https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

Coastal North America In Situ Chlorophyll and Optical Dataset

Gianna Milton[1, 2], Kimberly Hyde[1], Ivona Cetinic[3, 4], Maxwell Beal[3, 5], Ryan Vandermeulen[6], Haley Synan[1, 2]

[1] Northeast Fisheries Science Center, National Marine Fisheries Service, 28 Tarzwell Drive, Narragansett, RI, USA
[2] IBSS Corp, 1110 Bonifant St. Suite #315, Silver Spring, MD, USA
[3] NASA Goddard Space Flight Center, Greenbelt, Maryland, USA
[4] GESTAR II, Morgan State University, Baltimore, Maryland, USA
[5] Science Systems and Applications Inc., 10210 Greenbelt Road, Suite 600, Lanham, MD 20706
[6] NOAA Fisheries Office of Science and Technology, 1315 East-West Highway, 12th Floor Silver Spring MD 20910

Correspondence to: Gianna Milton (giannamilton@gmail.com)

Abstract. Chlorophyll is a crucial climate variable and remotely sensed satellite chlorophyll estimates are essential to understanding large-scale biological processes. Current ocean color chlorophyll algorithms, however, are less accurate in optically complex inland and coastal waters. As a result, remote sensing algorithms are not able to fully resolve chlorophyll concentrations in optically complex coastal regions leading to greater uncertainty in regions of high productivity. This can further lead to errors in regional biogeochemical models and hinder fisheries and ecosystem monitoring and research. In order to develop more accurate coastal chlorophyll retrievals, *in situ* chlorophyll measurements are needed for algorithm calibration and validation. Here the methods used to develop a single cohesive standardized *in situ* chlorophyll, chromophoric dissolved organic matter (CDOM), and remote sensing reflectance (R_{rs}) dataset from multiple public data repositories are described. The final product is a North American dataset of quality controlled chlorophyll measurements and associated metadata. The dataset includes more than 700,000 chlorophyll values with coincidental CDOM and R_{rs} data when available. Methodological flags such as chlorophyll measurement method, triplicate samples, and *in vitro/in vivo* flags distinguish between data collection methods. The dataset includes regularly formatted metadata across all sources to ensure due credit to the programs collecting the samples. The resulting dataset is then matched to satellite chlorophyll data to examine the difference in chlorophyll retrievals across multiple water types and regions. This dataset is currently being used to validate a new coastal chlorophyll algorithm in development and to assess chlorophyll retrievals. All methods and data are publicly available at doi.org/10.5281/zenodo.20140178 (Milton et al., 2026).

1 Introduction

Chlorophyll a (hereinafter referred to as chlorophyll) is the primary pigment in phytoplankton and is commonly used as a proxy for phytoplankton biomass (O'Reilly et al., 1998). This proxy is used to assess global changes and trends in primary production, nutrient uptake, and carbon export (Brewin et al., 2017). Phytoplankton in coastal environments are influenced by

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

nutrient inputs from runoff waters, coastal upwelling, light availability and ocean temperature, grazers and viruses, and as primary producers of aquatic ecosystems are foundational to the food web in many coastal areas (Nezlin et al., 2018). As such, chlorophyll concentrations are positively correlated with fisheries yields and are a key variable when researching coastal environments and ecosystem dynamics (Friedland et al., 2012). Chlorophyll concentrations tend to be higher in coastal zones due to greater nutrient supplies, mixing, and more productivity (Dai et al., 2023). Despite making up ~22% of global ocean waters by area, ~95% of marine resources and commercial fishing occurs in coastal marine ecosystems due in part due to the high phytoplankton biomass and corresponding chlorophyll values (Jänes et al., 2020, Stock et al., 2017, Van Oostende et al., 2018).

Chlorophyll concentration can be remotely estimated from ocean color satellites, delivering reliable surface chlorophyll a concentration for the majority of open ocean waters. Chlorophyll estimates on such large spatial and temporal scales are possible through the use of bio-optical algorithms applied to the remote sensing reflectance (Rrs), which use Rrs ratios to calculate the corresponding chlorophyll concentrations in surface waters (Hu et al., 2019, O'Reilly & Werdell, 2019). The spatial and temporal resolutions of these remote sensing based estimates of phytoplankton biomass are helpful to assessing large-scale oceanographic interactions within fisheries science, and aiding researchers in understanding and monitoring seasonal and long-term change in chlorophyll and phytoplankton (Hyde et al., 2007).

While satellites can accurately estimate open ocean chlorophyll concentrations, this accuracy is reduced in coastal waters (Wei et al., 2022). Coastal regions are more optically complex than open ocean waters due to sediment and chromophoric dissolved organic matter (CDOM) transport and turbidity caused by mixing in shallow waters. High concentrations of CDOM and non-algal particles can create bias in empirical chlorophyll algorithms due to overlapping spectral signals (Morel & Prieur, 1977). As distance from the coast increases, satellite-based chlorophyll estimates tend to improve. This is due to the dispersal of terrestrial material and a stronger coupling of optical variability with phytoplankton biomass. Creating an algorithm to correct the optical biases introduced by high sediments and CDOM concentrations becomes more complex when considering the broad range of coastal environments in North America, as each region has unique coastal dynamics (Dai et al., 2023). For example, the coastal zone in the Northeastern United States is more optically complex with more turbid waters than the California coastline (Mélin & Vantrepotte, 2015). An algorithm tailored towards correctly estimating chlorophyll from a wide range of coastal environments and optical water types is needed to account for this deficit.

Algorithms can be used to improve chlorophyll retrieval from satellites, but they need *in situ* data for validation and training; acting as ground truth for the algorithms to be built off of (O'Reilly & Werdell, 2019). *In situ* data with high quality assurance and quality control (QA/QC) is more valuable than raw, unprocessed data and is held to higher standards for satellite validation as it can more accurately show the discrepancy between satellite and *in situ* values. However, it is difficult to aggregate and organize *in situ* chlorophyll measurements from multiple projects and dataset repositories, as projects can use different sampling methods and different metadata and file formats. These inconsistencies compound and make the creation of a single, formatted dataset cumbersome and time intensive, but no less important to the creation or validation of an algorithm. By combining and standardizing multiple public databases into a single, easily accessible dataset, algorithm developers can

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

more easily integrate the *in situ* samples into algorithms and models, resulting in better performance and more accurate chlorophyll retrievals (Lehmann et al., 2023, O'Reilly & Werdell, 2019, Valente et al., 2016).

This manuscript presents the methods used to develop the coastal optical dataset, a merged dataset of *in situ* chlorophyll, CDOM, and Rrs data compiled from 11 unique public oceanographic data repositories, manuscripts, and research programs. The primary purpose of the dataset is to create a comprehensive set of chlorophyll measurements to aid in the development of a multi-sensor coastal chlorophyll algorithm. Existing multi-sensor datasets rely on *in situ* reference datasets to inform empirical components of chlorophyll-a algorithms. The widely used Ocean Colour Climate Change Initiative (OC-CCI) merged chlorophyll product uses a global *in situ* dataset developed and updated by Valente et al. (2022) to identify the algorithm best suited to each optical water type. While OC-CCI is produced globally, the dataset is focused primarily on case 1 waters (Sathyendranath et al., 2019). Thus there is a need to develop a chlorophyll dataset specifically for coastal waters.

This dataset was developed, in part, to support the development of a long-term satellite based chlorophyll dataset applicable to North American coastal waters. Aligning satellite radiometric inputs across ocean color sensors is key to development of an unbiased climate-scale chlorophyll timeseries. However, successful merging and chlorophyll algorithm selection depend on spatially and temporally representative *in situ* data to ensure any modifications improve product performance. Establishing a standardized *in situ* chlorophyll dataset allows for robust algorithm assessment over the 30-year satellite record, and can inform any necessary statistical corrections for coastal waters. However, this dataset can be used for other research purposes, such as quantifying current differences in satellite chlorophyll to *in situ* values and analysing how chlorophyll correlates to water classes.

2 Methods

Chlorophyll, CDOM, and Rrs measurements for this dataset were gathered from a variety of publicly available datasets, ranging from data repositories, individual projects, and data manuscripts. Data were locally downloaded or remotely fetched from these websites and subset to standard dates, depths, coordinates, and variables, before conducting QA/QC on each individual dataset depending on the refinement needed. Corresponding metadata were retained to ensure each sample was correctly cited without ambiguity to align with Findability, Accessibility, Interoperability, and Reuse (FAIR) metadata standards, including datetime, latitude, longitude, experiment, affiliations, investigators, DOI and contact when possible (Neeley et al., 2025). If samples did not have essential metadata values such as datetime, latitude, longitude, or experiment, they were not included (Milton et al., 2026).

All data were subset to North American waters by combining two shapefiles: The Marine Ecoregions from the North American Environmental Atlas (www.cec.org/north-american-environmental-atlas/marine-ecoregions/), which includes all Exclusive Economic Zones within North American countries, and the U.S. regional Fishery Management Councils Region Boundaries for more inland areas (www.arcgis.com/home/item.html?id=84cbbc49011b49e1b959a7b6a7a0d339) (Fig. 1). These shapefiles were selected due to these regions being well sampled, representing a wide range of optical conditions, and

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

to their connection to fisheries management regions. Subsetting measurements to these shapefiles ensures only relevant data were collected. This dataset does not include estuaries or rivers and instead focused on coastal waters. Additionally, only data within the top 150 m of the water column were retained in this dataset.

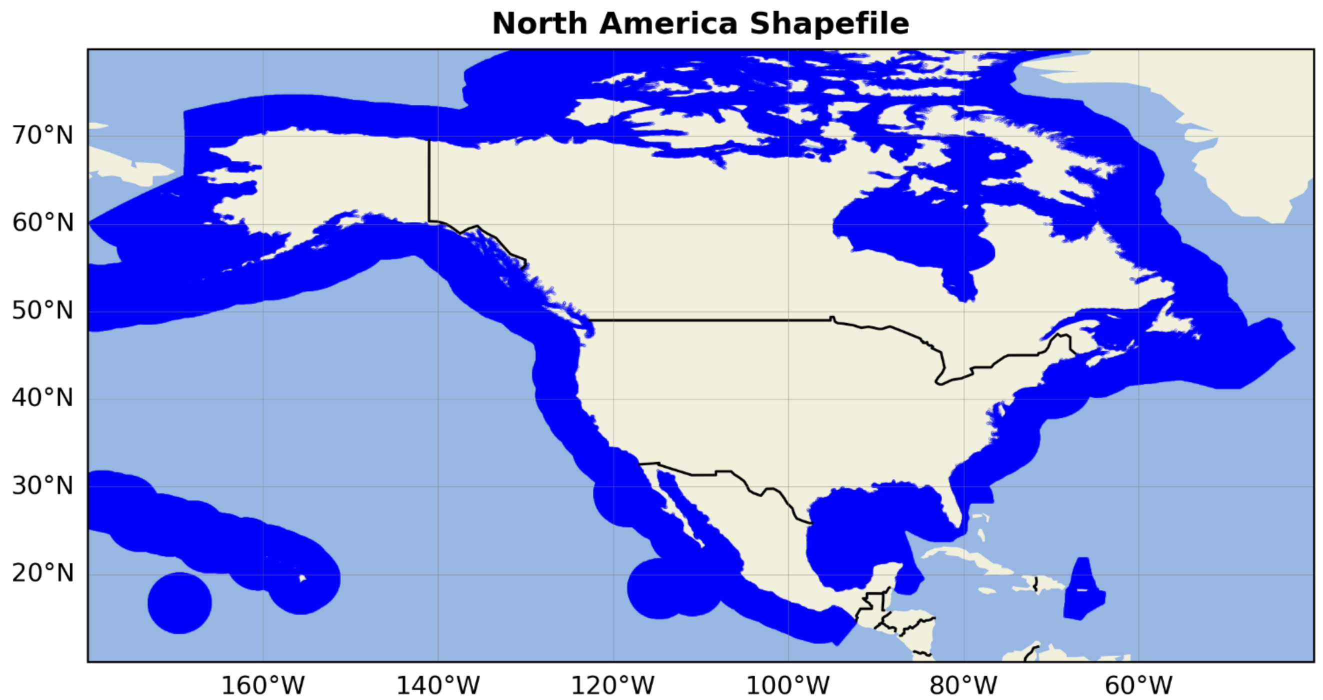


Figure 1: North American Coastal waters shapefile as described by The Marine Ecoregions from the North American Environmental Atlas and the U.S. regional Fishery Management Councils Region Boundaries. All chlorophyll, CDOM, and Rrs data measurements are subset to these boundaries.

2.1 Dataset Methods and Methodological Flags

In situ datasets used for algorithm development or remote sensing product validation should strive to have the highest quality data. Therefore, due to the diversity of methods in measuring chlorophyll, a flagging system was added to the dataset to aid in distinguishing between the different methods. First, chlorophyll can be measured discretely (in vitro), where samples of water are collected onto a filter and then processed using chemical analysis to estimate concentrations of chlorophyll. These methods include High Performance Liquid Chromatography (HPLC) (here referred to as chlorophyll_HPLC), fluorometric approaches, and spectrophotometric approaches. These methods are considered to be true measurements of chlorophyll concentrations. Conversely, chlorophyll fluorescence (in vivo) measurements are often used for estimation of phytoplankton biomass, but due to the physiological aspect of the measurements, should not be used for validation of development of the remote sensing algorithms. If a repository or dataset distinguished between in vivo and in vitro, then only in vitro chlorophyll values were retained. However, some large data repositories do not fully distinguish between these methods and it becomes difficult to parse out which method was used. Therefore, an additional flag, 'data_type_flag' was used for these datasets to

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

distinguish between chlorophyll measured in vitro and in vivo by inspecting the sample resolution of the measurements (Table 1, Fig. B3). The supplementary section details the flag application method (Appendix C).

Based on previous studies, superior chlorophyll validation data should be measured using HPLC methods, and collected in triplicates (Neeley et al., 2025). While HPLC is considered to have the highest accuracy, with high-quality benchmark standards and strict standardized protocols used across the field, triplicates also add a dimension of precision and allows for a view into the variability of the chlorophyll concentrations. And so all chlorophyll samples in this dataset have 2 additional flags: an HPLC flag to indicate if chlorophyll is measured with HPLC methods, and a triplicate flag to indicate if chlorophyll is measured as part of triplicates (Table 1). While many algorithm developers prefer triplicate HPLC measurements over standard fluorometric and triplicates over single samples, this can greatly reduce the number of available measurements as these methods are more expensive and time consuming. In this dataset only ~2.1% of the measurements used HPLC methods and only 0.6% are triplicates. Due to the low amount of triplicate and HPLC measurements, all methods were included and not reduced to ensure spatial and temporal coverage.

Some chlorophyll datasets did not indicate if samples were measured in triplicate, therefore unless otherwise noted, triplicates were determined by counting the number of samples recorded at unique datetime, location, and depth (Fig. B1). If three samples were taken at identical locations and times, then the triplicate flag was turned to True (0) (Table 1). The determination of HPLC flags depended on the source of the dataset as described in the Supplement (Appendix C, Fig. B2). It should be noted that some HPLC and triplicate flags come from an assumption that the data used these methods, not a guarantee. While some projects have extensive metadata and explicitly dictate methods, other repositories and older datasets lack comprehensive metadata records and do not include the type of analysis method used. The processing flags used in this dataset are based on dataset metadata assumptions and number of samples per station/depth, therefore the processing flags could contain errors when the metadata is sparse. If a user wants the highest quality chlorophyll data, then the flags should be subset to HPLC = 0, triplicate = 0, and data_type_flag ≠1 and 2. Data_type_flag should not be subset to 0 as this flag is only applied to large data repositories that do not distinguish between in vivo and in vitro, and is not applied to other datasets or manuscripts. If a user only wants extracted chlorophyll but has no preference towards HPLC or triplicate methods, then the dataset only needs to be subset to data_type_flag ≠1 and 2.

Table 1: Flag definitions for chlorophyll samples to distinguish methods used

Flags	**0**	**1**	**2**
HPLC	HPLC methods used	HPLC methods not used	HPLC methods potentially used but mislabeled (SeaBASS data only)
Triplicate	Triplicates recorded	Triplicates not recorded	-----------------------
data_type_flag	*In vitro* methods used	*In vivo* methods used	data_type indeterminate (SeaBASS data only)

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

The CDOM measurements gathered for this dataset are exclusively CDOM concentrations. Specifically, CDOM measured via fluorescence and not absorption. Fluorescence CDOM samples have concentration units of milligram/cubic meter (mg m^-3) typically derived *in situ* from quinine sulfate standards. This dataset does not distinguish between CDOM fluorescence methods as long as they are in the standard units (mg m^3). Other CDOM measurement methods such as absorption coefficient or CDOM with Relative Fluorescence Units (RFU) or meter^-1 (1/m) units were excluded. R_{rs} measurements range from multispectral to hyperspectral wavelengths depending on the project and are standardized such that the wavelength value is appended to the column name (Table C1). Only wavelengths from 305 to 800 nm at 1 nm intervals were saved and any fine scale wavelengths that measured in the tenths of nanometers were omitted (ex: wavelength = 465.1, 465.2). All chlorophyll and CDOM data are in units of mg m^-3, and all R_{rs} data are in 1/steradian (1/sr) with wavelengths in nm.

2.2 Sources

In situ chlorophyll measurements from 2000 through 2025 were gathered from 5 oceanographic data repositories: SeaWiFS Bio-optical Archive and Storage System (SeaBASS), The Integrated Ocean Observing System (IOOS), World Ocean Database (WOD), The Canadian Integrated Ocean Observing System (CIOOS), and Biological and Chemical Oceanography Data Management Office (BCO-DMO). Additionally, project specific chlorophyll data were aggregated from their unique project websites, including California Cooperative Oceanic Fisheries Investigations (CalCOFI), The Southeast Area Monitoring & Assessment Program (SEAMAP), the Hawaii Ocean Time Series (HOTS), Aquamatch, the Consolidated Arctic Pigments dataset (Pan-Arctic), and The GLObal Reflectance community dataset for Imaging and optical sensing of Aquatic environments (GLORIA), resulting in 11 unique data sources (Table 2). R_{rs} data were acquired from SeaBASS and GLORIA, and CDOM data were obtained from SeaBASS, IOOS, and BCO-DMO.

Table 2: Data sources with date, latitude, and longitude ranges, variables collected, number of raw samples (chl: non-HPLC chlorophyll, chl_HPLC: HPLC chlorophyll, rrs: remote sensing reflectance, cdom: chromophoric dissolved organic matter), and reference links ordered by number of chlorophyll samples.

Data Source	**Latitude Range (degrees North)**	**Longitude Range (degrees East)**	**Temporal Range**	**Variables (raw count)**	**Source url**
SeaBASS	-0.59 → 78.30	149.0 → -59.14	2000-01-07 → 2025-09-17	chl = 248,983 chl_HPLC = 5,333 rrs = 50,477 cdom = 523,704	https://seabass.gsfc.nasa.gov/
IOOS	19.73 → 59.60	-157.84 → -72.64	2000-01-11 → 2026-04-17	chl = 184,882 cdom = 5,041	https://sensors.ioos.us/

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

CalCOFI	29.83 → 37.85	-125.8 → -117.2	2000-01-07 → 2021-05-13	chl = 100,975	https://calcofi.org/data/oceanographic-data/bottle-database/
SEAMAP	17.93 → 30.37	-97.36 → -81.32	2000-06-13 → 2025-11-24	chl = 54,255	https://seamapdata.gsmfc.org/
WOD	17.07 → 73.97	-168.1 → -47.02	2000-01-04 → 2022-07-09	chl = 55,235	https://www.ncei.noaa.gov/access/world-ocean-database-select/dbsearch.html
AquaMatch	24.57 → 35.85	-118.8 → -75.49	2000-01-04 → 2024-02-21	chl = 49,944	doi.org/10.6073/pasta/2f750544112e5408928dd9a61e6ace30
CIOOS	41.0 → 77.07	-175.2 → -53.26	2000-01-05 → 2024-12-31	chl = 34,363 chl_HPLC=4,747	https://cioos.ca/
BCO-DMO	27.09 → 49.13	-126.9 → -63.48	2000-02-01 → 2023-01-28	chl = 17,078 chl_HPLC = 1,626 cdom = 450,202	https://www.bco-dmo.org/
HOTS	21.34 → 22.75	-158.4 → -157.9	2000-02-02 → 2023-12-31	chl = 9,979 chl_HPLC = 2,131	https://hahana.soest.hawaii.edu/hot/hot-dogs/interface.html
Pan-Arctic	47.39 → 77.28	-174.0 → -46.89	2002-05-12 → 2024-10-28	chl_HPLC = 2,602	https://doi.org/10.11583/DTU.29445104
GLORIA	24.48 → 40.44	-122.3 → -73.45	2000-01-11 → 2016-09-21	chl = 787 rrs = 1,452	https://doi.pangaea.de/10.1594/PANGAEA.948492

The QA/QC steps taken to turn the inputted chlorophyll, CDOM, and R_{rs} data into standardized datasets for algorithm development varied slightly depending on if the data came from a repository, project, or manuscript (Fig. 2, Appendix C). In general, data were locally downloaded and metadata appended to each sample. Once a datetime column was made, column names standardized, and any samples flagged or missing datetime, depth, or coordinates removed, the methodological flags were added. Finally, the dataset was subset to the top 150 meters of the water column, years 2000-2025, and to the shapefile of North America (Fig. 1). While these are the general steps taken, each source had nuances that needed to be addressed to best create a standard, compiled chlorophyll dataset. Detailed methodology on how each unique source was processed is available in the Supplementary section: Description of Methods used for Sources (Appendix C).

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

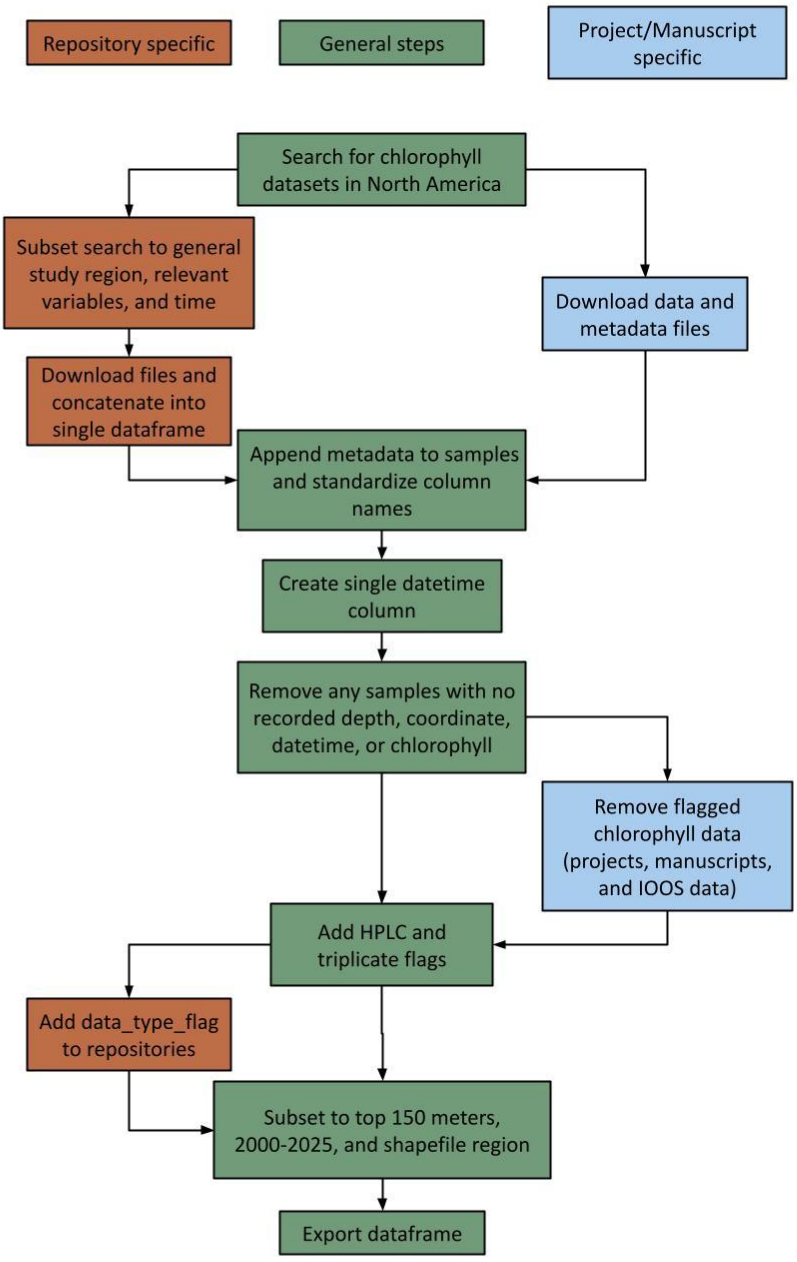


Figure 2: General workflow for compiling data from repositories, projects, and manuscripts. Green directions (center boxes) refer to actions taken for all sources, brown directions (left boxes) refer to actions taken specifically for repositories (SeaBASS, IOOS, WOD), and blue directions (right boxes) refer to actions taken for data from projects and manuscripts.

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

2.3 Sources

The data were concatenated into 3 datasets by variable: one for chlorophyll-only measurements, one for CDOM-only measurements, and one for R_{rs}-only measurements. Coincident data in the three datasets were matched to create a final merged dataset. All measurements that were not coincidental were concatenated. If a project measures both ‘chlorophyll’ and ‘chlorophyll_HPLC’, the two measurements are saved in the same row. This results in a final dataset with all chlorophyll, CDOM, and R_{rs} measurements, with 1,607,285 rows, 17 metadata columns, and 481 data variable columns. Datasets also have metadata entries including experiment (project), station, affiliations, investigators, contact, cruise, cast, DOI_url, and source when relevant (Table 3). Each row of each dataset has a unique ID code per sample added. This ID code was created by combining the source of data (e.g. SeaBASS, CalCOFI, IOOS) plus the project, datetime, latitude, longitude, and depth. Each code has an additional sample count value (0, 1, 2, ect) to account for replicated data. For example, SeaBASS chlorophyll measurements from the project SFMBON, collected on 10/16/2018 21:55:00, at 28.0093°N and -82.9613°W, at depth of 2 meters have the following unique IDs:

SB_SFMBON_20181016-215500_28.0093_-82.9613_2.0m_0

SB_SFMBON_20181016-215500_28.0093_-82.9613_2.0m_1.

Table 3: Metadata availability for final appended dataset. The HPLC, triplicate, and data_type_flag are only available for chlorophyll measurements.

Column	definition	Data availability (%)
ID_code	Unique identification code for sample	100
source	Repository or project data was gathered from	100
experiment	Project or experiment name	100
affiliations	Institution, university, or affiliation responsible	97.09
investigators	Names of P.I's or contributing scientists	81.77
cruise	Name of Cruise	59.02
station	Name of Station	28.9
cast	The specific cast number	5.51
DOI_url	DOI or website of raw dataset	94.69
contact	Email address of head investigator(s)	18.55

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

HPLC	Flag indicating if chlorophyll sample used HPLC methods (0=Yes, 1=No, 2= indeterminate)	100
triplicate	Flag indicating if chlorophyll sample measured triplicates (0=Yes, 1=No)	100
data_type_flag	Flag indicating if sample was extracted or in vivo/ fluorescence (0=extracted, 1 = in vivo, 2 = indeterminate)	63.58

2.4 Satellite Matchup

To assess how the coastal chlorophyll dataset compares to satellite chlorophyll estimates, the *in situ* measurements were matched to the OC-CCI chlorophyll imagery. The *in situ* data within the upper 10 meters were temporally (±12 hours) and spatially (a 3x3 averaged array around the nearest pixel) matched to the OC-CCI global 4km daily chlorophyll data. The *in situ* data were reduced to the single, shallowest sample per date/coordinate to reduce sample bias. This satellite matchup deviates from standard matchup techniques such as the Bailey & Werdell (2006) methods as this project aims to inspect the preliminary matchup results on a broader scale. Additional analyses on how chlorophyll concentrations vary with optical properties were done by examining the Optical Water Types (OWTs) of coastal waters (Wei et al., 2022). The Wei et al. (2022) OWT model uses k-means clustering of R_{rs} to characterize water classes (1-23) with respect to their reflectance signature, bio-optical properties, and biogeochemical variabilities. The OWTs used were derived from daily 4km global R_{rs} data from OC-CCI (wavelengths (nm) = 412, 443, 490, 510, 560, 665). OWTs 1-11 primarily correspond with Case 1, open ocean, and blue water while OWTs 12-23 are Case 2, coastal waters with more complex optical signals (Wei et al., 2022). The OWT matchup was conducted using the same methods as the satellite matchup, with the exception that the mode of the 3x3 pixels was used rather than the mean as OWTs are characterized as discrete classes.

3 Results

The final dataset has 709,758 chlorophyll measurements (mg m^-3) and includes chlorophyll (extracted or *in vivo*) and chlorophyll_HPLC (Fig. 3) (Milton et al., 2026). Chlorophyll ranges from 0 to 1157.1 mg m^-3, with 99.001% of the measurements less than 100 mg m^3. Chlorophyll_HPLC ranges from 0 mg m^-3 to 105.01 mg m^-3, with 99.99973% of the measurements below 100 mg m^3. The Western and Eastern North American coastline have the most data points at 273,359 (West) and 232,583 (East), while Northern Canadian coastlines and Hawaii have the fewest at 34,896 (North Canada) and 16,622 (Hawaii) (Fig. 4). The CDOM dataset has 851,581 CDOM measurements and ranges from 0 mg m^-3 to 617.1 mg m^-3, with 99.9% of the measurements below 100 mg m^3 (Fig. 5). CDOM measurements were found in the southern Californian

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

coastline and along the Gulf and East coastlines. The R_{rs} dataset has 51,085 measurements, 97.2% of which comes from SeaBASS, and spatially covers most of North American coastlines aside from Canadian waters (Fig. 6). The wavelengths range from 305 to 800 nm and the exact wavelengths vary by project. There are 4,250 chlorophyll samples with coincidental CDOM measurements, and 888 chlorophyll samples with coincidental R_{rs} measurements (Fig. B4).

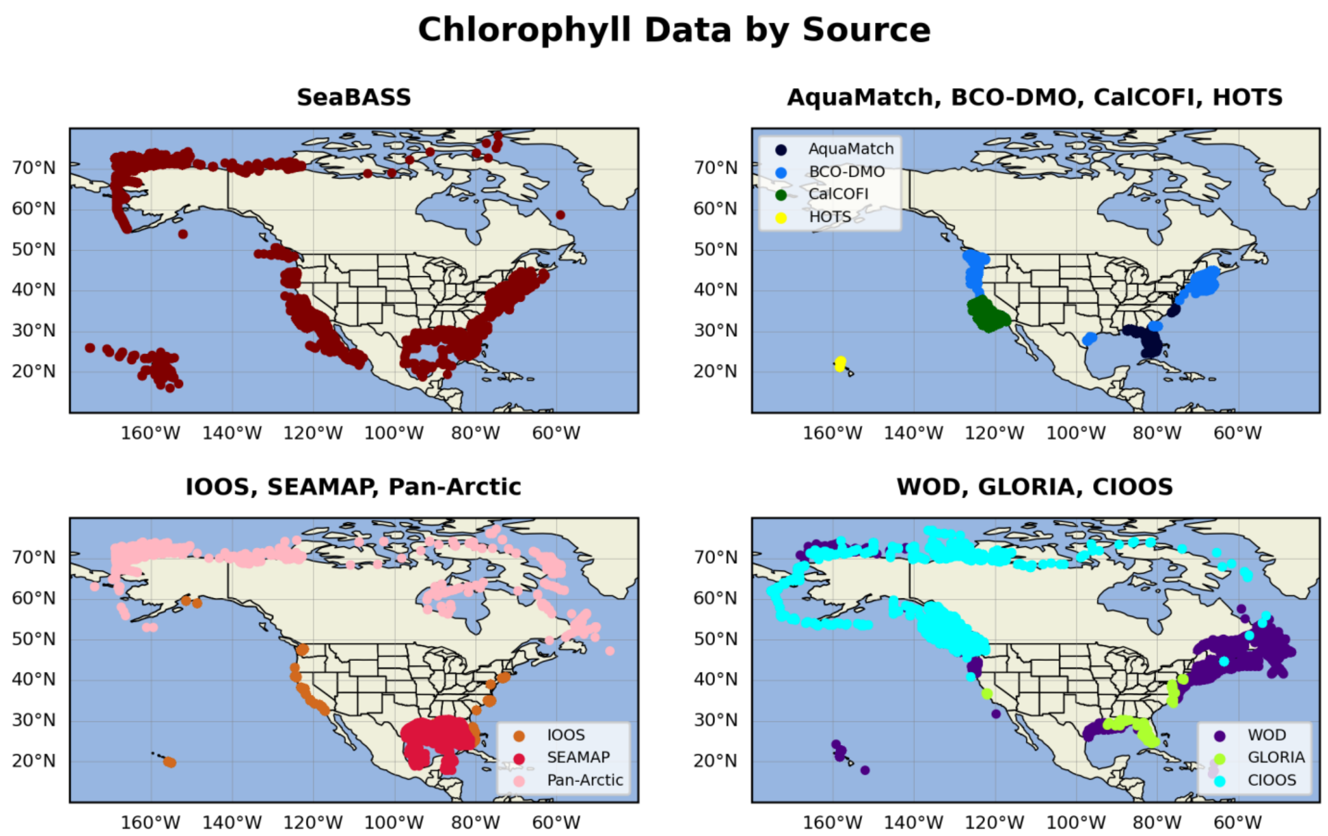


Figure 3: All chlorophyll samples in North America colored by data source, separated by data repository (Table 2).

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

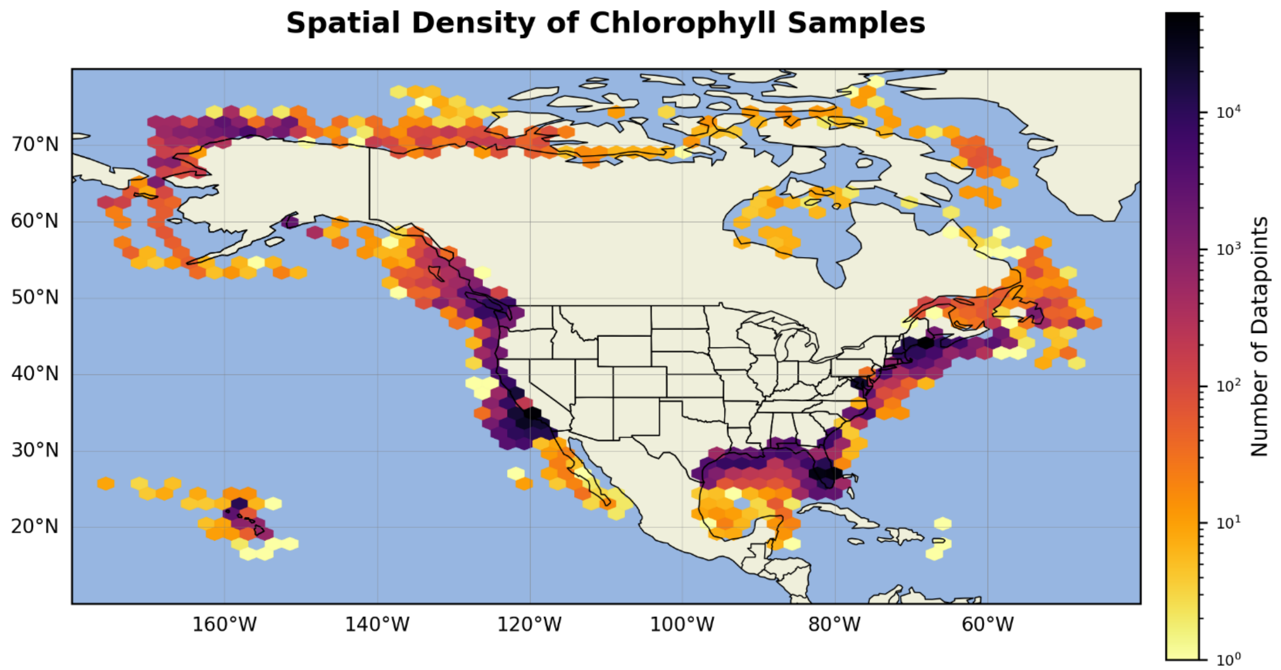


Figure 4: The spatial density of all chlorophyll samples in the North American dataset, with the colorbar as the number of datapoints on a log scale.

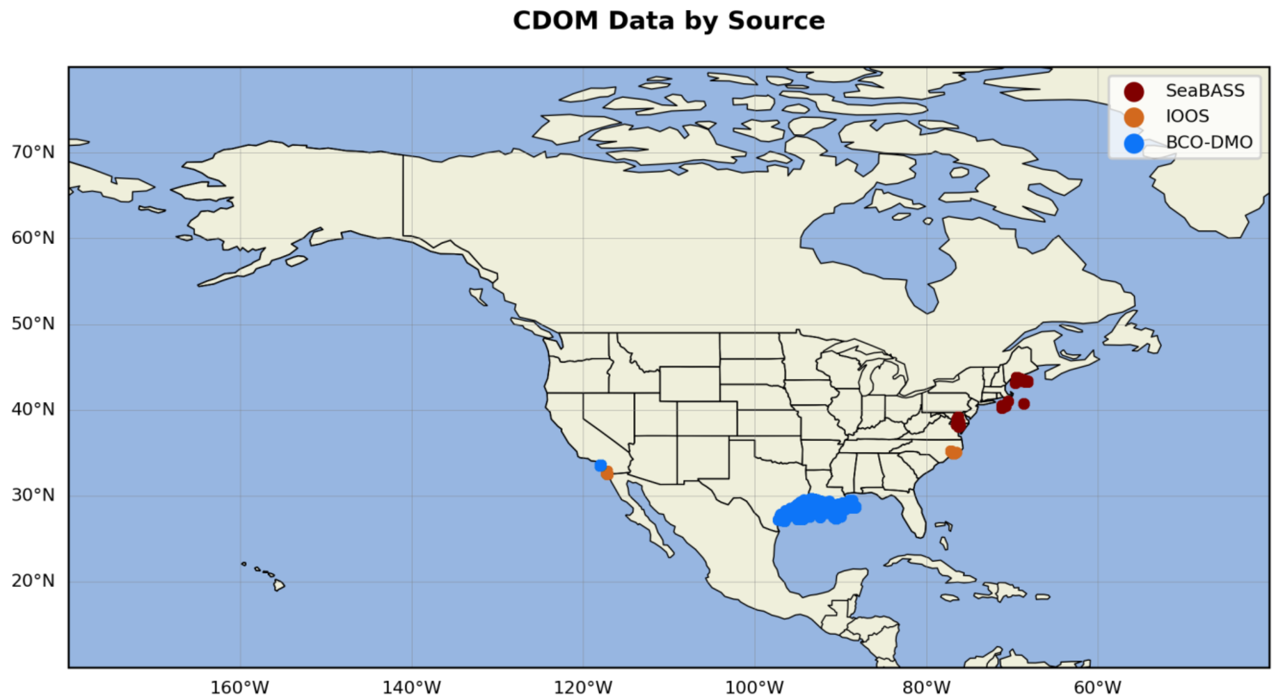


Figure 5: All CDOM measurements in North America colored by dataset source (Table 2).

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

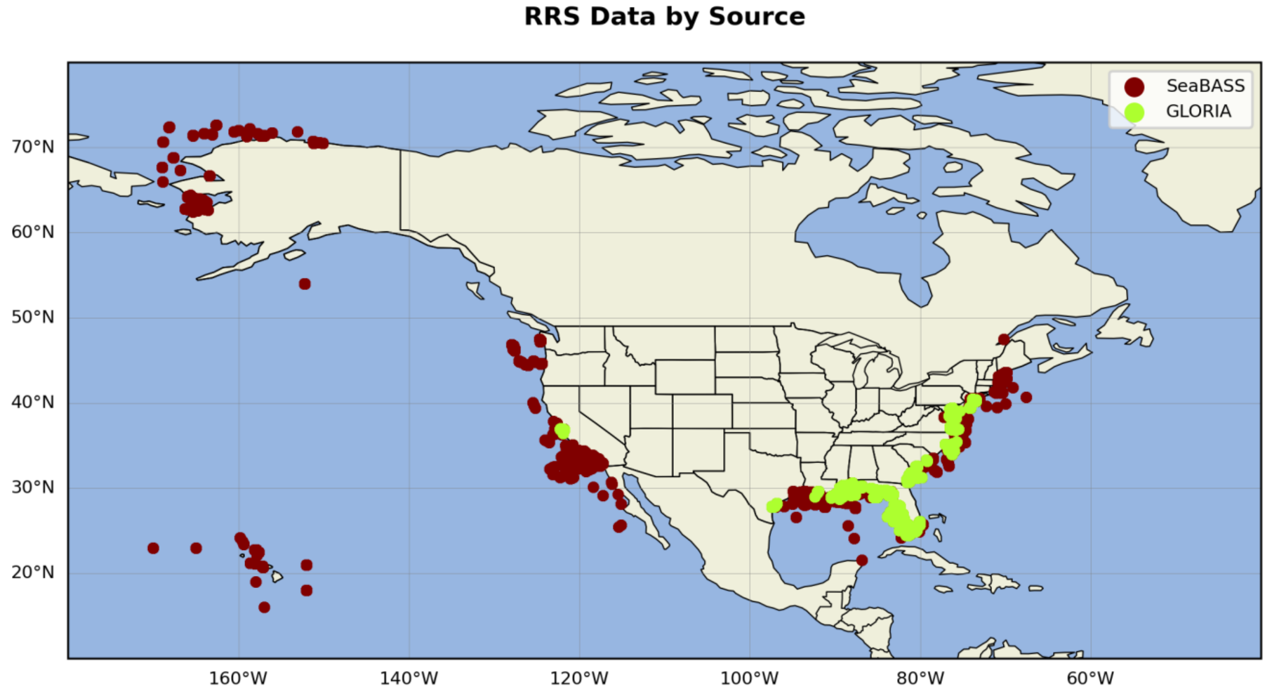


Figure 6: All R_{rs} measurements in North America by source (Table 2).

The temporal distribution of the chlorophyll dataset shows that all years have at least 15,000 data points, with 2000, 2005, 2006, and 2011 having the highest number of samples and 2001, 2020, and 2021 having the fewest (Fig. 7). Chlorophyll measurements from SeaBASS make up the majority of samples from 2000-2011, and IOOS makes up the majority from 2012-2025. Seasonally, summer months (June, July, August) record the most number of samples while winter months (December, January, February) record the least (Fig. 7). The majority of CDOM data was recorded in March, April, May, and June and in 2009 and 2010, and the majority of R_{rs} data was recorded in 2001 and 2011, with July being the most sampled month (Figs. B5-B8).

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

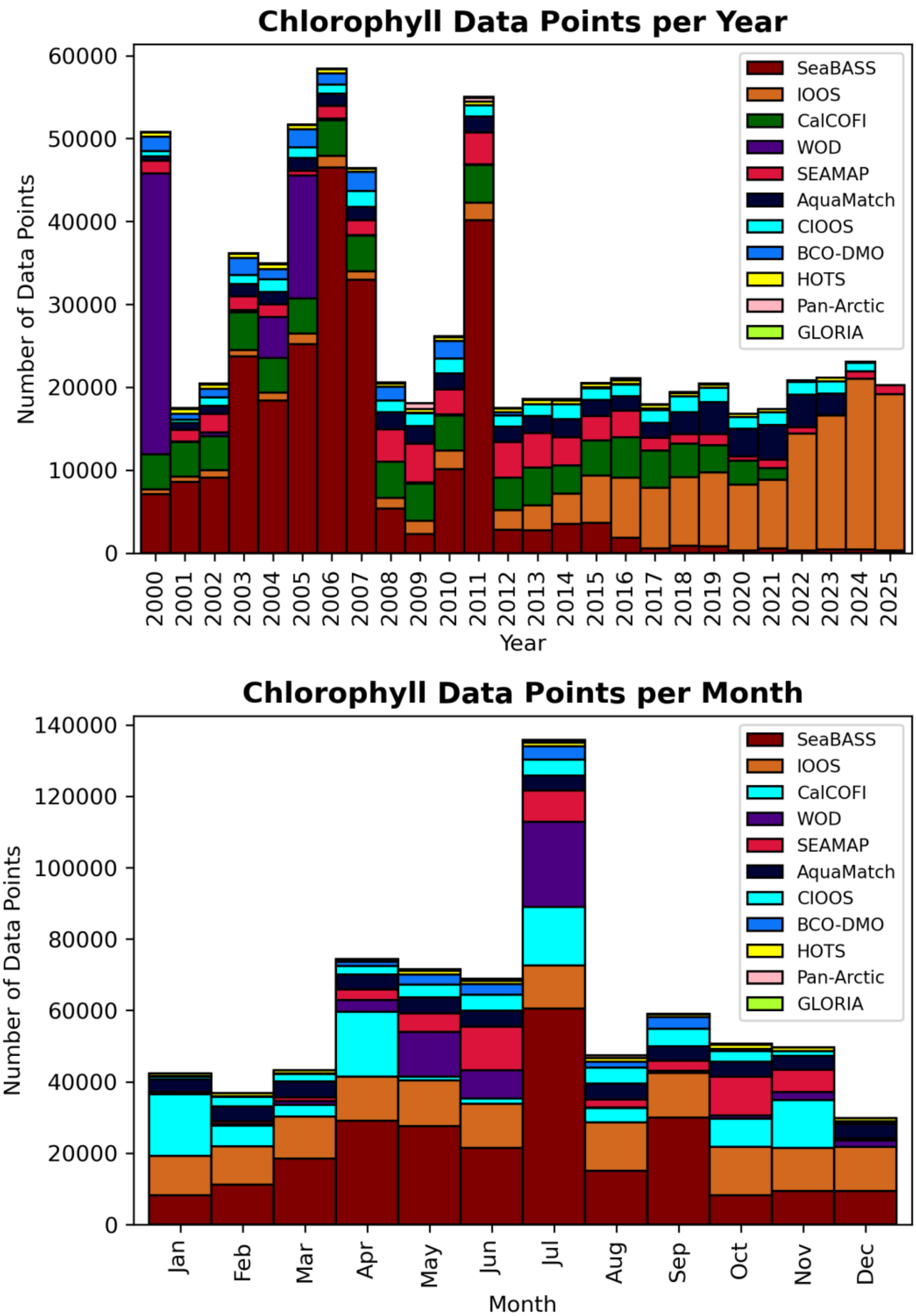


Figure 7: Histogram by number of Data Points per project or repository per year (top) and by month (bottom).

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

The satellite matchup between OC-CCI chlorophyll and *in situ* chlorophyll resulted in 70,837 data points matching (Fig. 8). This is reduced from the chlorophyll dataset due to pre-matchups data reductions (i.e. only using 1 of the replicated samples) and cloud coverage / data availability in OC-CCI data. Out of those, 40,076 were extracted chlorophyll (*in vitro)*, 768 were HPLC and 126 were triplicate. The majority of the resulting points were sampled along the contiguous United States coastline, with the Eastern North American region having the most matched points (Fig. 8). This in part could be due to Northern regions having more cloud coverage than Southern regions, resulting in more satellite data coverage in southern North America.

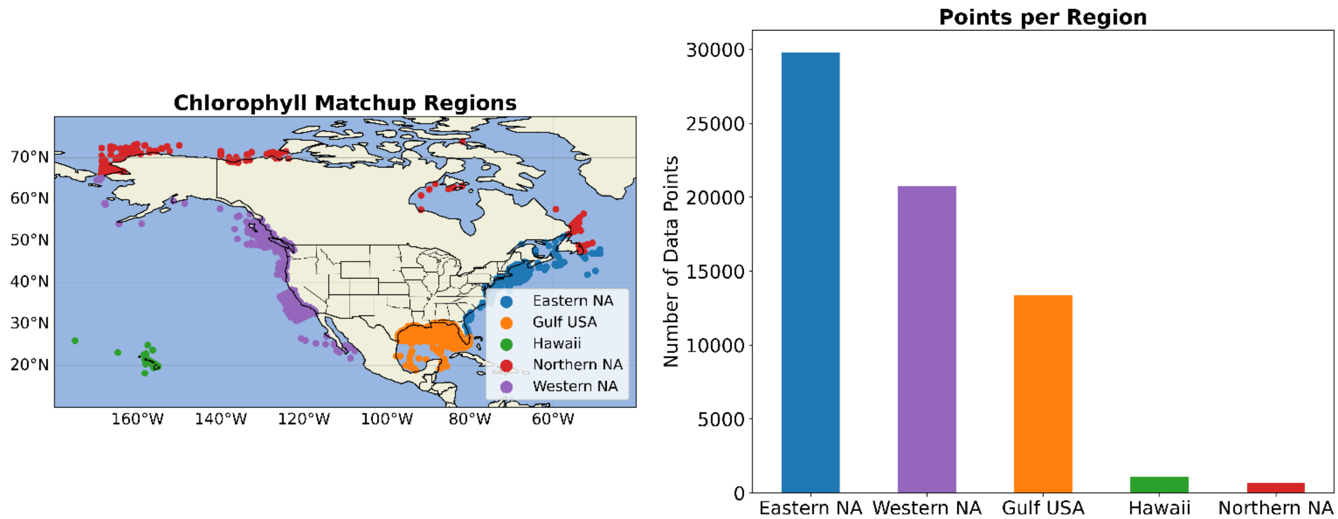


Figure 8: Map of chlorophyll samples in North America (NA) that matched to the OC-CCI dataset, colored by region (left) and distributed by the number of points per region (right).

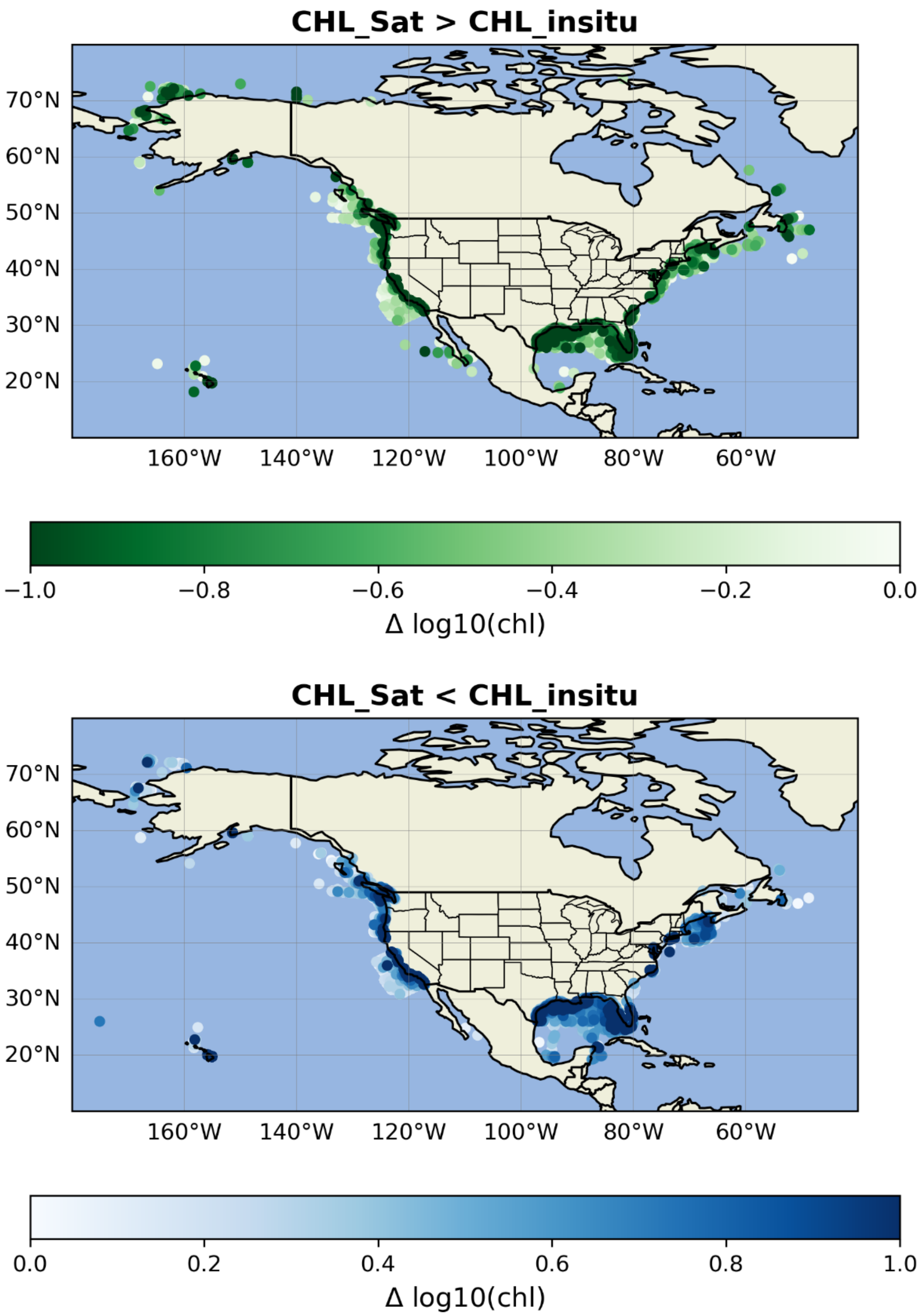
4 Discussion

4.1 In Situ matchup to OC-CCI chlorophyll

A preliminary, simple satellite matchup to the *in situ* dataset was conducted to broadly inspect how current satellite chlorophyll products differ from observational measurements. To conduct this, the log of the *in situ* chlorophyll (CHL_insitu) was subtracted from the log of OC-CCI chlorophyll (CHL_sat) to inspect the difference between the satellite estimates and the chlorophyll values (Fig. 9). This Δlog chlorophyll quantifies the difference between chlorophyll values and there is a noticeable onshore to offshore gradient in Δlog chlorophyll with larger changes closer to shore and smaller changes offshore (Fig. 9).

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions



260 **Figure 9: Plots of Δlog10 chlorophyll (CHL_insitu - CHL_sat), with the top plot (green) showing where the satellite value was greater than the *in situ* value and the bottom plot (blue) showing where *in situ* values were greater than satellite.**

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Not all samples had poor matchups (Fig. 10). When inspecting all matchups regardless of methods, the majority of the data falls within two standard deviations of the mean and have differences of <1 mg m^-3, with samples slightly skewing towards CHL_Sat>CHL_insitu. However, as mentioned in Section 2.1, many algorithms prefer extracted or *in vitro* chlorophyll over *in vivo*. The right subplot of Figure 10 shows the distribution of CHL_insitu - CHL_Sat of only extracted measurements. This reduces the matched dataset from 70,837 to 40,076, however chlorophyll fluorescence should not be used in validation of remote sensing products or in algorithm development (Sec. 2.1) as these methods do not accurately portray chlorophyll concentrations. And so the remaining analysis of the satellite to *in situ* chlorophyll matchups will only be of extracted chlorophyll measurements.

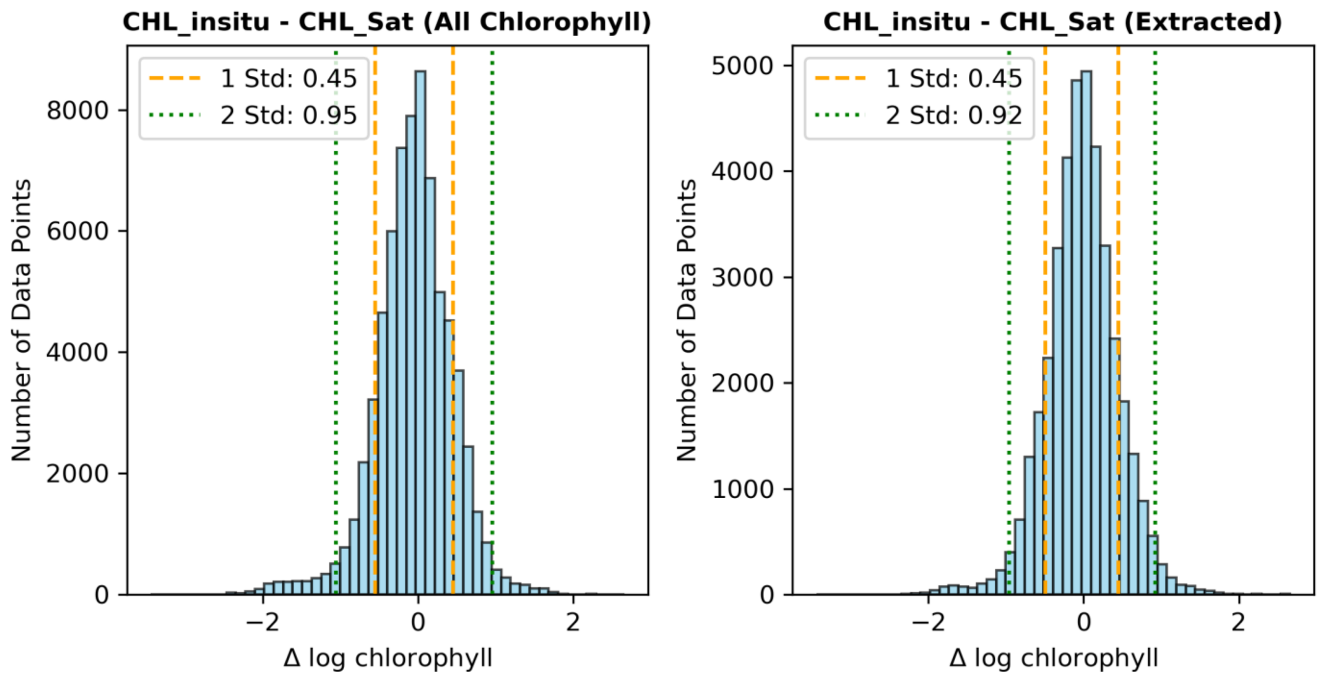


Figure 10: Data distribution of Δlog chlorophyll (Chl_insitu - CHL_Sat), with all chlorophyll samples plotted on the left and only extracted samples plotted on the right. Both histograms show a normal distribution favored slightly towards CHL_sat > CHL_insitu.

The correlation between CHL_Sat and CHL_insitu is further shown in the 2-dimensional histogram of CHL_insitu and CHL_Sat (Fig. 11, left), where 80% of the points are clustered along the 1:1 line with more data points on the right side of the line where satellite values are greater than *in situ*. The second subplot in Figure 11 (right) shows the same values colored by the distance the sample was from shore in kilometers. Here, sample points further from shore (brown-yellow) tend to be closer to the 1:1 reference line while the measurements recorded closer to shore (green-blue) have a wider distribution away from the reference line. This shows that lower chlorophyll concentrations tend to occur further from shore and are more aligned with satellite values, while inshore measurements have a wider distribution.

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

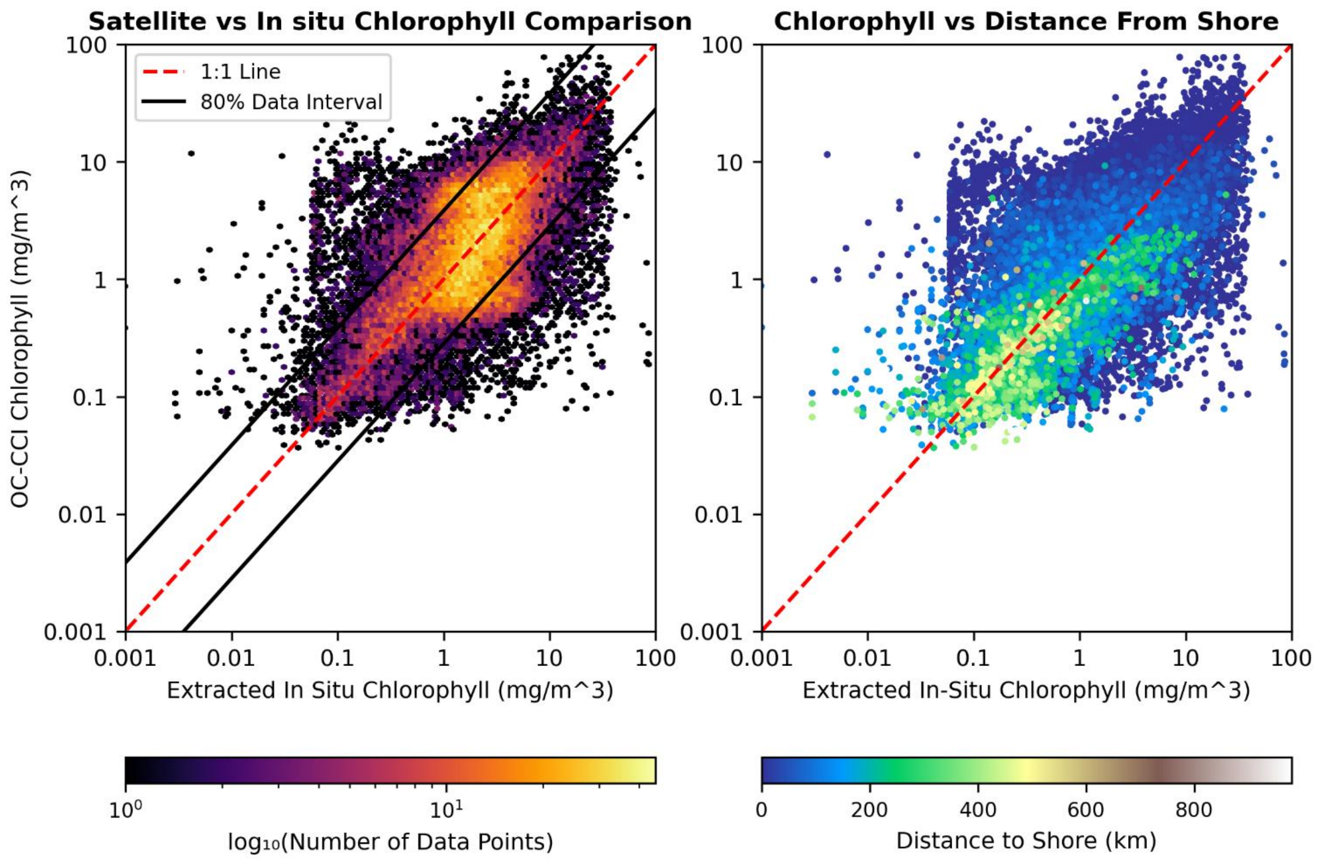


Figure 11: 2-dimensional histogram of extracted chlorophyll vs in situ chlorophyll (left), and distribution of chlorophyll per distance from shore (right). The dashed line in both plots is the 1:1 reference line.

4.2 Optical Water Types

285 The *in situ* chlorophyll dataset was compiled for satellite validation and algorithm development, but can be applied to a wide variety of marine topics including water classifications. By inspecting the OWT matchup, an understanding of how CHL_insitu (subset to extracted chlorophyll measurements) and CHL_Sat interact in different water types can be made. Figure 12 shows a gradient of onshore to offshore OWTs with higher OWTs in coastal waters and lower OWTs in open ocean waters. Given this gradient, it is not unexpected that the chlorophyll matchups show higher chlorophyll concentrations occurring in higher OWTs and lower chlorophyll in lower OWTs (Fig. 13). This gradient occurs both in the *in situ* chlorophyll data and the OC-CCI data, though the *in situ* chlorophyll data has a wider spread of data variation per OWT than the OC-CCI data, possibly due to both the OWT and the satellite chlorophyll data being derived from OC-CCI data.

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

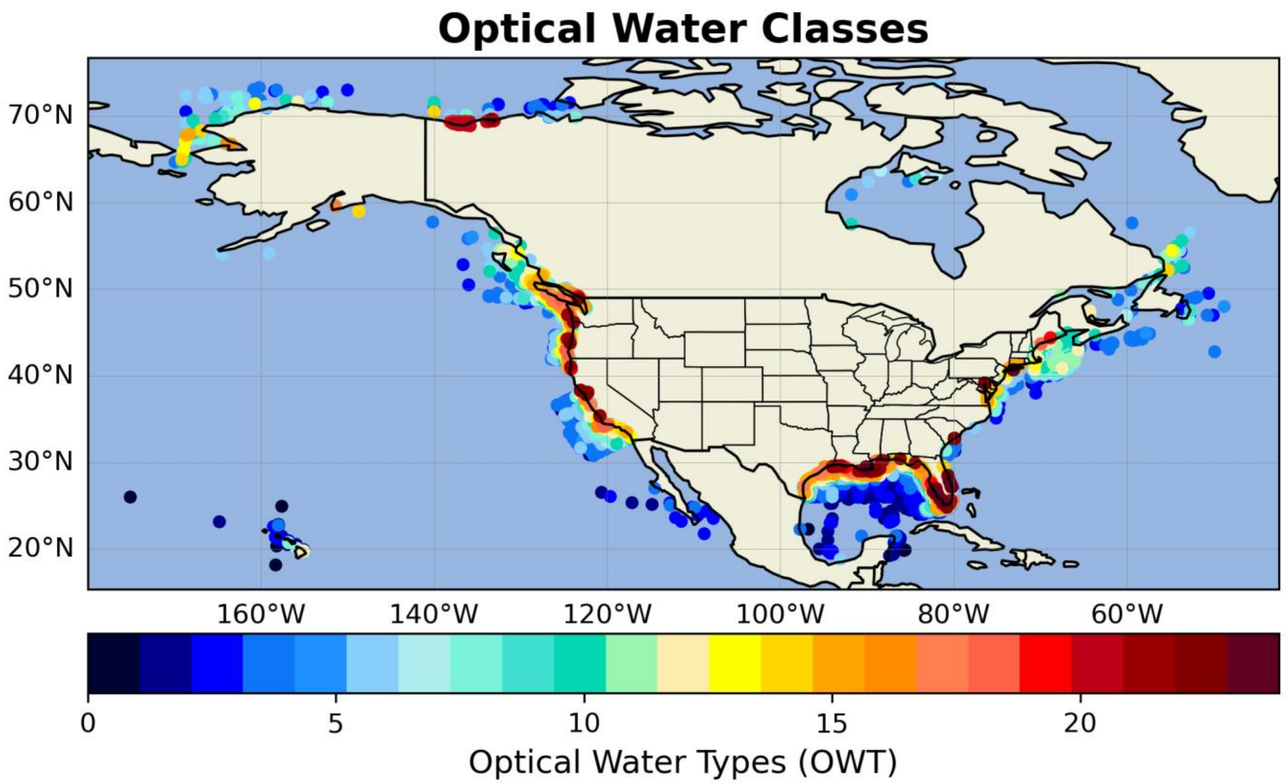


Figure 12: Optical Water Types of chlorophyll matched samples in North America, calculated from OC-CCI global daily reflectance
295 **data, where blue points are Case 1 waters and red points are Case 2 waters.**

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

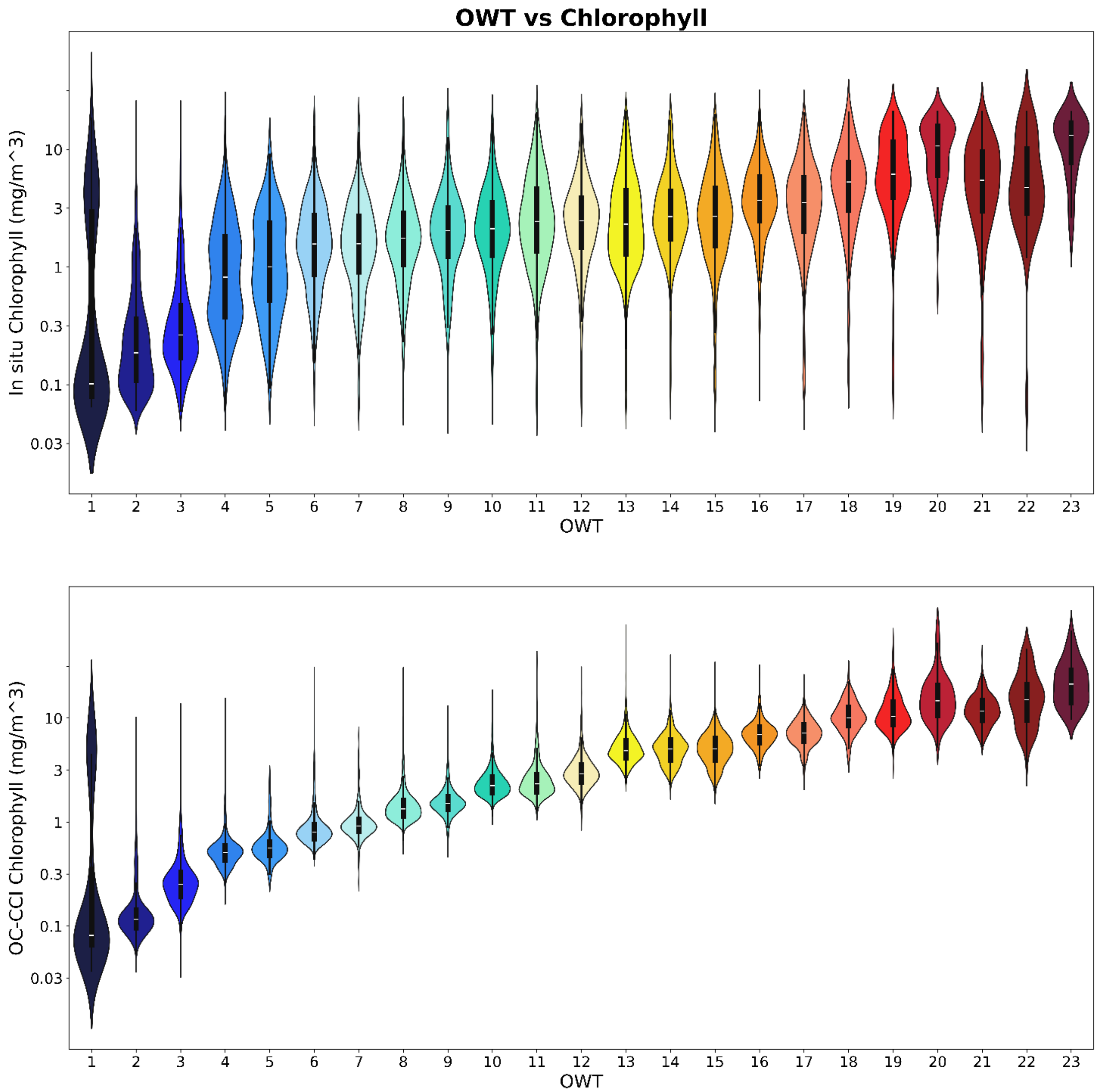
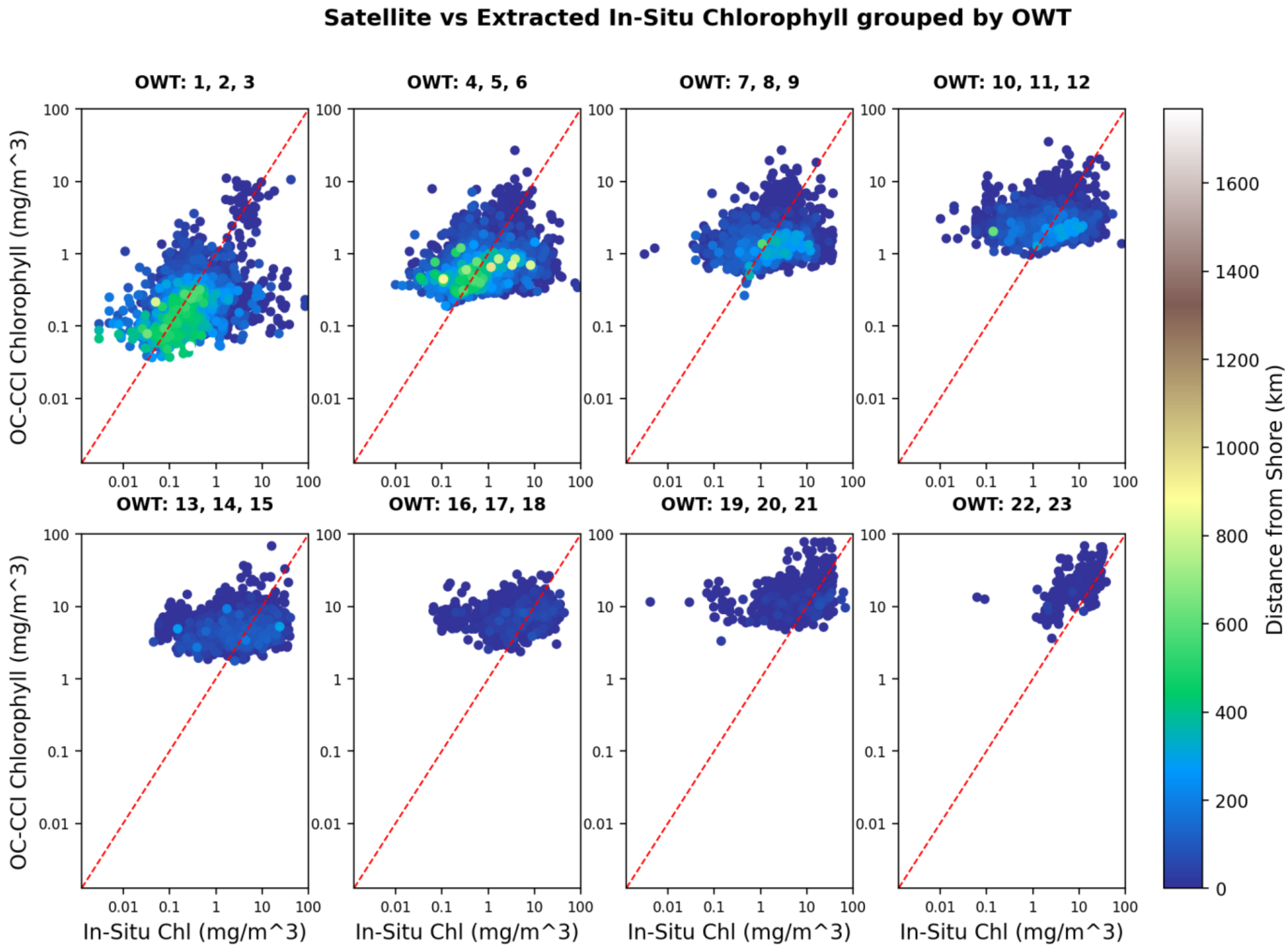


Figure 13: Violin plot of Optical Water Type vs extracted *in situ* chlorophyll (top) and OC-CCI chlorophyll (bottom), colored by OWT. In both violin plots OWT 1 has the largest spread in chlorophyll data.

OWTs are tools to understand the bio-optical/biogeochemical characteristics of the water, where high OWTs tend to have more in-water constituents such as CDOM, suspended sediments, and chlorophyll as shown in Figure 13. They can be used as indicators of water quality in coastal oceans and inland waters (Wei et al., 2022). In addition, they can indicate regions where open ocean chlorophyll algorithms are likely to have reduced accuracy. Lower OWTs tend to occur offshore where

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

satellite and *in situ* measurements are more closely aligned, while higher OWTs occur closer to shore where the chlorophyll concentrations between satellite and *in situ* vary more (Fig. 14).



305

Figure 14: Violin plot of Optical Water Type vs extracted *in situ* chlorophyll (top) and OC-CCI chlorophyll (bottom), colored by OWT. In both violin plots OWT 1 has the largest spread in chlorophyll data.

Data Availability Statement

Data sets are freely available on the Zenodo data repository "Coastal North America In Situ Chlorophyll and Optical Dataset"
310 at [doi.org/10.5281/zenodo.20140178 (Milton et al., 2026)](https://doi.org/10.5281/zenodo.20140178).

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Code availability

Python scripts used for the aggregation of data are found on the GitHub repository github.com/NEFSC/READ-EDAB-GMilton-CoastalChlorophyll with step-by-step directions found at nefsc.github.io/READ-EDAB-GMilton-CoastalChlorophyll/intro.html

5 Conclusion

In situ data is essential for satellite validation and model training, due to the high accuracy and standardized methods. This chlorophyll dataset, along with CDOM and R_{rs}, has 709,758 data points collected from 2000 to 2025 encompassing coastal North America and can aid in understanding complex coastal environments. The HPLC, data_type_flag, and triplicate flags help end users decipher sample methods and determine for themselves the specific types of samples that best fits their needs. The match-up conducted between *in situ* chlorophyll and OC-CCI chlorophyll shows that while some points match well, there are inconsistent results in coastal regions that need to be addressed.

The hope for this dataset is for it to be applied not only to satellite validation or model development, but to a wide array of applications. Chlorophyll is an essential variable in understanding phytoplankton and productivity, and thus is important to both scientific research and fisheries management. This *in situ* dataset could be used to help show the changes in phytoplankton over time or be appended onto existing chlorophyll datasets to increase sample size.

This coastal *in situ* and optical dataset was created by merging data from multiple publicly available repositories with varying degrees of standardization and metadata availability. The result is a uniform dataset that is well documented and easily reproducible. While tedious and time consuming, the creation of this standardized chlorophyll dataset from disparate and unique sources with uniform methodological flags has the potential to greatly aid in algorithm development. This dataset is considered final with no plans for updates, however the processing code and methods are publicly available such that any future datasets uploaded to repositories could easily be incorporated. Additional coding tutorials have been created to walk users through the steps taken to create this dataset (see Data Availability Statement).

Appendix A: Citations of Data Sources

Seabass:

- Werdell, P. J., Bailey, S., Fargion, G., Pietras, C., Knobelspiesse, K., Feldman, G., & McClain, C. (2003). Unique data repository facilitates ocean color satellite validation. *Eos, Transactions American Geophysical Union*, *84*(38), 377–387. https://doi.org/10.1029/2003EO380001

CalCOFI:

- TheCalCOFI observations are available at https://calcofi.org/ccdata.html

HOTS:

- HOTS data was obtained via the Hawaii Ocean Time-series HOT-DOGS application; University of Hawai'i at Mānoa. National Science Foundation Award #2241005 https://hahana.soest.hawaii.edu/hot/hot-dogs/interface.html

IOOS:

- Integrated Ocean Observing System (U.S.). (2017). *Manual for real-time quality control of phytoplankton data: A guide to quality control and quality assurance of phytoplankton observations.* https://doi.org/10.7289/V56D5R6S

WOD:

- Mishonov, A. V. (2024). World Ocean Database 2023. *NOAA Atlas NESDIS ; 97.* https://doi.org/10.25923/Z885-H264

BCO-DMO:

- Cai, W.-J., Rabalais, N., & Fennel, K. (2020). *DIC, TA, pH from R/V Pelican cruise conducted in the northern Gulf of Mexico in April and July 2017* [Dataset]. Biological and Chemical Oceanography Data Management Office. https://doi.org/10.1575/1912/bco-dmo.772513.2
- Campbell, L., & Knap, A. (2019). *Hydrographic, nutrient and oxygen data from CTD bottles during R/V Pelican cruises from 2016-06-29 to 2018-06-24* [Dataset]. Biological and Chemical Oceanography Data Management Office. https://doi.org/10.1575/1912/bco-dmo.753882.1
- Campbell, L., Knap, A., DiMarco, S., & Henrichs, D. W. (2021). *CTD profile data from R/V Pt. Sur PS 18-09 Legs 01 and 03, Sept. - Oct. 2017* [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/10.26008/1912/BCO-DMO.809428.1
- Hansel, C., & Lamborg, C. (2019). *Coastal water biogeochemistry collected aboard the R/V Endeavor along the North Atlantic coast from 2017-08-20 to 2017-08-28* [Dataset]. Biological and Chemical Oceanography Data Management Office. https://doi.org/10.1575/1912/bco-dmo.765327.1
- Joye, S. (2019). *Bottle data (including temperature, salinity, density, O2) from CTD rosettes from R/V Atlantis and R/V F.G. Walton Smith cruises AT18-02 and WS1010 in the Gulf of Mexico Macondo wellhead area in 2010 (DWH_Deep_Microbes project)* [Dataset]. Biological and Chemical Oceanography Data Management Office. https://doi.org/10.1575/1912/bco-dmo.3727.1
- Joye, S. B. (2012a). *CTD data from R/V Atlantis cruise AT18-02 in the Gulf of Mexico Macondo wellhead area in Nov-Dec 2010 (DWH_Deep_Microbes project)* (Lod. bco-dmo.Org/id/dataset/3728; Version Version 1) [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/lod.bco-dmo.org/id/dataset/3728
- Joye, S. B. (2012b). *CTD data from R/V F.G. Walton Smith cruise WS1010 in the Gulf of Mexico Macondo wellhead area in May-June 2010 (DWH_Deep_Microbes project)* (Lod. bco-dmo.Org/id/dataset/3729; Version Version 1) [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/lod.bco-dmo.org/id/dataset/3729

- Kieber, D. J., & Xue, L. (2024). *Chlorophyll-a concentrations in seawater collected near the BATS station during R/V Atlantic Explorer cruises AE2113 (July 2021) and AE2303 (January 2023)* [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/10.26008/1912/BCO-DMO.929873.1
- Kudela, R. M., Bruland, K. W., & Peterson, T. D. (2015). *CTD sample bottle water chemistry (macronutrients, chlorophyll) measurements from R/V Wecoma multiple cruises in the Northeast Pacific coastal waters off states of Washington and Oregon from 2004 to 2006 (RISE project)* (Lod. bco-dmo.Org/id/dataset/3246; Version Version 14 July 2015) [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/lod.bco-dmo.org/id/dataset/3246
- Kudela, R. M., Bruland, K. W., Peterson, T. D., & Hickey, B. M. (2015a). *Surface primary productivity from R/V Wecoma multiple cruises in the Northeast Pacific coastal waters off states of Washington and Oregon from 2004 to 2006 (RISE project)* (Http://lod.bco-dmo.Org/id/dataset/3249; Version Version 14 July 2015) [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). http://lod.bco-dmo.org/id/dataset/3249
- Kudela, R. M., Bruland, K. W., Peterson, T. D., & Hickey, B. M. (2015b). *Vertical profiles of primary productivity from R/V Wecoma multiple cruises in the Northeast Pacific coastal waters off states of Washington and Oregon from 2004 to 2006 (RISE project)* (Lod. bco-dmo.Org/id/dataset/3248; Version Version 14 July 2015) [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/lod.bco-dmo.org/id/dataset/3248
- Kudela, R. M., & Lucas, A. J. (2014a). *CTD casts from R/V Nerissa in the Southern California Bight off Huntington Beach in 2012 (SoCalPlumeEx2012 project)* (Lod. bco-dmo.Org/id/dataset/537627; Version Version 04 November 2014) [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/lod.bco-dmo.org/id/dataset/537627
- Kudela, R. M., & Lucas, A. J. (2014b). *YellowFin CTD Data from fishing vessels in the Southern California Bight off Huntington Beach in 2012 (SoCalPlumeEx2012 project)* (Lod. bco-dmo.Org/id/dataset/537818; Version Version 04 November 2014) [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/lod.bco-dmo.org/id/dataset/537818
- Kurtay, G., Stauffer, B., Robinson, K. L., Geist, S., & Schnetzer, A. (2021). *CTD data collected on three NSF RAPID Plankton cruises on R/V Point Sur and R/V Pelican in the Gulf of Mexico during 2017-2018* [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/10.26008/1912/BCO-DMO.827969.1
- Lee, P. (2020). *HPLC pigment data from samples collected during R/V Savannah cruises conducted in the South Atlantic Bight off the coast of Georgia from 2015-2017* [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/10.26008/1912/BCO-DMO.816216.1

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

- McGillicuddy, D., & Kosnyrev, O. (2013). *Bottle Data collected from R/V Oceanus, R/V Tioga and R/V Endeavor cruises from the Gulf of Maine, Mass Bay to Bay of Fundy, Cape Cod Bay and Georges Bank from May 28, 2003 to August 04, 2010* [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/10.1575/1912/5824
- Richardson, T., Condon, R., & Neuer, S. (2019). *Pigment concentrations (HPLC) from R/V Atlantic Explorer cruises AE1102, AE1118, AE1206, AE1219 in the Sargasso Sea, Bermuda Atlantic Time-Series Station (BATS) from 2011-2012 (Trophic BATS project)* [Dataset]. Biological and Chemical Oceanography Data Management Office. https://doi.org/10.1575/1912/bco-dmo.3884.1
- Roberts, B. (2022). *Discrete data on hydrography (from CTD casts) and chemical analyses of dissolved nutrients and organic carbon on the Louisiana-Texas shelf in the Gulf of Mexico from R/V Pelican cruise 28 September – 11 October 2017* [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/10.26008/1912/BCO-DMO.844721.1
- Roman, M., Boicourt, W., Pierson, J., & Huebert, K. (2018). *ScanFish Optical Plankton Counter (OPC) data from R/V Pelican cruises PE03-NGOMEX, PE04-NGOMEX, PE06-NGOMEX, PE07-NGOMEX, PE09-05, and PE11-06 in the Northern Gulf of Mexico between 2003 and 2010* [Dataset]. Biological and Chemical Oceanography Data Management Office. https://doi.org/10.1575/1912/bco-dmo.749700
- Trainer, V. L., Hickey, B. M., Cochlan, W. P., & Trick, C. (2009). *All bottle data collected from cruises W0306A, W0308C, AT11-17, AT11-30, TUIM14MV, TN200 from the Coastal Waters off Washington State and Vancouver Island; 2004-2006 (ECOHAB-PNW project)* (Lod. bco-dmo.Org/id/dataset/3228; Version Version 30 January 2009) [Dataset]. Biological and Chemical Oceanography Data Management Office (BCO-DMO). https://doi.org/lod.bco-dmo.org/id/dataset/3228
- Wetz, M. (2020). *Discrete hydrographic measurements—Effects of Hurricane Harvey on Estuarine Water Quality in the Lavaca-Colorado, Guadalupe and Nueces Estuaries between August 2017 and January 2018.* [Dataset]. Biological and Chemical Oceanography Data Management Office. https://doi.org/10.1575/1912/bco-dmo.787316.1

AquaSat/ AquaMatch:

- Brousil, M. R., Meyer, M. F., Willi, K., Steele, B. G., De La Torre, J., & Ross, M. R. V. (2024). *AquaMatch Chlorophyll a Data from Water Quality Portal: ~1970-2024* [Dataset]. Environmental Data Initiative. https://doi.org/10.6073/PASTA/2F750544112E5408928DD9A61E6ACE30
- Ross, M. R. V., Topp, S. N., Appling, A. P., Yang, X., Kuhn, C., Butman, D., Simard, M., & Pavelsky, T. M. (2019). AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters. *Water Resources Research*, *55*(11), 10012–10025. https://doi.org/10.1029/2019WR024883

GLORIA:

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

- Lehmann, M. K., Gurlin, D., Pahlevan, N., Alikas, K., Conroy, T., Anstee, J., Balasubramanian, S. V., Barbosa, C. C. F., Binding, C., Bracher, A., Bresciani, M., Burtner, A., Cao, Z., Dekker, A. G., Di Vittorio, C., Drayson, N., Errera, R. M., Fernandez, V., Ficek, D., … Yue, L. (2023). GLORIA - A globally representative hyperspectral in situ dataset for optical sensing of water quality. *Scientific Data*, *10*(1), 100. https://doi.org/10.1038/s41597-023-01973-y

SEAMAP:

- SEAMAP data obtained from https://seamapdata.gsmfc.org/pages/download-form.php

CIOOS:

- Stewart, A., deYoung, B., Smit, M., Donaldson, K., Reedman, A., Bastien, A., Brunsting, R., Carter, B., Covey, B., Kelly, R., Peterson, E., Pirenne, B., Plourde, A., Ste-Marie, A.-S., Tollefsen, C., Wallace, D., & Whoriskey, F. (2019). The Development of a Canadian Integrated Ocean Observing System (CIOOS). *Frontiers in Marine Science*, *6*, 431. https://doi.org/10.3389/fmars.2019.00431

Pan-Artic:

- Heidemann, A. C., Hayward, A., Assmy, P., Basu, A., Bracher, A., Castellani, G., Ditullio, G., Dragańska-Deja, K., Fujiwara, A., Fragoso, G. M., Høyer, J., Hwang, J., Iversen, M., Von Jackowski, A., Juul-Pedersen, T., Kowalczuk, P., Lee, Y., Matsuoka, A., Merz, A., … Gonçalves-Araujo, R. (2026). *A Pan-Arctic Pigment Database for Phytoplankton and Sea-Ice Algae*. https://doi.org/10.5194/essd-2026-144

Appendix B: Additional Figures

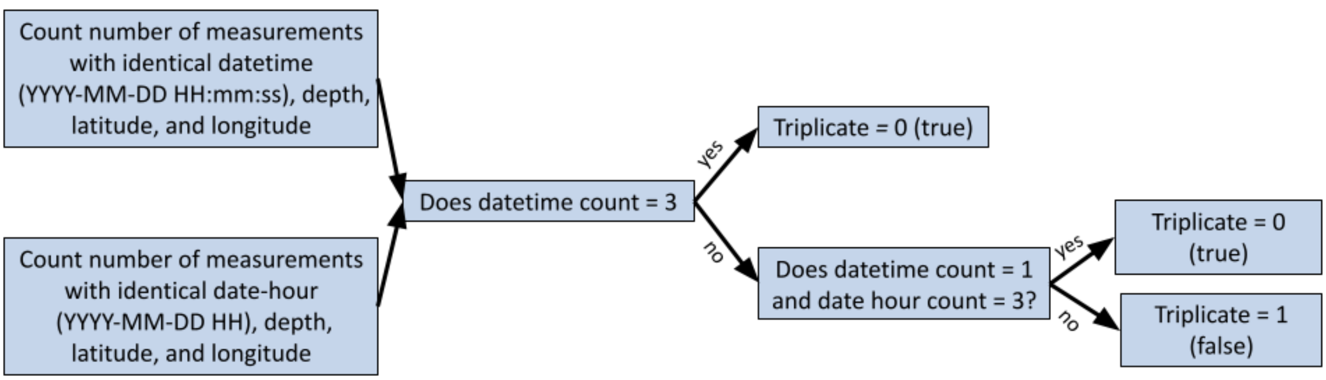
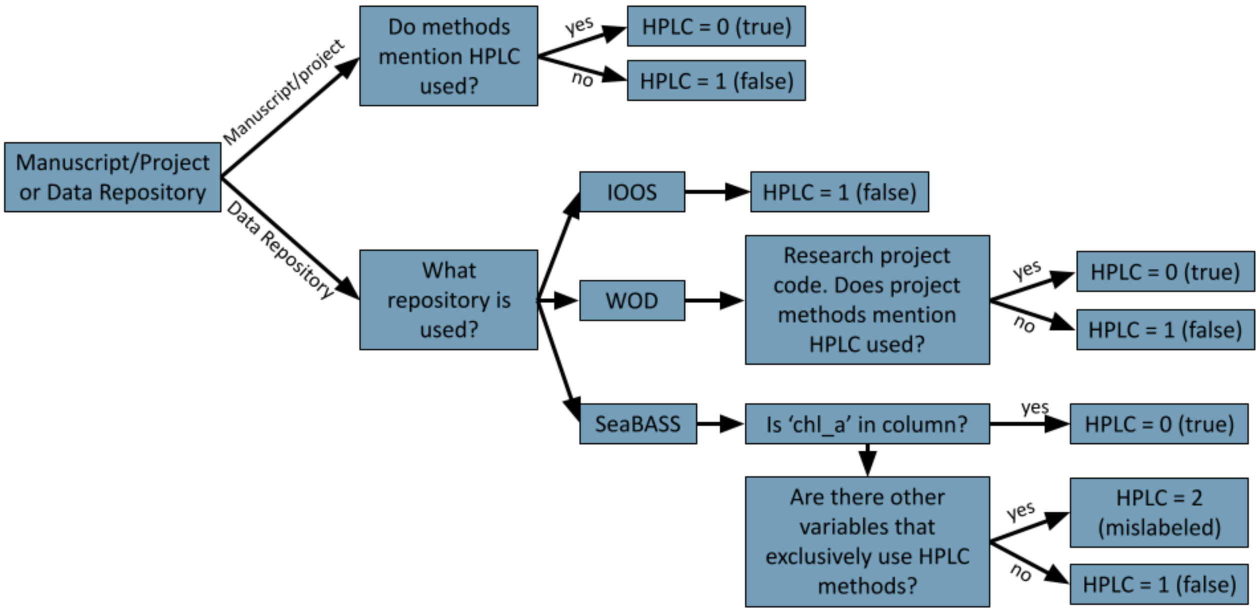


Figure B1: General workflow for creating triplicate flags, where flag = 0 means samples are recorded in triplicates and flag = 1 means no triplicates recorded.

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026



460 **Figure B2: General workflow for creating HPLC flags, where flag = 0 means samples are recorded using HPLC methods and flag = 1 means no HPLC methods used. HPLC flag = 2 is unique to SeaBASS measurements and means HPLC was used in other variables in the dataset, but the chlorophyll column was mislabeled to 'chl' (extracted) rather than 'chl_a' (HPLC).**

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

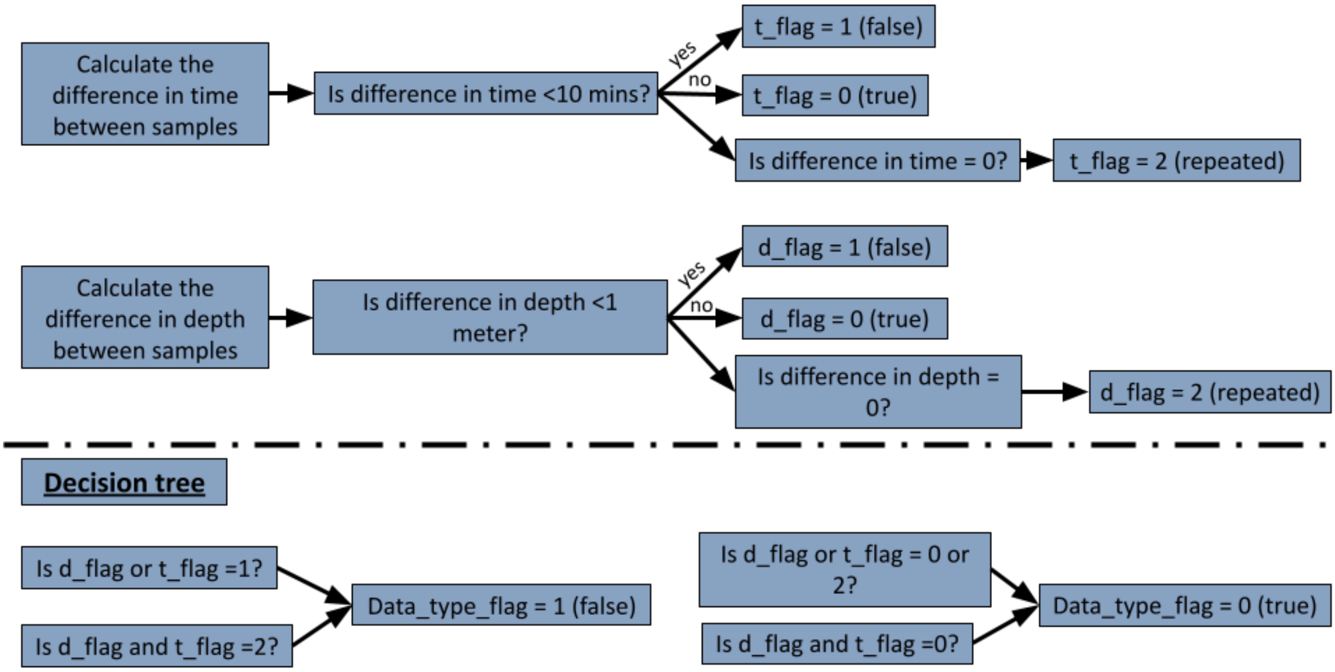


Figure B3: General workflow for creating data_type_flag, where flag = 0 means chlorophyll samples are extracted and flag = 1 means they're measured *in vivo*. T_flag and d_flag = 2 refers to when a sample's change in time or depth is 0 i.e. these values repeat. This typically occurs in older data where specific time and depth values per sample were not recorded but rather the start time or depth were repeated for all samples. In these cases, the difference in time and depth is not indicative of the methods used for chlorophyll, and so if both flags = 2 then the data_type_flag = 1 (false), but if one of the flags is 0 (true) and the other 2, then data_type_flag = 0 (true).

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

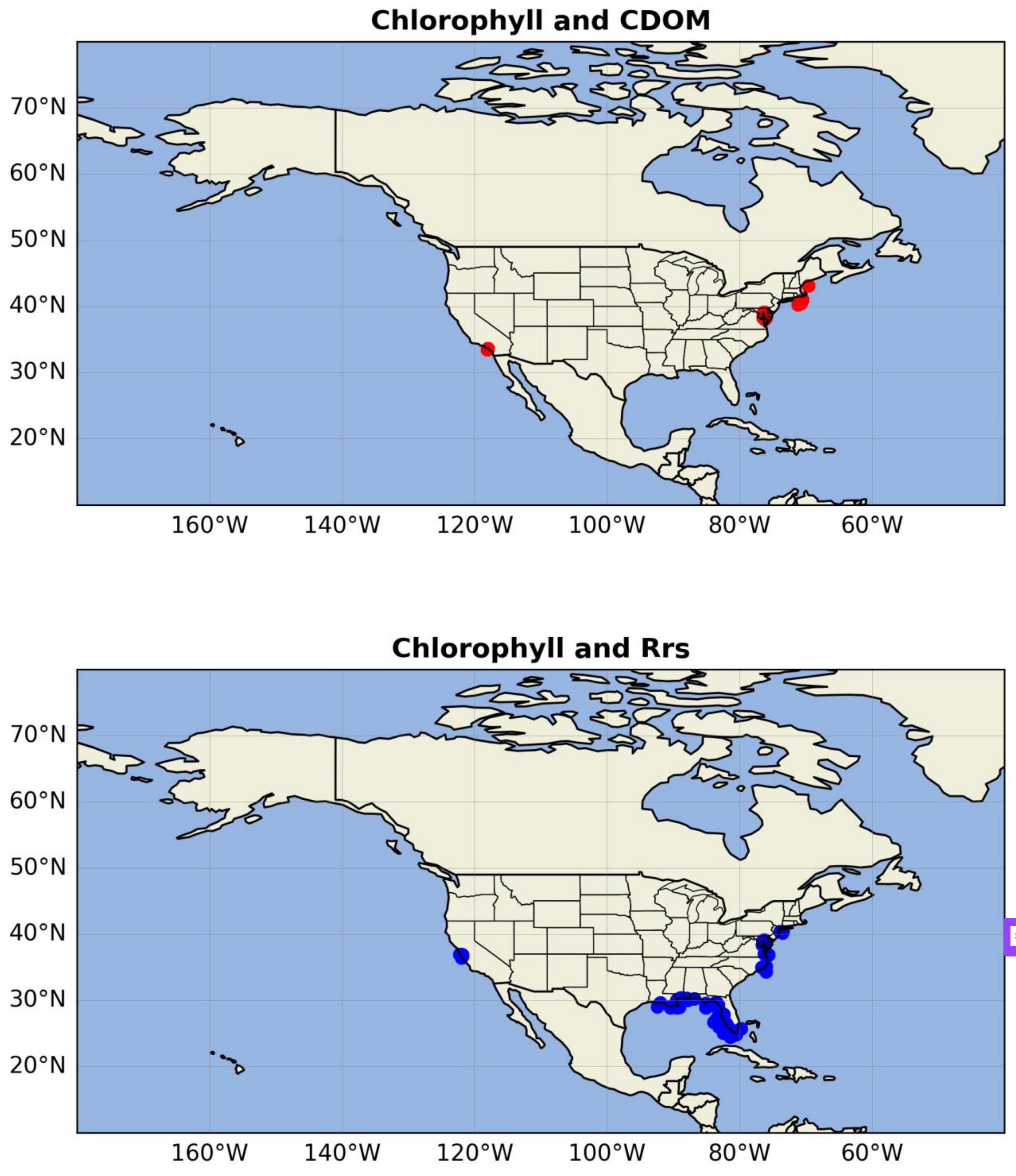
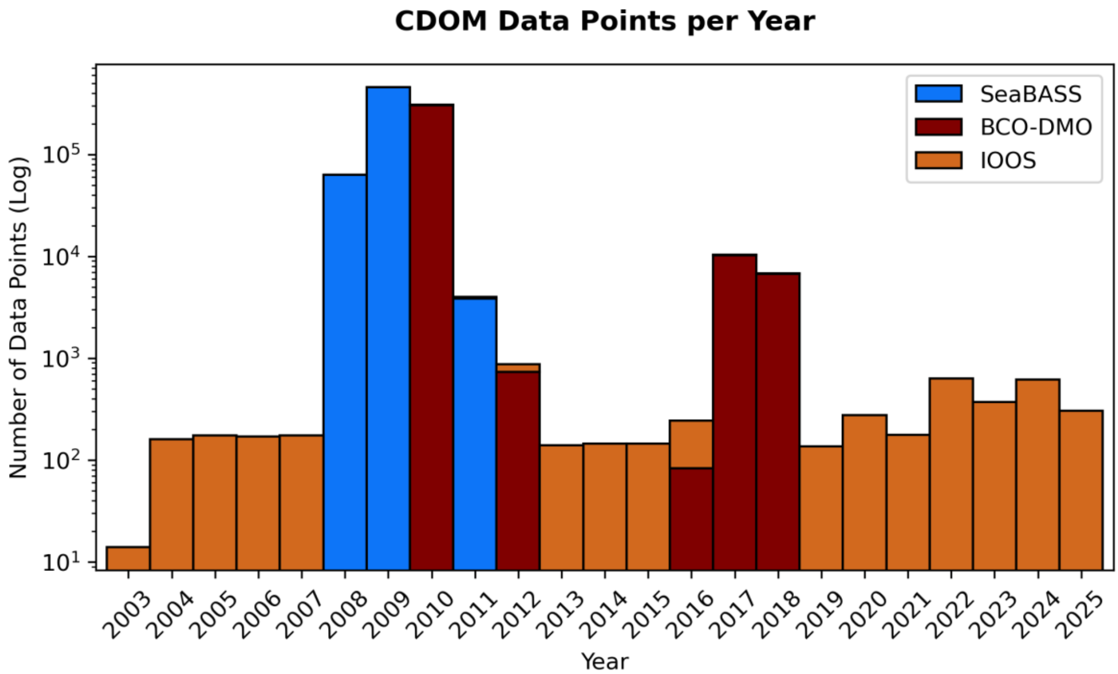


Figure B4: All data points with coincidental variables to chlorophyll. The top plot (red points) are measurements with coincidental chlorophyll and CDOM values, and the bottom plot (blue points) are measurements with coincidental chlorophyll and R_{rs}values.

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026



475 **Figure B5: Histogram by number of Data Points per project or repository per year of cdom**

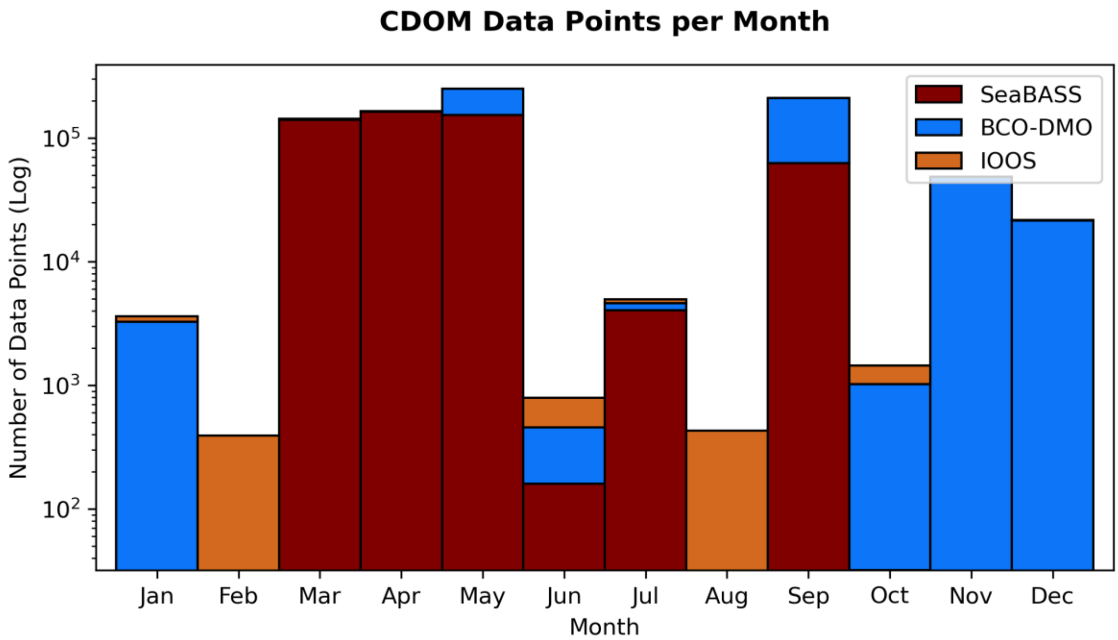
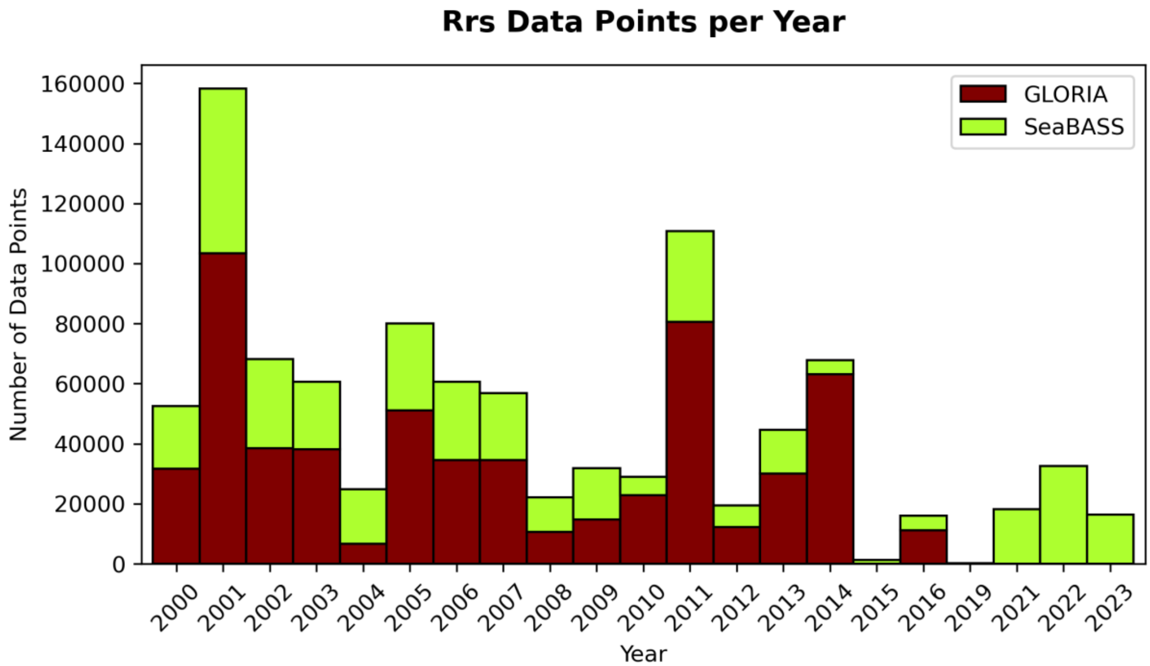


Figure B6: Histogram by number of Data Points per project or repository per month of cdom

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026



480 **Figure B7: Histogram by number of Data Points per project or repository per year of R_{rs}**

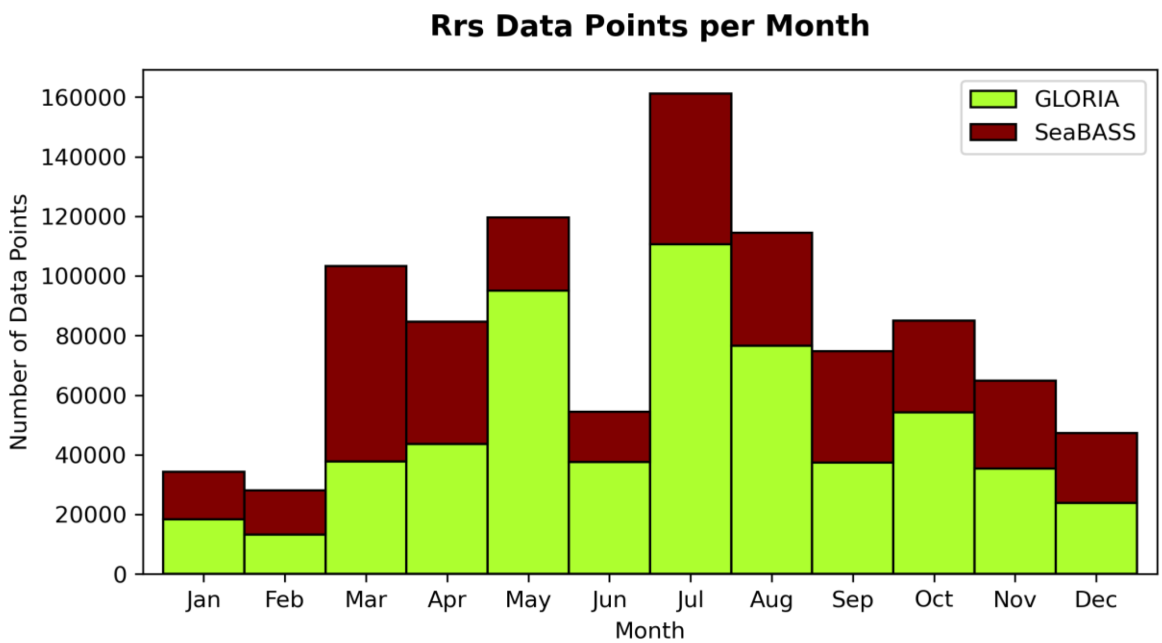


Figure B8: Histogram by number of Data Points per project or repository per month of R_{rs}

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Appendix C: Description of Methods used for Sources

1. SeaWiFS Bio-optical Archive and Storage System (SeaBASS)

The SeaBASS repository (Table 2) is a publicly available archive of global *in situ* environmental data, maintained by the National Aeronautics and Space Administration (NASA) Ocean Biology Processing Group (Werdell et al., 2003). The primary purpose of SeaBASS is to create a repository of data products specifically for satellite validation and algorithm development. The motivation of SeaBASS and this project are closely aligned; thus the SeaBASS naming conventions and metadata formatting are used as the basis of the final compiled dataset. The SeaBASS data makes up a substantial portion of the final dataset; ~35% of the chlorophyll dataset, 61% of the CDOM dataset, and ~97% of the R_{rs} dataset.

Due to the high volume of data, the SeaBASS data were downloaded in specific regions to reduce processing times; East Coast, West Coast, Gulf Coast, Alaska, and Hawaii before concatenating regions together. Additional chlorophyll data from PACE validation cruises in 2025 were also included. These data are available as individual files via SeaBASS, or can be accessed through the PACE matchup site (https://pace.oceansciences.org/pace_bgc_oci.htm). Finally, SeaBASS data that were duplicated from the CalCOFI and GLORIA data by datetime, depth, and coordinate were removed to ensure that the CalCOFI bottle data and GLORIA data were not repeated.

According to the SeaBASS's "Standardized Fields and Units" (https://seabass.gsfc.nasa.gov/wiki/stdfields), the variable 'chl' corresponds to chlorophyll derived fluorometrically/spectrophotometrically, while 'chl_a' corresponds to HPLC measured chlorophyll *a*. Therefore, SeaBASS HPLC flags were determined by the presence or absence of the variable 'chl_a'. However, the SeaBASS repository has evolved over time. Variable names and metadata structures are less consistent in some of the earlier datasets. As a result, 'chl_a' is sometimes mislabeled as 'chl'. There are a couple hints that point towards this mislabeling: a file could have variables exclusively measured by HPLC methods but label the chlorophyll samples as chl, or a file name contains the word 'HPLC' (Fig. B2). Measurements that were suspected to use HPLC but are potentially mislabeled have the HPLC flag = 2 and comprise ~1.5% of the SeaBASS chlorophyll dataset (Table 1).

The SeaBASS variable 'chl_stimf' is defined as chlorophyll measured *in vivo* but could at times be mislabeled as 'chl'. For algorithm development, chlorophyll measured fluorometrically is preferred over *in vivo* fluorescence, and so SeaBASS data has the additional flag 'data_type_flag' to help indicate the method used in chlorophyll sampling (Table 1). This flag was decided by detecting the data method (cast, bottle, flow_thru, ect.), then measuring the temporal and spatial resolution of the samples i.e. the time between samples and vertical/horizontal distance between samples. High spatial and temporal resolution may indicate *in vivo* chlorophyll sampling, as fluorescence measurements are often recorded continuously. Therefore, if a sample's spatial or temporal resolution was high, then it was assumed that the chlorophyll was not measured from a bottle or cast but rather estimated *in vivo*. For example if chlorophyll was sampled continuously every 10 seconds or every 0.5 meters, it was assumed the measurement was *in vivo* and the 'data_type_flag' turned to 1 (False) to indicate chl not sampled fluorometrically (Fig. B3). Other large repositories (IOOS, WOD) with high resolution sample rates mixed in with discrete samples also have the data_type_flag, which is described in the Supplementary Sections 1.2 and 1.5. Unique to the

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

SeaBASS dataset is the option for data_type_flag to equal 2. This flag is used to designate samples that had no change in temporal or spatial resolution i.e. the time between samples and the change in depth between samples was 0. This typically occurred in older data that did not record the exact time or depth a sample was taken, but instead used the start time or average depth for all samples. In these cases, the sample resolution is not indicative of how chlorophyll was measured, so data_type_flag was set to 2 to indicate the sample method as indeterminate (Fig. B3).

CDOM and R_{rs} were also retrieved from the SeaBASS repository. R_{rs} data are saved in two different formats in SeaBASS: ‘long’ and ‘wide’ (Table C1). Long format saves the R_{rs} and wavelength in two separate columns, creating multiple rows for a single measurement. Wide format appends the wavelength onto the column name, creating multiple columns per measurement. Some projects save R_{rs} in long format and others in wide. To best standardize SeaBASS R_{rs} data, all projects were transposed into long format for data processing ease. Any fine scale wavelengths that measured in the tenths of nanometers were omitted (ex: wavelength = 465.1, 465.2).

Table C1: Examples of the two formats R_{rs} can be found, with long format repeating a single sample over multiple wavelengths through rows and wide format appending the wavelength to the rrs columns.

Long format				**Wide format**			
date	rrs	wavelength			rrs_384	rrs_385	rrs_443
2021-01-01	0.00304	384	<—>	2021-01-01	0.00304	0.0293	0.006639
2021-01-01	0.0293	385					
2021-01-01	0.006639	443					

2. Integrated Ocean Observing System

The Integrated Ocean Observing System (IOOS) is a repository of publicly available oceanographic data made up of multiple regional associations, including historic and near real-time data (Table 2) (Integrated Ocean Observing System (U.S.), 2017). The IOOS Environmental Sensor Map integrates data from regional associations, as well as local data providers onto a near-realtime archive that can be used to bulk download chlorophyll and CDOM data through the IOOS Environmental Research Division's Data Access Program (ERDDAP) server. For this dataset, the data were retrieved by searching the variable: ‘mass_concentration_of_chlorophyll_in_sea_water’ and processing each dataset or project that contains the keyword. Many IOOS datasets sample regularly at high temporal and vertical resolutions (ex: every hour, every 2 meters, since 2000), which is too finescale for satellite validation and algorithms and can negatively impact data processing speeds. Therefore, IOOS data was reduced to the top 10 meters and 1 day averages. Additionally, since the IOOS ERDDAP does not distinguish between *in vivo* and extracted, a ‘data_type_flag’ was added to detect high sample resolution to aid in distinguishing between these methods.

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

Some IOOS stations recorded extreme outliers in chlorophyll concentrations. For example, the station ‘gov_usgs_nwis_01304562’ in Long Island, New York has multiple chlorophyll recordings > 200,000 mg m^-3. Due to the high volume of data collected, sometimes suspect measurements pass the IOOS QA/QC, so additional quality control was conducted on the IOOS dataset. Only data within the lower 95% quantile were retained and the top 5% of the chlorophyll data (chlorophyll >38.5 mg m^3) were removed (Fig. C1). Additionally, any miscalculated negative chlorophyll values were also omitted. Finally, the IOOS data has no HPLC indicator, therefore it was assumed HPLC is not used and the triplicate flag was turned to “False” since only the daily averages are included. CDOM data from IOOS were also organized and processed in a similar way to the IOOS chlorophyll data. The term ‘CDOM’ was searched on their ERDDAP server and any flagged samples removed. Then, the datasets were reduced to the top 10 meters and 1 day averages. Similar to chlorophyll, the IOOS CDOM data repository does not distinguish between methods used to collect CDOM so disparate methods could be included.

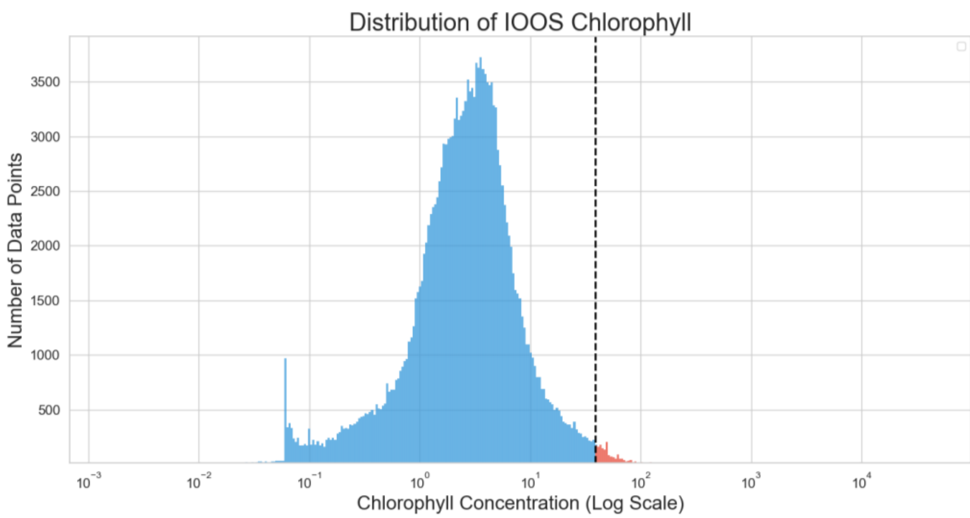


Figure C1: Distribution of the IOOS chlorophyll values, where the black dotted line is the 95% quantile cutoff (at 38.5 mg m^3), the blue bars are values included in the dataset and the red bars are excluded values. The x-axis represents IOOS chlorophyll values on a log scale and the y axis represents the number of samples.

3. California Cooperative Oceanic Fisheries Investigations

The California Cooperative Oceanic Fisheries Investigations (CalCOFI) is an ongoing oceanographic monitoring program that samples quarterly off the coast of Southern and Central California and all data are publicly available (Table 2). CalCOFI measures a wide variety of variables, including chlorophyll from CTD rosette bottles. The raw chlorophyll bottle data were first matched to the cast and station dataset to append appropriate metadata. Chlorophyll datapoints flagged by CalCOFI were removed and the HPLC flag turned to False as CalCOFI bottle data did not use HPLC methods. With regularly scheduled cruises and consistent station locations, the CalCOFI dataset makes up ~12.5% of this dataset.

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Open Access Earth System Science Data Discussions

4. Southeast Area Monitoring and Assessment Program

The Southeast Area Monitoring and Assessment Program (SEAMAP) regularly collects fishery-independent data for the Southeastern regions of the United States and Caribbean (Table 2). The program has three components: SEAMAP-Gulf of America, SEAMAP-South Atlantic, and SEAMAP-Caribbean. SEAMAP-Caribbean does not appear to collect chlorophyll data, and the SEAMAP-South Atlantic data are not available for public distribution. Therefore, only the SEAMAP-Gulf of America chlorophyll dataset is used for this dataset. The SEAMAP raw data are organized into unique single surface, mid-depth, and bottom depth chlorophyll measurement per sample, so the chlorophyll columns were transformed from 3 columns to a standardized depth column and chlorophyll column. According to the SEAMAP metadata, no HPLC methods were used and no triplicates recorded.

5. World Ocean Database

The World Ocean Database (WOD) is an oceanographic archive of vertical profile data. The WOD repository consists of data archived by the NOAA National Centers for Environmental Information (NCEI) and follows quality control flagging methods used in the World Ocean Atlas (Table 2) (Garcia et al., 2024). Chlorophyll data were bulk downloaded using the WOD select query from Ocean Stations, Moored Buoys, Surface-Only buckets, High Resolution CTDs, and Profiling floats. Additionally, data from the NOAA Northeast Fisheries Science Center Ecosystem Monitoring (EcoMon) program in WOD were gathered separately and organized to ensure all EcoMon data were included at the behest of principal investigators. The raw WOD data were processed by concatenating each cast block into a single dataframe with consistent column names. Any projects already incorporated into our database, such as CalCOFI and HOTS, were removed through inspection of the 'projects' column to avoid duplication. The sample rate of each cast was also measured to create a 'data_type_flag' for end users to better determine if the chlorophyll was extracted or estimated *in vivo*. Additionally, samples that recorded data at less than 1 meter intervals were removed to aid in data processing, since this sample rate is highly indicative of *in vivo* methods.

Each cast with a project code was researched for HPLC methods, since WOD does not include HPLC indicators or other phytoplankton pigments when bulk downloading chlorophyll from their data service. If the project specifically stated HPLC was implemented, the HPLC flag was turned to True, otherwise it was assumed HPLC was not used. It should be noted that WOD does not consistently include all metadata entries, resulting in missing or unlabeled Cruise IDs, Stations, Project, and Institutes. However, all datapoints do have 'Accession' values which correspond to the original data on the NCEI website (www.ncei.noaa.gov) as well as information about when the data were submitted, the data submitter, and the institute.

6. AquaMatch

AquaSat is a dataset of matched datapoints between *in situ* variables (chlorophyll, dissolved organic carbon, and suspended sediments) and satellite spectral reflectance from Landsat 5, 7, and 8 in inland American waters (Table 2) (Ross et al., 2019). The motivation behind AquaSat is similar to this chlorophyll dataset's; having comprehensive, regularly formatted *in situ* data points to be used for satellite validation and models. One component of AquaSat is a chlorophyll dataset

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

'AquaMatch Chlorophyll a Data from Water Quality Portal' (AquaMatch) with robust data quality tiering and relevant metadata (Brousil et al., 2024). While AquaMatch focuses more on inland waters (lakes, rivers, and estuaries), there are some points that fall within the study region (Fig. 1). Affiliations responsible for the chlorophyll samples were manually added to the dataset for metadata retention. Due to the extensive metadata and documentation, the AquaMatch flagging system was used to remove suspect chlorophyll points. The AquaSat dataset also includes CDOM measurements, but none in coastal waters, and so chlorophyll is the only variable from this dataset included.

While the AquaMatch documentation shows at one point the existence of a HPLC tag dictating if the method was used, this variable was dropped in the final dataset. Since HPLC methods cannot be determined it is assumed no HPLC methods were used, and the HPLC flag was assigned 1. This chlorophyll dataset only uses AquaSat's chlorophyll data subset to the coastal regions, though it should be noted that AquaSat has extensive inland data for total suspended solids, dissolved organic carbon, CDOM, and Secchi DiskDepth not used in this paper.

7. The Canadian Integrated Ocean Observing System

The Canadian Integrated Ocean Observing System (CIOOS) is the Canadian equivalent of IOOS and a hub of observational oceanographic data standardized across instruments and projects (Table 2) (Stewart et al., 2019). Established in 2019, the CIOOS repository holds fisheries, biogeochemical, and physical data including chlorophyll. For this dataset, the keyword 'chl' was searched on the CIOOS Atlantic and Pacific ERDDAP. Then, each resulting project was individually researched to ensure only extracted chlorophyll data were retained to avoid downloading gridded satellite data or fluorescence data. This resulted in 2 CIOOS datasets in the Pacific and one in the Atlantic. The Pacific datasets did not record HPLC data, but the Atlantic dataset regularly recorded bottle chlorophyll and HPLC chlorophyll, and so the HPLC flag was turned to True when applicable.

8. Biological and Chemical Oceanography Data Management Office

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) is a data repository funded by the National Science Foundation (NSF) with the purpose of housing and publishing oceanographic data from NSF funded projects (Table 2). Unlike other data repositories, BCO-DMO archives data in the format the project leaders dictate, resulting in irregularly formatted datasets. This makes bulk chlorophyll downloads difficult as projects save chlorophyll under different names. Additionally, the BCO-DMO website states that not all data is available on their ERRDAP. Therefore, it is not guaranteed that the BCO-DMO chlorophyll data compiled here is the total sum of all chlorophyll samples available. The HPLC flags were determined by researching each project's methods and variable definitions.

For this project, four methods were used to collect BCO-DMO data:

a. Unique chlorophyll keywords were searched on BCO-DMO's ERDDAP page (erddap.bco-dmo.org/erddap), including chl, chl_a, chlor, chlor_a, chlorophyll, chlorophyll_a, and Chl. The resulting projects were researched and, if it was determined that the chlorophyll samples were not *in vivo*, downloaded from the ERDDAP server.

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

b. Unique chlorophyll keywords were searched on the BCO-DMO's archive website (bco-dmo.org/search/dataset). For these projects, metadata including project names, principal investigators, and DOIs were manually saved to each project based on the project's landing page.

c. Specific projects, such as the Ecology and Oceanography of Harmful Algal Blooms (ECOHAB), were separately researched for on the BCO-DMO website. ECOHAB is a competitive research funding program focused on studying and understanding harmful algae blooms for developing models and preventative strategies. Since the project is focused on algae blooms, many datasets have well documented chlorophyll samples. Any projects with the key word 'ECOHAB' on BCO-DMO were researched and, if relevant, included in the dataset.

d. Data from the U.S. GLOBal ocean ECosystems dynamics (GLOBEC) were also organized from the database by searching the keyword 'GLOBEC' on BCO-DMO. All relevant datasets were downloaded and organized.

CDOM data on BCO-DMO were gathered by searching the keyword 'CDOM' and 'cdom' on the website and all results researched and processed if relevant. Again, due to the fractionated nature of the BCO-DMO repository, it is not guaranteed that this dataset holds a complete compilation of all CDOM data on BCO-DMO, but rather a compilation of the well documented, clearly labeled, and easily accessible CDOM data.

9. Hawaii Ocean Time Series

The Hawaii Ocean Time Series (HOTS) is a long term, ongoing monitoring program that samples at regular stations around the Hawaiian islands (Table 2). The motivation behind HOTS is to observe the physical and biochemical changes in the North Pacific subtropical gyre and document interannual variability. All chlorophyll data were downloaded from the Hawaii Ocean Time-series Data Organization & Graphical System (HOT-DOGS; hahana.soest.hawaii.edu/hot/hot-dogs). Latitude and longitude were manually appended based on the HOTS station information and converted from degrees and decimal minutes to decimal degrees to ensure consistent formatting. Pressure (units of decibars) was also converted to depth by dividing pressure by 1.02, derived from standard seawater density. HOTS records both HPLC and non-HPLC chlorophyll data, so the HPLC flag was determined depending on the presence or absence of 'chl_a' and converted to mg m^-3 for consistent units across datasets.

10. A Pan-Arctic Pigment Database for Phytoplankton and Sea-Ice Algae

The Pan-Arctic Pigment Database (Pan-Arctic) is a compilation of *in situ* pigment data exclusively using HPLC methods (Table 2) (Heidemann et al., 2026, manuscript under review). This includes HPLC chlorophyll for parts of Northern Alaska and Northern Canada from 2000-2024. As mentioned in the Section 2.1, HPLC samples are preferential to extracted chlorophyll and so these samples are crucial in understanding chlorophyll in this region. For this dataset, all measurements flagged as 'ice' are excluded as only seawater samples are wanted, and the HPLC flagged turned to True. It should be noted that as of writing, the manuscript is under review and so some details may have changed.

11. GLObal Reflectance community dataset for Imaging and optical sensing of Aquatic environments

The GLObal Reflectance community dataset for Imaging and optical sensing of Aquatic environments (GLORIA) project gathered hyperspectral R_{rs} and pigment data globally from a variety of water bodies to aid in remote sensing algorithm development (Table 2) (Lehmann et al., 2023). The GLORIA dataset consists of a wide array of variables, including chlorophyll, R_{rs}, and CDOM , however all CDOM data was recorded inland and so is not included here. For chlorophyll measurements, the HPLC flag was turned to "False" as methods for the remaining samples were not included.

The raw GLORIA R_{rs} dataset does not include metadata, so the reflectance data were matched to the station data to append datetime, location, and relevant project variables. The dataset was then subset to North America and R_{rs} transposed from 'wide format' to 'long format' (Table C1) for better data processing speeds and to be consistent with the SeaBASS R_{rs} dataset.

Author contributions

GM, KH, IC, and RV conceptualized this project, with methodology conducted by GM, KH, IC, MB, RV, and HS. Data curation was done by GM and KD, with investigation, validation, visualization, resources, and writing of the original draft conducted by GM. Supervision of project is through KH, IC, RV, with funding acquisition and administration by KH. GM, KH, IC, MB, RV, and HS conducted review and editing.

Competing interests

The authors declare that they have no conflict of interest.

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

We acknowledge all public data repositories, datasets, and data manuscripts for providing the data used in this paper, along with the team and researchers behind the organizations. We also thank the reviewers for their valuable comments and suggestions.

Financial support

This work was made possible by funding Inflation Reduction Act funding to the NOAA Fisheries Remote Sensing Strategic Initiative.

References

Bailey, S. W., & Werdell, P. J. (2006). A multi-sensor approach for the on-orbit validation of ocean color satellite data products. Remote Sensing of Environment, 102(1–2), 12–23. https://doi.org/10.1016/j.rse.2006.01.015

Brewin, R. J. W., Ciavatta, S., Sathyendranath, S., Jackson, T., Tilstone, G., Curran, K., Airs, R. L., Cummings, D., Brotas, V., Organelli, E., Dall'Olmo, G., & Raitsos, D. E. (2017). Uncertainty in Ocean-Color Estimates of Chlorophyll for Phytoplankton Groups. Frontiers in Marine Science, 4(104). https://doi.org/10.3389/fmars.2017.00104

Brousil, M. R., Meyer, M. F., Willi, K., Steele, B. G., De La Torre, J., & Ross, M. R. V. (2024). AquaMatch Chlorophyll a Data from Water Quality Portal: ~1970-2024 [Dataset]. Environmental Data Initiative. https://doi.org/10.6073/PASTA/2F750544112E5408928DD9A61E6ACE30

Dai, Y., Yang, S., Zhao, D., Hu, C., Xu, W., Anderson, D. M., Li, Y., Song, X.-P., Boyce, D. G., Gibson, L., Zheng, C., & Feng, L. (2023). Coastal phytoplankton blooms expand and intensify in the 21st century. Nature, 615(7951), 280–284. https://doi.org/10.1038/s41586-023-05760-y

Friedland, K. D., Stock, C., Drinkwater, K. F., Link, J. S., Leaf, R. T., Shank, B. V., Rose, J. M., Pilskaln, C. H., & Fogarty, M. J. (2012). Pathways between Primary Production and Fisheries Yields of Large Marine Ecosystems. PLoS ONE, 7(1), e28945. https://doi.org/10.1371/journal.pone.0028945

Heidemann, A. C., Hayward, A., Assmy, P., Basu, A., Bracher, A., Castellani, G., Ditullio, G., Dragańska-Deja, K., Fujiwara, A., Fragoso, G. M., Høyer, J., Hwang, J., Iversen, M., Von Jackowski, A., Juul-Pedersen, T., Kowalczuk, P., Lee, Y., Matsuoka, A., Merz, A., … Gonçalves-Araujo, R. (2026). A Pan-Arctic Pigment Database for Phytoplankton and Sea-Ice Algae. https://doi.org/10.5194/essd-2026-144

Hu, C., Feng, L., Lee, Z., Franz, B. A., Bailey, S. W., Werdell, P. J., & Proctor, C. W. (2019). Improving Satellite Global Chlorophyll a Data Products Through Algorithm Refinement and Data Recovery. Journal of Geophysical Research: Oceans, 124(3), 1524–1543. https://doi.org/10.1029/2019JC014941

Hyde, K. J. W., O'Reilly, J. E., & Oviatt, C. A. (2007). Validation of SeaWiFS chlorophyll a in Massachusetts Bay. Continental Shelf Research, 27(12), 1677–1691. https://doi.org/10.1016/j.csr.2007.02.002

Integrated Ocean Observing System (U.S.). (2017). Manual for real-time quality control of phytoplankton data: A guide to quality control and quality assurance of phytoplankton observations. https://doi.org/10.7289/V56D5R6S

Jänes, H., Macreadie, P. I., Zu Ermgassen, P. S. E., Gair, J. R., Treby, S., Reeves, S., Nicholson, E., Ierodiaconou, D., & Carnell, P. (2020). Quantifying fisheries enhancement from coastal vegetated ecosystems. Ecosystem Services, 43, 101105. https://doi.org/10.1016/j.ecoser.2020.101105

Lehmann, M. K., Gurlin, D., Pahlevan, N., Alikas, K., Conroy, T., Anstee, J., Balasubramanian, S. V., Barbosa, C. C. F., Binding, C., Bracher, A., Bresciani, M., Burtner, A., Cao, Z., Dekker, A. G., Di Vittorio, C., Drayson, N., Errera, R. M., Fernandez, V., Ficek, D., … Yue, L. (2023). GLORIA - A globally representative hyperspectral in situ dataset for optical sensing of water quality. Scientific Data, 10(1), 100. https://doi.org/10.1038/s41597-023-01973-y

Mélin, F., & Vantrepotte, V. (2015). How optically diverse is the coastal ocean? Remote Sensing of Environment, 160, 235–251. https://doi.org/10.1016/j.rse.2015.01.023

Milton, G., Hyde, K., Cetinic, I., Beal, M., Vandermeulen, R., & Synan, H. (2026). Coastal North America In Situ Chlorophyll and Optical Dataset [Dataset]. Zenodo. https://doi.org/10.5281/ZENODO.20140178

Mishonov, A. V. (2024). World Ocean Database 2023. NOAA Atlas NESDIS ; 97. https://doi.org/10.25923/Z885-H264

Morel, A., & Prieur, L. (1977). Analysis of variations in ocean color1. Limnology and Oceanography, 22(4), 709–722. https://doi.org/10.4319/lo.1977.22.4.0709

Neeley, A. R., Cetinić, I., & Thomas, C. (2025). Evaluation of fluorometrically-derived chlorophyll a as a satellite ocean color validation product using statistical metrics. Optics Express, 33(5), 10212. https://doi.org/10.1364/OE.549547

Nezlin, N. P., McLaughlin, K., Booth, J. A. T., Cash, C. L., Diehl, D. W., Davis, K. A., Feit, A., Goericke, R., Gully, J. R., Howard, M. D. A., Johnson, S., Latker, A., Mengel, M. J., Robertson, G. L., Steele, A., Terriquez, L., Washburn, L., & Weisberg, S. B. (2018). Spatial and Temporal Patterns of Chlorophyll Concentration in the Southern California Bight. Journal of Geophysical Research: Oceans, 123(1), 231–245. https://doi.org/10.1002/2017JC013324

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., Kahru, M., & McClain, C. (1998). Ocean color chlorophyll algorithms for SeaWiFS. Journal of Geophysical Research-Oceans, 103(C11), 24937–24953. https://doi.org/10.1029/98jc02160

O'Reilly, J. E., & Werdell, P. J. (2019). Chlorophyll algorithms for ocean color sensors—OC4, OC5 & OC6. Remote Sensing of Environment, 229, 32–47. https://doi.org/10.1016/j.rse.2019.04.021

Ross, M. R. V., Topp, S. N., Appling, A. P., Yang, X., Kuhn, C., Butman, D., Simard, M., & Pavelsky, T. M. (2019). AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters. Water Resources Research, 55(11), 10012–10025. https://doi.org/10.1029/2019WR024883

Sathyendranath, S., Brewin, R., Brockmann, C., Brotas, V., Calton, B., Chuprin, A., Cipollini, P., Couto, A., Dingle, J., Doerffer, R., Donlon, C., Dowell, M., Farman, A., Grant, M., Groom, S., Horseman, A., Jackson, T., Krasemann, H., Lavender, S., … Platt, T. (2019). An Ocean-Colour Time Series for Use in Climate Studies: The Experience of the Ocean-Colour Climate Change Initiative (OC-CCI). Sensors, 19(19), 4285. https://doi.org/10.3390/s19194285

Stewart, A., deYoung, B., Smit, M., Donaldson, K., Reedman, A., Bastien, A., Brunsting, R., Carter, B., Covey, B., Kelly, R., Peterson, E., Pirenne, B., Plourde, A., Ste-Marie, A.-S., Tollefsen, C., Wallace, D., & Whoriskey, F. (2019). The Development of a Canadian Integrated Ocean Observing System (CIOOS). Frontiers in Marine Science, 6, 431. https://doi.org/10.3389/fmars.2019.00431

https://doi.org/10.5194/essd-2026-599
Preprint. Discussion started: 24 August 2026

Stock, C. A., John, J. G., Rykaczewski, R. R., Asch, R. G., Cheung, W. W. L., Dunne, J. P., Friedland, K. D., Lam, V. W. Y., Sarmiento, J. L., & Watson, R. A. (2017). Reconciling fisheries catch and ocean productivity. Proceedings of the National Academy of Sciences, 114(8). https://doi.org/10.1073/pnas.1610238114

Valente, A., Sathyendranath, S., Brotas, V., Groom, S., Grant, M., Jackson, T., Chuprin, A., Taberner, M., Airs, R., Antoine, D., Arnone, R., Balch, W. M., Barker, K., Barlow, R., Bélanger, S., Berthon, J.-F., Beşiktepe, Ş., Borsheim, Y., Bracher, A., … Zibordi, G. (2022). A compilation of global bio-optical in situ data for ocean colour satellite applications – version three. Earth System Science Data, 14(12), 5737–5770. https://doi.org/10.5194/essd-14-5737-2022

Valente, A., Sathyendranath, S., Brotas, V., Groom, S., Grant, M., Taberner, M., Antoine, D., Arnone, R., Balch, W. M., Barker, K., Barlow, R., Bélanger, S., Berthon, J. F., Besiktepe, S., Brando, V., Canuti, E., Chavez, F., Claustre, H., Crout, R., … Zibordi, G. (2016). A compilation of global bio-optical in situ data for ocean-colour satellite applications. Earth Syst. Sci. Data Discuss., 2016, 1–41. https://doi.org/10.5194/essd-2015-37

Van Oostende, N., Dussin, R., Stock, C. A., Barton, A. D., Curchitser, E., Dunne, J. P., & Ward, B. B. (2018). Simulating the ocean's chlorophyll dynamic range from coastal upwelling to oligotrophy. Progress in Oceanography, 168, 232–247. https://doi.org/10.1016/j.pocean.2018.10.009

Wei, J., Wang, M., Mikelsons, K., Jiang, L., Kratzer, S., Lee, Z., Moore, T., Sosik, H. M., & Van Der Zande, D. (2022). Global satellite water classification data products over oceanic, coastal, and inland waters. Remote Sensing of Environment, 282, 113233. https://doi.org/10.1016/j.rse.2022.113233

Werdell, P. J., Bailey, S., Fargion, G., Pietras, C., Knobelspiesse, K., Feldman, G., & McClain, C. (2003). Unique data repository facilitates ocean color satellite validation. Eos, Transactions American Geophysical Union, 84(38), 377–387. https://doi.org/10.1029/2003EO380001