-244. World Meteorological Organization*.

Wu, Q., Ke, L., Wang, J., Pavelsky, T. M., Allen, G. H., Sheng, Y., . . . Song, C. (2023). Satellites reveal hotspots of global river extent change. *Nature Communications*, *14*, 1587. https://doi.org/10.1038/s41467-023-37061-3

Wurtsbaugh, W. A., Miller, C., Null, S. E., DeRose, R. J., Wilcock, P., Hahnenberger, M., . . . Moore, J. (2017). Decline of the world's saline lakes. *Nature Geoscience*, *10*(11), 816-821. https://doi.org/10.1038/ngeo3052

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. *Water Resources Research*, *55*(6), 5053-5073. https://doi.org/10.1029/2019wr024873

Yang, X., O'Reilly, C. M., Gardner, J. R., Ross, M. R. V., Topp, S. N., Wang, J. D., & Pavelsky, T. M. (2022). The Color of Earth's Lakes. *Geophysical Research Letters*, *49*(18), e2022GL098925. https://doi.org/10.1029/2022GL098925

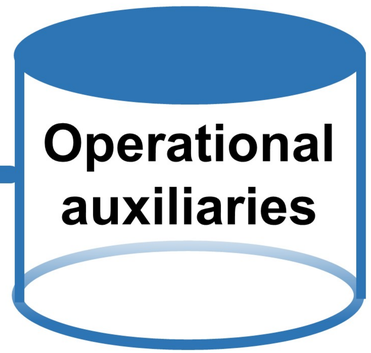
Yang, X., Pavelsky, T. M., Bendezu, L. P., & Zhang, S. (2022). Simple method to extract lake ice condition from Landsat images. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 4202010. https://doi.org/10.1109/Tgrs.2021.3088144

Yao, F., Livneh, B., Rajagopalan, B., Wang, J., Crétaux, J.-F., Wada, Y., & Berge-Nguyen, M. (2023). Satellites reveal widespread decline in global lake water storage. *Science*, *380*, 743-749. https://doi.org/10.1126/science.abo2812

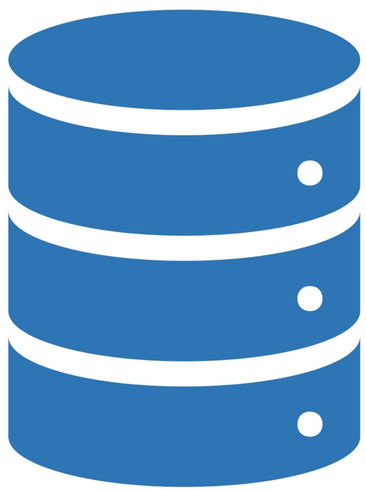
Zimnitskaya, H., & Geldern, J. v. (2011). Is the Caspian Sea a sea; and why does it matter? *Journal of Eurasian Studies*, *2*(1), 1-14. https://doi.org/10.1016/j.euras.2010.10.009

Figure 1.

Operational PLD
(paper I)



Scientific PLD
(paper II)



1. High-resolution global lake and reservoir polygons

- 6 million prior lakes >1 ha (SWOT science goal)
- Representative inundation extents based on the UCLA Circa-2015 Global Lake Dataset
- Appended by recently constructed reservoirs

2. Prior information to ease SWOT lake data production

- To link SWOT observations to prior lakes
- To enable computation of lake storage changes
- To ease lake processing and populate the lake products

3. Scientific metadata to facilitate SWOT data product applications

- Lake topology and catchments
- Typology
- Evaporation
- Ice phenology
- …

Figure 2.

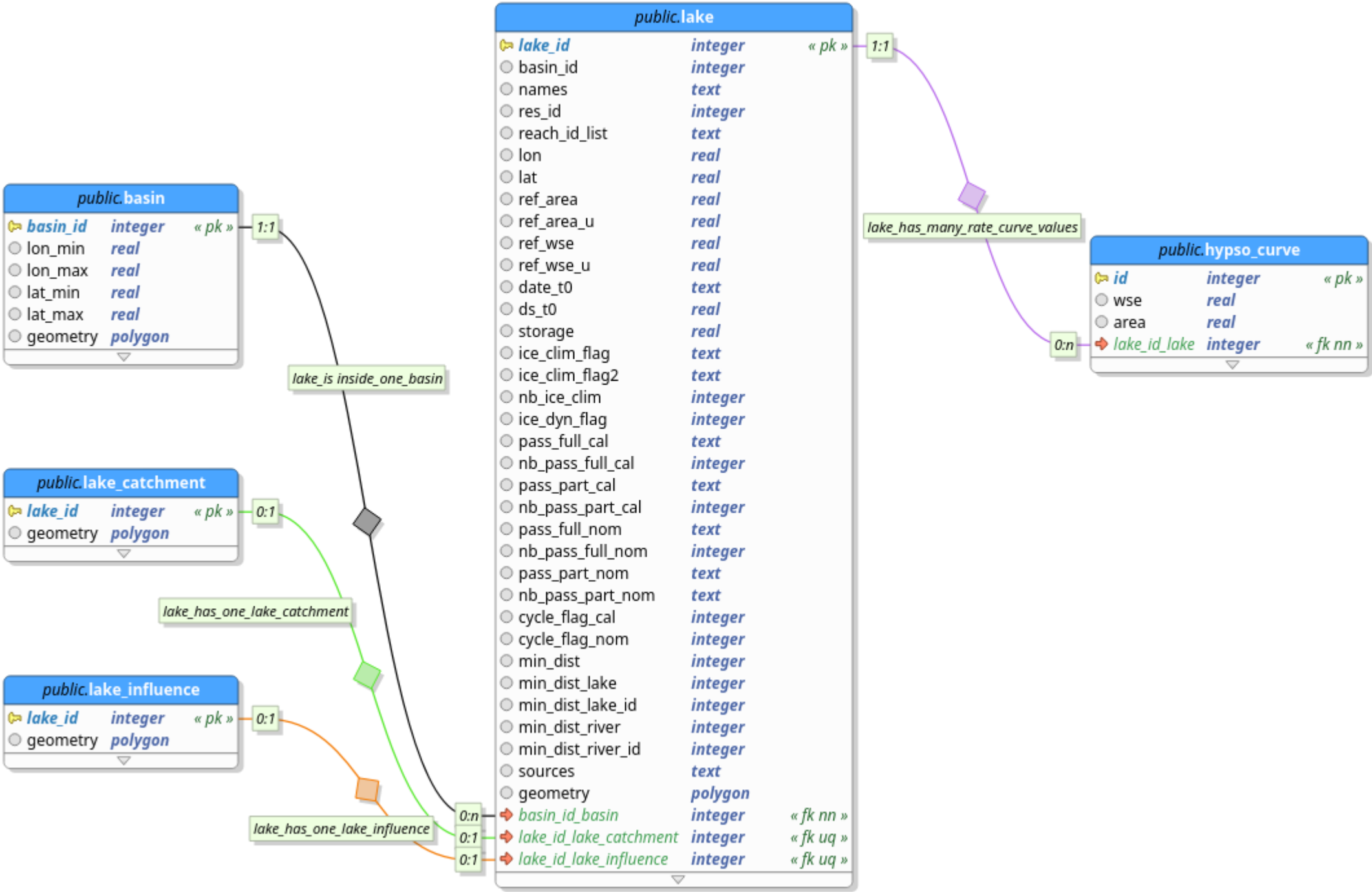
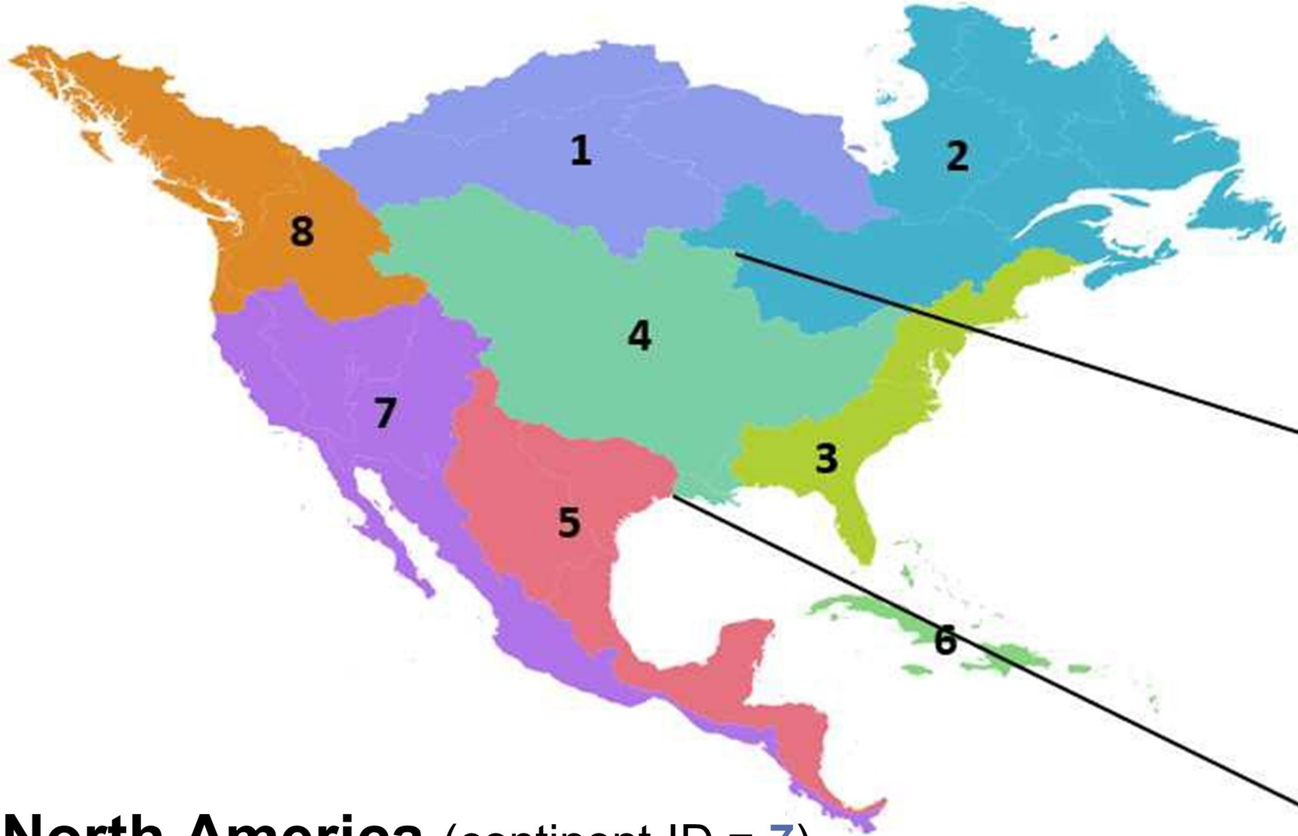

public.lake
lake_id integer « pk »
basin_id integer
names text
res_id integer
reach_id_list text
lon real
lat real
ref_area real
ref_area_u real
ref_wse real
ref_wse_u real
date_t0 text
ds_t0 real
storage real
ice_clim_flag text
ice_clim_flag2 text
nb_ice_clim integer
ice_dyn_flag integer
pass_full_cal text
nb_pass_full_cal integer
pass_part_cal text
nb_pass_part_cal integer
pass_full_nom text
nb_pass_full_nom integer
pass_part_nom text
nb_pass_part_nom text
cycle_flag_cal integer
cycle_flag_nom integer
min_dist integer
min_dist_lake integer
min_dist_lake_id integer
min_dist_river integer
min_dist_river_id integer
sources text
geometry polygon
basin_id_basin integer « fk nn »
lake_id_lake_catchment integer « fk uq »
lake_id_lake_influence integer « fk uq »
1:1
lake_has_many_rate_curve_values
0:n
public.hypso_curve
id integer « pk »
wse real
area real
lake_id_lake integer « fk nn »
public.basin
basin_id integer « pk »
lon_min real
lon_max real
lat_min real
lat_max real
geometry polygon
1:1
lake_is inside_one_basin
0:n
public.lake_catchment
lake_id integer « pk »
geometry polygon
0:1
lake_has_one_lake_catchment
0:1
public.lake_influence
lake_id integer « pk »
geometry polygon
0:1
lake_has_one_lake_influence
0:1

Figure 3.

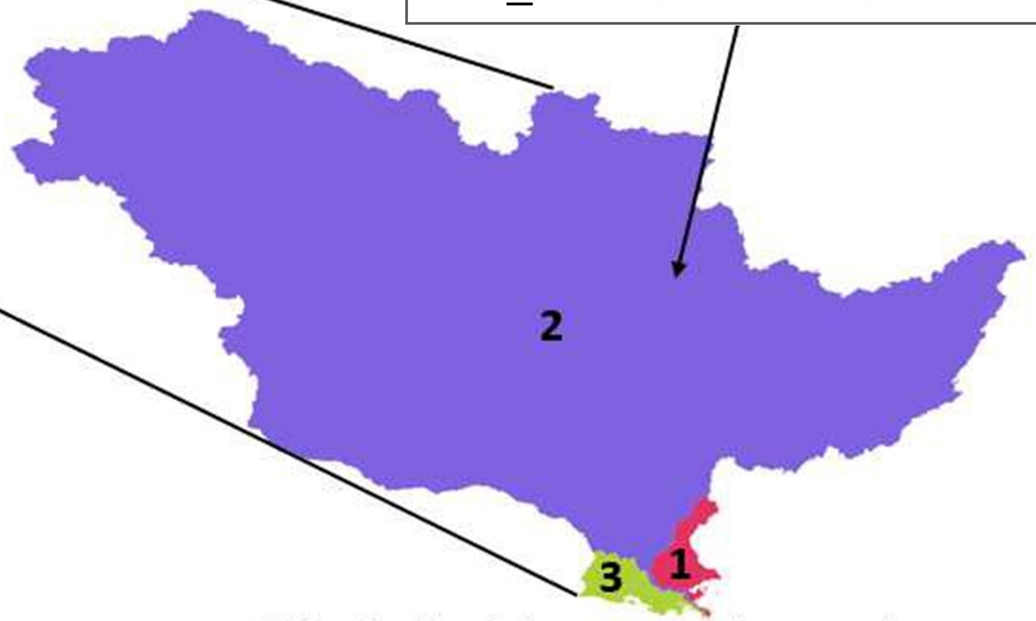


North America (continent ID = 7)
Sub-basins (level 2)

CBBNNNNNNT (*lake_id*)

- **C – Continent**
- **B – Basin code**
- **N – Lake index in the basin**
- **T – Type**

Example for a lake in this basin:
lake_id = 7420469602



Mississippi (basin 1st digit = 4)
Sub-basins (level 3)

Figure 4.

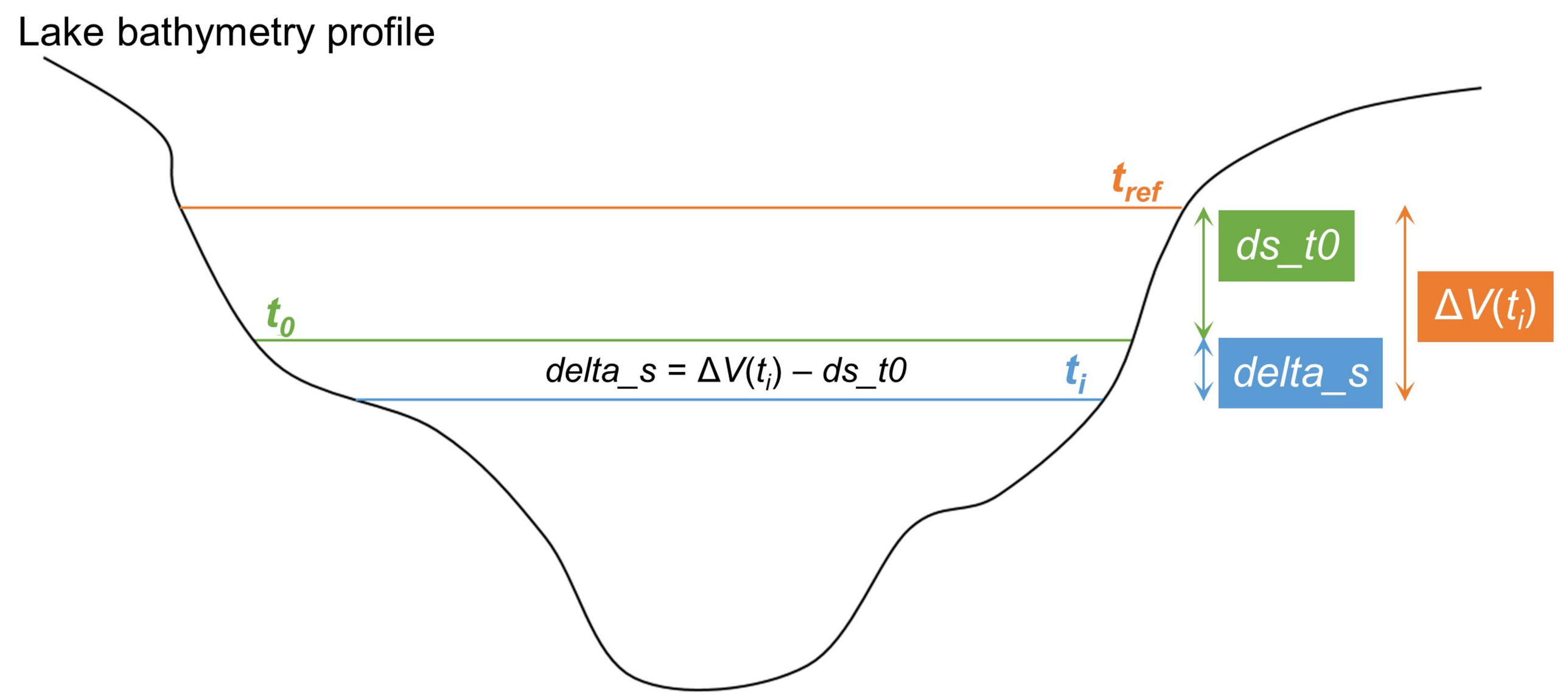

Lake bathymetry profile
t_{ref}
t_0
t_i
$delta_s = \Delta V(t_i) - ds_t0$
ds_t0
$delta_s$
$\Delta V(t_i)$

Figure 5.

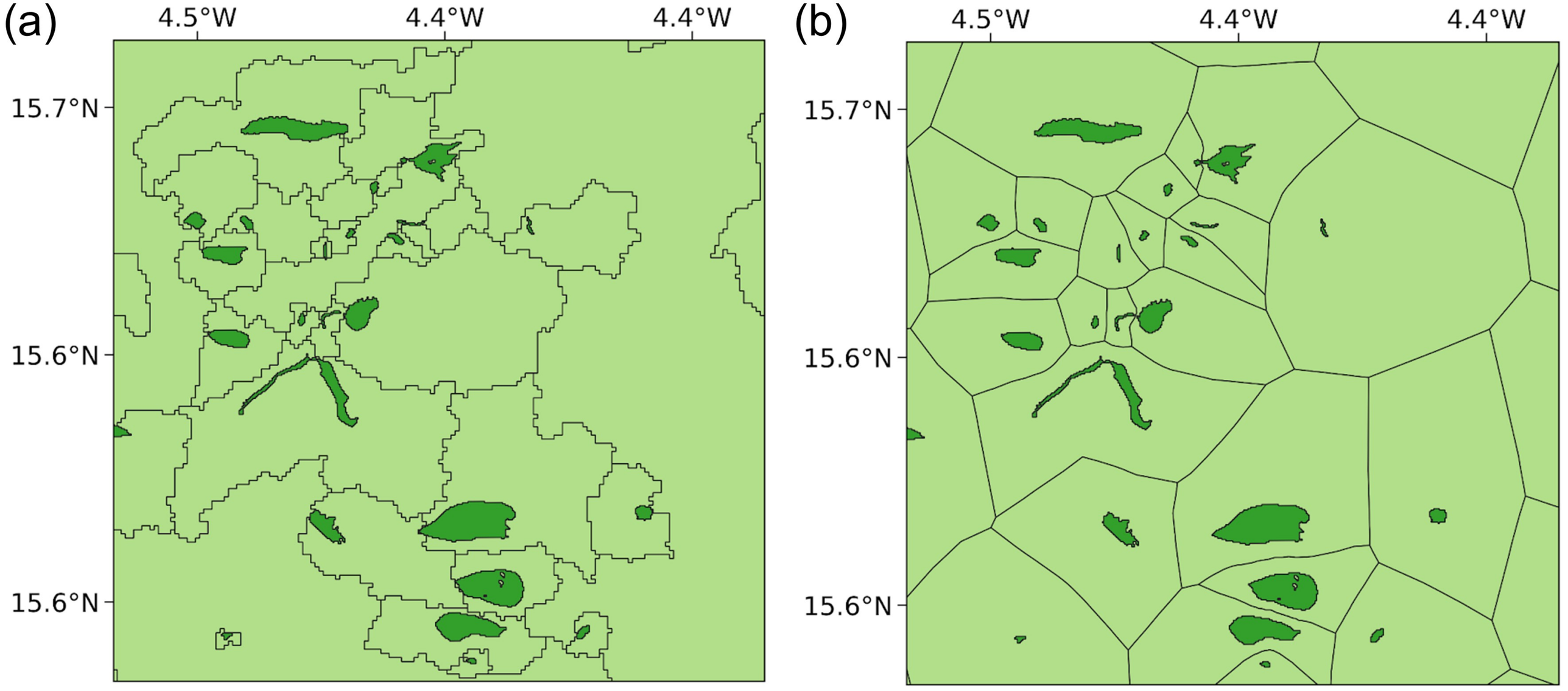

(a)
4.5°W
4.4°W
4.4°W
15.7°N
15.6°N
15.6°N
(b)
4.5°W
4.4°W
4.4°W
15.7°N
15.6°N
15.6°N

Figure 6.

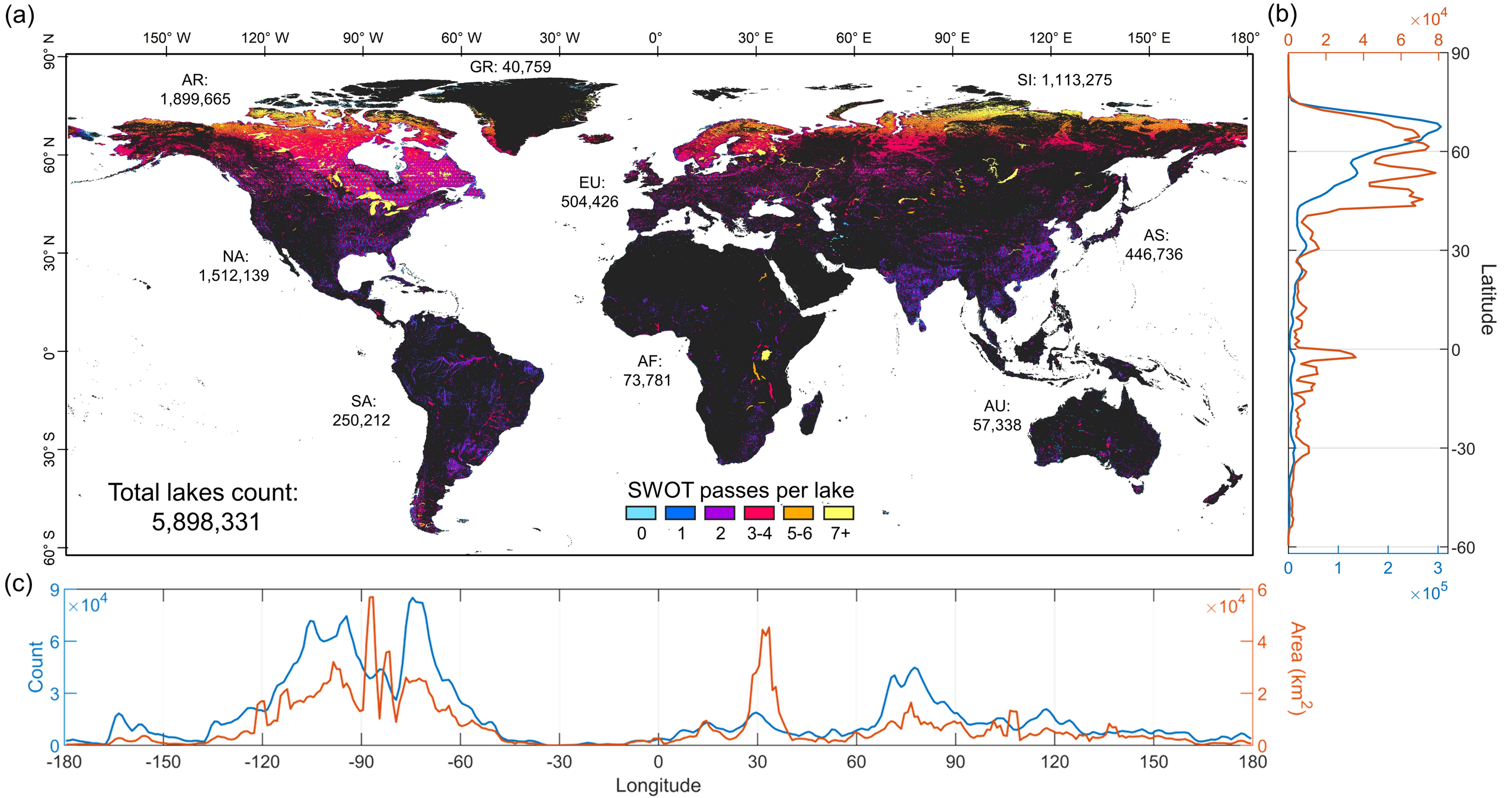

(a)
150° W
120° W
90° W
60° W
30° W
0°
30° E
60° E
90° E
120° E
150° E
180°
90° N
60° N
30° N
0°
30° S
60° S
GR: 40,759
AR:
1,899,665
SI: 1,113,275
EU:
504,426
NA:
1,512,139
AS:
446,736
AF:
73,781
SA:
250,212
AU:
57,338
Total lakes count:
5,898,331
SWOT passes per lake
0
1
2
3-4
5-6
7+
(b)
×10^4
0
2
4
6
8
90
60
30
0
-30
-60
Latitude
0
1
2
3
×10^5
(c)
×10^4
9
6
3
0
Count
×10^4
6
4
2
0
Area (km^2)
-180
-150
-120
-90
-60
-30
0
30
60
90
120
150
180
Longitude

Figure 7.

(a)

Cumulative count

10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7

Tail exponent (τ)
in $N \sim A^{-\tau}$ ($A \geq 30$ ha)
1.00
1.01
0.99 (fitting shown)
0.99

Requirement
6.25 ha

Goal
1 ha

10^{-1} 10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7 10^8

Size (ha)

HydroLAKES
A ≥ 10 ha

GLAKES
A ≥ 3 ha

SWOT PLD
A ≥ 1 ha

UCLA Circa-2015
(main PLD source)
A ≥ 0.4 ha

(b)

Perimeter (100 m)

10^1 10^2 10^3 10^4 10^5 10^6

Fractal dimension (d)
in $L \sim A^{d/2}$ ($A \geq 30$ ha)
1.25
1.33
1.30 (fitting shown)
1.30

Circular
(d = 1)

10^{-1} 10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7 10^8

Size (ha)

(c)

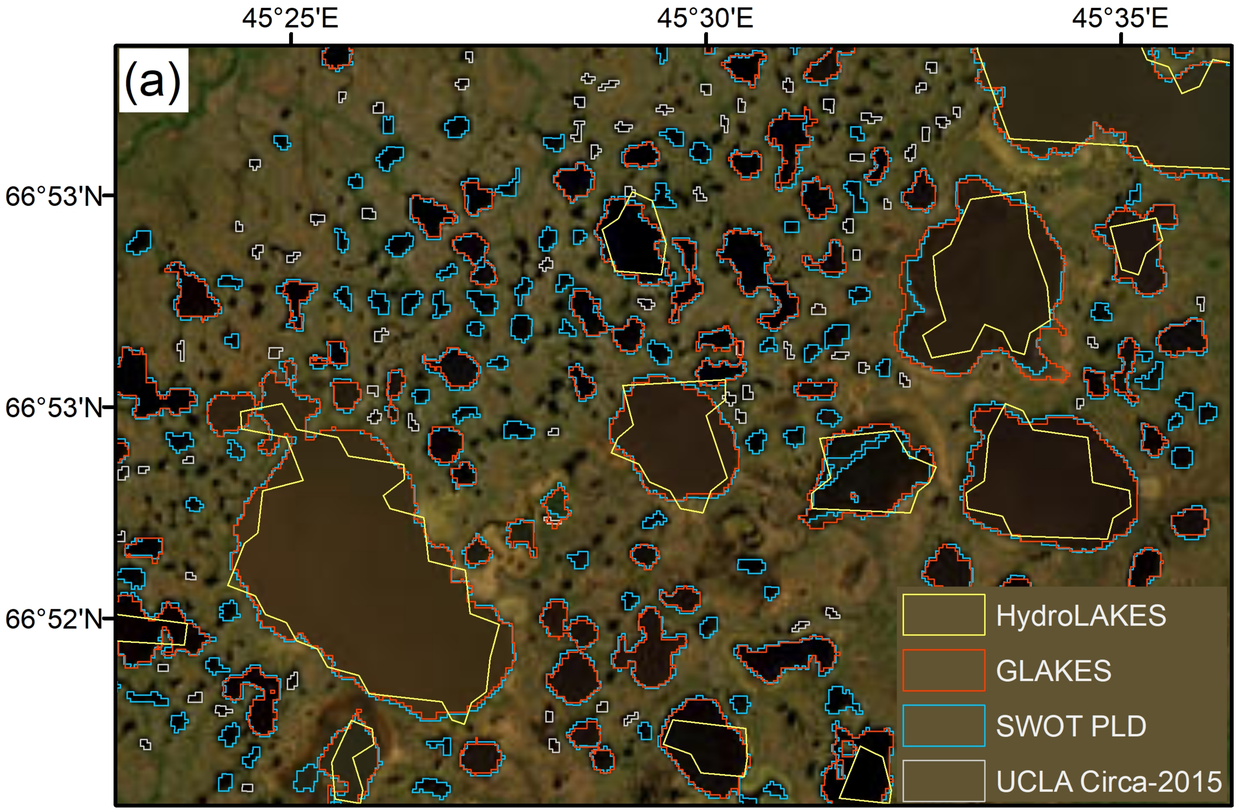
Area (ha)

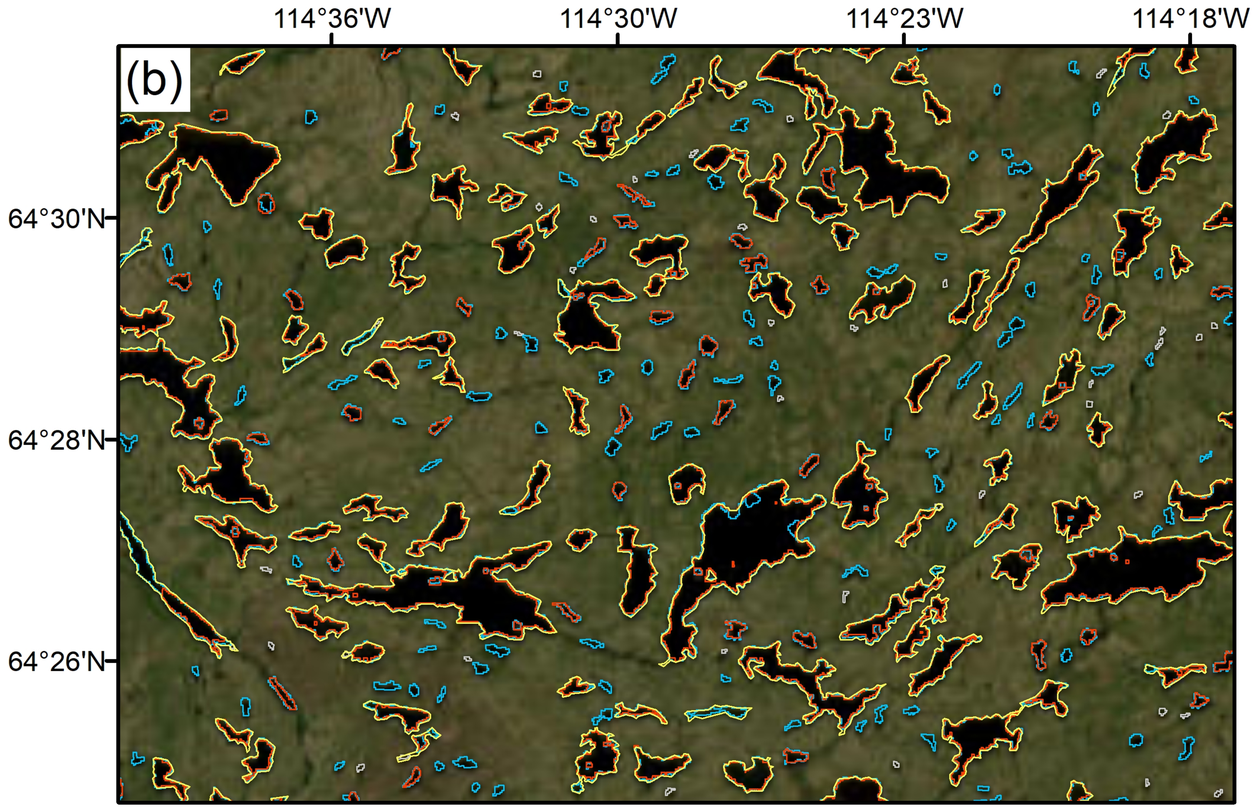
10^5 10^6 10^7

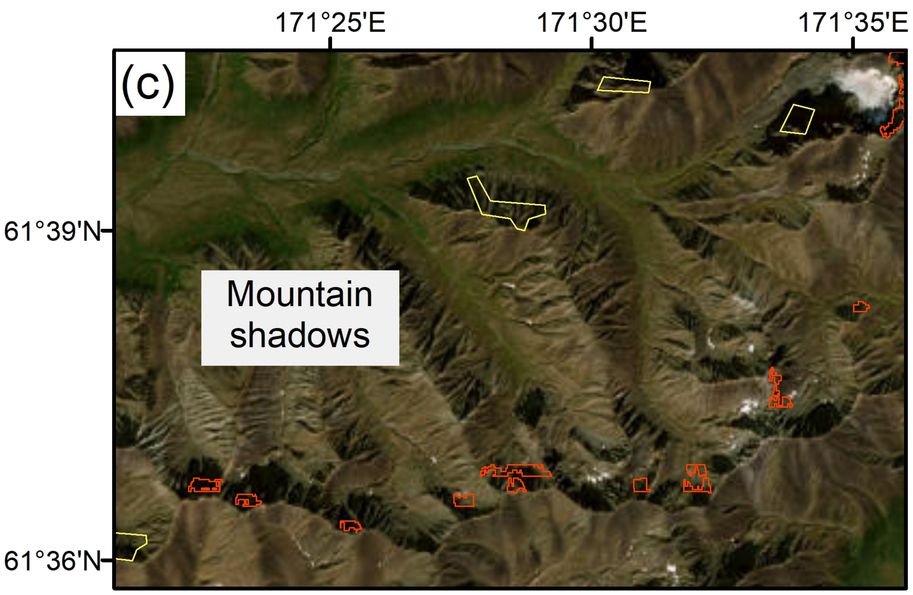
10^{-1} 10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7 10^8

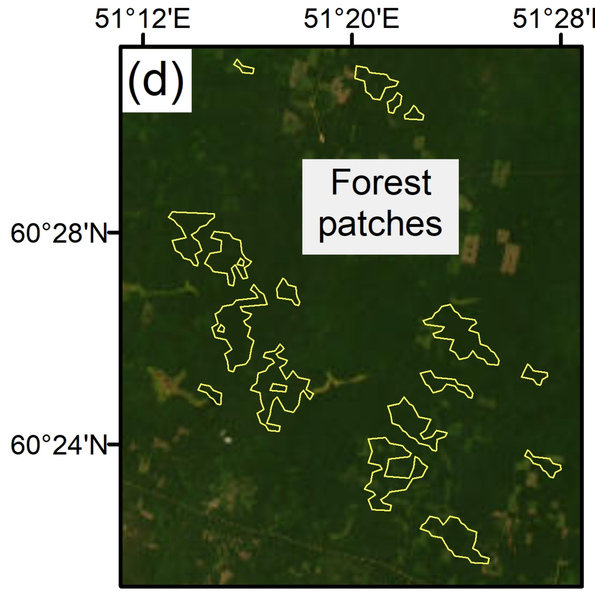
Size (ha)

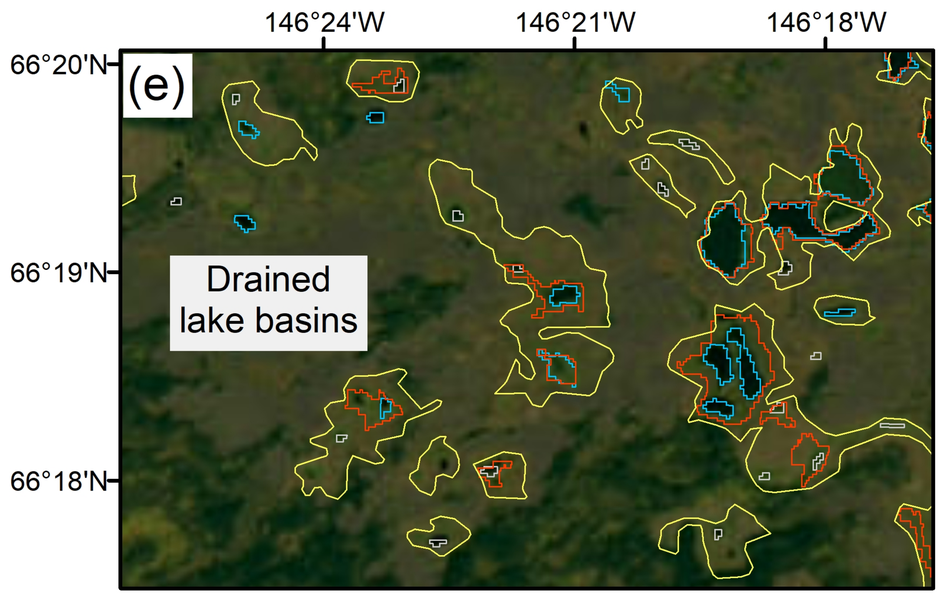
Figure 8.

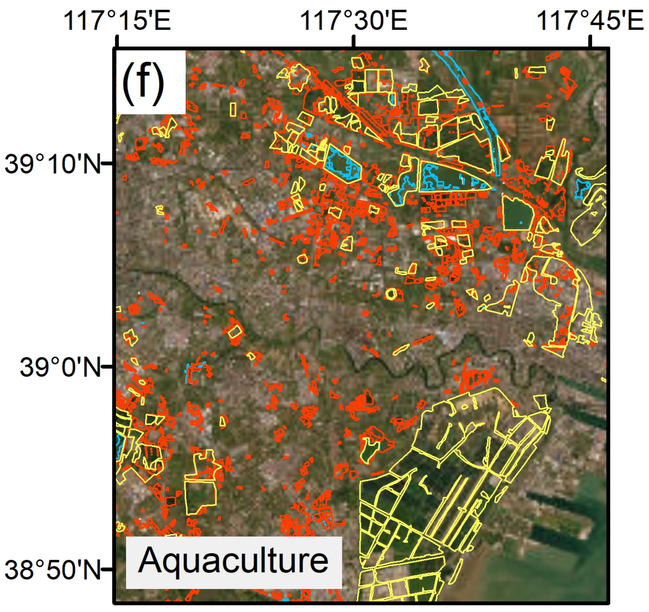

(a)
45°25'E
45°30'E
45°35'E
66°53'N
66°53'N
66°52'N
HydroLAKES
GLAKES
SWOT PLD
UCLA Circa-2015

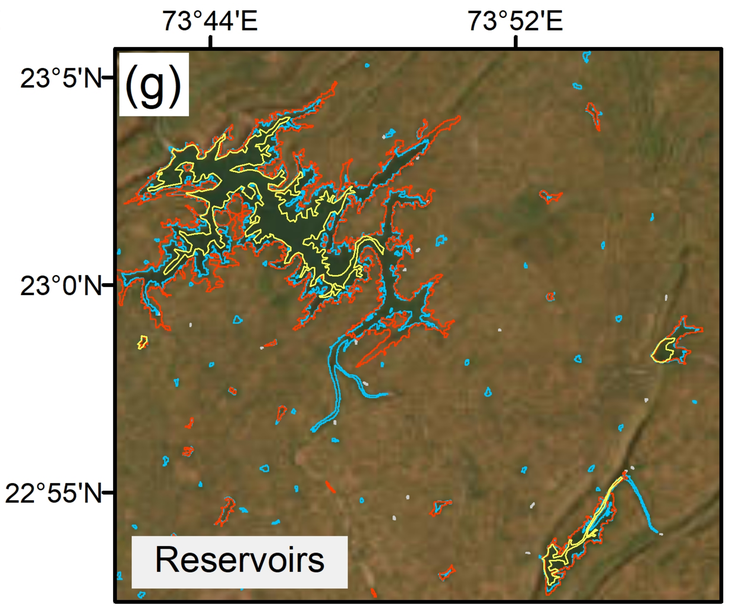

(b)
114°36'W
114°30'W
114°23'W
114°18'W
64°30'N
64°28'N
64°26'N


(c)
171°25'E
171°30'E
171°35'E
61°39'N
61°36'N
Mountain shadows


(d)
51°12'E
51°20'E
51°28'E
60°28'N
60°24'N
Forest patches


(e)
146°24'W
146°21'W
146°18'W
66°20'N
66°19'N
66°18'N
Drained lake basins


(f)
117°15'E
117°30'E
117°45'E
39°10'N
39°0'N
38°50'N
Aquaculture


(g)
73°44'E
73°52'E
23°5'N
23°0'N
22°55'N
Reservoirs

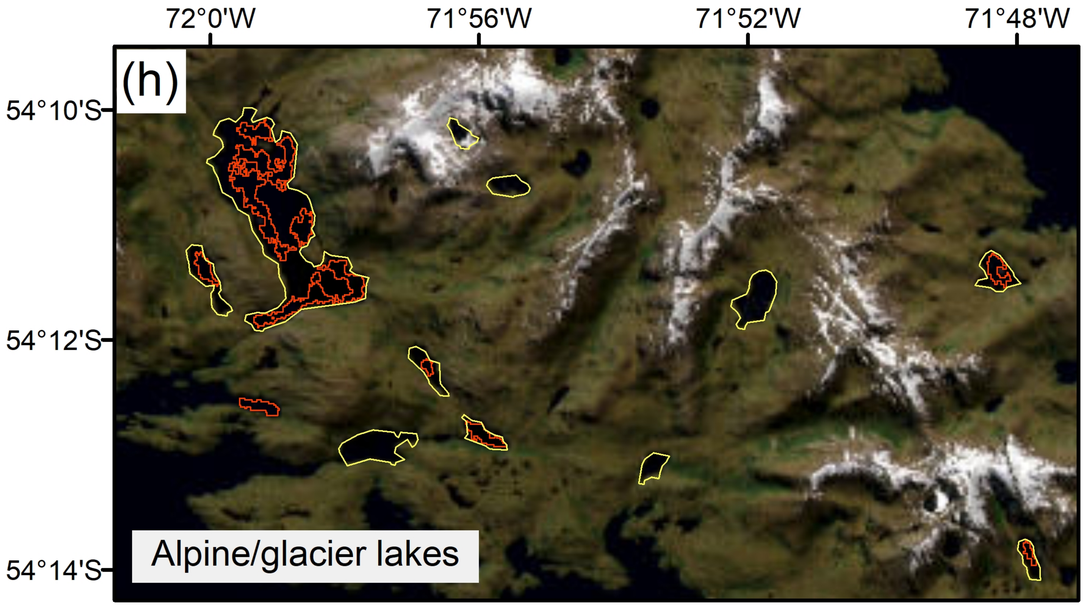
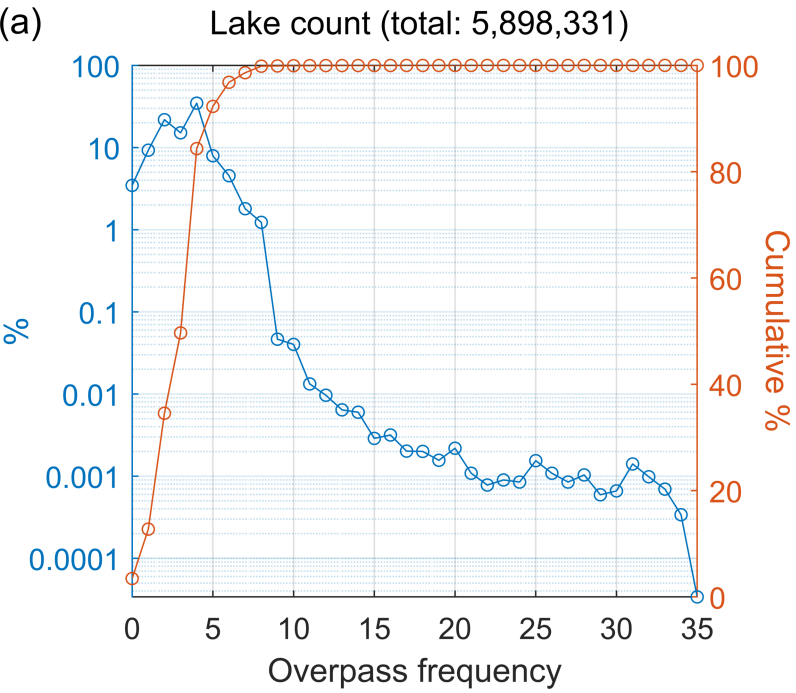

(h)
72°0'W
71°56'W
71°52'W
71°48'W
54°10'S
54°12'S
54°14'S
Alpine/glacier lakes

Figure 9.


(a)
Lake count (total: 5,898,331)
100
10
1
0.1
0.01
0.001
0.0001
%
100
80
60
40
20
0
Cumulative %
0
5
10
15
20
25
30
35
Overpass frequency

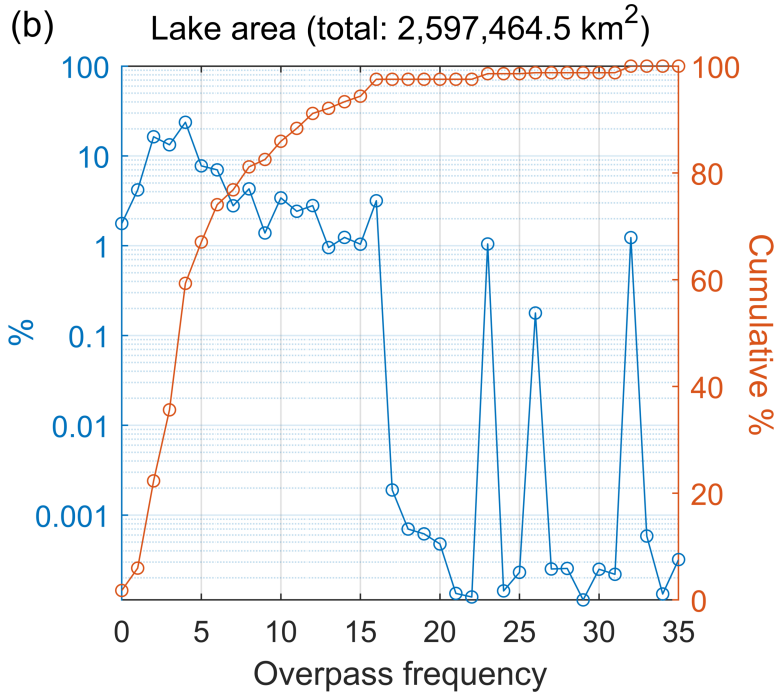

(b)
Lake area (total: 2,597,464.5 km2)
100
10
1
0.1
0.01
0.001
%
100
80
60
40
20
0
Cumulative %
0
5
10
15
20
25
30
35
Overpass frequency

Figure 10.

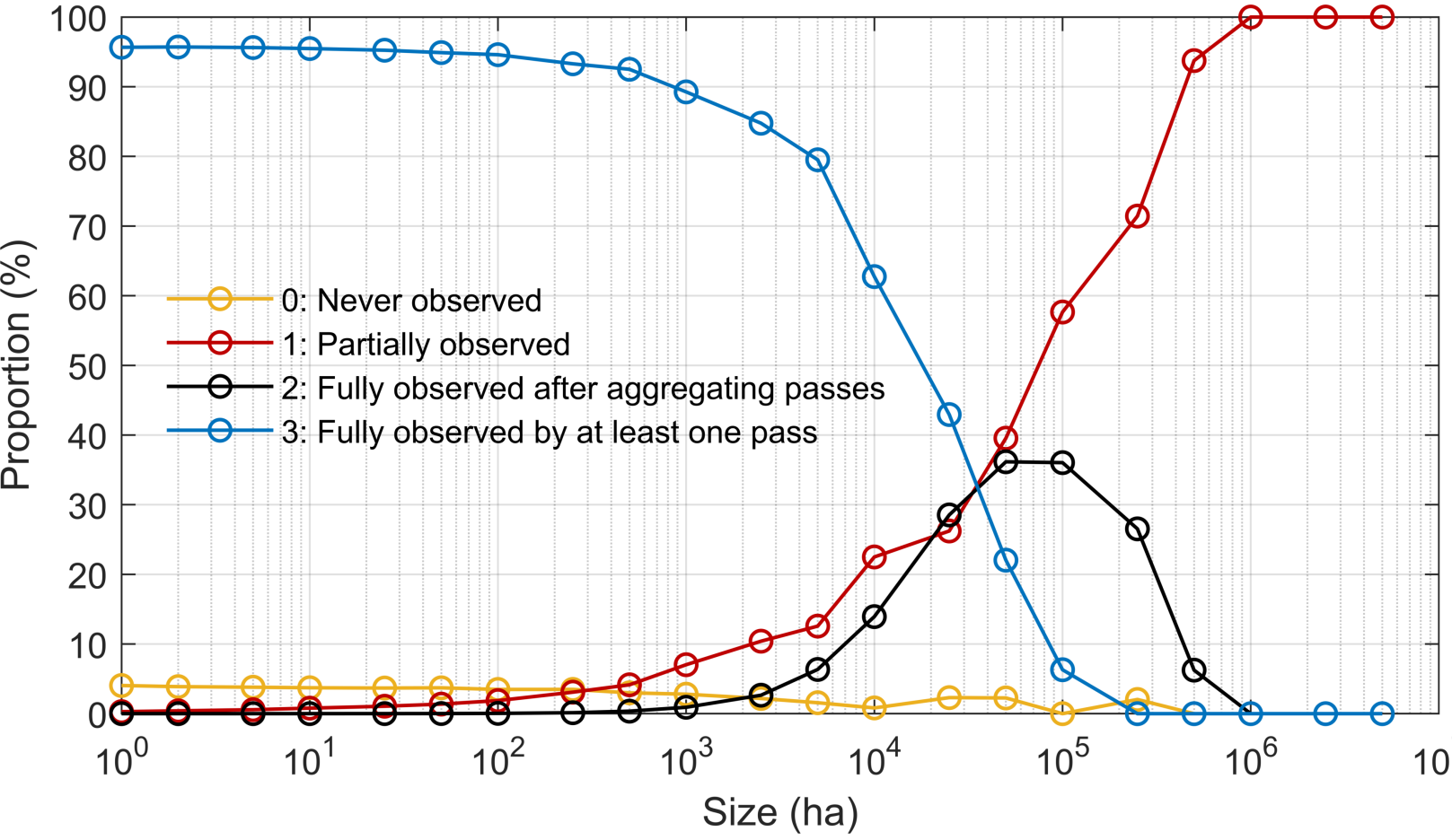
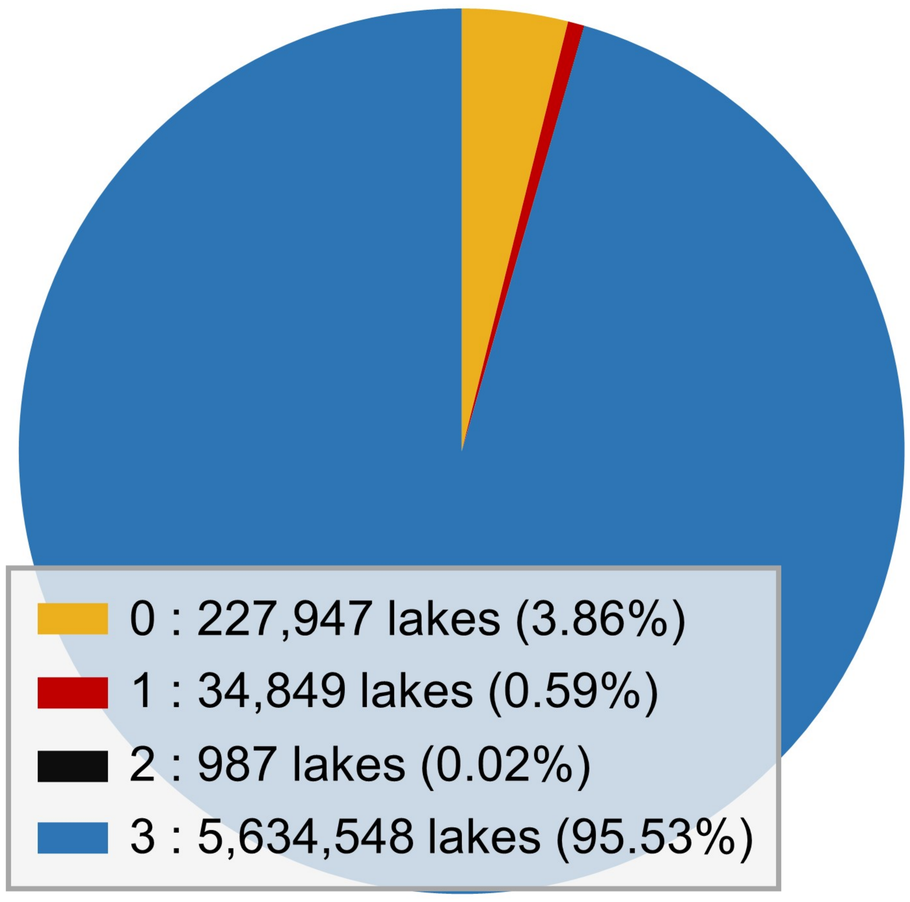

100
90
80
70
60
50
40
30
20
10
0
Proportion (%)
0: Never observed
1: Partially observed
2: Fully observed after aggregating passes
3: Fully observed by at least one pass
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
Size (ha)

Figure 11.

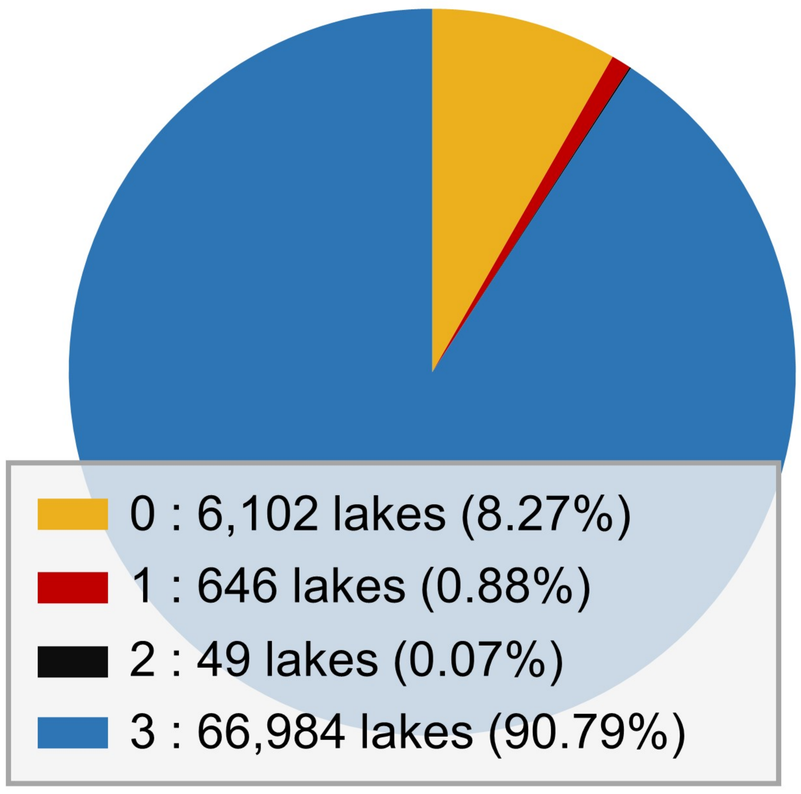
SWOT cycle flag for prior lakes

- 0 : Never observed
- 1 : Partially observed
- 2 : Fully observed after aggregating passes
- 3 : Fully observed by at least one pass

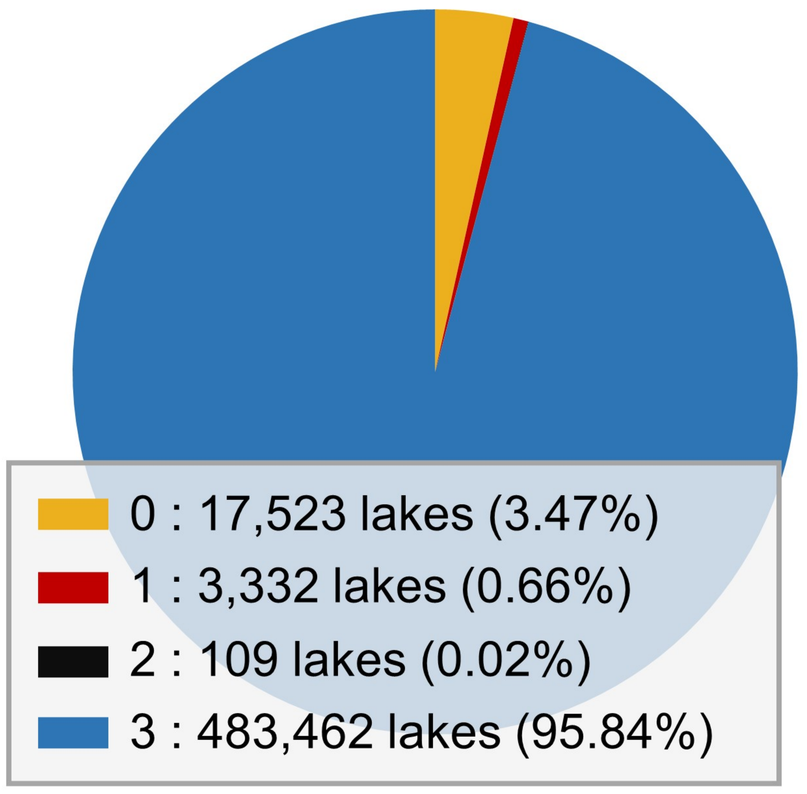
(a) Global composition of each cycle flag



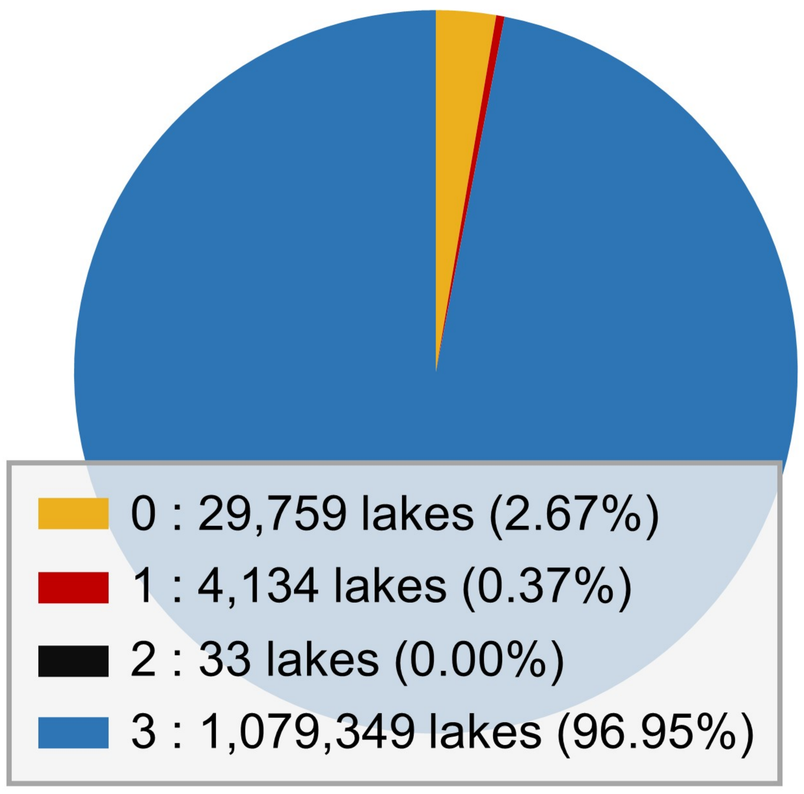
(b) Continent AF (code 1)



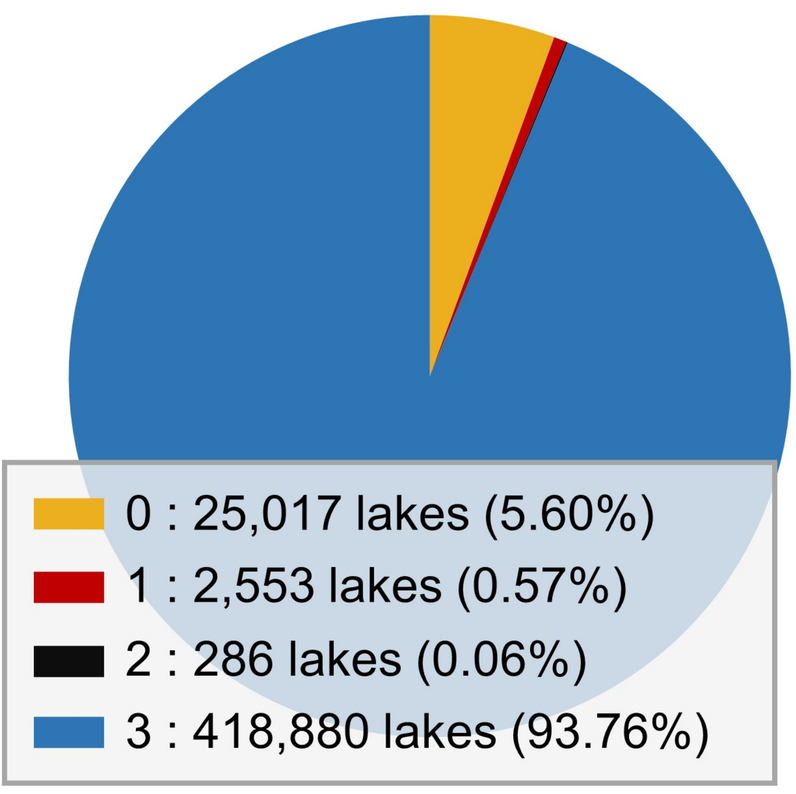
(c) Continent EU (code 2)



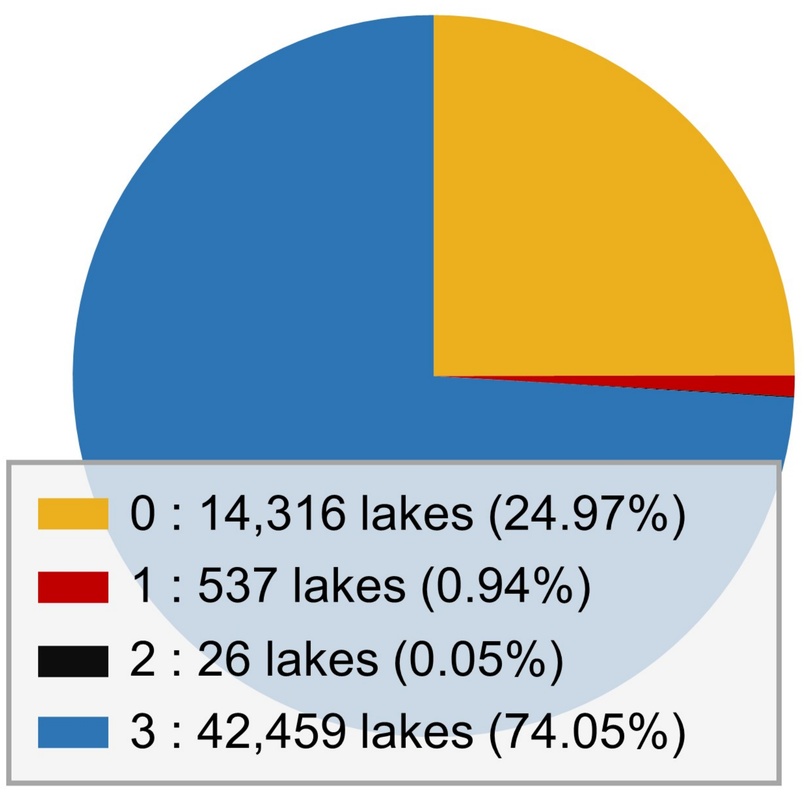
(d) Continent SI (code 3)



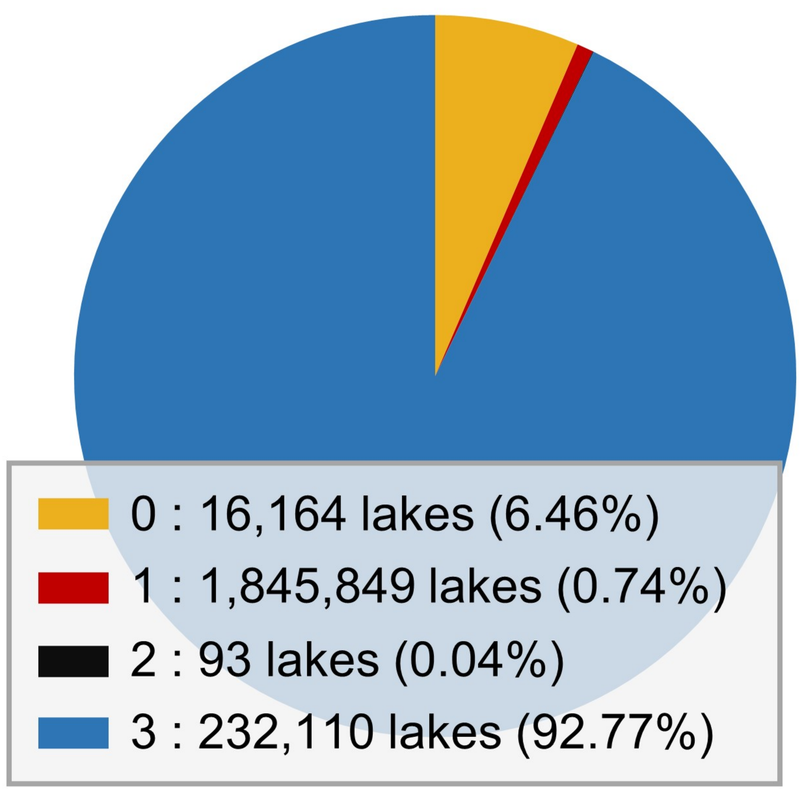
(e) Continent AS (code 4)



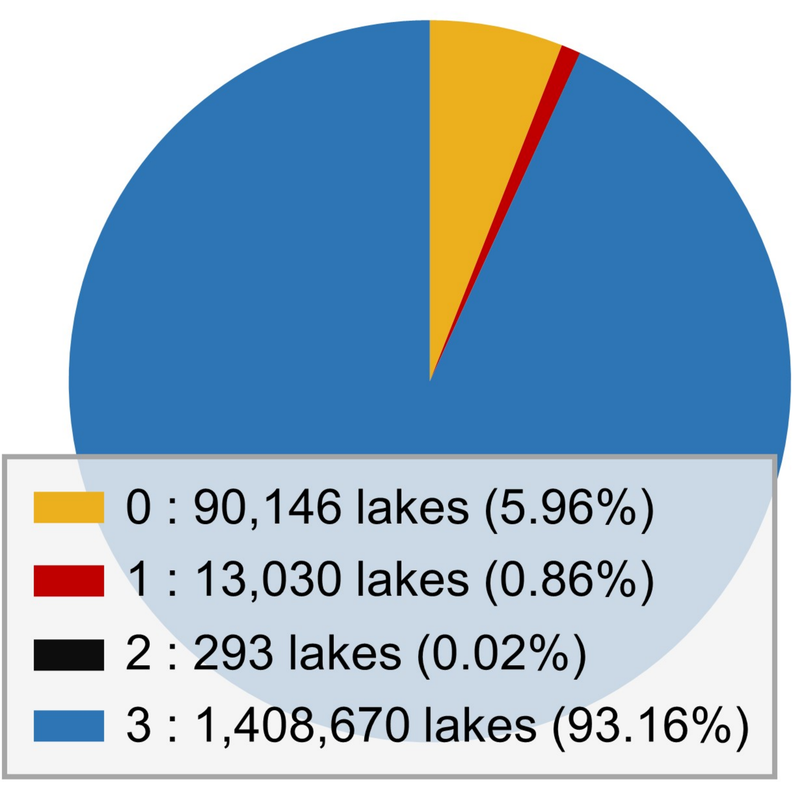
(f) Continent AU (code 5)



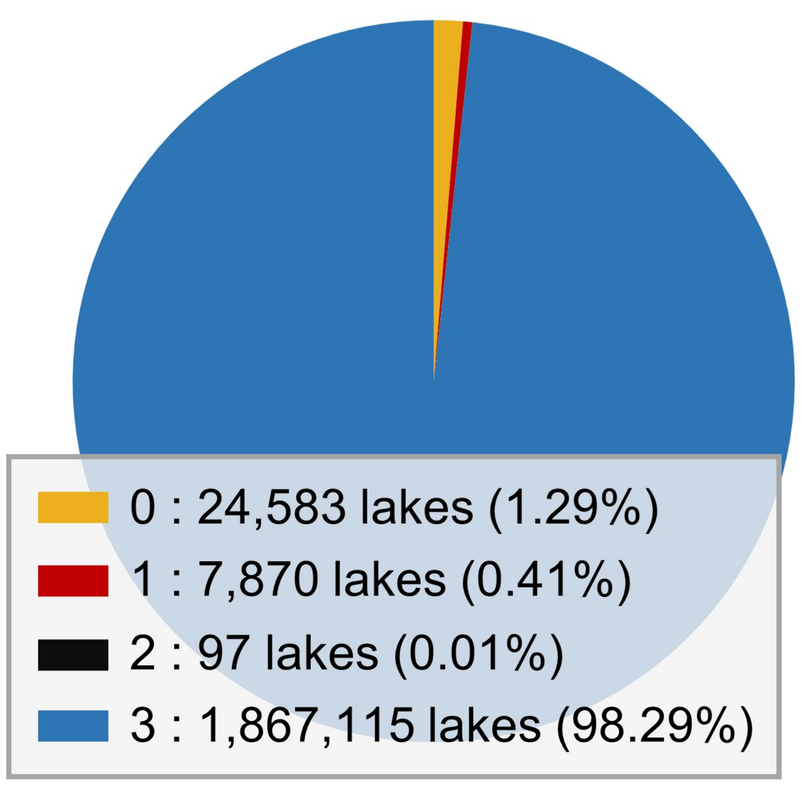
(g) Continent SA (code 6)



(h) Continent NA (code 7)



(i) Continent AR (code 8)



(j) Continent GR (code 9)

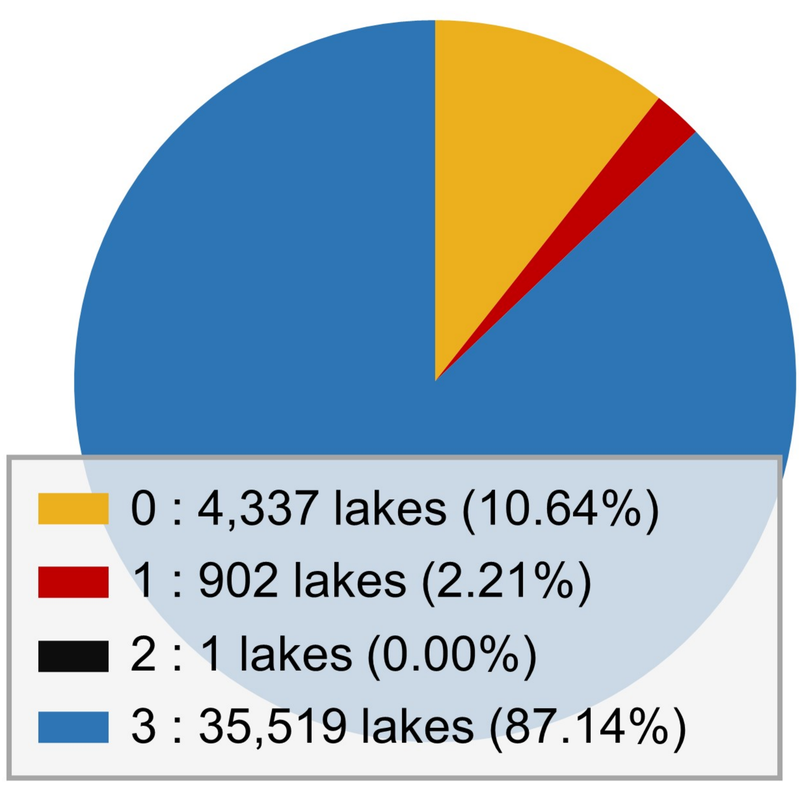


Figure 12.

(a)

(b)

_obs

(c)

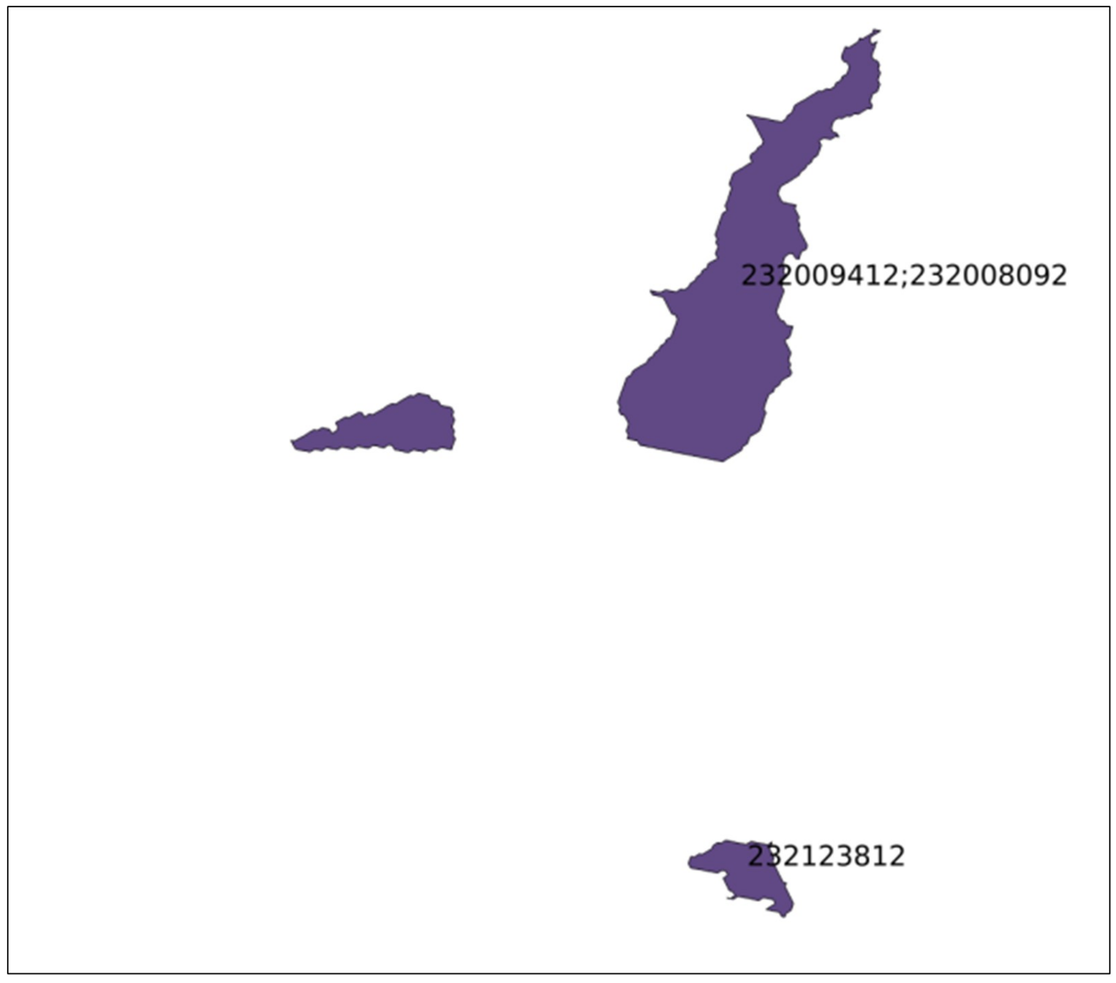
_Prior

(d)

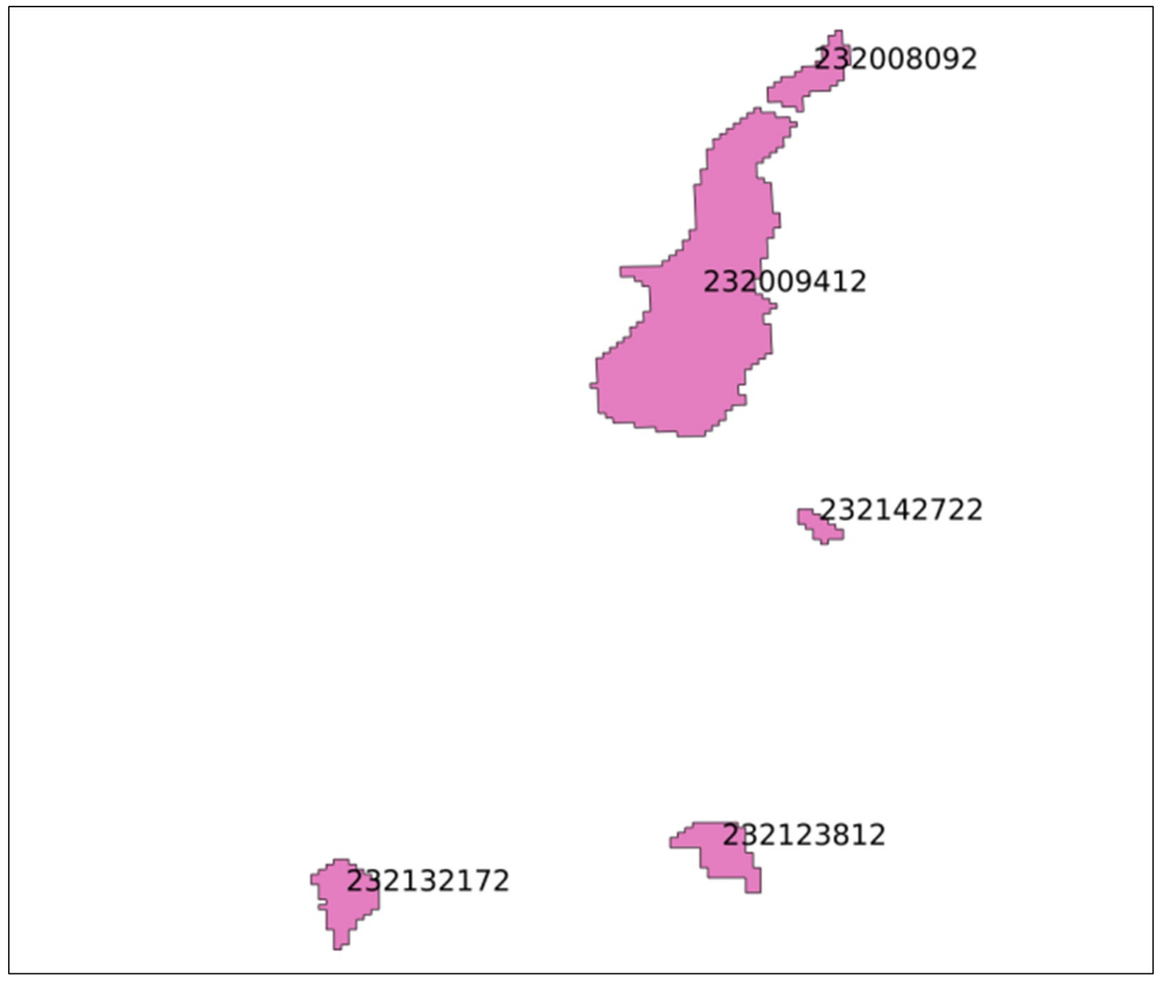
_Unassigned

Figure 13.

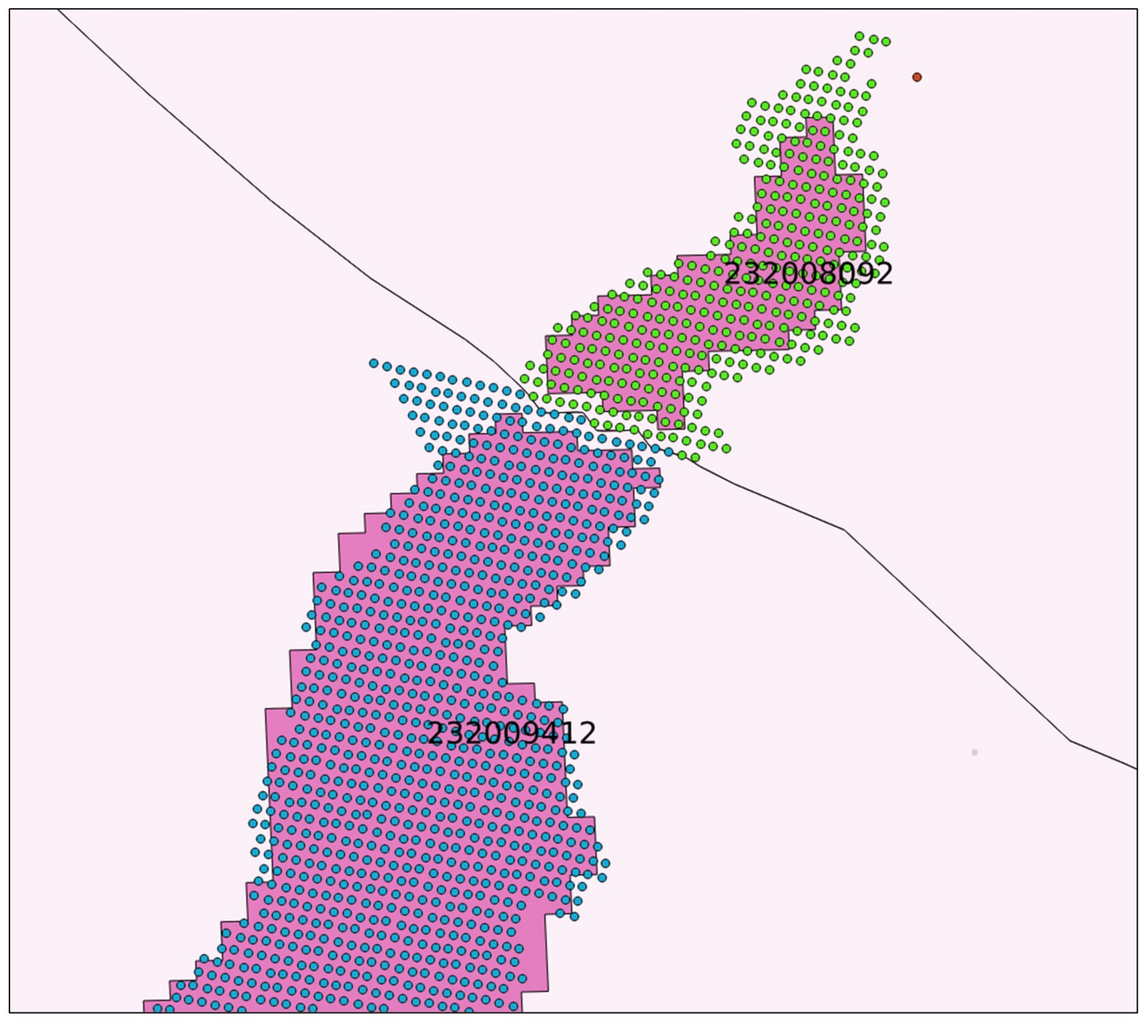
(a)


232009412;232008092
232123812

(b)


232008092
232009412
232142722
232123812
232132172

(c)


232008092
232009412