https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Open Access Earth System Science Data Discussions

Global daily satellite-derived lake surface water quality dataset (2002-2022)

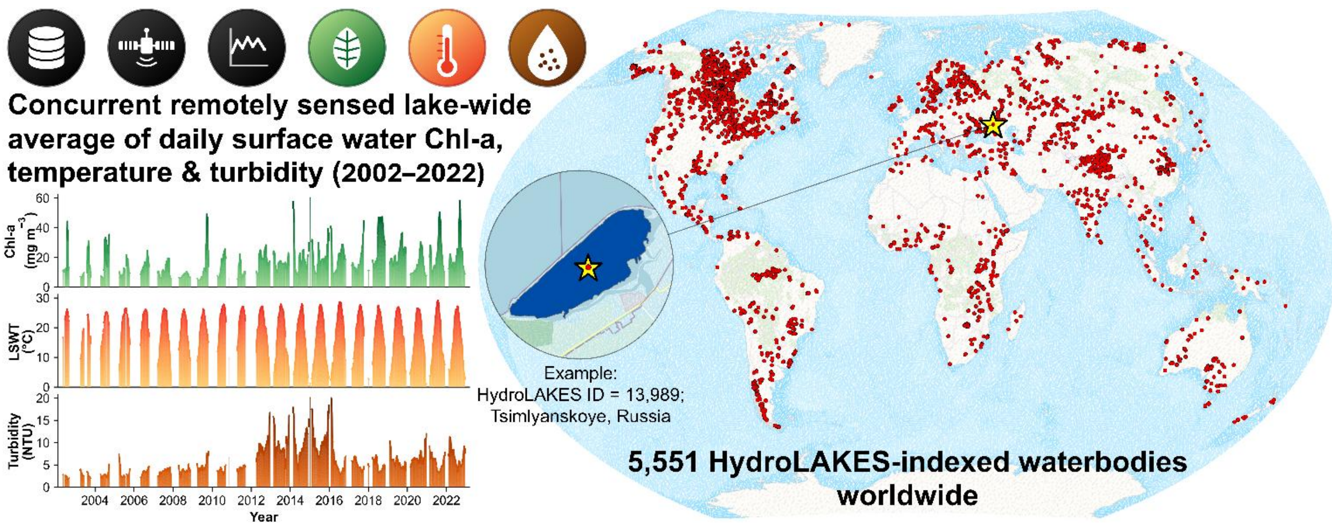
Ali Reza Shahvaran[1,2,3], Roohollah Noori[4,5,6,*], Mohammed Basheer[1], Kaveh Madani[4], Chenxi Mi[5,6], Kun Shi[5,6], YongQiang Zhou[5,6], R. Iestyn Woolway[7], YunLin Zhang[5,6], BoQiang Qin[5,6], Amir AghaKouchak[8], and Michelle T.H. van Vliet[9]

[1]Water Resources and Hydrology (WRH) Research Group, Department of Civil & Mineral Engineering, University of Toronto, Toronto, Ontario M5S 1A4, Canada.
[2]Ecohydrology Research Group, Department of Earth and Environmental Sciences, University of Waterloo, Ontario N2L 3G1, Canada.
[3]Remote Sensing of Environmental Change (ReSEC) Research Group, Department of Geography and Environmental Studies, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada.
[4]United Nations University Institute for Water, Environment and Health (UNU-INWEH), Richmond Hill, Ontario L4B 3P4, Canada.
[5]State Key Laboratory of Lake and Watershed Science for Water Security, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing 210008, China.
[6]University of Chinese Academy of Sciences, Nanjing (UCASNJ), Nanjing 211135, China.
[7]School of Ocean Sciences, Bangor University, Menai Bridge, Anglesey LL59 5AB, UK.
[8]Department of Civil and Environmental Engineering, University of California, Irvine, CA, USA.
[9]Department of Physical Geography, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands.

**Correspondence to*: Roohollah Noori (roohollah.noori@unu.edu)

Abstract. Satellite remote sensing is critical for monitoring global inland water quality, yet the complexity and volume of raw data archives often hinder their widespread application. Here, we present a standardized, accessible dataset of *concurrent* daily water quality statistics for 5,551 lakes worldwide from 2002 to 2022. Derived from the European Space Agency's Lakes Climate Change Initiative (ESA Lakes_cci), this dataset provides lake-wide statistics—including estimates of mean, ranges and extremes, and spatial coverage—for three relevant water quality parameters: chlorophyll-*a* (Chl-*a*), lake surface water temperature (LSWT), and turbidity. Distinctively, these records are harmonized to retain only days on which all three parameters are simultaneously valid, facilitating robust multivariate analysis. The data were extracted using HydroLAKES polygon boundaries and introduce a novel multi-tiered quality assurance flagging scheme to identify spatial inhomogeneities and temporal anomalies, providing a new analytical dimension. By transforming massive, multidimensional NetCDF files into easily-accessible tabular formats, this resource eliminates steep computational barriers, supporting researchers and other end users in conducting macro-scale limnological studies, training machine learning models, and assessing climate change impacts on lake water quality and aquatic ecosystems. The dataset and developed codes are openly accessible at https://doi.org/10.5281/zenodo.20347115 (Shahvaran *et al.*, 2026).

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026



1 Introduction

Lakes and reservoirs are vital components of the global hydrological cycle, providing essential freshwater resources, supporting biodiversity, and regulating regional climates (Tranvik *et al.*, 2009; Woolway *et al.*, 2020). Because their physical and biogeochemical properties respond rapidly to atmospheric and radiative forcing (Grant *et al.*, 2021; O'Reilly *et al.*, 2015; Persaud *et al.*, 2025), lakes act as critical sentinels of climate change (Woolway *et al.*, 2022b, 2020). Continuous monitoring of lake water quality is necessary for managing aquatic ecosystems (Carvalho *et al.*, 2019; Ho *et al.*, 2019), tracking eutrophication (Ho *et al.*, 2019), and supporting decision-making for global/local initiatives such as the United Nations Sustainable Development Goals (Chapman and Sullivan, 2022; Kavvada *et al.*, 2020; Saravani *et al.*, 2024). In recent years, satellite remote sensing has become a fundamental tool for large-scale limnological monitoring, providing long-term observations of essential climate variables such as chlorophyll-*a* (Chl-*a*) concentration (*i.e.*, a proxy of eutrophication and algal blooms (Carlson, 1977)), lake surface water temperature (LSWT), and turbidity (Calamita *et al.*, 2024; Carrea *et al.*, 2023). Among the most robust sources of these observations is the European Space Agency's Lakes Climate Change Initiative (ESA Lakes_cci) (Carrea *et al.*, 2023), which offers a homogenized, multi-sensor global archive of satellite imagery and processed water quality products (Carrea *et al.*, 2023). Yet, despite their scientific value, accessing and utilizing these remote sensing data present substantial technical barriers, particularly for non-technical end-users such as applied limnologists, ecologists, and water resource managers (Schaeffer *et al.*, 2013). First, the sheer volume of ESA Lakes_cci's global daily archive (approximately 768 GB) requires specialized knowledge of command-line retrieval protocols (*e.g.*, wget or API scripting) and significant computational storage. Second, the data are distributed in multidimensional file formats (*i.e.*, NetCDF). While standard in the climate sciences, these file formats often pose steep learning curves for researchers and other end-users accustomed to tabular data and can present challenges in conventional geographic information system (GIS) software,

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

frequently leading to spatial misalignment and inaccurate map projections. Furthermore, these source files contain numerous intermediate bands such as remote sensing reflectance; while crucial for algorithm development and atmospheric correction, these bands are largely unnecessary for researchers seeking only the final derived water quality parameters.

Beyond format and storage constraints, transforming raw global pixel arrays into usable, lake-specific time-series data requires computationally intensive geospatial processing (Chen *et al.*, 2022; Sagan *et al.*, 2020). End-users must independently mask, intersect, and aggregate georeferenced grid cells against complex geometric boundaries (*e.g.*, lake polygons) to extract relevant statistics for their specific water bodies of interest. For researchers conducting multivariate modeling or time-series analysis (Ho *et al.*, 2019; Hou *et al.*, 2022; Woolway *et al.*, 2024), or training machine learning or deep learning algorithms (Duan *et al.*, 2025; Mahdian *et al.*, 2026; Shen, 2018), aligning concurrent observations of Chl-*a*, LSWT, and turbidity is a major operational hurdle that demands complex temporal harmonization and extensive data cleaning (Woolway *et al.*, 2021b).

To bridge these gaps, we present a highly accessible, tabular dataset providing concurrent daily lake-wide water quality statistics (minimum, maximum, mean, standard deviation, valid pixel count, and spatial coverage ratio) for 5,551 lakes globally, spanning the period from 2002 to 2022. By integrating the ESA Lakes_cci Level-3S (L3S) global remote sensing products (Carrea *et al.*, 2024, 2023) with explicit lake boundaries from the widely used HydroLAKES v1.0 database (Messager *et al.*, 2016), this dataset significantly expands the number of monitored water bodies beyond the ~2,000 lakes provided in the standard ESA predefined time-series (Carrea *et al.*, 2023), and eliminates the need for extensive geospatial data extraction and processing by the end-user. We systematically filtered the raw archive to isolate only the essential water quality parameters, retaining days where Chl-*a*, LSWT, and turbidity statistics were simultaneously valid. This strict temporal harmonization was deliberately chosen to bypass the complex data-cleaning and gap-handling steps typically required for multivariate time-series modeling or training machine learning algorithms, providing end-users with a robust, uninterrupted analytical baseline. Packaged in standard, wide-format comma-separated values (CSV) files and accompanied by comprehensive quality assurance/quality control (QA/QC) flags, this dataset is designed to bypass the technical bottlenecks of raw satellite data processing. Crucially, the development of this multi-tiered QA/QC flagging framework represents the primary scientific contribution of this dataset, empowering users to dynamically isolate complex spatial artifacts from genuine, extreme limnological phenomena. Overall, the presented dataset offers immediate reuse value for a wide range of applications, including macro-scale lake water quality assessments (Caroni *et al.*, 2025; Hanly *et al.*, 2025; Topp *et al.*, 2020), climate change impact studies (Calamita *et al.*, 2024; Woolway et al., 2020), and data-driven ecological modeling (Meyer *et al.*, 2024; Read *et al.*, 2019).

2 Methods

2.1 Input data description

Global daily lake water quality products were obtained from the ESA Lakes_cci version 2.1 dataset (Carrea et al., 2024), specifically the L3S daily merged products on a common 1/120° (= $0.008\overline{3}$°; ~1 km resolution at the equator) WGS84 grid

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

covering 1992–2022 (though our extracted dataset spans 2002–2022 due to the absence of concurrent multiparameter observations prior to the Envisat/MERIS mission).

To ensure long-term consistency, the ESA Lakes_cci developers derived Chl-*a* and turbidity from optical sensors (Envisat/MERIS, Sentinel-3A/B OLCI, and Aqua/MODIS) where water pixels were identified using the Idepix algorithm (Wevers *et al.*, 2022), atmospherically corrected via POLYMER v4.13 (Steinmetz *et al.*, 2011), and processed through a fuzzy optical water type (OWT) framework (Moore *et al.*, 2014; Spyrakos *et al.*, 2018). This framework dynamically blends the outputs of specific empirical (*e.g.*, blue-green band ratios for clear water) and semi-analytical (*e.g.*, near-infrared/red models for turbid water) algorithms to minimize discontinuities across diverse water types (Carrea *et al.*, 2023; Neil *et al.*, 2019). These indices and retrieval frameworks are widely established in the literature and have demonstrated high reliability in capturing complex aquatic biogeochemical dynamics across diverse geographic scales (Pahlevan *et al.*, 2022; Shahvaran *et al.*, 2025, 2024). On the other hand, LSWT was retrieved from thermal infrared sensors (Envisat/AATSR, MODIS, AVHRR, and Sentinel-3 SLSTR) using an optimal estimation (OE) scheme inverted against the RTTOV radiative transfer model (Saunders *et al.*, 2018), followed by cross-sensor harmonization to adjust for platform differences.

2.2 Data extraction and processing

From this global archive, we accessed the complete dataset, totaling approximately 768 GB of daily merged NetCDF-4 files, downloaded via the Centre for Environmental Data Analysis (CEDA) archive (Carrea *et al.*, 2024) using wget. These files contain essential climate variables, from which we extracted three primary variables: Chl-*a* concentration (*chla_mean* in mg m^{-3}), LSWT (*lake_surface_water_temperature* in Kelvin (K)), and turbidity (*turbidity_mean* in Nephelometric Turbidity Units (NTU)). All processing was conducted using Python (v3.10), utilizing the *xarray* (Hoyer and Hamman, 2017), *rasterio* (Gillies, 2013), *numpy* (Harris *et al.*, 2020), *pandas* (McKinney, 2011, 2010), and *geopandas* (Jordahl *et al.*, 2021) libraries for high-performance array handling and tabular data structuring.

The processing pipeline began by converting the raw NetCDF files into georeferenced TIFFs (EPSG:4326) to ensure consistent spatial alignment; this step involved vertically mirroring data arrays to correct coordinate origin discrepancies and masking invalid/missing pixels defined by the source *_FillValue* as *NaNs*. To extract lake-specific time series, we utilized the HydroLAKES v1.0 polygon dataset (Messager *et al.*, 2016), which indexes ~1.43 million geospatially explicit distinct water bodies worldwide. By utilizing this comprehensive global framework rather than relying solely on the native ESA Lakes_cci predefined mask (which contains only ~2,000 lakes), we significantly expanded the number of monitored water bodies while ensuring end-users can seamlessly cross-reference our dataset with essential physical lake attributes (*e.g.*, mean depth, volume, and watershed area) provided in the HydroLAKES database. Crucially, to preserve geometric fidelity and capture vital nearshore water quality information, and ensure smaller water bodies were not entirely excluded, we employed a raster-based masking approach that included all pixels intersecting the HydroLAKES polygon boundaries (*all_touched=True*), rather than only those strictly contained within. While the source ESA Lakes_cci archive is already rigorously water-masked, prioritizing

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

shoreline pixel inclusion in this manner may still increase susceptibility to optical adjacency effects or sub-pixel land contamination, particularly for narrow or morphologically complex lakes.

Daily lake-wide statistics (minimum, maximum, mean, standard deviation, valid pixel count, and spatial coverage ratio) were then computed via vectorized zonal statistics (using *numpy.bincount* for memory management) across every unique HydroLAKES ID (*Hylak_id*). During this extraction phase, LSWT values were converted from the source unit of Kelvin to degrees Celsius (°C = K − 273.15), and all floating-point statistics were rounded to one decimal place to standardize the final tabular outputs. While aggregating data to the lake-wide level inherently compresses fine-scale spatial patterns (*e.g.*, the gradient between eutrophic nearshore zones and oligotrophic open waters in large lakes), the inclusion of spatial minimum, maximum, and standard deviation statistics for each daily observation explicitly allows users to quantify this intra-lake heterogeneity. Ultimately, the vast majority of the initial ~1.43 million water bodies were excluded because their surface areas were too small to be resolved by the 1 km satellite grid, or they lacked sufficient simultaneous valid retrievals for all three parameters, resulting in the final robust subset of 5,551 lakes. The subsequent filtering, temporal transposition, and application of QA/QC flags (Phases 3–5 of our workflow; Fig. 1) are detailed comprehensively in the next section.

2.3 Technical validation

2.3.1 Source data validation

The reliability of the ESA Lakes_cci variables was previously established by the dataset developers through formal validation under the ESA CCI Product Validation Plan (ESA CCI, 2023), utilizing direct satellite-field matchups and inter-product consistency checks. For LSWT, their published validation against 66,407 matchups (constructed within 3 km and 3 h) across 79 lakes demonstrated high accuracy for quality levels (QL) 4 and 5, yielding small median biases (−0.22 K for QL4; −0.15 K for QL5) and sub-Kelvin robust standard deviations (0.756 K and 0.504 K, respectively) (Carrea *et al.*, 2023); comprehensive validation figures and discussions regarding skin-temperature depth offsets are detailed extensively in their original data paper (Carrea *et al.*, 2023). For biogeochemical variables, the dataset developers validated reflectance data across 71 lakes against the LIMNADES database (Spyrakos *et al.*, 2018), showing strong bandwise agreement (*e.g.*, $r \approx 0.86$ for MERIS at 560 nm), while their Chl-*a* and turbidity retrievals employed an OWT-weighted blending strategy to mitigate optical complexity errors (Moore *et al.*, 2014); notably, while this strategy ensures global applicability, they noted that their evaluation is not fully independent as the *in-situ* sources also supported algorithm tuning. Given this extensive pre-existing validation of the source products by the ESA CCI, we therefore did not perform independent *in-situ* matchups, focusing our technical validation entirely on the rigorous data filtering and harmonization protocols detailed below.

2.3.2 Data filtering and harmonization

Building upon this established global validation on the original data, we performed the extraction of daily lake-wide statistics for Chl-*a*, LSWT, and turbidity clipped to HydroLAKES boundaries, and conducted a series of post-processing QC checks to

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

ensure data integrity and temporal synchronicity. We first harmonized the dataset by strictly retaining only those daily records that possessed simultaneous valid statistics for all three water quality parameters, discarding any incomplete temporal snapshots (*e.g.*, days where thermal data was retrieved but optical parameters were obscured by clouds or sunglint, which constitutes a substantial fraction of the original raw archive). Subsequently, we applied logical and physical validation filters to the remaining data: entries containing physically impossible negative values (*i.e.*, for Chl-*a* or turbidity) were dropped and the internal consistency of the statistics was verified by enforcing the mathematical hierarchy $Min \leq Mean \leq Max$. Finally, we ensured data reliability by requiring a minimum valid pixel count of ≥ 1 for all parameters, thereby eliminating artifacts derived from empty or null aggregations.

2.3.3 Quality assurance flagging scheme

To assist future users in assessing the reliability of individual data points without removing potentially valuable extreme events, a multi-tiered flagging system was implemented for all three water quality parameters. Rather than excluding data, we appended a quality assurance flag (**_qa_flag* columns) and a spatial coverage ratio (**_coverage_ratio* columns) to each daily time step. The flagging scheme assigns a concatenated string of integer codes (*e.g.*, "134" indicating the presence of flags 1, 3, and 4, but not 2) to any observation violating specific statistical or physical criteria, where "0" indicates high-quality data passing all checks. Importantly, while these thresholds are grounded in standard statistical confidence intervals and established oceanographic guidelines, they are designed strictly as cautionary indicators rather than definitive classifiers; distinguishing a genuine extreme event from a sensor artifact ultimately requires context-specific user discretion. The flags were defined as follows:

Flag 1: spatial inhomogeneity (hot pixel artifacts). This flag identifies dates where the maximum pixel value within the lake is disproportionately high compared to the lake-wide mean, suggesting sensor glint, cloud edge effects, or processing artifacts rather than a natural gradient. A flag is triggered if the spatial *Z*-score of the maximum value exceeds 4 as depicted in Eq. (1):

$$Z_{spatial} = \frac{Max_t - \mu_t}{\sigma_t} > 4 \tag{1}$$

where Max_t, μ_t, and σ_t represent the maximum, mean, and standard deviation of pixels within the lake for day t, respectively. A threshold of 4σ was selected to target extreme outliers (> 99.99% confidence in a normal distribution) while accommodating the natural non-Gaussian variability often observed in aquatic remote sensing (Campbell, 1995).

Flag 2: temporal anomaly (climatological outliers). To detect values that deviate significantly from the lake's long-term behavior, we calculated a temporal *Z*-score based on the lake's entire processed time series. A flag is assigned if the daily mean deviates by more than three standard deviations from the long-term mean as shown in Eq. (2):

$$Z_{temporal} = \frac{|\mu_t - \mu_{global}|}{\sigma_{global}} > 3 \tag{2}$$

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

where μ_{global} and σ_{global} are the mean and standard deviation of the parameter over the full study period, respectively. This flag highlights potential extreme events (*e.g.*, severe algal blooms) or atmospheric correction failures (*e.g.*, unmasked thin clouds or heavy aerosols artificially inflating retrievals), allowing users to exercise caution with statistically rare observations.

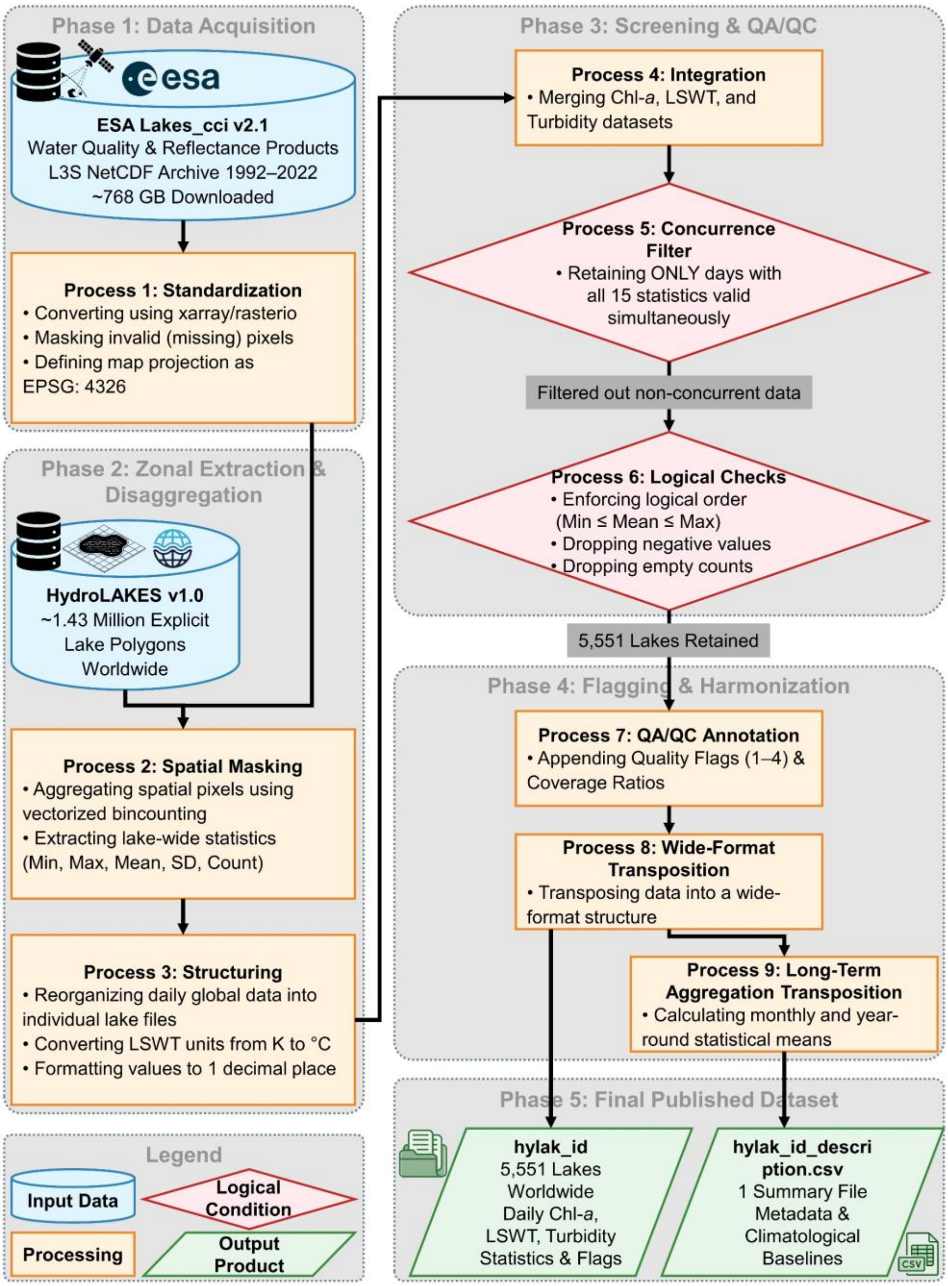


Figure 1: Schematic workflow of the methodology, from input data acquisition to the final output dataset. Abbreviations: ESA (European Space Agency); Lakes_cci (Lakes Climate Change Initiative); L3S (Level-3S / Super-collated); NetCDF (network common data form); EPSG (European Petroleum Survey Group); SD (Standard Deviation); LSWT (Lake Surface Water Temperature); K (Kelvin); QA/QC (Quality Assurance/Quality Control); CSV (Comma-Separated Values).

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Flag 3: rapid rate of change (temporal consistency). Following the quality assurance of real-time oceanographic data (QARTOD) Gross Rate of Change guidelines (Integrated Ocean Observing System, 2014), we evaluated the physical plausibility of changes between consecutive observations. A flag is triggered if the absolute difference between two consecutive data points exceeds three times the long-term standard deviation ($3\sigma_{global}$), provided the time gap between observations is ≤ 30 days. This constraint ensures that natural seasonal shifts occurring over long data gaps (*e.g.*, winter months) are not falsely flagged, focusing detection on physically improbable short-term spikes.

Flag 4: low spatial representativeness. To quantify how well the available satellite pixels represent the entire lake surface, we calculated a coverage ratio for each time step. This ratio is defined as the number of valid pixels divided by the theoretical maximum pixel count for that lake (derived from HydroLAKES polygon geometry and harmonized with ESA Lakes_cci's native spatial resolution). Observations are flagged if their spatial coverage ratio falls below 10% (or, as a fallback for water bodies where the theoretical maximum pixel count is undefined, if fewer than 3 valid pixels are retrieved, establishing a basic statistical floor to mitigate single-pixel sensor noise). This relative 10% threshold ensures that even in massive lakes where absolute pixel counts remain high, observations restricted to a single, isolated clear-sky region are appropriately flagged. This indicates that the derived statistics may be biased by shoreline adjacency effects or insufficient spatial sampling (Jiang *et al.*, 2023; Kiselev *et al.*, 2015).

2.4 Processing hardware

This extensive geospatial processing, including generating daily statistics for the three water quality parameters (*i.e.*, Chl-*a*, LSWT, and turbidity) across ~1.43 million lakes over the full 31-year raw data period (1992–2022, prior to applying our temporal filters), requires over one month of continuous computational runtime on a dedicated high-performance workstation (System specifications: a 24-core CPU (Intel Core Ultra 9 285K), 128 GB of RAM, and a dedicated GPU with 16 GB of VRAM (GeForce RTX 4080 SUPER)).

3 Results and discussion

3.1 Dataset structure and metadata

The generated dataset is publicly available and hosted on the Zenodo repository under the DOI: https://doi.org/10.5281/zenodo.20347115. To facilitate ease of access, targeted data retrieval, and methodological reproducibility, the repository is organized into three primary components: a summary metadata file (*hylak_id_description.csv*), a compressed directory of the temporal data (*hylak_id.zip*), and a compressed directory of the processing scripts (*code.zip*). At the global, lake-wide level, the standalone *hylak_id_description.csv* file provides a general overview and climatological baselines for all 5,551 processed lakes. To prevent redundancy, we do not duplicate the attribute tables inherent to the source

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

HydroLAKES v1.0 database. However, this summary file includes the central latitude and longitude coordinates for each water body which allows end-users to easily import the CSV into any standard GIS software or programming interface, enabling them to map the dataset spatially and rapidly verify if their specific lake of interest is included in the processed archive. A
220 complete metadata description of the variables within this summary file is provided in Table 1.

Table 1: Metadata description of the columns for the hylak_id_description.csv summary file. This 47-column wide-format file provides spatial attributes (central coordinates, country), Köppen-Geiger Climate Classifications (KGCC) zones, total valid observation counts, and long-term monthly and year-round parameter averages. The first row serves as the header, with each subsequent row representing one of the 5,551 processed lakes.

Column name	Description	Unit	Data type
hylak_id	Unique numerical identifier for the lake, corresponding to the HydroLAKES v1.0 database	Dimensionless	Integer
central_lat_wgs84_dd	Central latitude coordinate of the lake	Decimal Degrees	Float
central_lon_wgs84_dd	Central longitude coordinate of the lake	Decimal Degrees	Float
country	Primary country in which the lake is geographically located	Dimensionless	String
valid_daily_data_count	Total number of valid, concurrent daily observations available for the lake from 2002–2022	Dimensionless	Integer
kgcc_value	Numeric code representing the Köppen-Geiger Climate Classification (KGCC)	Dimensionless	Integer
kgcc_abbreviation	Standard alphabetical abbreviation of the assigned KGCC	Dimensionless	String
kgcc_full_name	Full descriptive name of the assigned KGCC	Dimensionless	String
chla_mean_mean_[month]	Long-term concurrent average chlorophyll-*a* (Chl-*a*) concentration for a specific month (12 distinct columns representing monthly climatologies from *_jan to *_dec)	mg m^{-3}	Float
chla_mean_mean_year-round	Overall long-term mean Chl-*a* concentration, calculated as the unweighted arithmetic average of all concurrently valid daily observations available for the lake.	mg m^{-3}	Float
lake_surface_water_temperature_mean_[month]	Long-term concurrent average lake surface water temperature (LSWT) for a	°C	Float

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Column name	Description	Unit	Data type
	specific month (12 distinct columns representing monthly climatologies from *_jan to *_dec)		
lake_surface_water_temperature_mean_year-round	Overall long-term mean LSWT, calculated as the unweighted arithmetic average of all concurrently valid daily observations available for the lake.	°C	Float
turbidity_mean_mean_[month]	Long-term concurrent average turbidity for a specific month (12 distinct columns representing monthly climatologies from *_jan to *_dec)	NTU	Float
turbidity_mean_mean_year-round	Overall long-term mean turbidity, calculated as the unweighted arithmetic average of all concurrently valid daily observations available for the lake.	NTU	Float

At the individual lake level, the *hylak_id.zip* file contains 5,551 individual CSV files, each corresponding to a single lake. To ensure seamless cross-referencing with global hydrological frameworks, each file is named according to its unique identifier from the HydroLAKES database (*Hylak_id* column). Every file contains the harmonized, daily time-series records of lake-wide statistics for Chl-*a*, LSWT, and turbidity spanning the 2002–2022 period. These files also house the multi-tiered QA/QC tracking flags—detailed extensively in the Technical Validation section—allowing users to filter the time-series data based on their specific research tolerances for spatial inhomogeneities or temporal anomalies. The structural metadata and exact column definitions for these individual time-series files are outlined in Table 2. To maintain direct traceability to the source ESA Lakes_cci NetCDF files, the original variable nomenclature (*e.g.*, *chla_mean*, representing the daily 1 km pixel average) was preserved; consequently, headers such as *chla_mean_mean* indicate the lake-wide spatial average of those individual pixel-level means.

Finally, the *code.zip* file contains the suite of Python scripts utilized throughout the methodology. Providing these scripts ensures full transparency of the extraction, spatial masking, and data aggregation pipelines utilized to transform the raw ESA Lakes_cci NetCDF files into the final processed tabular formats.

Table 2: Metadata description of the columns for the individual lake time-series files. Data are provided as 5,551 individual CSV files within the *hylak_id.zip* file, named via the *{hylak_id}.csv* convention. Each file shares an identical 22-column wide-format structure, with the first row as the header and each subsequent row representing a distinct daily satellite observation based on the ESA Lakes_cci's image acquisition dates.

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Column name	Description	Unit	Data type
date	Date of the satellite observation	YYYY-MM-DD	String
chla_mean_min	Minimum chlorophyll-*a* (Chl-*a*) concentration within the lake	mg m^{-3}	Float
chla_mean_max	Maximum Chl-*a* concentration within the lake	mg m^{-3}	Float
chla_mean_mean	Mean Chl-*a* concentration within the lake	mg m^{-3}	Float
chla_mean_sd	Standard deviation of Chl-*a* concentration within the lake	mg m^{-3}	Float
chla_mean_count	Number of valid pixels used for Chl-*a* aggregation	Dimensionless	Integer
chla_mean_coverage_ratio	Ratio of valid Chl-*a* pixels to the theoretical maximum pixel count	Ratio % (0 to 100)	Integer
chla_mean_qa_flag	Quality assurance flag indicating potential anomalies for Chl-*a* (see Technical Validation section for more details)	Dimensionless	Integer
lake_surface_water_temperature_min	Minimum lake surface water temperature (LSWT)	°C	Float
lake_surface_water_temperature_max	Maximum LSWT	°C	Float
lake_surface_water_temperature_mean	Mean LSWT	°C	Float
lake_surface_water_temperature_sd	Standard deviation of LSWT	°C	Float
lake_surface_water_temperature_count	Number of valid pixels used for LSWT aggregation	Dimensionless	Integer
lake_surface_water_temperature_coverage_ratio	Ratio of valid LSWT pixels to the theoretical maximum pixel count	Ratio % (0 to 100)	Integer
lake_surface_water_temperature_qa_flag	Quality assurance flag indicating potential anomalies for LSWT (see Technical Validation section for more details)	Dimensionless	Integer
turbidity_mean_min	Minimum turbidity within the lake	NTU	Float
turbidity_mean_max	Maximum turbidity within the lake	NTU	Float
turbidity_mean_mean	Mean turbidity within the lake	NTU	Float
turbidity_mean_sd	Standard deviation of turbidity within the lake	NTU	Float
turbidity_mean_count	Number of valid pixels used for turbidity aggregation	Dimensionless	Integer

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Column name	Description	Unit	Data type
turbidity_mean_coverage_ratio	Ratio of valid turbidity pixels to the theoretical maximum pixel count	Ratio % (0 to 100)	Integer
turbidity_mean_qa_flag	Quality assurance flag indicating potential anomalies for turbidity (see Technical Validation section for more details)	Dimensionless	Integer

3.2 Spatial distribution of the extracted dataset

The geographical distribution of the 5,551 processed HydroLAKES exhibits a pronounced concentration in the Northern Hemisphere, reflecting the natural global distribution of inland water bodies. Furthermore, the volume of valid, concurrent daily observations varies significantly by region due to localized climatological constraints on satellite visibility; mid-latitude lakes generally yield the highest data counts (indicated in red in Fig. 2), whereas high-latitude and equatorial lakes yield fewer valid retrievals (indicated in blue to yellow in Fig. 2) due to prolonged winter ice cover, polar darkness, or persistent seasonal cloudiness (Sharma *et al.*, 2019).

3.3 Dataset temporal coverage

To provide users with an inventory of temporal data gaps and sensor coverage, the daily availability of concurrent records is visualized in Fig. 3. Although the source ESA Lakes_cci archive dates back to 1992, our final dataset commences in 2002 due to the absence of simultaneously valid pixels for the three water quality parameters prior to the launch of the Envisat/MERIS optical mission. The density of available observations exhibits a strong seasonal pattern, with summer months yielding the highest proportion of valid data across the 5,551 lakes due to optimal clear-sky conditions and the absence of winter ice cover—considering that most of the lakes are located in the Northern Hemisphere, where summer months denote the warmest period. Most notably, a substantial reduction in simultaneous data availability is evident between April 2012 and May 2016. This interruption is an inherent hardware limitation of the source ESA Lakes_cci dataset; the high-quality biogeochemical parameters (Chl-*a* and turbidity) are primarily derived from the MERIS sensor aboard the Envisat satellite, which ceased operations in April 2012, and the OLCI sensor aboard Sentinel-3, which commenced operations in 2016 (Liu *et al.*, 2024). Direct scans of the raw NetCDF arrays confirm that global concurrent pixel availability during this specific four-year interim drop by roughly 70% compared to the subsequent Sentinel-3 era (averaging ~21,600 concurrent pixels annually between 2012–2015, versus ~62,800 annually between 2017–2022 in our systematic spatial sampling). While the Aqua/MODIS sensor acts as a bridge during this four-year gap, it yields significantly fewer valid high-quality retrievals over complex inland waters, resulting in the observed sparse—but non-zero—concurrent data coverage. Because our extraction methodology enforces a strict harmonizing filter—requiring all three parameters (Chl-*a*, LSWT, and turbidity) to be concurrently valid on any given day—the absence of these primary optical missions during this four-year interval results in the observed severe drop in

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

available harmonized records. Finally, as a minor detail regarding the temporal visualization, the faint, isolated vertical line appearing at the end of February simply represents data acquired on February 29 during leap years.

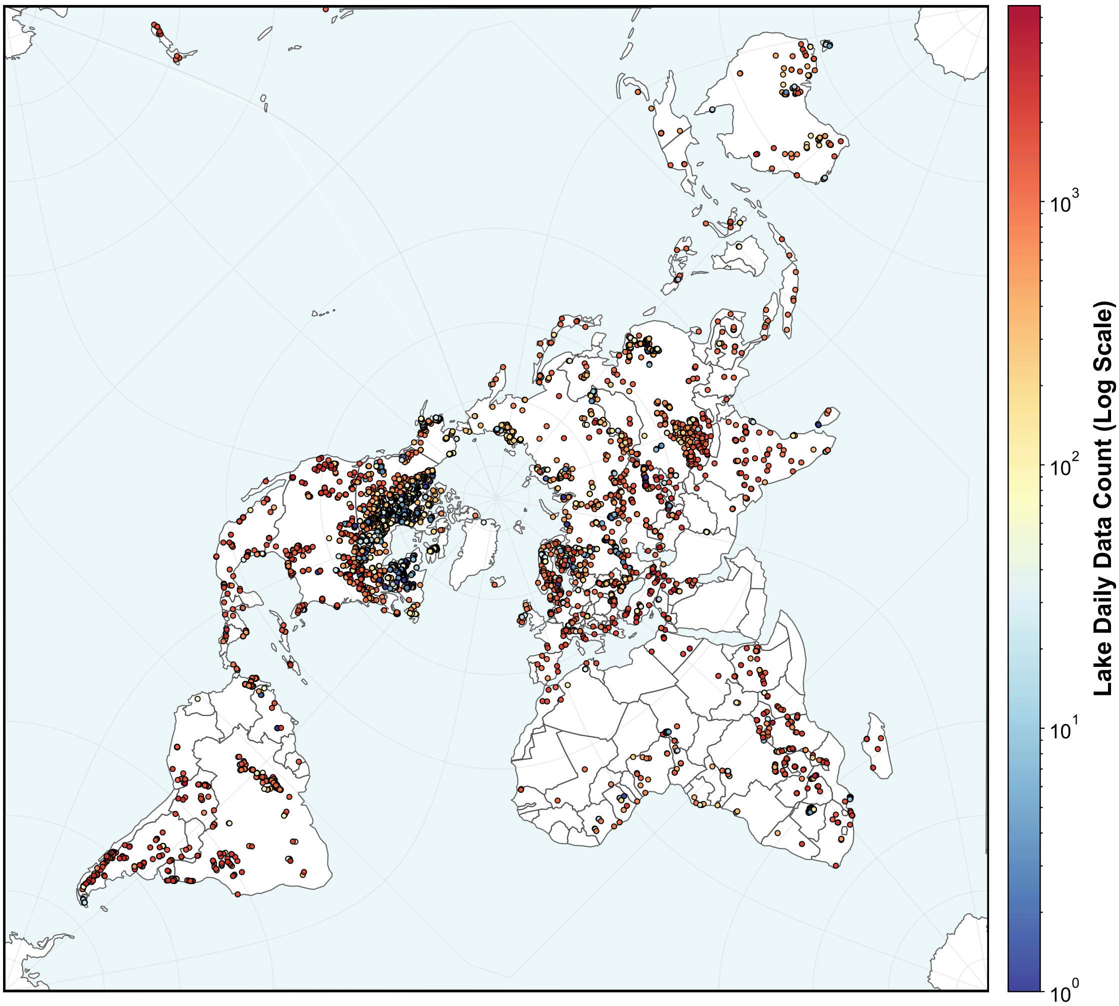


Figure 2: Spatial distribution of the 5,551 processed lakes. The map displays lakes integrated from the ESA Lakes_cci dataset containing valid, simultaneous daily statistics for chlorophyll-*a* (Chl-*a*), lake surface water temperature (LSWT), and turbidity over the 2002–2022 period. The color scale indicates the total count of daily data entries per lake on a logarithmic scale.

3.4 Dataset administrative and climatic coverage

As a descriptive metadata overview, the generated dataset encompasses 5,551 lakes distributed across 124 countries, providing extensive geographic and climatic coverage (Fig. 4). Geographically, the dataset is predominantly concentrated in the Northern Hemisphere, with Canada (33.9%), China (9.8%), the United States (9.7%), and Russia (9.3%) being the only four nations containing more than 500 processed lakes. Climatically, the lakes span 27 distinct Köppen-Geiger Climate Classification (KGCC) zones (Beck et al., 2023). Continental climates (Group D) represent the clear majority, accounting for 50.1% of the

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

monitored water bodies, followed by Arid (Group B; 15.4%), Tropical (Group A; 14.8%), Temperate (Group C; 12.9%), and Polar (Group E; 6.7%) groups. At the specific sub-zone level, the Dfc (continental, no dry season, cold summer) climate is the most heavily represented, containing 1,693 lakes (30.5%), followed by the Dfb (cold, no dry season, warm summer) sub-zone containing 724 lakes (13.0%).

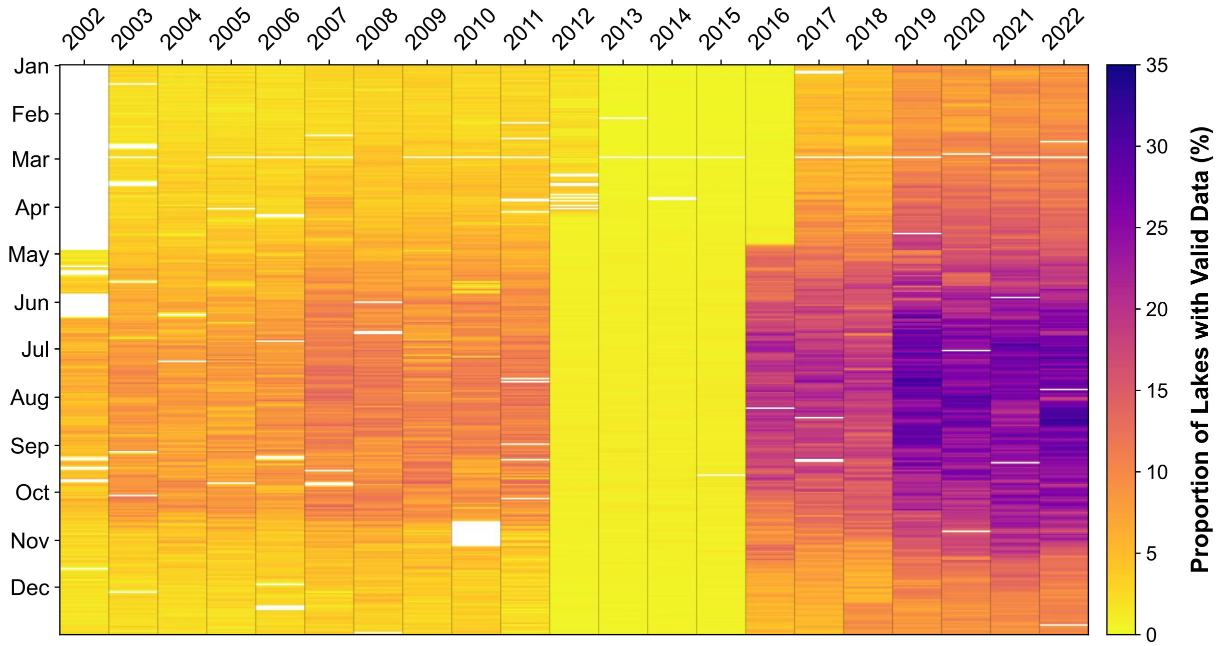


Figure 3: Temporal data availability for the 5,551 processed lakes (2002–2022). The heatmap illustrates the daily proportion of lakes possessing valid, simultaneous records for extracted water quality data. The *y*-axis represents the day of the year (labeled by month), and the *x*-axis represents the acquisition year. Blank (white) pixels denote days with no valid observations.

3.5 Dataset utility and analytical potential

To demonstrate the dataset's utility for both macro-scale limnological modeling and high-resolution temporal analysis, we summarize the long-term, year-round distributions alongside representative time-series profiles. Across the 5,551 processed lakes, the dataset successfully captures global variability and distinct environmental clustering across major climate zones, while preserving fundamental, statistically significant limnological correlations between parameters such as Chl-*a*, LSWT, and turbidity (Fig. 5). These underlying statistical distributions confirm that the harmonized dataset retains the complex environmental signals necessary to support robust multivariate modeling and data-driven applications. Furthermore, the embedded QA/QC flagging scheme adds a new dimension for scientific inquiry, enabling researchers to systematically evaluate statistical anomalies without discarding valuable extreme events. Moreover, to illustrate the dataset's temporal fidelity and seasonal resolution, the locations and spatial extents of a geographically diverse subset of 17 representative lakes spanning six continents are shown in Fig. 6, with their corresponding continuous multi-parameter dynamics presented in Fig. 7. These time series demonstrate the dataset's capacity to consistently capture distinct intra-annual phenology, such as seasonal thermal gradients and algal bloom cycles, alongside inter-annual dynamics over the 2002–2022 observation period. This offers a reliable, ready-to-use foundation for broad-scale ecological modeling and climate impact research.

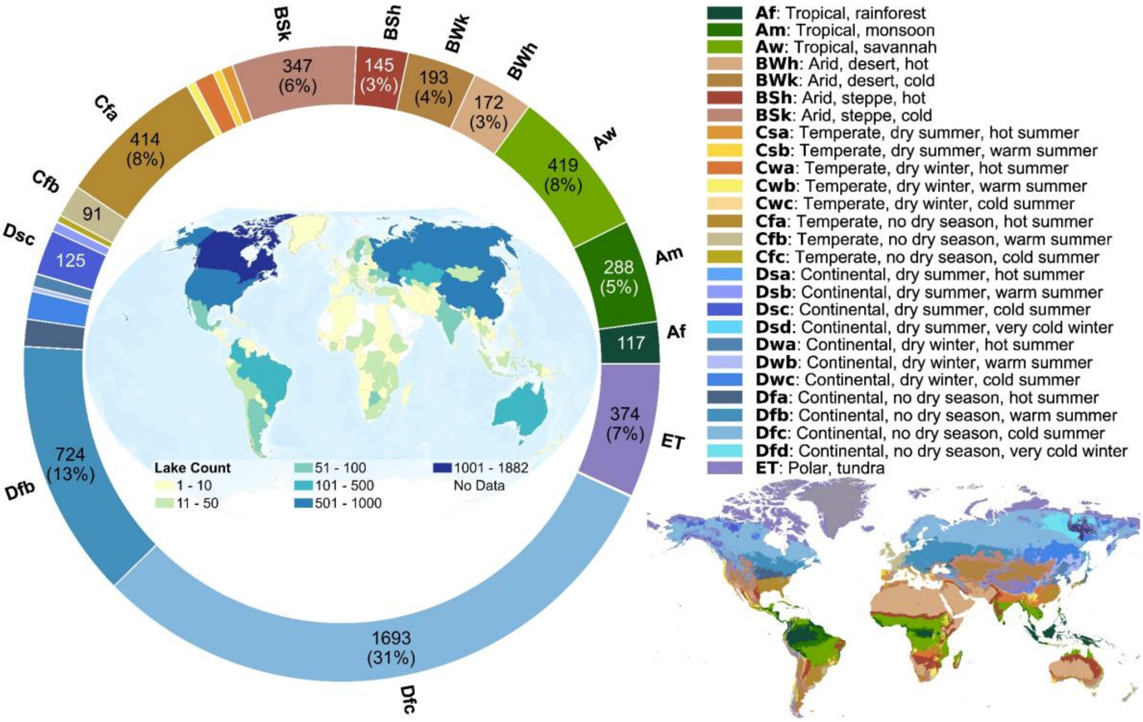


Figure 4: Administrative and climatic distribution of the 5,551 processed lakes. The central choropleth map illustrates the total number of processed lakes per country. The surrounding ring chart quantifies the distribution of these lakes across Köppen-Geiger Climate Classification (KGCC) zones (Beck et al., 2023), with exact counts per slice. The legend provides standard color codes, and full climate descriptions for each represented KGCC zone. A bottom-right reference map displays the global spatial extent of these zones.

3.6 Data use and recommendations

The provided dataset is structured in standardized, wide-format CSV files to maximize accessibility across various programming environments and standard GIS software. Because each lake is stored as an individual file named by its HydroLAKES ID (*{hylak_id}.csv* naming convention), users can easily use iterative loops in Python (*e.g.*, using *pandas.concat*) or R (*e.g.*, *data.table::rbindlist()*) to programmatically open, filter, and merge the 5,551 files into a single dataframe for large-scale analysis. The dataset is particularly suited for training machine learning or deep learning algorithms, conducting macro-scale environmental clustering, and calibrating regional lake water quality or ecological models.

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

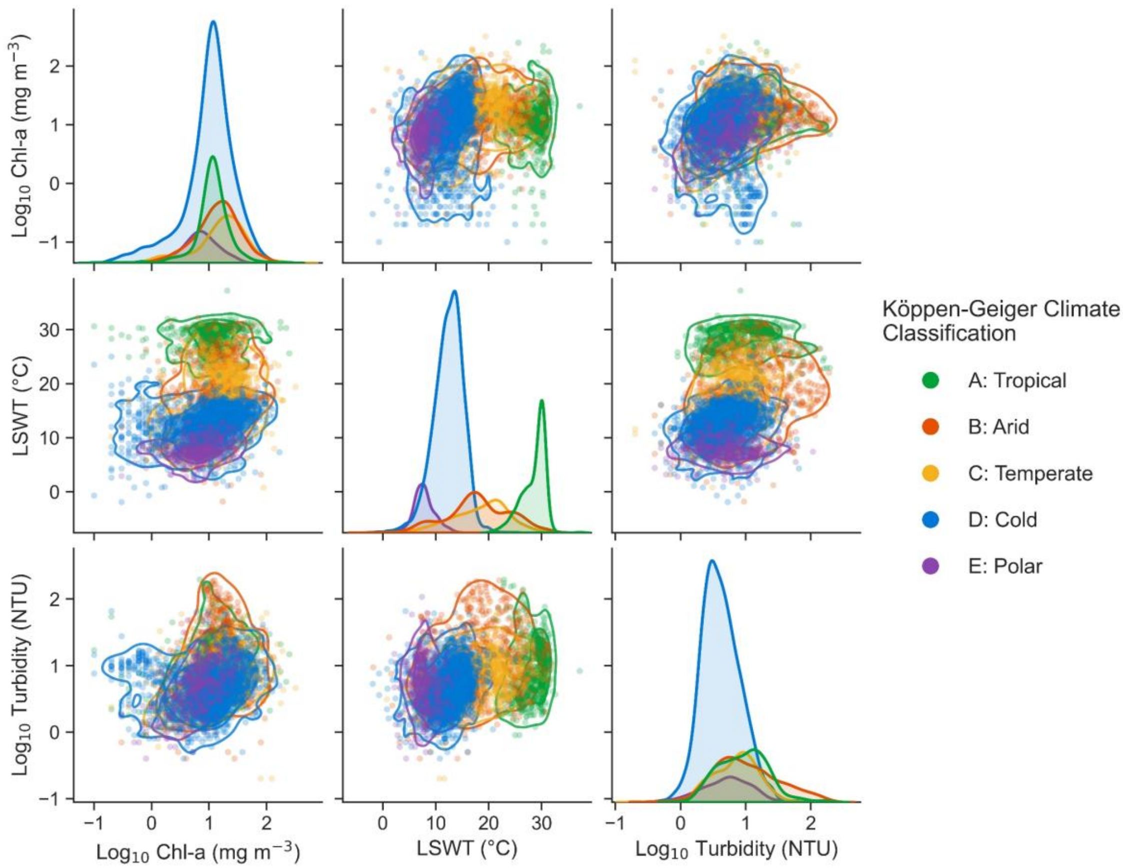


Figure 5: Pairwise relationships and univariate distributions of long-term year-round mean water quality parameters for the 5,551 processed lakes. Diagonal panels display kernel density distributions for Log$_{10}$-transformed chlorophyll-*a* (Chl-*a*) and turbidity, and raw lake surface water temperature (LSWT). Off-diagonal scatter plots illustrate bivariate relationships with overlaid density contours. Individual lake observations are color-coded by their Köppen-Geiger Climate Classification (KGCC) group (Beck *et al.*, 2023).

3.6.1 Handling QA/QC flags and extreme values

A critical technical feature of this dataset is its non-destructive approach to QC. The novel QA/QC flagging scheme serves as the core scientific contribution of this work, moving beyond simple data extraction to provide a robust analytical framework for end-users. Extreme values and statistical outliers have been deliberately retained to preserve natural environmental variability (Seegers *et al.*, 2021), such as short-lived, severe algal blooms (Hallegraeff *et al.*, 2021; Handler *et al.*, 2023; Michalak *et al.*, 2013), intense lake heatwaves (Woolway *et al.*, 2025, 2022a, 2021a), or flood/storm-driven sediment resuspension/plumes (McKinney *et al.*, 2019; Stockwell *et al.*, 2020). Therefore, it is imperative that end-users actively filter the time-series data using the provided *_qa_flag* columns before conducting analyses. For researchers conducting strict, long-term climate trend analysis, we highly recommend subsetting the data to exclude observations triggering Flag 1 (Spatial Inhomogeneity) and Flag 2 (Temporal Anomaly) to ensure pristine climatological baselines. Conversely, researchers

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Open Access Earth System Science Data Discussions

specifically investigating extreme biogeochemical events should carefully evaluate observations triggering Flag 2 or Flag 3 (Rapid Rate of Change) rather than discarding them outright, as these may represent genuine, albeit rare, limnological phenomena (Stockwell *et al.*, 2020).

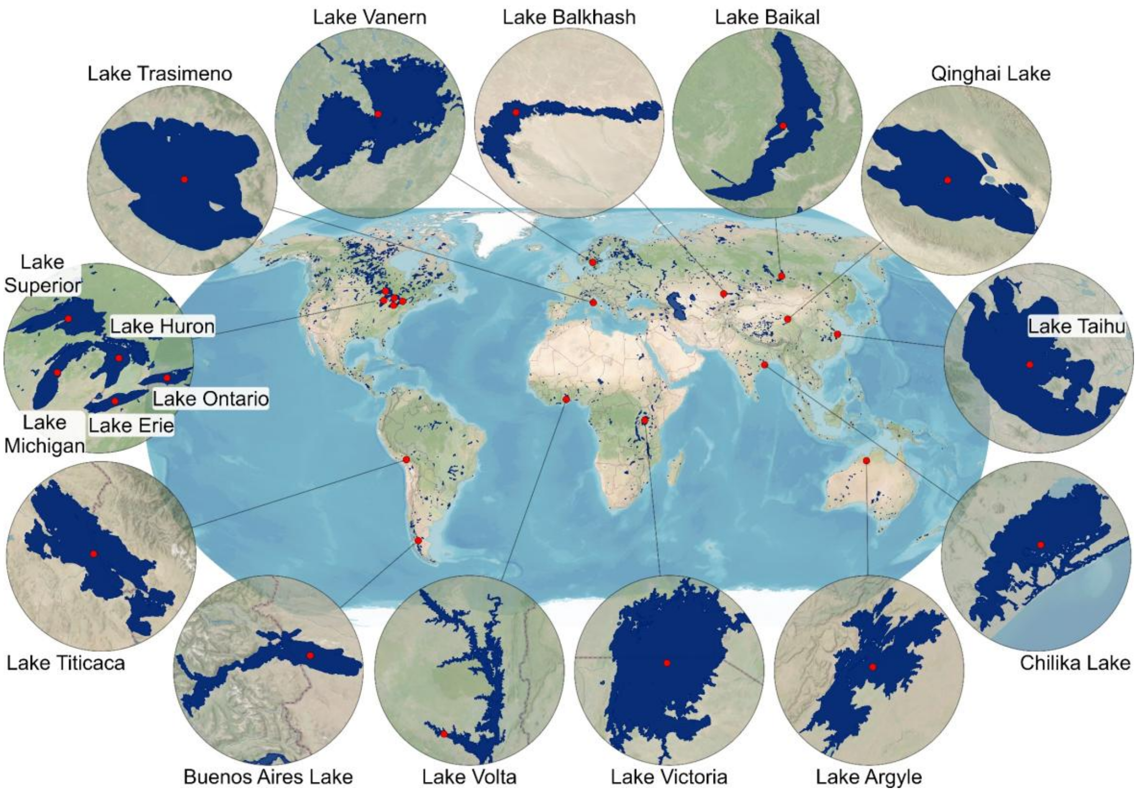


Figure 6: Geographical locations and spatial extents of the 17 representative lakes. The central global map illustrates the distribution of the selected lakes across six continents. Surrounding circular insets provide high-resolution views of the explicit lake boundaries. The corresponding multi-parameter time-series dynamics for these lakes are presented in Fig. 7.

3.6.2 Limitations and temporal gaps

Users must account for inherent temporal observation gaps when designing continuous forecasting models (*e.g.*, autoregressive integrated moving average (ARIMA) or recurrent neural networks) (Che *et al.*, 2018; Khosravi *et al.*, 2025; Su *et al.*, 2024). First, the dataset strictly represents open-water conditions; high-latitude and high-altitude lakes will exhibit systematic data gaps during winter months due to ice-cover masking, while tropical or coastal lakes may experience prolonged data gaps due to persistent seasonal cloudiness. Furthermore, our strict harmonization criterion—requiring simultaneous validity across all three parameters—inherently biases the temporal coverage toward clear-sky conditions where atmospheric corrections succeed, meaning extreme high-turbidity events or severe blooms that cause optical retrieval failures may be underrepresented. Additionally, these observations exclusively reflect satellite-observable near-surface thermal and optical conditions, and therefore should not be interpreted as representing full water-column dynamics or whole-lake processes, particularly in deep,

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Open Access Earth System Science Data Discussions

stratified, or strongly regulated systems. Second, as detailed in Section 3.3, a notable reduction in concurrent data availability exists between early 2012 and early 2016 due to source satellite mission transitions. Researchers conducting continuous, uninterrupted time-series modeling spanning the full 2002–2022 period are advised to apply appropriate temporal interpolation, gap-filling algorithms, or seasonal decomposition techniques to bridge this specific instrumental interim. Furthermore, for robust long-term trend analysis, anomaly detection, or machine learning model training, we strongly recommend employing segmented treatment of the time series or rigorous sensitivity analyses to ensure results are not biased by this transition interval.

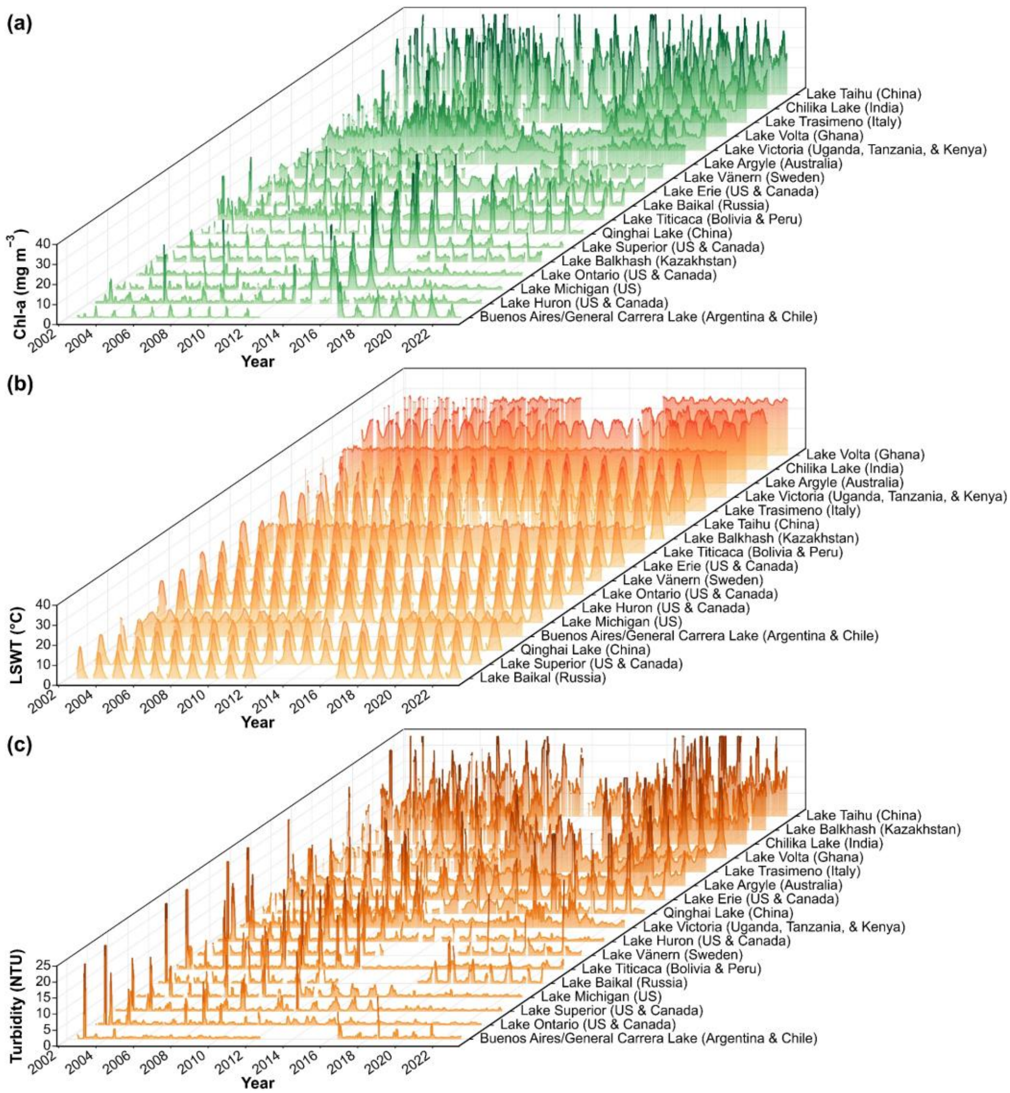


Figure 7: Long-term multi-parameter water quality dynamics across a representative subset of 17 global lakes (2002–2022). Three-dimensional ridgeline plots illustrate continuous temporal variations in mean (a) chlorophyll-*a* (Chl-*a*), (b) lake surface water temperature (LSWT), and (c) turbidity. Daily time-series data were smoothed using a 30-day rolling mean, with lakes ranked in descending order along the depth axis by their long-term overall averages

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

4 Conclusion

This study introduces a globally harmonized, daily lake-wide water quality dataset to provide concurrent records of Chl-*a*, LSWT, and turbidity for 5,551 lakes from 2002-2022. By transforming the extensive ESA Lakes_cci archive into standardized, tabular formats and aligning observations across parameters and time, the dataset removes major technical barriers that have historically limited the use of satellite products in large-scale limnological and climate research. The incorporation of HydroLAKES boundaries substantially expands spatial coverage beyond existing predefined lake masks, while the multi-tiered QA/QC flagging framework—the central methodological contribution of this work—provides an innovative mechanism for identifying spatial artifacts, temporal anomalies, and low-representativeness conditions without discarding potentially meaningful extremes. Collectively, these features offer a robust, transparent, and ready-to-use resource for ecological modeling, machine learning applications, and climate impact assessments.

By making complex remote-sensing products accessible to a broad community of researchers, practitioners, and decision-makers, this dataset supports reproducible science and strengthens global efforts to monitor inland water quality under intensifying environmental change. The open availability of both the data and the processing code ensures full traceability and provides a foundation for future methodological extensions, including the integration of additional sensors, variables, and higher-resolution products.

Data availability

The datasets are openly accessible at https://doi.org/10.5281/zenodo.20347115 (Shahvaran *et al.*, 2026). The repository is structured into three components: a compressed directory (*hylak_id.zip*) containing the daily water quality time-series statistics and QA/QC flags across all 5,551 lakes, a standalone metadata summary file (*hylak_id_description.csv*) providing spatial and climatological attributes, and a compressed directory (*code.zip*) containing the data processing scripts.

Code availability

All custom Python (v3.10) scripts used for data extraction, processing, and QA/QC are openly accessible at https://doi.org/10.5281/zenodo.20347115 (Shahvaran *et al.*, 2026). The processing pipeline relies on standard open-source libraries, primarily requiring *xarray*, *pandas*, *geopandas*, *rasterio*, and *numpy* to be pre-installed in the working environment.

Supplement link

The link to the supplement will be included by Copernicus, if applicable.

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Author contributions

Conceptualization: ARS, RN and MB; Methodology: ARS, RN and MB; Investigation: ARS, RN and MB; Formal analysis: ARS and RN; Visualization: ARS, RN and MTHV; Software: ARS; Validation: ARS, RN, MB, CM, YQZ, RIW, BQ and MTHV; Supervision: RN; Writing—original draft: ARS; Writing—review & editing: RN, MB, KM, CM, KS, YQZ, RIW, YLZ, BQ, AA and MTHV.

Competing interests

Authors declare that they have no competing interests.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

No applicable

Financial support

Not applicable

References

Beck, H.E., McVicar, T.R., Vergopolan, N., Berg, A., Lutsko, N.J., Dufour, A., Zeng, Z., Jiang, X., van Dijk, A.I.J.M., Miralles, D.G., 2023. High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections. Sci. Data 10, 724. https://doi.org/10.1038/s41597-023-02549-6

Calamita, E., Lever, J.J., Albergel, C., Woolway, R.I., Odermatt, D., 2024. Detecting climate-related shifts in lakes: A review of the use of satellite Earth Observation. Limnol. Oceanogr. 69, 723–741. https://doi.org/10.1002/lno.12498

Campbell, J.W., 1995. The lognormal distribution as a model for bio-optical variability in the sea. J. Geophys. Res. Ocean. 100, 13237–13254. https://doi.org/10.1029/95JC00458

Carlson, R.E., 1977. A trophic state index for lakes. Limnol. Oceanogr. 22, 361–369.

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

https://doi.org/10.4319/lo.1977.22.2.0361

Caroni, R., Greife, A.J., Bresciani, M., Giardino, C., Tellina, G., Carrea, L., Liu, X., Simis, S., Albergel, C., Pinardi, M., 2025. Examining global trends of satellite-derived water quality variables in shallow lakes. Remote Sens. Appl. Soc. Environ. 38, 101565. https://doi.org/10.1016/j.rsase.2025.101565

Carrea, L., Crétaux, J.-F., Liu, X., Wu, Y., Calmettes, B., Duguay, C.R., Merchant, C.J., Selmes, N., Simis, S.G.H., Warren, M., Yesou, H., Müller, D., Jiang, D., Embury, O., Bergé-Nguyen, M., Albergel, C., 2023. Satellite-derived multivariate world-wide lake physical variable timeseries for climate studies. Sci. Data 10, 30. https://doi.org/10.1038/s41597-022-01889-z

Carrea, L., Crétaux, J.F., Liu, X., Wu, Y., Bergé-Nguyen, M., Calmettes, B., Duguay, C.R., Jiang, D., Merchant, C.J., Mueller, D., Selmes, N., Simis, S., Spyrakos, E., Stelzer, K., Warren, M., Yesou, H., Zhang, D., 2024. ESA Lakes Climate Change Initiative (Lakes_cci): Lake products, Version 2.1. https://doi.org/10.5285/7FC9DF8070D34CACAB8092E45EF276F1

Carvalho, L., Mackay, E.B., Cardoso, A.C., Baattrup-Pedersen, A., Birk, S., Blackstock, K.L., Borics, G., Borja, A., Feld, C.K., Ferreira, M.T., Globevnik, L., Grizzetti, B., Hendry, S., Hering, D., Kelly, M., Langaas, S., Meissner, K., Panagopoulos, Y., Penning, E., Rouillard, J., Sabater, S., Schmedtje, U., Spears, B.M., Venohr, M., van de Bund, W., Solheim, A.L., 2019. Protecting and restoring Europe's waters: An analysis of the future development needs of the Water Framework Directive. Sci. Total Environ. 658, 1228–1238. https://doi.org/10.1016/j.scitotenv.2018.12.255

Chapman, D. V., Sullivan, T., 2022. The role of water quality monitoring in the sustainable use of ambient waters. One Earth 5, 132–137. https://doi.org/10.1016/j.oneear.2022.01.008

Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y., 2018. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Sci. Rep. 8, 6085. https://doi.org/10.1038/s41598-018-24271-9

Chen, J., Chen, S., Fu, R., Li, D., Jiang, H., Wang, C., Peng, Y., Jia, K., Hicks, B.J., 2022. Remote Sensing Big Data for Water Environment Monitoring: Current Status, Challenges, and Future Prospects. Earth's Futur. 10, 1–33. https://doi.org/10.1029/2021EF002289

Duan, H., Cao, Z., Luo, J., Shen, M., 2025. AI-driven opportunities and challenges in lake remote sensing. Inf. Geogr. 1, 100014. https://doi.org/10.1016/j.infgeo.2025.100014

ESA CCI, 2023. Lakes_cci+ - Phase 2 Product validation Plan (PVP).

Gillies, S., 2013. Rasterio: geospatial raster I/O for {Python} programmers. Mapbox.

Grant, L., Vanderkelen, I., Gudmundsson, L., Tan, Z., Perroud, M., Stepanenko, V.M., Debolskiy, A. V., Droppers, B., Janssen, A.B.G., Woolway, R.I., Choulga, M., Balsamo, G., Kirillin, G., Schewe, J., Zhao, F., del Valle, I.V., Golub, M., Pierson, D., Marcé, R., Seneviratne, S.I., Thiery, W., 2021. Attribution of global lake systems change to anthropogenic forcing. Nat. Geosci. 14, 849–854. https://doi.org/10.1038/s41561-021-00833-x

Hallegraeff, G.M., Anderson, D.M., Belin, C., Bottein, M.-Y.D., Bresnan, E., Chinain, M., Enevoldsen, H., Iwataki, M., Karlson, B., McKenzie, C.H., Sunesen, I., Pitcher, G.C., Provoost, P., Richardson, A., Schweibold, L., Tester, P.A., Trainer, V.L., Yñiguez, A.T., Zingone, A., 2021. Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts. Commun. Earth Environ. 2, 117. https://doi.org/10.1038/s43247-021-00178-8

Handler, A.M., Compton, J.E., Hill, R.A., Leibowitz, S.G., Schaeffer, B.A., 2023. Identifying lakes at risk of toxic cyanobacterial blooms using satellite imagery and field surveys across the United States. Sci. Total Environ. 869, 161784. https://doi.org/10.1016/j.scitotenv.2023.161784

Hanly, P.J., Webster, K.E., Soranno, P.A., 2025. LAGOS-US LANDSAT: Remotely sensed water quality estimates for U.S. lakes over 4 ha from 1984 to 2020. Sci. Data 12, 1315. https://doi.org/10.1038/s41597-025-05600-w

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Ho, J.C., Michalak, A.M., Pahlevan, N., 2019. Widespread global increase in intense lake phytoplankton blooms since the 1980s. Nature 574, 667–670. https://doi.org/10.1038/s41586-019-1648-7

Hou, X., Feng, L., Dai, Y., Hu, C., Gibson, L., Tang, J., Lee, Z., Wang, Y., Cai, X., Liu, J., Zheng, Y., Zheng, C., 2022. Global mapping reveals increase in lacustrine algal blooms over the past decade. Nat. Geosci. 15, 130–134. https://doi.org/10.1038/s41561-021-00887-x

Hoyer, S., Hamman, J., 2017. xarray: N-D labeled Arrays and Datasets in Python. J. Open Res. Softw. 5, 10. https://doi.org/10.5334/jors.148

Integrated Ocean Observing System, 2014. Manual for Real-Time Quality Control of In-situ Temperature and Salinity Data. https://doi.org/https://doi.org/10.25923/x02m-m555

Jiang, D., Scholze, J., Liu, X., Simis, S.G.H., Stelzer, K., Müller, D., Hunter, P., Tyler, A., Spyrakos, E., 2023. A data-driven approach to flag land-affected signals in satellite derived water quality from small lakes. Int. J. Appl. Earth Obs. Geoinf. 117, 103188. https://doi.org/10.1016/j.jag.2023.103188

Jordahl, K., den Bossche, J., Wasserman, J., McBride, J., Fleischmann, M., Gerard, J., Tratner, J., Perry, M., Farmer, C., Hjelle, G.A., others, 2021. geopandas/geopandas: v0.8.1. Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.3946761

Kavvada, A., Metternicht, G., Kerblat, F., Mudau, N., Haldorson, M., Laldaparsad, S., Friedl, L., Held, A., Chuvieco, E., 2020. Towards delivering on the Sustainable Development Goals using Earth observations. Remote Sens. Environ. 247, 111930. https://doi.org/10.1016/j.rse.2020.111930

Khosravi, K., Farooqe, A.A., Shahvaran, A.R., Daggupati, P., Heddam, S., Hatamiafkoueieh, J., 2025. Beyond conventional modeling: A cutting-edge hybrid IAER-AMT decision-tree-based algorithm for high-resolution river turbidity prediction. Ain Shams Eng. J. 16, 103511. https://doi.org/10.1016/j.asej.2025.103511

Kiselev, V., Bulgarelli, B., Heege, T., 2015. Sensor independent adjacency correction algorithm for coastal and inland water systems. Remote Sens. Environ. 157, 85–95. https://doi.org/10.1016/j.rse.2014.07.025

Liu, X., Warren, M., Selmes, N., Simis, S.G.H., 2024. Quantifying decadal stability of lake reflectance and chlorophyll-a from medium-resolution ocean color sensors. Remote Sens. Environ. 306, 114120. https://doi.org/10.1016/j.rse.2024.114120

Mahdian, M., Noori, R., Saravani, M.J., Shahvaran, A.R., Shahmohammad, M., Gaffney, P.P.J., Salamattalab, M.M., Anboohi,

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

M.S., Hosseinzadeh, M., Xia, F., Zhou, Y., Zhang, Y., Kolehmainen, M., Abolfathi, S., 2026. Linking Hypolimnion to Epilimnion in a Stratified Arctic Lake: Machine Learning-Based Estimation of Hypolimnetic Water Quality from Epilimnetic Measurements. Water Res. 294, 125367. https://doi.org/10.1016/j.watres.2026.125367

McKinney, P., Austin, J., Fai, G., 2019. The wind-driven formation of cross-shelf sediment plumes in a large lake. Limnol. Oceanogr. 64, 1309–1322. https://doi.org/10.1002/lno.11117

McKinney, W., 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. Python High Perform. Sci. Comput. 14, 1–9.

McKinney, W., 2010. Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference. pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

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. Nat. Commun. 7, 13603. https://doi.org/10.1038/ncomms13603

Meyer, M.F., Topp, S.N., King, T. V., Ladwig, R., Pilla, R.M., Dugan, H.A., Eggleston, J.R., Hampton, S.E., Leech, D.M., Oleksy, I.A., Ross, J.C., Ross, M.R. V., Woolway, R.I., Yang, X., Brousil, M.R., Fickas, K.C., Padowski, J.C., Pollard, A.I., Ren, J., Zwart, J.A., 2024. National-scale remotely sensed lake trophic state from 1984 through 2020. Sci. Data 11, 77. https://doi.org/10.1038/s41597-024-02921-0

Michalak, A.M., Anderson, E.J., Beletsky, D., Boland, S., Bosch, N.S., Bridgeman, T.B., Chaffin, J.D., Cho, K., Confesor, R., Daloğlu, I., DePinto, J. V., Evans, M.A., Fahnenstiel, G.L., He, L., Ho, J.C., Jenkins, L., Johengen, T.H., Kuo, K.C., LaPorte, E., Liu, X., McWilliams, M.R., Moore, M.R., Posselt, D.J., Richards, R.P., Scavia, D., Steiner, A.L., Verhamme, E., Wright, D.M., Zagorski, M.A., 2013. Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. Proc. Natl. Acad. Sci. 110, 6448–6452. https://doi.org/10.1073/pnas.1216006110

Moore, T.S., Dowell, M.D., Bradt, S., Ruiz Verdu, A., 2014. An optical water type framework for selecting and blending retrievals from bio-optical algorithms in lakes and coastal waters. Remote Sens. Environ. 143, 97–111. https://doi.org/10.1016/j.rse.2013.11.021

Neil, C., Spyrakos, E., Hunter, P.D.D., Tyler, A.N.N., 2019. A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types. Remote Sens. Environ. 229, 159–178. https://doi.org/10.1016/j.rse.2019.04.027

O'Reilly, C.M., Sharma, S., Gray, D.K., Hampton, S.E., Read, J.S., Rowley, R.J., Schneider, P., Lenters, J.D., McIntyre, P.B., Kraemer, B.M., Weyhenmeyer, G.A., Straile, D., Dong, B., Adrian, R., Allan, M.G., Anneville, O., Arvola, L., Austin, J., Bailey, J.L., Baron, J.S., Brookes, J.D., de Eyto, E., Dokulil, M.T., Hamilton, D.P., Havens, K., Hetherington, A.L., Higgins, S.N., Hook, S., Izmest'eva, L.R., Joehnk, K.D., Kangur, K., Kasprzak, P., Kumagai, M., Kuusisto, E., Leshkevich, G., Livingstone, D.M., MacIntyre, S., May, L., Melack, J.M., Mueller-Navarra, D.C., Naumenko, M., Noges, P., Noges, T., North, R.P., Plisnier, P., Rigosi, A., Rimmer, A., Rogora, M., Rudstam, L.G., Rusak, J.A., Salmaso, N., Samal, N.R., Schindler, D.E., Schladow, S.G., Schmid, M., Schmidt, S.R., Silow, E., Soylu, M.E., Teubner, K., Verburg, P., Voutilainen, A., Watkinson, A., Williamson, C.E., Zhang, G., 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophys.

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Res. Lett. 42, 10773–10781. https://doi.org/10.1002/2015GL066235

Pahlevan, N., Smith, B., Alikas, K., Anstee, J., Barbosa, C., Binding, C., Bresciani, M., Cremella, B., Giardino, C., Gurlin, D., Fernandez, V., Jamet, C., Kangro, K., Lehmann, M.K., Loisel, H., Matsushita, B., Hà, N., Olmanson, L., Potvin, G., Simis, S.G.H., VanderWoude, A., Vantrepotte, V., Ruiz-Verdù, A., 2022. Simultaneous retrieval of selected optical water quality indicators from Landsat-8, Sentinel-2, and Sentinel-3. Remote Sens. Environ. 270, 112860. https://doi.org/10.1016/j.rse.2021.112860

Persaud, B.D., Shahvaran, A.R., Attiah, G., Chasmer, L.E., English, M.C., Pour, H.K., Wolfe, B.B., 2025. Trend Analysis of Surface Shortwave (Solar) Radiation in Continental Northwest Territories, Canada (1980–2023). J. Clim. 38, 5241–5263. https://doi.org/10.1175/JCLI-D-25-0123.1

Read, J.S., Jia, X., Willard, J., Appling, A.P., Zwart, J.A., Oliver, S.K., Karpatne, A., Hansen, G.J.A., Hanson, P.C., Watkins, W., Steinbach, M., Kumar, V., 2019. Process-Guided Deep Learning Predictions of Lake Water Temperature. Water Resour. Res. 55, 9173–9190. https://doi.org/10.1029/2019WR024922

Sagan, V., Peterson, K.T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B.A., Maalouf, S., Adams, C., 2020. Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. Earth-Science Rev. 205, 103187. https://doi.org/10.1016/j.earscirev.2020.103187

Saravani, M.J., Saadatpour, M., Shahvaran, A.R., 2024. A web GIS based integrated water resources assessment tool for Javeh Reservoir. Expert Syst. Appl. 252, 124198. https://doi.org/10.1016/j.eswa.2024.124198

Saunders, R., Hocking, J., Turner, E., Rayer, P., Rundle, D., Brunel, P., Vidot, J., Roquet, P., Matricardi, M., Geer, A., Bormann, N., Lupu, C., 2018. An update on the RTTOV fast radiative transfer model (currently at version 12). Geosci. Model Dev. 11, 2717–2737. https://doi.org/10.5194/gmd-11-2717-2018

Schaeffer, B.A., Schaeffer, K.G., Keith, D., Lunetta, R.S., Conmy, R., Gould, R.W., 2013. Barriers to adopting satellite remote sensing for water quality management. Int. J. Remote Sens. 34, 7534–7544. https://doi.org/10.1080/01431161.2013.823524

Seegers, B.N., Werdell, P.J., Vandermeulen, R.A., Salls, W., Stumpf, R.P., Schaeffer, B.A., Owens, T.J., Bailey, S.W., Scott, J.P., Loftin, K.A., 2021. Satellites for long-term monitoring of inland U.S. lakes: The MERIS time series and application for chlorophyll-a. Remote Sens. Environ. 266, 112685. https://doi.org/10.1016/j.rse.2021.112685

Shahvaran, A.R., Noori, R., Basheer, M., Madani, K., Mi, C., Shi, K., Zhou, Y.Q., Woolway, R.I., Zhang, Y.L., Qin, B.Q., AghaKouchak, A., van Vliet, M.T.H., 2026. Global daily satellite-derived lake surface water quality dataset (2002-2022), Zenodo [data set], https://doi.org/10.5281/zenodo.20347115

Shahvaran, A.R., Kheyrollah Pour, H., Binding, C., Van Cappellen, P., 2025. Mapping satellite-derived chlorophyll-a concentrations from 2013 to 2023 in Western Lake Ontario using Landsat 8 and 9 imagery. Sci. Total Environ. 968, 178881. https://doi.org/10.1016/j.scitotenv.2025.178881

Shahvaran, A.R., Kheyrollah Pour, H., Van Cappellen, P., 2024. Comparative Evaluation of Semi-Empirical Approaches to Retrieve Satellite-Derived Chlorophyll-a Concentrations from Nearshore and Offshore Waters of a Large Lake (Lake Ontario). Remote Sens. 16, 1595. https://doi.org/10.3390/rs16091595

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Sharma, S., Blagrave, K., Magnuson, J.J., O'Reilly, C.M., Oliver, S., Batt, R.D., Magee, M.R., Straile, D., Weyhenmeyer, G.A., Winslow, L., Woolway, R.I., 2019. Widespread loss of lake ice around the Northern Hemisphere in a warming world. Nat. Clim. Chang. 9, 227–231. https://doi.org/10.1038/s41558-018-0393-5

Shen, C., 2018. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. Water Resour. Res. 54, 8558–8593. https://doi.org/10.1029/2018WR022643

Spyrakos, E., O'Donnell, R., Hunter, P.D., Miller, C., Scott, M., Simis, S.G.H.H., Neil, C., Barbosa, C.C.F.F., Binding, C.E., Bradt, S., Bresciani, M., Dall'Olmo, G., Giardino, C., Gitelson, A.A., Kutser, T., Li, L., Matsushita, B., Martinez-Vicente, V., Matthews, M.W., Ogashawara, I., Ruiz-Verdú, A., Schalles, J.F., Tebbs, E., Zhang, Y., Tyler, A.N., Martinez-Vicente, V., Matthews, M.W., Ogashawara, I., Ruiz-Verdú, A., Schalles, J.F., Tebbs, E., Zhang, Y., Tyler, A.N., 2018. Optical types of inland and coastal waters. Limnol. Oceanogr. 63, 846–870. https://doi.org/10.1002/lno.10674

Steinmetz, F., Deschamps, P.-Y., Ramon, D., 2011. Atmospheric correction in presence of sun glint: application to MERIS. Opt. Express 19, 9783. https://doi.org/10.1364/OE.19.009783

Stockwell, J.D., Doubek, J.P., Adrian, R., Anneville, O., Carey, C.C., Carvalho, L., De Senerpont Domis, L.N., Dur, G., Frassl, M.A., Grossart, H., Ibelings, B.W., Lajeunesse, M.J., Lewandowska, A.M., Llames, M.E., Matsuzaki, S.S., Nodine, E.R., Nõges, P., Patil, V.P., Pomati, F., Rinke, K., Rudstam, L.G., Rusak, J.A., Salmaso, N., Seltmann, C.T., Straile, D., Thackeray, S.J., Thiery, W., Urrutia-Cordero, P., Venail, P., Verburg, P., Woolway, R.I., Zohary, T., Andersen, M.R., Bhattacharya, R., Hejzlar, J., Janatian, N., Kpodonu, A.T.N.K., Williamson, T.J., Wilson, H.L., 2020. Storm impacts on phytoplankton community dynamics in lakes. Glob. Chang. Biol. 26, 2756–2784. https://doi.org/10.1111/gcb.15033

Su, J., Lin, Z., Xu, F., Fathi, G., Alnowibet, K.A., 2024. A hybrid model of ARIMA and MLP with a Grasshopper optimization algorithm for time series forecasting of water quality. Sci. Rep. 14, 23927. https://doi.org/10.1038/s41598-024-74144-7

Topp, S.N., Pavelsky, T.M., Jensen, D., Simard, M., Ross, M.R. V., 2020. Research Trends in the Use of Remote Sensing for Inland Water Quality Science: Moving Towards Multidisciplinary Applications. Water 12, 169. https://doi.org/10.3390/w12010169

Tranvik, L.J., Downing, J.A., Cotner, J.B., Loiselle, S.A., Striegl, R.G., Ballatore, T.J., Dillon, P., Finlay, K., Fortino, K., Knoll, L.B., Kortelainen, P.L., Kutser, T., Larsen, S., Laurion, I., Leech, D.M., McCallister, S.L., McKnight, D.M., Melack, J.M., Overholt, E., Porter, J.A., Prairie, Y., Renwick, W.H., Roland, F., Sherman, B.S., Schindler, D.W., Sobek, S., Tremblay, A., Vanni, M.J., Verschoor, A.M., von Wachenfeldt, E., Weyhenmeyer, G.A., 2009. Lakes and reservoirs as regulators of carbon cycling and climate. Limnol. Oceanogr. 54, 2298–2314. https://doi.org/10.4319/lo.2009.54.6_part_2.2298

Wevers, J., Müller, D., Kirches, G., Quast, R., Brockmann, C., 2022. IdePix for Sentinel-3 OLCI Algorithm Theoretical Basis Document. Zenodo.

Woolway, R.I., Albergel, C., Frölicher, T.L., Perroud, M., 2022a. Severe Lake Heatwaves Attributable to Human-Induced Global Warming. Geophys. Res. Lett. 49. https://doi.org/10.1029/2021GL097031

Woolway, R.I., Jennings, E., Shatwell, T., Golub, M., Pierson, D.C., Maberly, S.C., 2021a. Lake heatwaves under climate change. Nature 589, 402–407. https://doi.org/10.1038/s41586-020-03119-1

https://doi.org/10.5194/essd-2026-528
Preprint. Discussion started: 25 August 2026

Woolway, R.I., Kayastha, M.B., Tong, Y., Feng, L., Shi, H., Xue, P., 2025. Subsurface heatwaves in lakes. Nat. Clim. Chang. 15, 554–559. https://doi.org/10.1038/s41558-025-02314-0

Woolway, R.I., Kraemer, B.M., Lenters, J.D., Merchant, C.J., O'Reilly, C.M., Sharma, S., 2020. Global lake responses to climate change. Nat. Rev. Earth Environ. 1, 388–403. https://doi.org/10.1038/s43017-020-0067-5

Woolway, R.I., Kraemer, B.M., Zscheischler, J., Albergel, C., 2021b. Compound hot temperature and high chlorophyll extreme events in global lakes. Environ. Res. Lett. 16, 124066. https://doi.org/10.1088/1748-9326/ac3d5a

Woolway, R.I., Sharma, S., Smol, J.P., 2022b. Lakes in Hot Water: The Impacts of a Changing Climate on Aquatic Ecosystems. Bioscience 72, 1050–1061. https://doi.org/10.1093/biosci/biac052

Woolway, R.I., Tong, Y., Feng, L., Zhao, G., Dinh, D.A., Shi, H., Zhang, Y., Shi, K., 2024. Multivariate extremes in lakes. Nat. Commun. 15, 4559. https://doi.org/10.1038/s41467-024-49012-7