www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

A bio-optical database for the remote sensing of water quality in BRAZil coAstal and inland waters (BRAZA)

Daniel Andrade Maciel[1✉], Claudio Clemente Faria Barbosa[1], Evlyn Márcia Leão de Moraes Novo[1], Raianny Leite do Nascimento Wanderley[1], Pedro Bacellar[1], Lohan Barbosa Baia[2], Henrique Bernini[1], Rafael G. Chasles[1], Aurea Ciotti[3], Alice Fassoni-Andrade[4], Rogério Flores Júnior[1], Thainara Lima[5], Fernando Lopes[6], Rogério Ribeiro Marinho[7], André Martinelli[8], Vitor S. Martins[5], Marcela Miranda[9], Gabriel Moiano Cesar[1], David da Motta Marques[10], Maurício Almeida Noernberg[11], Fernanda Helena Oliveira da Silva[6], Felipe Pacheco[9], Rejane S. Paulino[5], Waterloo Pereira Filho[12], Júlio César Pimenta dos Santos[1], Lino Augusto Sander de Carvalho[13], Marcio Sousa da Silva[14], Matheus Henrique Tavares[10], Marta Ummus[15], Aline de Matos Valério[1,16], Pedro Walfir M. e Souza Filho[2], Fernanda Sayuri Yoshino Watanabe[17] & Rafaela Cristine Zem[11]

Understanding water quality (WQ) is essential for grasping biogeochemical cycles and assessing human impacts such as deforestation, climate change, dam construction, illegal mining, and urbanization. However, monitoring WQ across Brazil's vast aquatic systems requires significant resources. Remote sensing improves this process by offering high-resolution and large-scale observations. Nevertheless, to produce reliable remote-sensing WQ products based on remote sensing demands comprehensive datasets with concurrent aquatic reflectance and *in situ* measurements (e.g., Chlorophyll-a, Secchi Disk Depth, Suspended Sediments). Such data enables advanced semi-analytical and machine learning models to capture Brazil's bio-optical water complexity. Collaborative efforts and open-data sharing are essential for building these datasets. Here, we introduce a new curated, high-quality bio-optical dataset across different aquatic systems in Brazil, called BRAZA (Bio-optical aquatic database for remote sensing of water quality in BRAZil coAstal and inland waters). By leveraging data from 17 institutions, covering 2,895 stations across +128 lakes, rivers, reservoirs, and coastal areas of Brazil's five administrative regions our dataset presents an important contribution to support remote sensing WQ-based analysis in Brazil.

[1]Earth Science Coordination of National Institute for Space Research (INPE), São José dos Campos, SP, Brazil. [2]Federal University of Pará, Belém, PA, Brazil. [3]CEBIMAR/USP, São Sebastião, SP, Brazil. [4]University of Brasília (UnB), Institute of Geosciences, Brasília, Brazil. [5]Department of Agricultural & Biological Engineering, Mississippi State University (MSU), Mississippi State, MS, 39762, USA. [6]Federal University of Ceará, Fortaleza, CE, Brazil. [7]Federal University of Amazonas, Manaus, AM, Brazil. [8]Geological Survey of Brazil, Manaus, Brazil. [9]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York, US. [10]Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil. [11]Center of Marine Studies, Federal University of Paraná (CEM/UFPR), Pontal do Paraná, PR, Brasil. [12]Federal University of Santa Maria (UFSM), Santa Maria, RS, Brasil. [13]Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil. [14]Vale Institute of Technology (ITV), Belém, PA, Brazil. [15]Embrapa Pesca e Aquicultura, Palmas, TO, Brazil. [16]O'Neill School of Public and Environmental Affairs, Indiana University, Bloomington, IN, USA. [17]Department of Cartography, São Paulo State University (UNESP), Presidente Prudente, SP, Brazil. ✉e-mail: daniel.maciel@inpe.br

Background & Summary

Brazil covers 8.51 million of km^2 with approximately 180,000 km^2 of open freshwater resources, including rivers, reservoirs, and floodplains[1], as well as 3.5 million km^2 of coastal waters. Although the country is rich in water resources, their distribution is uneven. While the Amazon region has abundant water, the northeast region faces challenges that lead to a range of environmental impacts[2]. Current human activities, including illegal gold mining, deforestation, wildfires, and extensive agricultural and industrial development, significantly threaten the quality of Brazil's water resources[3–5]. Therefore, it is essential to establish a comprehensive system for monitoring water quality (WQ) in Brazil to track the impacts of human activities and provide environmental monitoring of fresh and coastal aquatic ecosystems. Organizations such as the Brazilian Water Agency (ANA) and the São Paulo Environmental Company (CETESB) manage a network of *in-situ* stations distributed across the country and states, respectively[6,7]. These stations provide periodic observations of the water bodies, typically at intervals of one-to-three months with few spatial samples, resulting in insufficient temporal and spatial representativeness. In many cases of rivers with high turbidity, most of the sediment load (up to 90% of the annual load) is transported in just a fraction of time (sometimes less than 10%)[8], therefore, an alternative approach for obtaining WQ variables with better spatiotemporal coverage is by using remote sensing data.

Remote sensing provides a synoptic view of the aquatic ecosystems, by offering high-resolution spatial and temporal observations that complement traditional monitoring methods[9]. By leveraging high spatiotemporal data, remote sensing can provide near-real-time WQ maps across different scales (i.e., from global to local), which helps to reduce the costs of environmental management for these aquatic ecosystems. Technological and scientific advancements between 2013 and 2015 have addressed previous limitations (e.g., low Signal-To-Noise ratio, SNR) with the launch of new high-quality Earth-observing sensors, such as the Operational Land Imager (OLI) Landsat-8/9 and MultiSpectral Imager (MSI) Sentinel-2[10,11]. With improved spatiotemporal resolution and higher SNR, these sensors have become suitable for large-scale monitoring of aquatic resources. In addition, recent hyperspectral sensors are also expanding the applications of remote sensing for water quality-based studies, marking the beginning of a new era in water quality monitoring from space, enabling the development of more precise and robust algorithms for mapping optically active constituents (OACs)[12–15].

The use of remote sensing to monitor aquatic resources are connected to the understanding of the radiative transfer theory, which involves comprehending how optically active constituents (OACs) in water interact with light, ultimately affecting the water-leaving radiance (L_w). The main goal is to measure L_w, detected by remoted sensors[16–18]. However, L_w does not have stability along time, because it is dependent on the incident irradiance[16]. To overcome this issue, the Apparent Optical Property (AOP) known as Remote Sensing Reflectance (R_{rs}, sr^{-1}) could be calculated as the ratio of the L_w to the downwelling irradiance (E_s) just beneath the water's surface in each wavelength (λ). Unlike radiance, R_{rs} remains stable over time, allowing for temporal studies and remote sensing monitoring of aquatic systems. Since the 1970s, this approach has been critical for monitoring both the open ocean and coastal waters, particularly after the launch of sensors designed for this purpose, such as the Coastal Zone Color Scanner (CZCS)[19]. Although using remote sensing (RS) to monitor aquatic environments offers several benefits, there are challenges in adapting satellite sensors and algorithms originally designed for the open ocean to freshwater and coastal ecosystems. One of the main issues is the optical complexity of inland and coastal waters, which are characterized by a mixture of various OACs such as phytoplankton pigments (e.g., Chlorophyll-*a* concentration (Chl-*a*)), coloured dissolved organic matter (CDOM), and inorganic sediments[9]. Additionally, there is a lack of sensors with high radiometric, spectral, and spatial resolutions, and high Signal-to-Noise Ratio (SNR). These sensor characteristics are essential for detecting the low radiance typically found in many continental water bodies. Moreover, low R_{rs} values of coastal regions, aligned with a low temporal resolution could limit the monitoring of highly dynamic events, such as harmful algae blooms. Additionally, mapping small rivers and lakes has proven challenging due to the limited spatial resolution of sensors primarily designed for open-ocean applications, as well as challenges in atmospheric and adjacency corrections[20–22].

Two types of algorithms are used for retrieving the OACs: empirical and semi-analytical models[23]. Empirical models rely on the statistical relationship between R_{rs} and OACs as well as other water properties, such as transparency (e.g., Secchi Disk Depth). These models can be developed using different methods, including simple regressions (i.e., one-term regressions), multi-variable regressions, and machine learning and deep learning techniques[24,25]. All of these models rely on samples for training and validation, demanding a comprehensive dataset for effective spatiotemporal monitoring. Semi-analytical models, which simplify radiative transfer theory, have strong spatiotemporal applications[26], however, they require precise measurements of water Inherent Optical Properties (IOPs), which makes them less commonly used. This is largely due to the uncertainties in these measurements and sensors corrections in freshwater ecosystems[27]. Disregarding the algorithm applied, all of them demand large sample size with measurements encompassing a huge variability of optical properties and OACs for providing accurate results.

Recently, several initiatives have been conducted to create freely open-access datasets of remote sensing data focusing on global inland and coastal waters[28–33]. One of the recent examples is the GLORIA dataset[32], which provides more than 7,000 stations with concurrent measurements of AOPs and OACs. Other examples include the SeaSWIR dataset[31] and the HYPERMAQ dataset[28] focusing on moderate to extremely turbid waters. There are also other initiatives that provides *in-situ* water quality data aligned with satellite-derived R_{rs}, such as AQUASAT[34]. In addition, other studies have been providing synthetic datasets to support large-scale development of water quality models[33,35,36]. Despite the global distribution of these datasets, there is still a lack of the knowledge of the variation of the AOPs and OACs in Brazil, which has a wide variability in OACs[37].

Even though remote sensing is an essential tool for large scale water quality monitoring, its application to Brazilian aquatic ecosystems still has limitations. The vastness of Brazil's territory leads to a variety of environments[38], ranging from the Negro River[39] dissolved organic-matter-rich rich water in the Amazon to the eutrophic reservoirs in São Paulo and Rio de Janeiro[40]. Additionally, there are the turbid waters from the

Amazon River and the clear waters reservoirs, such as Três Marias in Minas Gerais[37,41], and the coastal waters of the Alcatrazes Archipelago. This high variability creates a mix of OACs, resulting in an optically complex environment that poses challenges for the development of algorithms[12].

Large-scale applications or even local algorithms require a representative dataset to allow the model to understand and adapt to the variability of optical parameters at a given aquatic system. Therefore, it is necessary to build *in-situ* databases to use the data collected in different locations. These initiatives are only possible through an inter-institutional collaboration network, based on the principles of free data sharing[32]. To support a bio-optical database availability for Brazilian freshwater and coastal ecosystems, this study describes the process used to collect, compile and organize BRAZA, a bio-optical dataset for Brazilian waters. With 17 collaborating institutions over Brazil and United States, we compiled 2,895 *in-situ* measurements of R_{rs} and other WQ parameters over the country's five regions. This paper describes the dataset, the variability in the parameters available in the dataset (i.e., water quality parameters) and their quality control.

Methods

Study area. The study area covers all the five Brazil's regions (Fig. 1). The development of this dataset was based on a national-scale collaboration among more than 30 researchers from Brazil and the United States. The Instrumentation Laboratory for Aquatic Ecosystems at the National Institute for Space Research (LabISA-INPE led the efforts to create a large-scale dataset for bio-optical modeling data in Brazil. Since the 1990s, INPE team has collected numerous measurements of water AOPs and IOPs across Brazil. The effort to organize the data began in 2015[38], with the compilation of approximately 500 stations, primarily collected in the Amazon region. In 2019, Silva *et al.*[37] organized the LabISA data into a comprehensive, SQL-like database, which was further updated in 2021 by Maciel *et al.*[12] and more recently by Maciel *et al.*[42]. Despite these efforts, the vast size of Brazil and the huge bio-optical diversity of its waters demands a wider collection of measurements. As a result, producing a comprehensive dataset for large-scale remote sensing of Brazilian waters requires significant effort and a multi-collaborative approach. Several researchers were invited to collaborate on the dataset's construction to address this need. The sections below provide detailed information on the measurement of each variable and the available datasets.

The dataset collected includes >128 different environments in Brazil, as coastal waters, reservoirs, small and large rivers (e.g., Amazon, Negro, Tapajós, and Guamá rivers), fishponds, and floodplain lakes. The project collaborators provided a total of 2,895 R_{rs} measurements across the five regions of Brazil. Apart from the radiometric measurements, the dataset contains 1,506 valid measurements of Chlorophyll-*a* and 161 of phycocyanin. For sediment concentration, 1,420, 1,042, and 1,028 contains valid measurements of total suspended sediments (TSS) and their organic (TSO) and inorganic (TSI) fractions, respectively. Secchi Disk Depth (Z_{sd}) is available for 2,374 stations, and turbidity for 1,246 stations. The absorption coefficient by CDOM at 440 nm (aCDOM(440)) is available for 837 stations. The Dissolved Total Carbon (DTC) and its organic (DOC) and inorganic (DIC) portions are available for 580, 801 and 733 stations, respectively. A description of the methods used for each variable measurement is available in Section 3.2.

Dataset organization. The data organization used a database-oriented structure following that which supports the integration with global datasets such as GLORIA. The remote sensing data is presented in a table with the R_{rs} data between 400 and 900 nm (see Section 2.2.1). A second table contains the metadata (i.e., station identifier) and other necessary information such as date and time, geographic coordinates, and water quality parameters. A *Metadata* sheet is available in excel file, as well as a sheet with associated R_{rs} measurement method. A third table includes the associated flags for each spectrum. Water quality parameters used in this dataset comprehend phytoplankton pigment concentrations (Chlorophyll-*a*, Pheophytin and Phycocyanin, ug L^{-1}), non-algal particles information (total suspended sediment and its organic/inorganic part concentrations, mg L^{-1}), aCDOM(440) (m^{-1}), Secchi Disk Depth (Z_{sd}) (m), turbidity (NTU), and Dissolved Total, Organic and Inorganic carbon concentrations (mgL^{-1}) (DTC, DOC, DIC, respectively).

***In-situ* measurements.** *Remote sensing reflectance.* The *in-situ* R_{rs} data was collected using different types of radiometers as identified in the *Metadata* table and can follow different settings. Measurements of water-leaving radiance (L_w), downwelling irradiance beneath water surface (E_s), and sky radiance (L_{sky}) were taken for most of the available datasets to calculate R_{rs}. Radiometric quantities were measured generally between 9:30 and 14:30 local time, to avoid low sun-zenith angle and glint occurrence. Although we expect differences in the measurements between radiometers from different manufacturers, we considered it as having second order of importance, as the uncertainties between the measurements are generally lower than 10%[43,44]. In addition, by integrating a large database, we reduce the effects and possible biases of systematic errors[45]. Indeed, we expect this effort will stimulate future site intercomparisons. Following, detailed information about each form of measurement is presented.

(1) **Satlantic HyperPRO II**. Measurements taken with HyperPRO II were slightly different from those of the other sensors. HyperPRO II measures the upwelling radiance below water surface (L_u) and E_s, in a buoy or free-falling mode. Therefore, it is necessary to extrapolate the L_u data to L_w. L_u data were extrapolated to the surface by calculating the Diffuse Attenuation Coefficient of upwelling radiance (KL_u) based on the empirical method proposed by Austin and Petzold[19] and Morel *et al.*[46] when measured in buoy mode or by calculating a KL_u based on L_u(0-) profiles when measured in profile mode. L_u(0-) data were converted to L_w using the Fresnel and water refraction index (1.34). Variable values necessary to these calculations were obtained from the HyperPRO II user manual. The bandwidth is 3.3 nm.

(2) **TRIOS-RAMSES** radiometers. In this setting, one radiometer is used to measure water leaving radiance (L_w), with an angle of 45° in relation to nadir and an azimuth angle in relation to the sun of 135°.

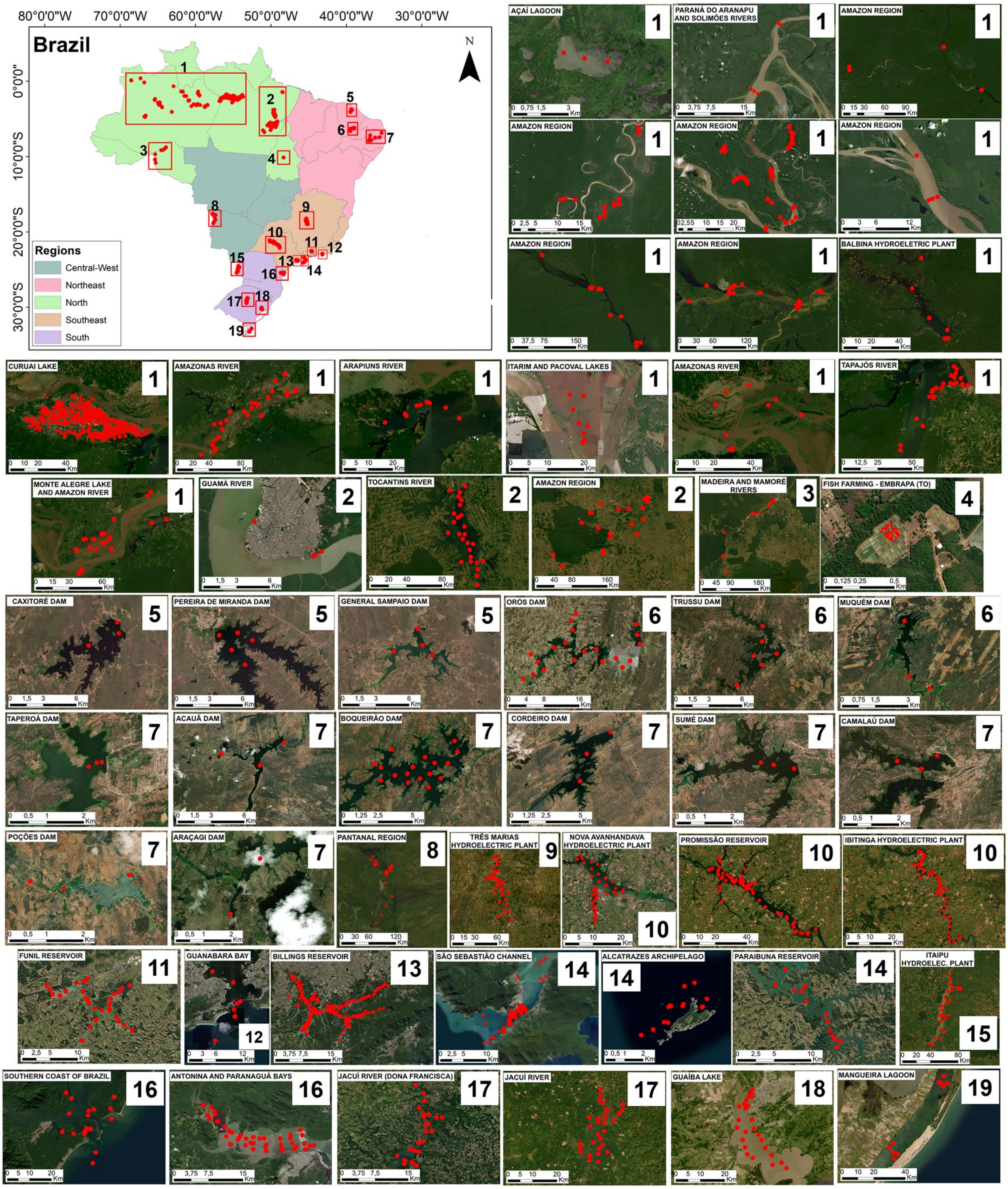

Fig. 1 Study area. Each red-point corresponds to *in-situ* measured station. The Amazon box was grouped due to the large number of unique lakes. Numbers in each box refer to the location highlighted in the Brazil map.

Simultaneously, another radiometer measured sky radiance (L_{sky}), with an view angle of 45° in relation to zenith and an azimuth angle of 135°. Finally, a cosine sensor TRIOS-RAMSES was used for measured downwelling irradiance just beneath the water surface (E_s). The nominal bandwidth of RAMSES is 3.3 nm. This setting follows the geometrical protocols of Mobley[47].

(3) **ASD FieldSpec**. ASD FieldSpec-4 also was used in radiance mode to collect L_w, L_{sky}, and the radiance of a Spectralon Lambertian plaque (L_s), which is then converted to irradiance by integrating it into the hemisphere (i.e., multiplying L_s by π). In some stations, they were collected in the Reflectance mode, measuring water reflectance and subtracting it by the sky reflectance multiplied by the ρ factor from Mobley[47]. The nominal bandwidth of this sensor is 2.2 nm in the visible-to-near-infrared (VNIR) channels.

(4) **ASD HandHeld-2**. In this setup, the HandHeld-2 instrument is used in radiance mode to collect water

leaving radiance (L_w), sky radiance (L_{sky}) and L_{plate}. In some stations (noted in the *Metadata* table), ASD HandHeld-2 data was collected in Reflectance mode without measuring L_{sky}. The spectral resolution of HandHeld-2 is defined by ASD as < 3 nm at 700 nm.

(5) **Spectron SE-590**. The measurements collected with the SPECTRON SE-590 followed the same pattern of the used for ASD's HandHeld-2 and FieldSpec-4 measurements. A radiance measurement of the water (L_w) and a Lambertian plaque (L_{plate}) was taken. In this setup, no L_{sky} measurements were taken. L_{plate} was extrapolated to E_s by integrating the L_{plate} throughout the hemisphere (i.e., multiplying L_{plate} by π). The half-band width is approximately 10 nm, with a 3-nm difference between each channel.

Except for Satlantic HyperPRO II, R_{rs} was calculated using Eq. 1, with the ρ calculated using Mobley[47] Look-up-Table and auxiliary parameters, such as wind speed, sun-zenith and view angle when available. For HyperPro II measurements, R_{rs} was calculated as L_w/E_s (i.e., without the L_{sky} effect). When wind speed is not available, we consider it less than $5\,m\,s^{-1}$.

$$R_{rs} = \frac{L_w - \rho\, L_{sky}}{E_s} \tag{1}$$

The resulting R_{rs} was also classified into specific Optical Water Types (OWTs) based on seven OWTs proposed by Pahlevan *et al.*[45]. These seven OWTs are based on a grouping of the 21 OWTs defined by Spyrakos *et al.*[48]. The Pahlevan *et al.*[45] OWTs classes were used instead of Spyrakos *et al.*[48] to reduce the number of classes and group similar spectra. In this classification, OWT-1 are related to more clear waters, OWT-6 is connected to eutrophic locations and OWT-7 to very turbid waters. To assign a class to each measured R_{rs} spectrum, we used the Spectral Angle Mapper (SAM)[49] with the seven OWTs as reference. The class for a specific spectrum was determined by the lowest SAM value between the spectrum and the seven references.

Water quality variables. A diverse set of water quality variables was compiled in this study to construct the database. These variables are: Chl-*a*; Pheophytin, Phycocyanin, Suspended Sediment (total, organic and inorganic – TSS, TSI, and TSO) concentration; aCDOM(440), Z_{sd}, Turbidity, DTC, DOC and DIC. The methods for measuring each of these WQ variables are defined below:

Chlorophyll-a (Chl-*a*) and Pheophytin (P) concentrations. In most of the measurements, Chl-*a* and P were measured by filtering a determined water volume (~100–1000 mL) in a 47 or 25 mm diameter and 0.7 pore size Whatmann Glass-Fiber Filter (GF/F) in duplicates. Chlorophyll-a concentrations (Chl-*a*) were determined following the protocol outlined by APHA[50], which involves pigment extraction using 90% acetone. The breakdown of algal cells was carried out by incubating the samples at 4 °C for 12 hours, followed by centrifugation at 3000 rpm for 20 minutes. The absorption of the extracted pigments was measured using a spectrophotometer. To account for phaeophytin correction, the samples were acidified with 0.1 M HCl and reanalysed on the spectrophotometer. Pigment concentrations were calculated using Lorenzen's equation, as described in APHA[50]. After that, the average Chl-*a* and P values were calculated. For SS e Alcatrazes the non-acidification method was used (Welshmeyer).

Phycocyanin (PC). PC concentration was measured by filtering a defined volume of water (100–500 mL) through a 47 mm diameter and 0.7 pore size Whatmann Glass-Fiber Filter (GF/F). Water samples were collected from the subsurface (~15 cm) in dark bottles and stored in cooled containers to prevent thermal and photodegradation. Filtration was performed using GF/F filters under low light conditions and low vacuum pressure to minimize pigment degradation. Filters were stored in liquid nitrogen until further analysis. PC extraction followed the protocols of Sarada *et al.*[51], with adaptations by Horváth *et al.*[52]. Filters were suspended in a phosphate buffer (100 mM and pH 7.2) and submitted to 3 freeze–thaw cycles. The samples were then sonicated for 90 seconds at 20 kHz and centrifuged for 15 minutes at 3,000 rpm. PC concentrations were determined spectrophotometrically applying the formula proposed by Bennett and Bogorad[53]. All samples were analysed in duplicate, and the mean value was used as the reference concentration for each sampling station.

Suspended Sediments (Total, Organic, Inorganic). Were measured using the Wetzel and Likens[54] or[55] methods in most of the stations. For that, water samples (generally 0.1 to 1 L) were filtered into previously weighted 0.7 μm pore size GF/F filters (for some campaigns in Pará state, a 0.4 μm filter was used). After filtering, samples were stored in refrigerated containers until further analysis in laboratory. For Total Suspended Sediments, the filters were dry for 24 hours at 60 °C and their weight was measured. Total Inorganic Sediments (TSI) were measured by placing these filters into a muffle at 480 °C for 1 hour, to eliminate organic portion, and then weighed again. Total Organic Sediments (TSO) were calculated by the difference between TSS and TSI.

aCDOM(440). Dissolved organic matter is composed of a complex set of dissolved substances that come from the decomposition of animal and vegetal tissues and from the synthetic activity of microorganisms. The chromophore fraction of dissolved organic matter is called Colored Dissolved Organic Matter (CDOM). To calculate the aCDOM(440) water samples were filtered in a 0.22 μm pore size filter for each station whose absorbance is measured by the spectrophotometer according to Bricaud *et al.*[56]. To carry out these measurements in inland waters, a 10 cm cuvette was used, whereas for the Alcatrazes and São Sebastião channel the pathlength was set to up to 2 m. To remove residual absorption and effects of temperature variation, dispersion and refraction, the aCDOM(440) values were subtracted from the average between 750 and 800 nm according to Green and

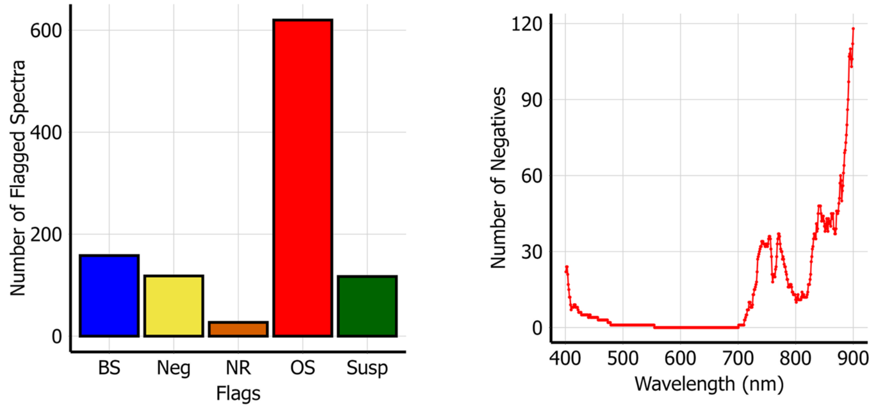

Fig. 2 Number of flagged R_{rs} spectra in each flag. (**A**) Total number of flagged spectra and (**B**) Total number of negative spectra per wavelength. BS: Baseline Shift, Neg: Number of Negatives; NR: Noisy Red; OS: Oxygen Signal; Susp: Suspicious.

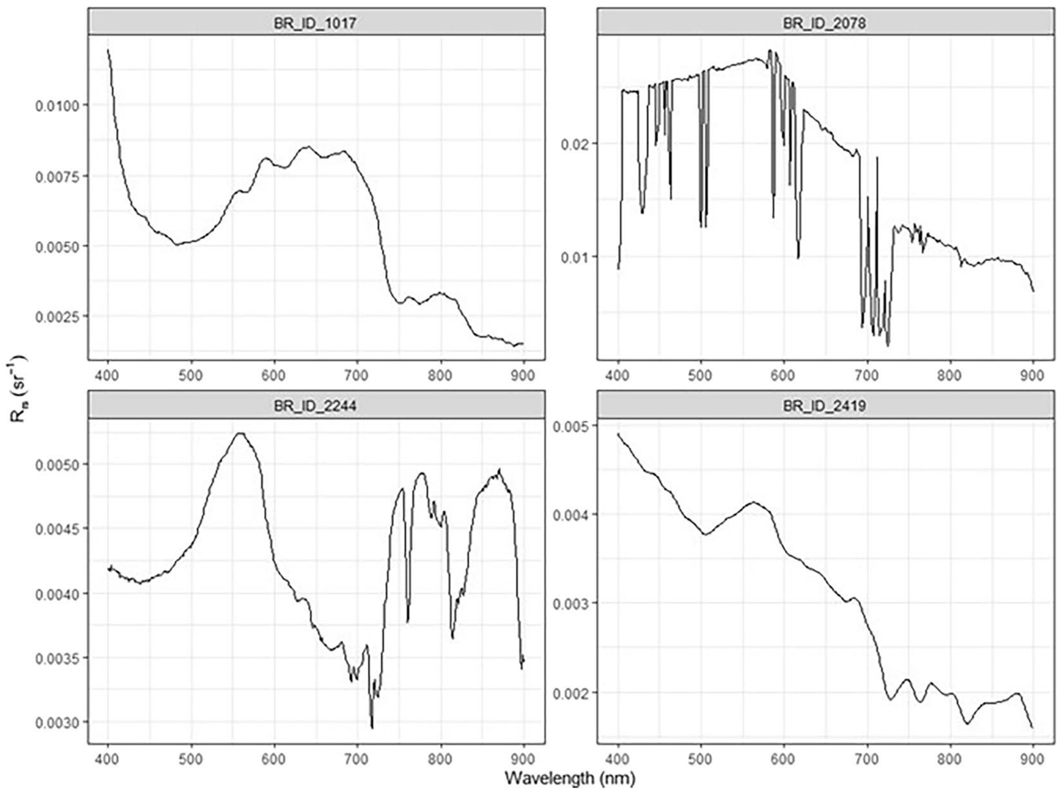

Fig. 3 Example of spectra flagged as "Suspicious" based on the interpretation of specialists.

Blough[57]. The calculation for aCDOM(440) follows Bricaud *et al.*[56]. The equipment and methods used to measure aCDOM(440) are detailed in the metadata table.

Secchi Disk Depth (Z_{sd}). Z_{sd} is a widely used and classic measurement of water transparency. It represents the depth in which a disk lowered in the water disappears in relation to an observer view[21,58]. In the datasets available in this study, the Secchi Disk Depth (Z_{sd}) was measured with a 30 cm diameter black-white or white disk until the depth of disappearance of the disk. White disk was used only in the Alcatrazes and São Sebastião field missions, and it is not expected a high difference between these two disks.

Dissolved Carbon (Total, Organic, Inorganic). To determine the DOC concentration, water samples were collected, filtered through a Whatman nylon membrane filter with a pore size of 0.22 μm and 47 mm in diameter, stored in sterilized polyethylene bottles, wrapped in aluminium foil and kept refrigerated (4 °C) until the time of laboratory analysis. The samples were left at room temperature before analysis. Concentration measurements

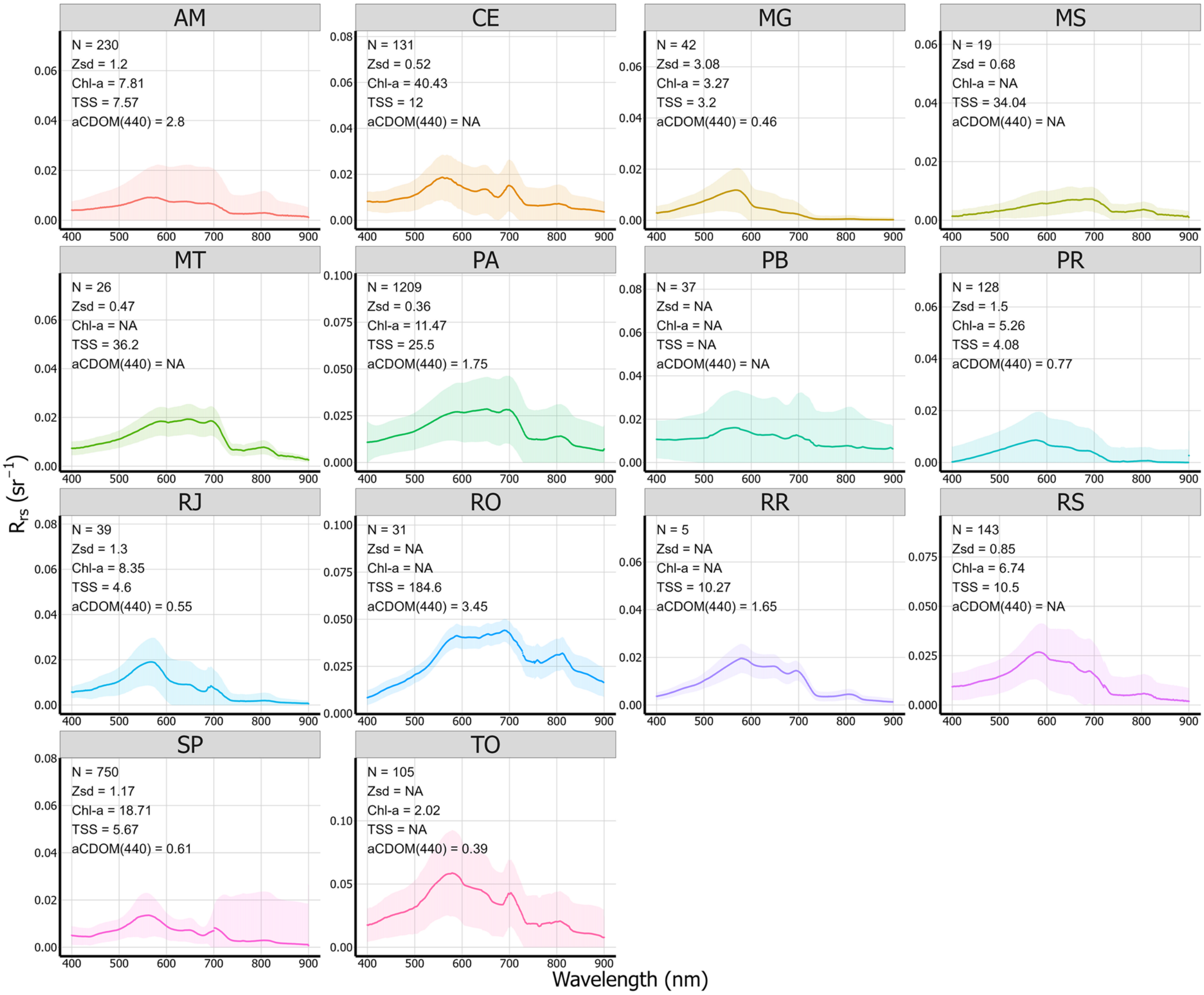

Fig. 4 Variability of R_{rs} for the available data. Each box-title represents the state in which the R_{rs} came from. Solid line are the median values, and shaded represents median ± standard deviation of R_{rs} values. Please, note that y-axis is in different scale for visualization purposes.

were carried out using a Total Organic Carbon (TOC) analyzer, allowing the direct measurement of DOC and DIC. The inorganic fraction is purged using phosphoric acid (H_3PO_4) and hydrochloric acid (HCl), transforming carbonates and bicarbonates into CO_2 and H_2O[59]. By separating these fractions, it is possible to determine the DOC concentration in water samples.

Turbidity. Turbidity is a measurement of light scattering in a determined angle and wavelength and are represented in Nephelometric Turbidity Unities (NTU). It is an important measurement as it is easy to be taken in the field and have a very high relationship with inorganic suspended sediments. The equipment and methods used to measure turbidity in the field are detailed in the Metadata table.

As the data were measured using different methods by each collaborator, each of these methods are identified in the *Metadata* table for each entry of the BRAZA dataset, when necessary.

Data Records

The BRAZA dataset is available at Figshare[60] and it contains three associated Excel files tables with the data. All these three tables contain a unique identifier (BR_ID) to connect each table. The first table (*rrs.xlsx*) contains the BR_ID and R_{rs} between 400 and 900 nm (e.g., Rrs_400 for 400 nm, in steradians, sr^{-1})). The second table (*stations.xlsx*) contains three sheets: the first sheet (*Data*) contains the unique station identifier (BR_ID), a station_name, ancillary data (e.g., latitude, longitude, date, hour (GMT), lake name)) and water quality related parameters, with associated unit (e.g., chla_ugl, for Chlorophyll-*a* in micrograms per liter). In this table, a second sheet (*Metadata*) contains the methods used for each measurement, when available, with associated references. The third sheet (*rrs_methods)* contains the explanation on each different method used to calculate the R_{rs}. Finally, a third table (*flags.xlsx*), contains the associate *flags* for the quality-control of the dataset, as stated in Section 42.3.

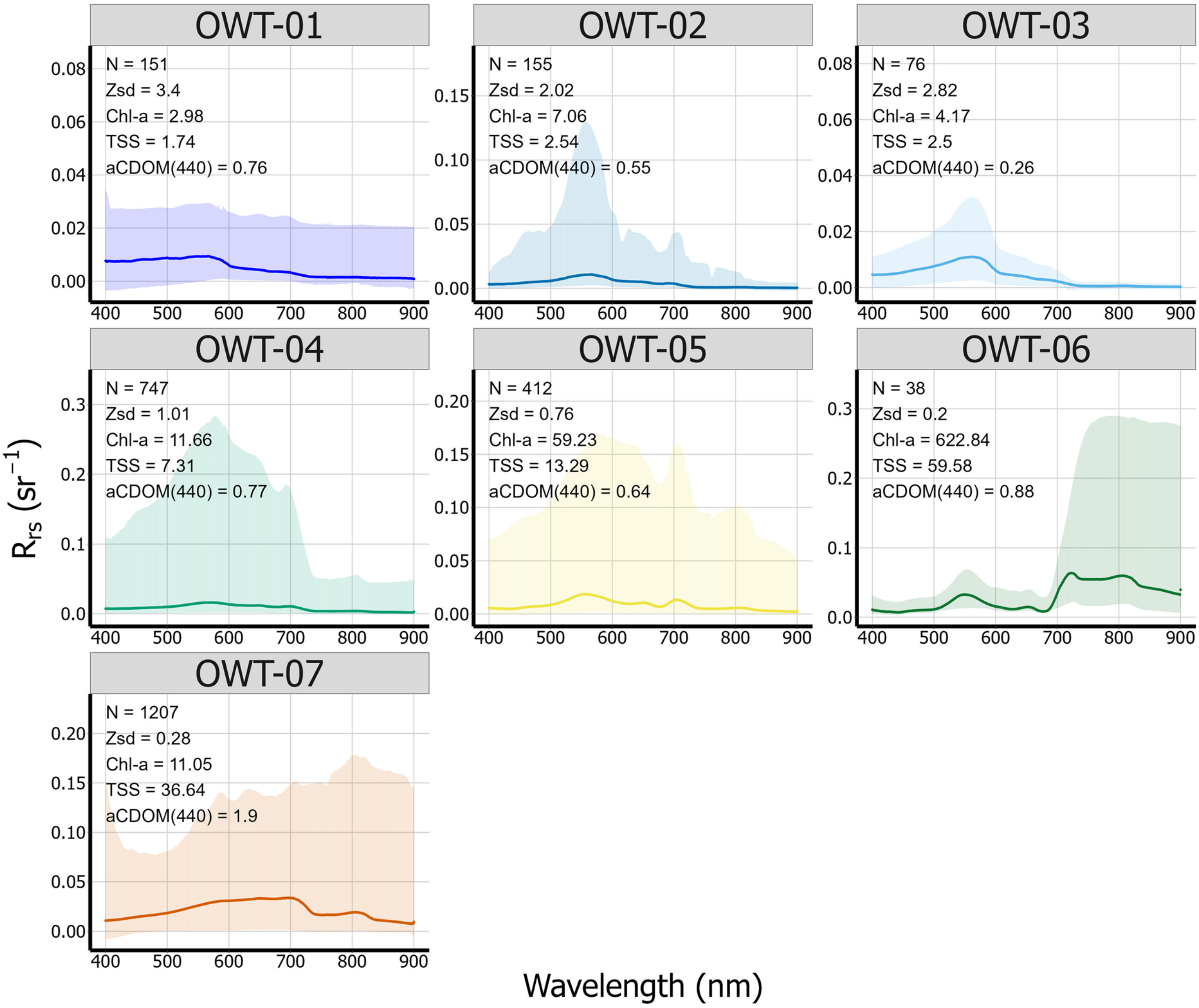

Fig. 5 Variability of R_{rs} for each OWT according Pahlevan *et al.*[45] classification. Please, note that y-axis has different scale for improving the visualization.

Technical Validation

Quality control. We assessed the quality of each spectrum in the dataset, as well as the WQ variables measured by using visual inspection and other metrics described below. Similar to the GLORIA[32] dataset, we keep all Rrs data allowing users to decide whether to use the flagged data. For the R_{rs} assessment, we provided a table ("*flags.csv*") with the flagged data in each of our assessment (Fig. 2a). The first flag is for negative values (Fig. 2b), which could happen due to uncertainties in glint correction (i.e., large L_{sky} values) or the low water signal at longest wavelengths. Among the 685 flagged spectra (25% of the dataset), at least one wavelength had negative values whose frequency increased at wavelengths larger than 700 nm. Dataset users should be cautioned to use or simply remove it from further analysis. We also flagged specific spectra by visual inspection. Unusually (i.e., larger values in the blue in high turbid waters) were flagged. That was manually done by three different specialists to ensure consistency (Fig. 3). An example of some flagged spectra is shown in the Fig. 3. We also flagged using the GLORIA dataset flags "*baseline_shift*", "*oxygen_signal*", and "*noyse_red_edge*". These flags were calculated in R programming language, using the scripts provided by Maciel *et al.*[61], which was also used in the GLORIA dataset to ensure consistency. Detailed explanation of these flags is available in[32]. Another flag used was "*suspicious*". We used this flag by visually inspection of each spectrum by specialized researchers to use or remove this data from further analysis. Spectra that did not seem to be realistic (e.g, high TSS concentration and very low Rrs values were flagged). Spectra that did not seem to be realistic (e.g, high TSS concentration and very low R_{rs} values were flagged).

In addition to the previous metrics, we used a quantitative score called *Quality Water Index Polynomial* (QWIP)[62] to assess the quality of the retrieved R_{rs} spectra. The QWIP is based on the Apparent Visible Wavelength (AVW), a one-dimensional metric of colour that is correlated to the spectral shape based on a weighted harmonic mean for visible wavelengths. This approach allowed us to identify R_{rs} data that falls outside a general trend observed for optically deep waters, ranging from clear ocean to turbid environments. The QWIP is based on a fourth-order polynomial, that was fitted between AVW and a Normalized Difference Index (NDI) between 492 and 665 nm bands. The $QWIP_{score}$ is the subtraction of the predicted QWIP based on the ANW (i.e.,

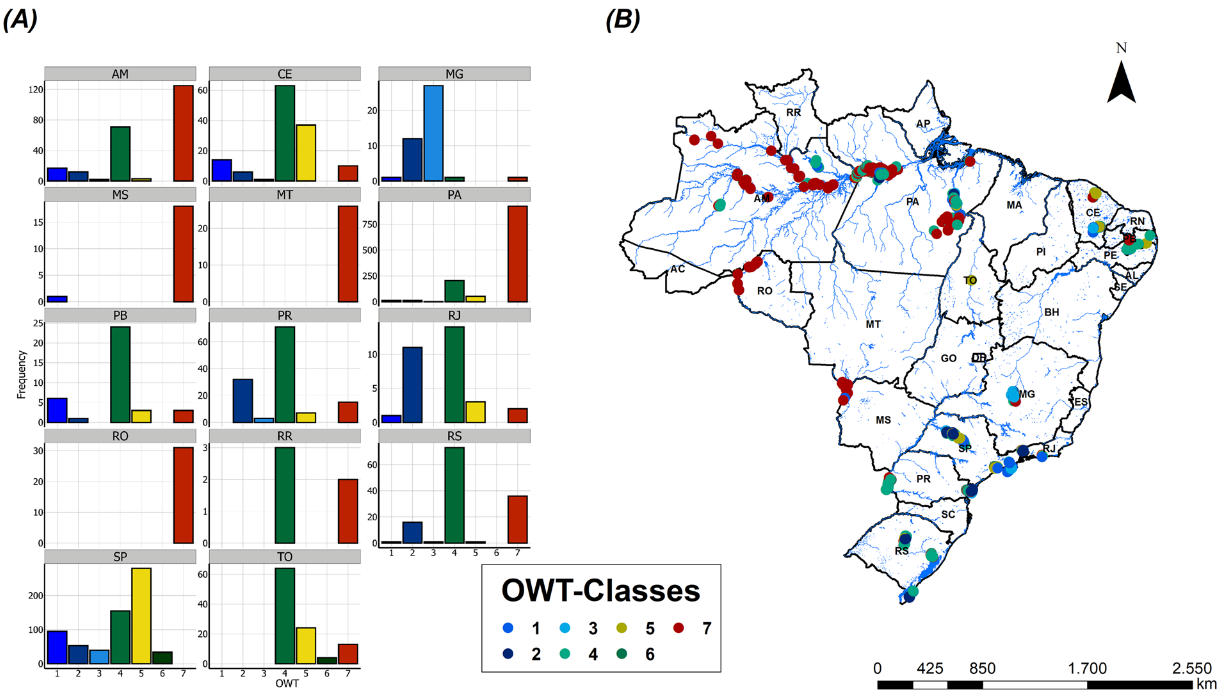

Fig. 6 (**A**) Frequency distribution of OWT's per state. Each colour represents its specific class based on x-axis. (**B**) Mapping with the locations of the OWT's.

the fourth order polynomial) and the real NDI ratio between 492 and 665 nm bands for a given spectrum. The process will ultimately provide a measure of deviation between the measured spectra and the central tendency of the ratio. If the difference between the NDI and the QWIP exceeds ± 0.2, the spectrum will be flagged. In our dataset, 669 spectra were flagged, and the QWIP was included in the flags.csv table.

R_{rs} variability. The compiled *in-situ* dataset represents the bio-optical properties across all five regions of Brazil (Fig. 4). In the dataset, the largest R_{rs} sample size is from Pará state (North Brazil) (N = 1,209 points), primarily from Curuai Lake (N = 929), which has been the subject of study for over 20 years by the LabISA team and other collaborators, involving more than ten field missions for data collection. Following Pará, São Paulo state (Southeast) has 750 unique stations, with data collected over more than 15 years in several reservoirs (i.e., Ibitinga, Promissão) and coastal waters of São Sebastião Channel. The dataset for São Paulo encompasses various eutrophic reservoirs (e.g., Billings (N = 353), Ibitinga (N = 89), and Promissão (N = 66)), clear water environments (e.g., Paraíbuna Reservoir (N = 13)), and coastal waters (São Sebastião Channel (N = 172)). This diverse range of environments, including rivers, reservoirs, floodplains, and coastal waters, allows for the creation of a database with significant optical variation. After Pará and São Paulo, the third most representative state is Amazonas state (North), with 230 unique stations, followed by Rio Grande do Sul (South Brazil) state, with 143 valid measurements, Ceará (N = 131) (Northeast), Paraná (N = 128) (South), and Tocantins (TO) (N = 105) (North) stands as states with N > 100. Minas Gerais State (MG) (N = 42) (Southeast), Rio de Janeiro (RJ) (N = 39) (Southeast), Paraíba (PB) (N = 37) (Northeast), Rondônia (RO) (North) (N = 31), Mato Grosso (MT) (N = 26) (Central-west), Mato Grosso do Sul (N = 19) and Roraima (RR) (N = 5) (North) presented a lower sample size.

The variation in the R_{rs} spectra is evident and further illustrated by the variability of the OACs. Despite this, we can see that the spectra capture high turbidity (e.g., the Amazon Region), eutrophic (e.g., RJ and SP states), and clear water environments (Três Marias reservoir in MG state). High sediment concentrations were observed in the Amazon Basin stations (e.g., Pará state, median TSS = 25.5 mgL^{-1}). Water transparency was lowest for the Pará stations (median Z_{sd} = 0.36 m), and the highest was obtained for MG state (Z_{sd} = 3.08 m). The aCDOM(440) was higher for the AM state (3.07 m^{-1}), much connected to the waters of Mamirauá Várzea[63] and Negro River[39,64].

The variability in R_{rs} observed in the state-grouped dataset is also illustrated in Fig. 5, plotted according to each respective OWTs (see Section 2.1). The large number of stations were classified as OWT-7, which is related to the turbid waters of Amazon Basin (e.g., Pará and Amazonas states) (N = 1,207). The second most common class was OWT-4, that represents medium turbid environment (N = 747). As commented, it is observed by the values of the OACs (e.g., mean Z_{sd} = 1.01 m, and mean TSS = 7.31 mgL^{-1} for OWT-4 versus mean Z_{sd} = 0.28 and mean TSS = 36.64 mgL^{-1} for OWT-7). It was followed by OWT-5 (N = 412), OWT-2 (N = 155), OWT-1 (N = 151), OWT-3 (N = 76) and OWT-6 (N = 38). Note that some stations were not classified as the wavelengths measured were only between 400–700 nm.

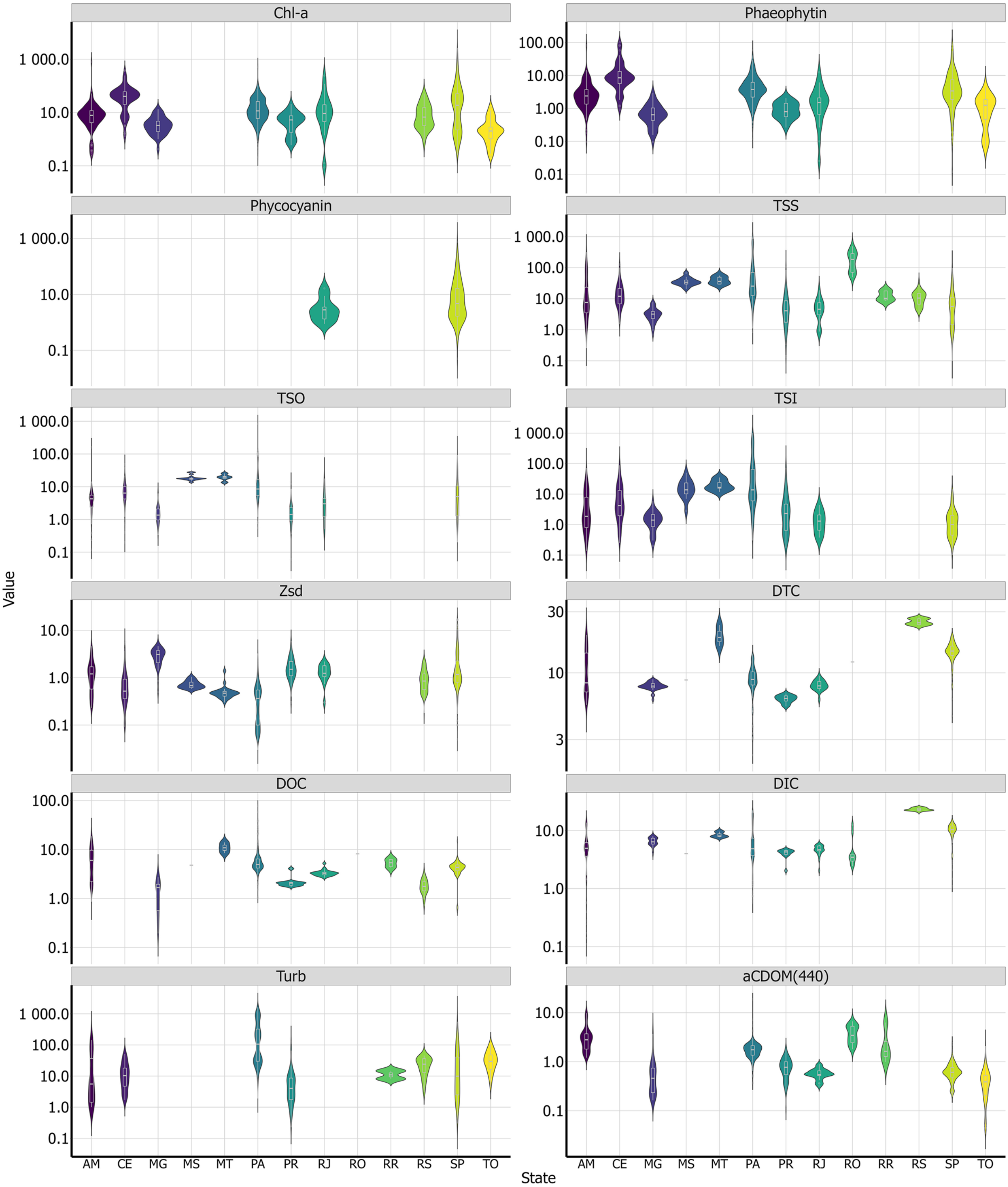

Fig. 7 Violin plot of water quality parameters variations under the dataset for each state. For the units of these measurements, please see Section 2.2.2.

This variability is also observed in the frequency of the OWTs per state (Fig. 6). For AM, OWT-4 and OWT-7 were the most available, indicating turbid environments, as expected. For MG data, OWT-3 was the most common, linked to the clear waters of the Três Marias reservoir (the unique sample location in this state). For Pará state, it followed what was obtained for Amazonas, with the most common OWTs being OWT-7 and OWT-3. For Paraná, Rio de Janeiro, Rio Grande do Sul and Roraima datasets, the most common OWT was OWT-4, indicating medium turbid locations. For SP, OWTs 3 and 5 were the most frequent, connected to the coastal waters of São Sebastião channel and eutrophic (OWT-5 and OWT-6) reservoirs in Billings, Ibitinga, Promissão and Nova Avanhandava.

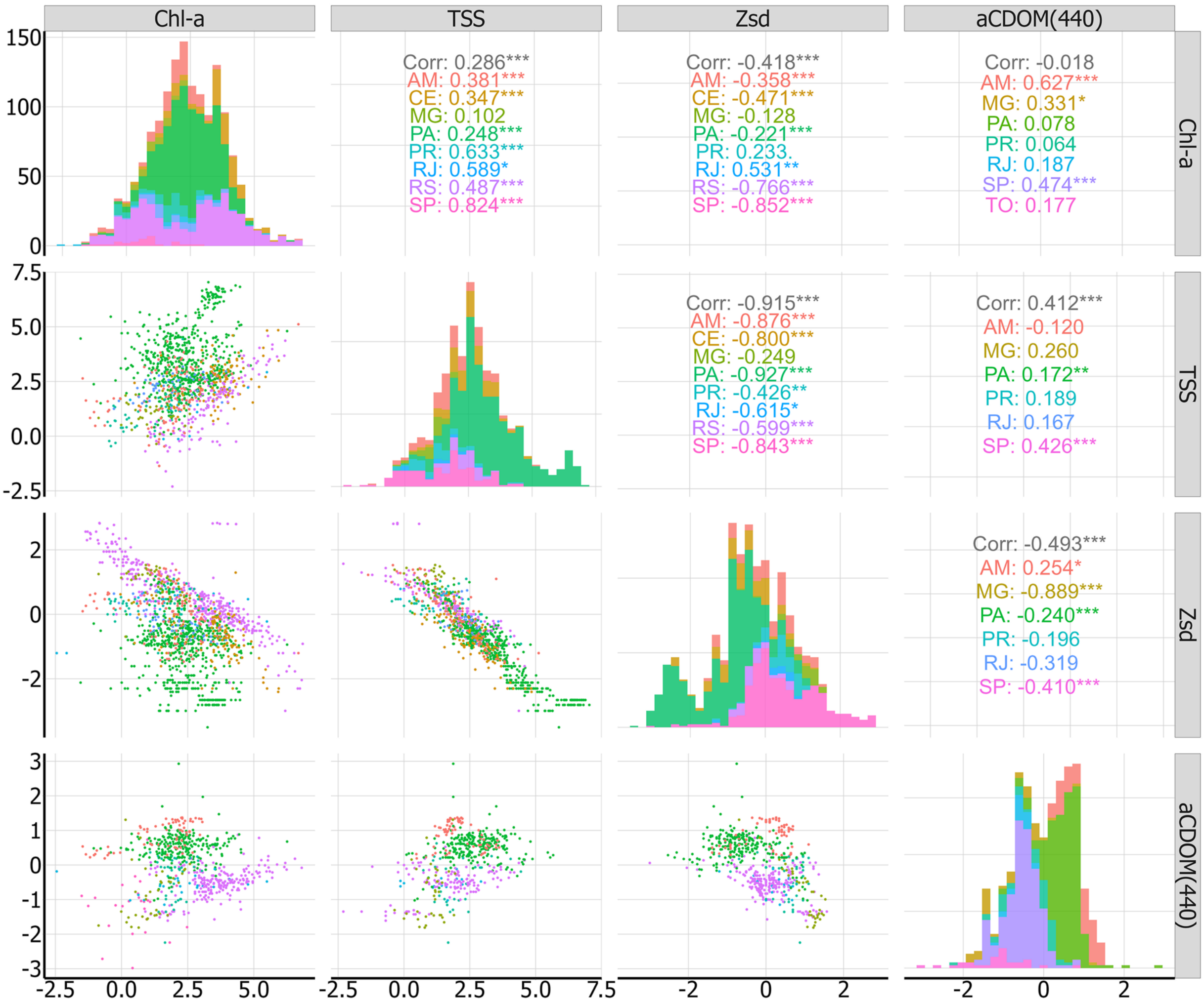

Fig. 8 Relationships between the Chl-*a*, TSS, Z_{sd} and aCDOM(440) in the available BRAZA dataset. Note that the values in this figure are in log-scale.

Water quality parameters variation. The variability of the WQ parameters is shown in Fig. 7. It is possible to observe that the Chl-*a* varies widely (~0 to up to 1,000 μgL^{-1}), but remains mainly between 1–100 μgL^{-1}, with the highest values obtained for the SP state, mostly in the Billings reservoir. Suspended Sediment presented its highest values for PA state, in the Curuai Lake (up to 1000 mg L^{-1}), which was also observed for TSI and TSO. Z_{sd} was higher (in average) for MG state (Três Marias reservoir), although the highest observed values were for the Alcatrazes Archipelago (coastal waters of the SP state, with Z_{sd} maximum of 18 m). DTC and DIC were highest for the RS data, on Lake Mangueira. DOC highest values were obtained in the AM dataset, mainly related to the Negro River (organic-matter rich area). It was also followed by the results obtained for aCDOM (440), in which the highest values were obtained for AM and PA data. Turbidity also followed results of TSS. Other validation performed was the analisis of the relationships between the WQ variables (Fig. 8). We identified a strong negative correlation between total suspended solids (TSS) and the Z_{sd} in all states, with a correlation coefficient of $R = -0.908$, which was expected. Additionally, a significant relationship was found between Chl-*a* and Zsd in certain states, such as São Paulo, where Chl-*a* significantly contributes to the absorption budgets in eutrophic reservoirs. Furthermore, both the aCDOM(440) and Z_{sd} exhibited a high negative correlation for the state of Minas Gerais (MG), with a correlation coefficient of $R = -0.889$.

Code availability

The scripts used for quality control of the data, as well as processing and generating all the figures of this paper, are available in Zenodo[65].

Received: 19 March 2025; Accepted: 10 July 2025;
Published online: 21 July 2025

References

1. MAPBIOMAS. MAPBIOMAS v. 4.1. (2020).
2. Getirana, A., Libonati, R. & Cataldi, M. Brazil is in water crisis — it needs a drought plan. *Nature* **600**, 218–220 (2021).

3. de Mello, K. *et al*. Multiscale land use impacts on water quality: Assessment, planning, and future perspectives in Brazil. *J Environ Manage* **270**, 110879 (2020).
4. Tundisi, J. G., Goldemberg, J., Matsumura-Tundisi, T. & Saraiva, A. C. F. How many more dams in the Amazon. *Energy Policy* **74**, 703–708 (2014).
5. Lobo, F. D. L., Costa, M. P. F. & de Moraes Novo, E. M. L. Time-series analysis of Landsat-MSS/TM/OLI images over Amazonian waters impacted by gold mining activities. *Remote Sens Environ* **157**, 170–184 (2015).
6. CETESB. InfoAguas. (2020).
7. ANA. Hidroweb: Hydrological Information System. http://hidroweb.ana.gov.br (2024).
8. Slabon, A. & Hoffmann, T. Uncertainties of Annual Suspended Sediment Transport Estimates Driven by Temporal Variability. *Water Resour Res* **60**, e2022WR032628 (2024).
9. Sagan, V. *et al*. Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth Sci Rev* **205**, 103187 (2020).
10. Kuhn, C. *et al*. Performance of Landsat-8 and Sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-a and turbidity. *Remote Sens Environ* **Accepted**, 104–118 (2019).
11. Pahlevan, N. *et al*. Simultaneous retrieval of selected optical water quality indicators from Landsat-8, Sentinel-2, and Sentinel-3. *Remote Sens Environ* **270**, 112860 (2022).
12. Maciel, D. A., Barbosa, C. C. F., de Moraes Novo, E. M. L., Flores Júnior, R. & Begliomini, F. N. Water clarity in Brazilian water assessed using Sentinel-2 and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing* **182**, 134–152 (2021).
13. Giardino, C. *et al*. First evaluation of PRISMA Level 1 data for water applications. *Sensors* (**SUBMITTED**, (2020).
14. O'Shea, R. E. *et al*. Advancing cyanobacteria biomass estimation from hyperspectral observations: Demonstrations with HICO and PRISMA imagery. *Remote Sens Environ* **266** (2021).
15. Bresciani, M. *et al*. Application of New Hyperspectral Sensors in the Remote Sensing of Aquatic Ecosystem Health: Exploiting PRISMA and DESIS for Four Italian Lakes. (2022).
16. Mobley, C. D. *Light and Water: Radiative Transfer in Natural Waters* (Academic press, 1994).
17. Preisendorfer, R. W. *Hydrologic Optics. Volume 1. Introduction*. (1976).
18. Mobley, C. D., Boss, E. S. & Roesler, C. Ocean optics web book. *OOWB/NASA* (2010).
19. Austin, R. W. & Petzold, T. J. The determination of the diffuse attenuation coefficient of sea water using the coastal zone color scanner. *Austin, R. W Petzold, T. J.* **112**, 211–212 (1981).
20. Fassoni-Andrade, A. C. *et al*. Amazon hydrology from space: scientific advances and future challenges. *Reviews of Geophysics* 1–97 https://doi.org/10.1029/2020rg000728 (2021).
21. Lee, Z.-P. *et al*. Secchi disk depth: A new theory and mechanistic model for underwater visibility. *Remote Sens Environ* **169**, 139–149 (2015).
22. Pahlevan, N. *et al*. Seamless retrievals of chlorophyll- a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. *Remote Sens Environ* **2**, 111604 (2019).
23. Lee, Z.-P., Carder, K. L. & Arnone, R. A. Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep waters. *Appl Opt* **41**, 5755 (2002).
24. Chen, X. *et al*. Long-Term water clarity patterns of lakes across China using Landsat series imagery from 1985 to 2020. *Hydrol Earth Syst Sci* **26**, 3517–3536 (2022).
25. Zhang, Y. *et al*. Improving remote sensing estimation of Secchi disk depth for global lakes and reservoirs using machine learning methods. *GIsci Remote Sens* **59**, 1367–1383 (2022).
26. Lee, Z.-P., Du, K. P. & Arnone, R. A model for the diffuse attenuation coefficient of downwelling irradiance. *Journal of Geophysical Research C: Oceans* **110**, 1–10 (2005).
27. Sander de Carvalho, L. A., Faria Barbosa, C. C., de Moraes Novo, E. M. L. & de Moraes Rudorff, C. Implications of scatter corrections for absorption measurements on optical closure of Amazon floodplain lakes using the Spectral Absorption and Attenuation Meter (AC-S-WETLabs). *Remote Sens Environ* **157**, 123–137 (2015).
28. Lavigne, H. *et al*. The HYPERMAQ dataset: bio-optical properties of moderately to extremely turbid waters. *Earth Syst Sci Data* **14**, 4935–4947 (2022).
29. Castagna, A. *et al*. Optical and biogeochemical properties of diverse Belgian inland and coastal waters. *Earth Syst Sci Data* **14**, 2697–2719 (2022).
30. Drayson, N. *et al*. Australian aquatic bio-optical dataset with applications for satellite calibration, algorithm development and validation. *Data Brief* **44**, 108489 (2022).
31. Knaeps, E. *et al*. The SeaSWIR dataset. *Earth Syst Sci Data* **10**, 1439–1449 (2018).
32. Lehmann, M. K. *et al*. GLORIA - A globally representative hyperspectral *in situ* dataset for optical sensing of water quality. *Sci Data* **10** (2023).
33. Zhai, M. *et al*. Satellite-ground synchronous *in-situ* dataset of water optical parameters and surface temperature for typical lakes in China. *Sci Data* **11**, 1–14 (2024).
34. Ross, M. R. V. *et al*. AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters. *Water Resour Res* **55**, 10012–10025 (2019).
35. Loisel, H., Jorge, D. S. F., Reynolds, R. A. & Stramski, D. A synthetic optical database generated by radiative transfer simulations in support of studies in ocean optics and optical remote sensing of the global ocean. *Earth Syst Sci Data* **15**, 3711–3731 (2023).
36. Pitarch, J. & Brando, V. E. A hyperspectral and multi-angular synthetic dataset for algorithm development in waters of varying trophic levels and optical complexity. *Earth Syst Sci Data* **17**, 435–460 (2025).
37. Silva, E. F. F. da *et al*. A machine learning approach for monitoring Brazilian optical water types using Sentinel-2 MSI. *Remote Sensing Applications: Society and Environment* vol. 23 https://doi.org/10.1016/j.rsase.2021.100577 (2021).
38. Barbosa, C. C. F. *et al*. Brazilian inland water bio-optical dataset to support carbon budget studies in reservoirs as well as anthropogenic impacts in Amazon floodplain lakes: Preliminary results. *International Archives of the Photogrammetry, Remote Sensing and Spatial. Information Sciences - ISPRS Archives* **40**, 1439–1446 (2015).
39. Marinho, R. R. *et al*. Estimating the Colored Dissolved Organic Matter in the Negro River, Amazon Basin, with *In Situ* Remote Sensing Data. *Remote Sens (Basel)* **16**, 613 (2024).
40. Lima, T. M. A. de *et al*. Assessment of Estimated Phycocyanin and Chlorophyll-a Concentration from PRISMA and OLCI in Brazilian Inland Waters: A Comparison between Semi-Analytical and Machine Learning Algorithms. *Remote Sens (Basel)* **15** (2023).
41. Curtarelli, V. P. *et al*. Diffuse attenuation of clear water tropical reservoir: A remote sensing semi-analytical approach. *Remote Sens (Basel)* 1–23 (2020).
42. Maciel, D. *et al*. Towards global long-term water transparency products from the Landsat archive. *Remote Sens Environ* **299**, 113889 (2023).
43. Zibordi, G. *et al*. *In situ* determination of the remote sensing reflectance: an inter-comparison. *Ocean Sci* **8**, 567–586 (2012).
44. Alikas, K. *et al*. Comparison of Above-Water Seabird and TriOS Radiometers along an Atlantic Meridional Transect. *Remote Sensing* **12**, 1669 (2020). *2020, Vol. 12, Page 1669*.
45. Pahlevan, N. *et al*. ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters. *Remote Sens Environ* **258** (2021).

46. Morel, A. & Maritorena, S. Bio-optical properties of oceanic waters: A reappraisal. *J Geophys Res Oceans* **106**, 7163–7180 (2001).
47. Mobley, C. D. Polarized reflectance and transmittance properties of windblown sea surfaces. *Appl. Opt.* **54**, 4828–4849 (2015).
48. Spyrakos, E. *et al*. Optical types of inland and coastal waters. *Limnol Oceanogr* **63**, 846–870 (2018).
49. Clark, R., Swayze, G., Boardman, J. & Kruse, F. Comparison of Three methods for Material Identification and mapping. *Summaries of 4th Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop* (1993).
50. APHA. Standard Methods for Examination of Water and Wastewater (Standard Methods for the Examination of Water and Wastewater). *Standard Methods* 5–16 ISBN 9780875532356 (1998).
51. Sarada, R., Pillai, M. G. & Ravishankar, G. A. Phycocyanin from Spirulina sp: influence of processing of biomass on phycocyanin yield, analysis of efficacy of extraction methods and stability studies on phycocyanin. *Process Biochemistry* **34**, 795–801 (1999).
52. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, 1–20 (2013).
53. Bennett, A. & Bogobad, L. Complementary chromatic adaptation in a filamentous blue-green alga. *J Cell Biol* **58**, 419–435 (1973).
54. Wetzel, R. G. & Likens, G. E. *Limnological Analysis*. (Springer Science & Business Media, 2013).
55. APHA. Standard methods for the examination of water and waste water. *Am J Public Health Nations Health* https://doi.org/10.2105/AJPH.56.3.387 (1999).
56. Bricaud, A., Morel, A. & Prieur, L. Absorption by dissolved organic matter of the sea (yellow substance) in the UV and visible domains. *Limnol Oceanogr* **26**, 43–53 (1981).
57. Green, S. A. & Blough, N. V. Optical absorption and fluorescence properties of chromophoric dissolved organic matter in natural waters. *Limnol Oceanogr* **39**, 1903–1916 (1994).
58. Pitarch, J. A review of secchi's contribution to marine optics and the foundation of secchi disk science. *Oceanography* **33**, 26–37 (2020).
59. Qian, J. & Mopper, K. Automated High-Performance, High-Temperature Combustion Total Organic Carbon Analyzer. *Anal Chem* **68**, 3090–3097 (1996).
60. Maciel, D. BRAZA - a bio-optical database for the remote sensing of water quality in BRAZil coAstal and inland waters. *figshare* https://doi.org/10.6084/m9.figshare.28566197 (2025).
61. Maciel, D. A., Lehmann, M., Gurlin, D. & Pahlevan, N. R code for GLORIA quality control flags. *Zenodo* https://doi.org/10.5281/ZENODO.7372445 (2022).
62. Dierssen, H. M. *et al*. QWIP: A Quantitative Metric for Quality Control of Aquatic Reflectance Spectral Shape Using the Apparent Visible Wavelength. *Frontiers in Remote Sensing* **3**, 869611 (2022).
63. Da Silva, M. P., De Carvalho, L. A. S., Novo, E., Jorge, D. S. F. & Barbosa, C. C. F. Use of optical absorption indices to assess seasonal variability of dissolved organic matter in Amazon floodplain lakes. *Biogeosciences* **17**, 5355–5364 (2020).
64. Marinho, R. R. & Zanin, P. R. & Filizola Junior, N. P. The Negro River in the Anavilhanas Archipelago: Streamflow and geomorphology of a complex anabranching system in the Amazon. *Earth Surf Process Landf* **47**, 1108–1123 (2022).
65. Maciel, D. dmaciel123/BRAZA: Version 0.1. https://doi.org/10.5281/zenodo.15690038 (2025).

Acknowledgements
This research is funded through the 2017-2018 Belmont Forum and BiodivERsA joint call for research proposals, under the BiodivScen ERANet COFUND program, and with the funding organizations French National Research Agency (ANR), São Paulo Research Foundation and National Science Foundation (NSF), the Research Council of Norway and the German Federal Ministry of Education and Research (BMBF). This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number #2013/09045-7; #2018/12083-1, #2020/14613-8, #2021/13367-6, 2022/08775-0, and #2023/13904-7), by the Coordenação Nacional de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001, by Fundação de Amparo à Pesquisa do Estado do Amazonas (Edital Universal FAPEAM 20 anos) and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant numbers 482605/2013-8, 140505/2020-2 and 444756/2024-8. This study was also funded in part by Paraíba State Research Support Foundation / FAPESQ, Amazon + 10 Initiative - São Paulo Research Foundation (FAPESP 2022/10443-6), Rondônia Research Foundation (FAPERO), and the Tocantins Research Foundation (FAPT). We thank ITV for supporting field data collections in Pará state. It was also partially funded by the Distrito Federal Research Foundation (FAPDF; grant number 00193-00001379/2024-02). Lino Sander would like to thank FAPERJ projects: Temático - 210.078/2023, APQ1 - e-26/210.202/2018, JCNE - E-26/201.406/2022, and the Sistema de Monitoramento da Costa Brasileira (SiMCosta) for the scientific funding. The authors also thank the Brazilian Space Agency (AEB) for supporting this research. Authors are in debt with Jean Ommeto (INPE) and José Etham de Lucena Barbosa (UEPB) for coordinating the projects that generated the *in-situ* data for Paraíba state. We also thank the AQUASENSE project (http://aquasense.igd.unb.br/) for providing the hydrological and water-quality data collected in the Madeira River basin. The authors also thank all the local people, boat drivers, crew members, students, and other people that were indispensable during the field mission collection. We would like to thank the anonymous reviewers for their careful evaluation and insights provided in the review process.

Author contributions
D. A. M. wrote the manuscript, contributed to data collection and organization, applied the quality control of the data, organize the data acquisition and the wrote the manuscript. C. B. and E. N. were responsible for project management, data maintenance and writing of the manuscript. D.A.M, J. C. P. S. and R. P. were responsible for the quality control of the dataset. All the other authors were responsible for data collection and sample organization.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to D.A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.