BIO-MATE: A biological ocean data reformatting effort

Biological ocean data collected from ships find reuse in aggregations of historical data. These data are heavily relied upon to document long term change, validate satellite algorithms for ocean biology and are useful in assessing the performance of autonomous platforms and biogeochemical models. There is a need to combine subsurface biological and physical data into one aggregate data product to support reproducible research. Existing aggregate products are dissimilar in source data, have largely been isolated to the surface ocean and most omit physical data. These products cannot easily be used to explore subsurface bio-physical relationships. We present the first version of a biological ocean data reformatting effort (BIO-MATE, <https://gitlab.com/KBaldry/BIO-MATE>). BIO-MATE uses R software that reformats openly sourced published datasets from oceanographic voyages. These reformatted biological and physical data from underway sensors, profiling sensors, pigments analysis and particulate organic carbon analysis are stored in an interoperable and reproducible BIO-MATE data product for easy access and use ([AODNlink]).

# Background & Summary

Marine phytoplankton blooms support ocean food-webs and influence global climate through the biological carbon pump (1;2:3). Ocean physics and other environmental drivers control the timing, magnitude and extent of phytoplankton blooms through complex bio-physical relationships (4;5;6;7). To study these relationships integrated data structures that link biological and physical ocean data are needed. Ship-based data are the gold standard for accurate biological oceanographic measurements (8). These data are often published separately to physical ocean data, stored across different repositories and in multiple formats. This makes it difficult and time-consuming to aggregate and link biological and physical data. The described data product attempts to make this task easier.

The biological ocean data reformatting effort (BIOMATE) works to link existing, open-access biological and physical datasets across oceanographic voyages and promote their re-use. This has been done by developing a BIOMATE R software package that not only reformats published datasets, but also cross-references between biological and physical data and allows access to citation information (<https://git> hub.com/KimBaldry/BIOMATE-Rpackage). The resulting BIOMATE data product allows users to easily access, manipulate and cite published ship-based datasets of different dimensions for multiple applications.

The BIO-MATE data product can be accessed via the Australian Ocean Data Network (<https://portal.aodn.org.au/>). The aggregation includes data collected from shipboard underway sensors, profiling sensors mounted on sampling rosettes, lab analysis for phytoplankton pigments and lab analysis for particulate organic carbon (POC). These data are stored as four data streams, cross-referenced by unique expedition codes (EXPOCODE) and profiling station identifications (CTD\_ID). An additional data stream contains supporting information for the data product including a list of oceanographic voyages, investigator contact information and data citations for reformatted datasets. We have also included an aggregated data table for biological data. Users are required to refer to supporting data and cite all data products accessed through BIO-MATE, as well as the BIO-MATE data product itself. We consulted the distribution licenses of all data sources to ensure that with this condition data are re-used lawfully.

The data product is currently has been used to address discrepancies between Southern Ocean satellite and in-situ measurements (Johnson *et al.* in prep), to understand how the response of in-situ fluorometers changes in the Southern Ocean (Baldry *et al*. in prep), and to investigate the role of ocean physics in mediating subsurface chlorophyll features (Baldry *et al.* in prep). These examples highlight the malleability of this data product to improve our understanding of biological oceanography in the Southern Ocean. Further uses include validating satellite observations (9;10), developing new ways to validate in-situ bio-optical observations collected by autonomous profiling platforms in the presence of dynamic fronts (11;12;8), training ocean state estimations (13), informing bio-physical models () and using multi-variate analyses to understand bio-physical relationships ().

We recognise the massive effort in producing the thousands of data records in this data product. This includes the investigators and data officers who have spent countless hours in ship time, project organisation, grant writing, laboratory analysis, data processing and report writing. Oceanographic data are often collected with regional studies in mind, but their value increases with publication and re-use. We encourage all investigators to publish their data for re-use through data products like BIO-MATE.

# Methods

#### *Published datasets in BIO-MATE*

The BIO-MATE aggregate data product brings together ship-based data that have been collected by a Principal Investigator (PI) and openly published by a to a publicly accessible data (Figure 2). The first version of BIO-MATE includes published datasets associated with four types of measurements: 1. sensors in the vessels underway seawater in-take (underway sensor data stream),  
2. profiling sensors mounted to sampling rosettes (profiling sensor data stream),  
3. pigments measured in the laboratory (pigment data stream),  
and 4. POC measured in the laboratory (POC data stream).

Data records from the pigment data stream were first identified in data repositories that host biological data (November 2019). Pigment data records were identified using the search term “chlorophyll” and a latitude bound of 30 S to 90 S from PANGAEA (<https://www.pangaea.de/>), SeaBASS (<https://seabass.gsfc.nasa.gov/>), the Australian Ocean Data Network (AODN, <https://portal.aodn.org.au/>), GLODAP (<https://www.glodap.info/>), the Palmer Long Term Ecological Record (Pal-LTER, <https://pal.lternet.edu/data>), the Biological and Chemical Oceanography Data Management Office (BCO-DMO, <https://www.bco-dmo.org/data>), the CSIRO Marlin Data Trawler (Marlin, <https://www.cmar.csiro.au/data/trawler/>) and the Australian Antarctic Data Center (AADC, <https://data.aad.gov.au/>). Data records from the profiling sensors and underway sensors data streams were then identified in these repositories and in the CCHDO (<https://cchdo.ucsd.edu/>) and MGDS (<https://www.marine-geo.org/>).

From pigment data records, 178 relevant voyages were identified using unique 12-digit expedition codes (EXPOCODES) assigned as follows; National Oceanographic Data Centre (NODC) platform codes followed by voyage 8 digit start dates (YYYYMMDD). NODC platform and country codes are recorded on Git Hub (<https://github.com/KimBaldry/BIO-MATE/product_data/supporting_information/codes>) and within the BIOMATE software (<https://github.com/KimBaldry/BIOMATE-Rpackage/inst/codes>). If the vessel name or voyage start/end dates were absent, this information was found using Google to discover voyage records. This voyage information was used to do a final Google keywords search (i.e. ship name, synonyms for voyages, year, “underway”/ “CTD”, “chlorophyll”, ”POC”, “cruise report” and “data”) to determine if any absent records were findable and find accompanying cruise reports.

#### *Semi-automated BIO-MATE workflow for reformatting datasets*

A semi-automated workflow and the BIO-MATE R software (<https://github.com/KimBaldry/BIOMATE-Rpackage>) were used to reformat published datasets, and produce the BIO-MATE data product (Figure 3). Downloaded data files were split by EXPOCODES if they recorded data within a larger dataset (e.g PALMER-LTER data records). Files for the profiling sensor data stream were further split into individual profiles. Processing metadata were manually entered into a table to inform the BIO-MATE R software and a bulk run of the software was performed to reformat data files. The workflow is described in more detail in the following subsections (Figure 2).

##### *Download of published datasets*

Published datasets were manually downloaded from open source repositories and stored locally in accordance with data policies. Some manual reformatting of a small portion of downloaded data had to be performed on old datasets, prior to the application of reformatting scripts, due to formatting irregularities. Downloaded data files, and their amendments, used to create the BIO-MATE data product are not published in BIO-MATE, but are available upon request to the corresponding author.

##### *Splitting large datasets with BIO-MATE software*

The BIO-MATE R software requires each file to only contain observations from a single voyage. Further, the profiling sensor data stream requires each file to only contain observations from a single profiling cast, held in a discrete directory for each voyage.

The *split\_delim\_file* function splits files using identified variables containing EXPOCODE synonyms and/or profiling station information. This function can be used to split a single, large data file into smaller files as required. For this version of the data product, a number of files had to be split to be ingested into the BIO-MATE core functions. A record of these can be found in Git Hub ().

##### *Processing metadata*

Information on file formats, dataset information, citation information, location data variables and ocean data variables are needed to reformat published datasets with BIO-MATE software. This information is called processing metadata herein, and was manually entered and stored as comma delimited text files. The processing metadata required to run BIO-MATE software is described in Table 1, and differs for each data stream. All processing metadata used to construct the BIOMATE aggregated data product is stored in Git Hub (<https://github.com/KimBaldry/BIO-MATE/tree/main/product_data/processing_metadata>).

##### *Dataset citation with BibTEX files*

Information is included in the BIO-MATE data product, for citing published datasets, laboratory analysis methodologies (for the PIG and POC data streams) and the data repositories through which published data records were accessed. Each citation was recorded as a BibTeX entry, compatible with EndNote, R and LaTeX. Each BibTeX entry has a tag that is referenced in the processing metadata. This tag is used to link citations to their corresponding data records when datasets are ingested in the BIO-MATE software. Citation information is then printed in the header information in reformatted files. Where possible BibTeX entries were sourced from data repositories. If BibTeX entries were not found, they were created manually.

All BibTeX entries are stored on Git Hub (<https://github.com/KimBaldry/BIO-MATE/product_data/supporting_information/citations>) and in the BIOMATE software (<https://github.com/KimBaldry/BIOMATE-Rpackage/inst/citations>). A look-up table is included in the BIO-MATE software to help users find relevant BibTeX entries needed to cite datasets appropriately (). A function *export\_ref* supports the export of a smaller BibTeX file based on user selections of EXPOCODES and data streams that they have accessed through the product.

##### *Reformatting and linking data streams with BIO-MATE R software*

The BIO-MATE R software was run to reformat data files to the WHP-Exchange format (<https://exchange-format.readthedocs.io/en/latest/index.html>), using the original or split data files, processing metadata and citation information as input. The software arranges reformatted WHPE files into four data streams in local directories that include separate WHPE files, for each EXPOCODE, and for underway sensors, profiling sensor casts, pigment measurements, and POC measurements.

Each data stream has its own reformatting function within the BIO-MATE R software (*UWY\_to\_WHPE*, *PROF\_to\_WHPE*, *PIG\_to\_WHPE*, *POC\_to\_WHPE*). The software requites physical (underway sensor and profiling sensor) data streams to be reformatted before biological (pigment and POC) data streams to accommodate a biological-physical matching algorithm within the PIG\_to\_WHPE and POC\_to\_WHPE functions. The algorithm links biological data in the pigment and POC data streams to the physical data in the profiling sensor and underway sensor data streams. Biological data records are given a profiling sensor identification tag (CTD\_ID) if matched to physical data in BIO-MATE.

To match biological data to physical data, the algorithm first uses EXPOCODES to find relevant physical data in profiling sensor data streams. It then matches biological and physical data records by comparing station number (STNBR) and cast number (CASTNO) records. If matches are detected using STNBR and CASTNO, the validity of these matches is checked by comparing time and position information. If position and time were not recorded in biological datasets, it is assumed that the STNNBR and CASTNO records are correct. Otherwise, a match is recorded if both the biological and physical data record data either within 24 hours of each other or within 8 km (14). After this routine, if unmatched biological data still exist, a database of time and position information from all profiling sensor data relating to the EXPOCODE is created. Matches are found for biological data, if it contains position information, by finding the closest profiling sensor record within 1km in the database. If time information exits, matches are identified as the closest profiling sensor record within 6 hours, otherwise only matching date information is required.

[underway matching info if developed]

#### *Quality assurance*

Limited quality assurance has been performed on the BIO-MATE data product and is variable across published datasets. The initial integrity of these data records lies with the Principal Investigators of the published data record. As a result, reformatted data have varying levels of quality control and post-processing. We have included cruise report citations in our product to aid in further data quality assurance efforts.

This allows a range of users to benefit from the BIO-MATE aggregate product and ensures data remains to the standard it was published at. The quality assurance required of physical and biological ocean data varies according to application, and is up to the user to confirm the data is suitable for their application. For example, when assessing basin-scale trends, a lower level of quality assurance is required, compared to when validating or training satellite algorithms or ocean models. Future versions of BIO-MATE could implement quality assurance metrics.

# Data Records

The four data streams are all stored on the IMAS data portal (<https://data.imas.utas.edu.au/portal/search>), linked through unique EXPOCODES. Supporting data contains a metadata table and BibTeX citation files. The spatial extent of the data records is confined largely to the Southern Ocean, and was collected from XXXX - XXXX (Figure 4). A summary of the data records in the BIO-MATE aggregate data product is presented in Table 2.

#### *Underway sensor data stream*

The underway sensor data stream contains a comma delimited WHP-Exchange file for each voyage ([EXPOCODE]\_UWY.csv). The format of this file consists of headers to store metadata, followed by a data table that reports records collected by underway sensors mounted on the vessel (Table ??).

#### *Profiling sensor data stream*

The profiling sensor data stream contains a comma delimited WHP-Exchange file for each unique profiling cast conducted on each voyage ([EXPOCODE]*[station number]*[cast number]\_ctd1.csv). The file is formatted to store metadata as headers which is followed by the data table that reports records from profiling sensors mounted on a sampling rosette (Table 3).

#### *Pigment data stream*

The pigment data stream contains a comma delimited WHP-Exchange file for each voyage (named [EXPOCODE]*PIG*[SOURCED\_FROM]\_[METHOD].csv). The format of this file consists of headers to store supporting information, followed by a data table that records measurements from the lab analysis of seawater samples for pigments performed by principle investigators (Table 4).

#### *Particulate Organic Carbon data stream*

The POC data stream contains a comma delimited WHP-Exchange file for each voyage (named [EXPOCODE]*POC*[SOURCED\_FROM]\_[METHOD].csv). The format of this file consists of headers to store supporting information, followed by a data table that records measurements from the lab analysis of seawater samples for particulate organic carbon performed by principle investigators (Table ??).

#### *Supporting data*

Supporting data are included in the BIO-MATE aggregate data product to support the correct citation of data and guide user access to data. This data includes 1. a BibTeX file, that contains information to reference all BIO-MATE data records 2. an index table indicating data availability and citation tags against data records listed by EXPOCODE, data stream, method and source, 3. a records table for all data repositories from which BIO-MATE data was sourced from and 4. a records table for all pigment and POC analysis methods used in BIO-MATE data.

# Technical Validation

We validated the quality of the BIO-MATE data compilation, by displaying a number of key data distributions and trends. This validation does not confirm the quality of individual data points, in which the authors have placed no additional quality assurance to the published datasets.

The location data associated with the published datasets has been interpreted correctly by the software. This is evident from the success of the biophysical matching algorithm, along with the spatial distribution of the data and recorded sampling depths (Figure 4). The data are spread over - - and predominantly collected in the month of January. This is consistent with the fact that ship-based sampling in the Southern Ocean is conducted during Austral summer, and displays a lag time in publishing most recent datasets to data repositories. All data are in the ocean, not on land, confirming the absence of spurious location data, and most samples are located in the Southern Ocean which is consistent with our search constraints. Finally information on sampling time of ship-based biological data is as expected, and CTD sampling times (start, bottom and end) are sequential and follow a trend with sampling depth (Figure 7).

The biological ocean data associated with the published datasets has been interpreted correctly by the software. Over-all fluorometrically derived chlorophyll (FCHLORA), HPLC derived chlorophyll a (Chl a) and HPLC derived total chlorophyll (TCHLA) measurements show a log-normal distribution, as expected. High values (>20 3BCg/l) are constrained to the coastal zones as expected (Figure 6). There is a linear relationship between chlorophyll-a derived from HPLC methods and chlorophyll derived from fluorometric methods (Figure 8). This is expected, although considerable variability is expected due to the influence of phaeopigments and other accessory pigments on fluorometrically derived chlorophyll measurements.

Five fluorometric methods to derive chlorophyll have coincident HPLC measurements. Briefly, the ANTXVIII\_2 and JGOFS method shows good correlation between the two. The PALMER\_LTER method shows considerable variability between methods. This may be due to the coastal location of most samples and the influence of accessory pigments, but further investigation is needed. Finally, only a small number of records were collected with Mueller *et al.* (2003) and unknown fluorometric methods (< X), making it difficult to assess the quality of the measurements.

The physical ocean data associated with published datasets has been interpreted correctly by the software. Temperature and salinity ranges fall within expected vales for the ocean, and display expected trends with latitude (Figure5).

# Usage Notes

The community is welcome to contribute to the development of BIO-MATE software and to contribute published data to the aggregation, by following a user guide (Figure 3).

#### *Contributing to BIO-MATE software development*

It is recommended that changes to BIO-MATE software be made through Git Hub. Contributors can fork the existing repository (<https://github.com/KimBaldry/BIOMATE-Rpackage>) and make changes directly to the source code. Once changes are made, they can be directed back to the BIOMATE R package repository and released as an updated version of the BIO-MATE software. If the BIO-MATE source code is to be significantly developed, we suggest that the corresponding author is contacted and a hand-over of the software is negotiated. We encourage the addition of new data streams to BIO-MATE, the expansion of BIO-MATE capabilities, the addition of quality assurances and increases in software efficiency.

#### *Contributing data to BIO-MATE*

Users can submit published biological ocean data to BIO-MATE using the R shiny app BIO-SHARE (Figure 3). Once data are submitted they can be downloaded by the user and automatically submitted to the BIO-MATE Git Hub repository for future addition into the product. We ask that all data submitted to BIO-MATE are published elsewhere and that users enter an accurate citation for the data they are submitting.

For large data submissions, users can create their own workflows using the BIO-MATE R package to reformat data and information (Figure 3). Once data have been reformatted, they can be submitted to the corresponding author via Git Hub or direct communication.

Currently, BIO-MATE only supports data files stored in text-delimited formats, with structured headers and columns in a data table, and NetCDF format. The user is required to enter in some metadata to inform the software on input formats.

#### *Recommended use in data analyses*

We encourage the use of the data aggregate product as a new integrated database of biological and physical data. Data files from selected voyages can be identified using unique EXPOCODES and CTD\_IDs. This makes it easy to use multiple data streams in analysis, by indexing files across these EXPOCODES. Alternatively, the selection tool on the IMAS repository helps users to select voyages using spatial bounds.

# Code Availability

All data processing was performed in R software (Version 1.1.423). The BIO-MATE R software is freely available (<https://github.com/KimBaldry/BIOMATE-Rpackage>). The semi-automated workflow and accompanying processing data used to construct the data product, along with the code used to create the data descriptor is freely accessible via Git Hub (<https://github.com/KBaldry/BIO-MATE>).

# Acknowledgements

This research was supported under Australian Research Council’s Special Research Initiative for Antarctic Gateway Partnership (Project ID SR140300001) and a 2019 Fellowship from the Scientific Committee of Antarctic Research. It was also partially supported by Geoscience Australia, in the form of funded study leave. Data included in the data product were made available by the following data repositories; PANGAEA, AODN/IMOS, SeaBASS, CCHDO, AADC, GLODAP, PAL-LTER, CSIRO, MDGS and BCO-DMO. Records of data access dates, source addresses and digital object identifiers are recorded as metadata, alongside appropriate data citations. We acknowledge the enormous community effort undertaken in the collection, analysis and publication of this data and thank principle investigators for publishing their data in open access repositories.

# Author Contributions

KB designed the data product, performed the data aggregation and wrote the manuscript. RJ contributed to the data product design and manuscript edits.

# Competing Interests

The authors of this manuscript declare no conflicts of interest.

# Figures

pig\_data$SA = unlist(lapply(1:nrow(pig\_data),FUN = function(x){  
 if(!is.na(pig\_data$DATE\_analyser[x]) & !is.na(pig\_data$TIME\_analyser[x]) & !is.na(pig\_data$LAT\_analyser[x]) & !is.na(pig\_data$LON\_analyser[x])){SA = sunAngle( t = as.POSIXct(paste(pig\_data$DATE\_analyser[x], pig\_data$TIME\_analyser[x]), tz = "UTC"), longitude = pig\_data$LON\_analyser[x], latitude = pig\_data$LAT\_analyser[x])$altitude}else{SA = NA}  
 SA  
}))  
pig\_data = pig\_data %>% mutate(dd = as.numeric(substr(DATE\_analyser, 9,10)), MM = as.numeric(substr(DATE\_analyser,6 ,7)), YYYY = as.numeric(substr(DATE\_analyser, 1, 4)), HH = as.numeric(substr(TIME\_analyser, 1,2)))  
pig\_data$DEPTH[pig\_data$DEPTH %in% c("uw","Underway", "underway","surface","ice","BUCKET")] = 0  
pig\_data$DEPTH = as.numeric(pig\_data$DEPTH)

## Warning: NAs introduced by coercion

pig\_data$DEPTH[pig\_data$STNNBR == "U" & is.na(pig\_data$DEPTH)] = 0   
pig\_data$DEPTH[pig\_data$DEPTH < 0] = 0  
  
CTD\_df = CTD\_df %>% mutate(dd = as.numeric(substr(DATE, 9,10)), MM = as.numeric(substr(DATE,6 ,7)), YYYY = as.numeric(substr(DATE, 1, 4)))

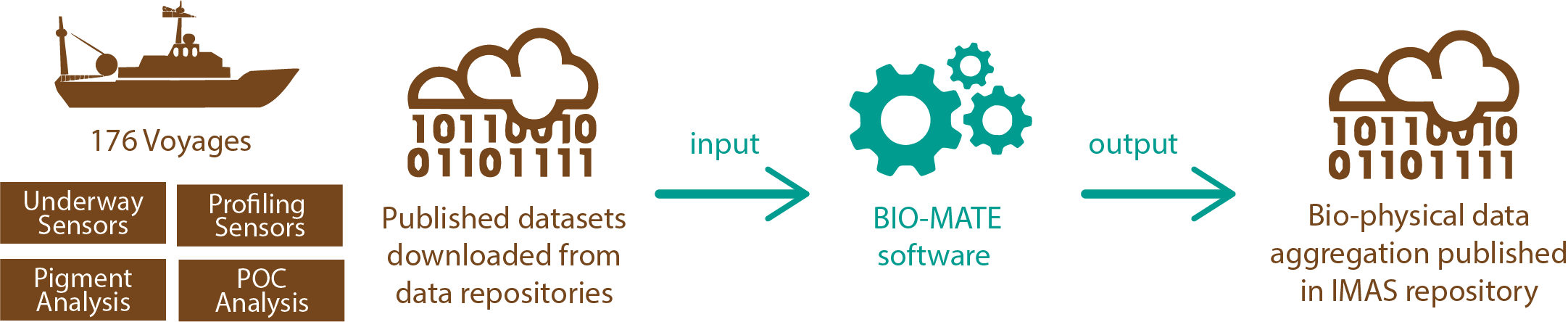


Figure 1: The BIO-MATE concept for creating a consistent data compilation from existing ship-based oceanographic data

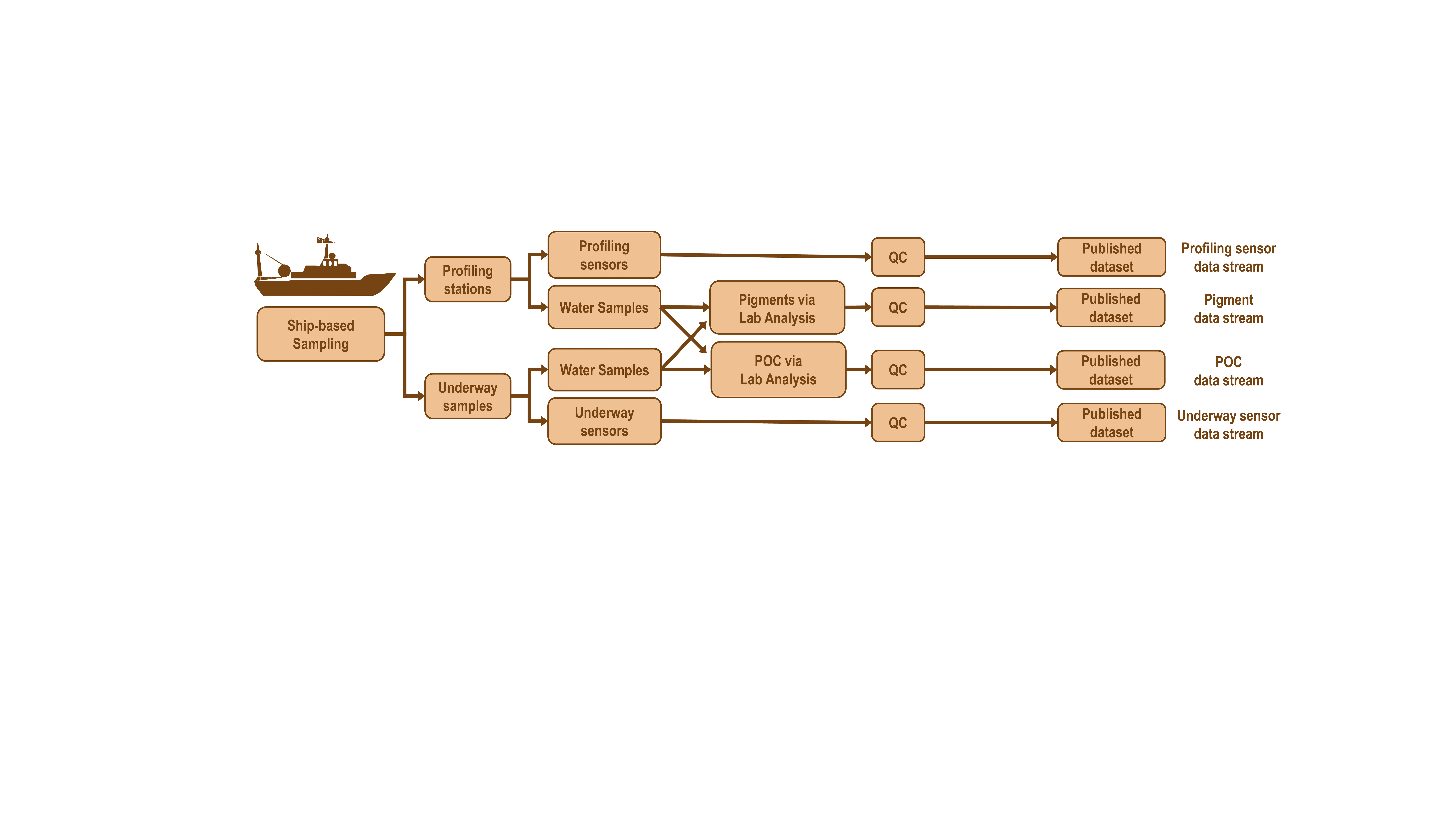


Figure 2: The origins of the data in the BIO-MATE data streams.

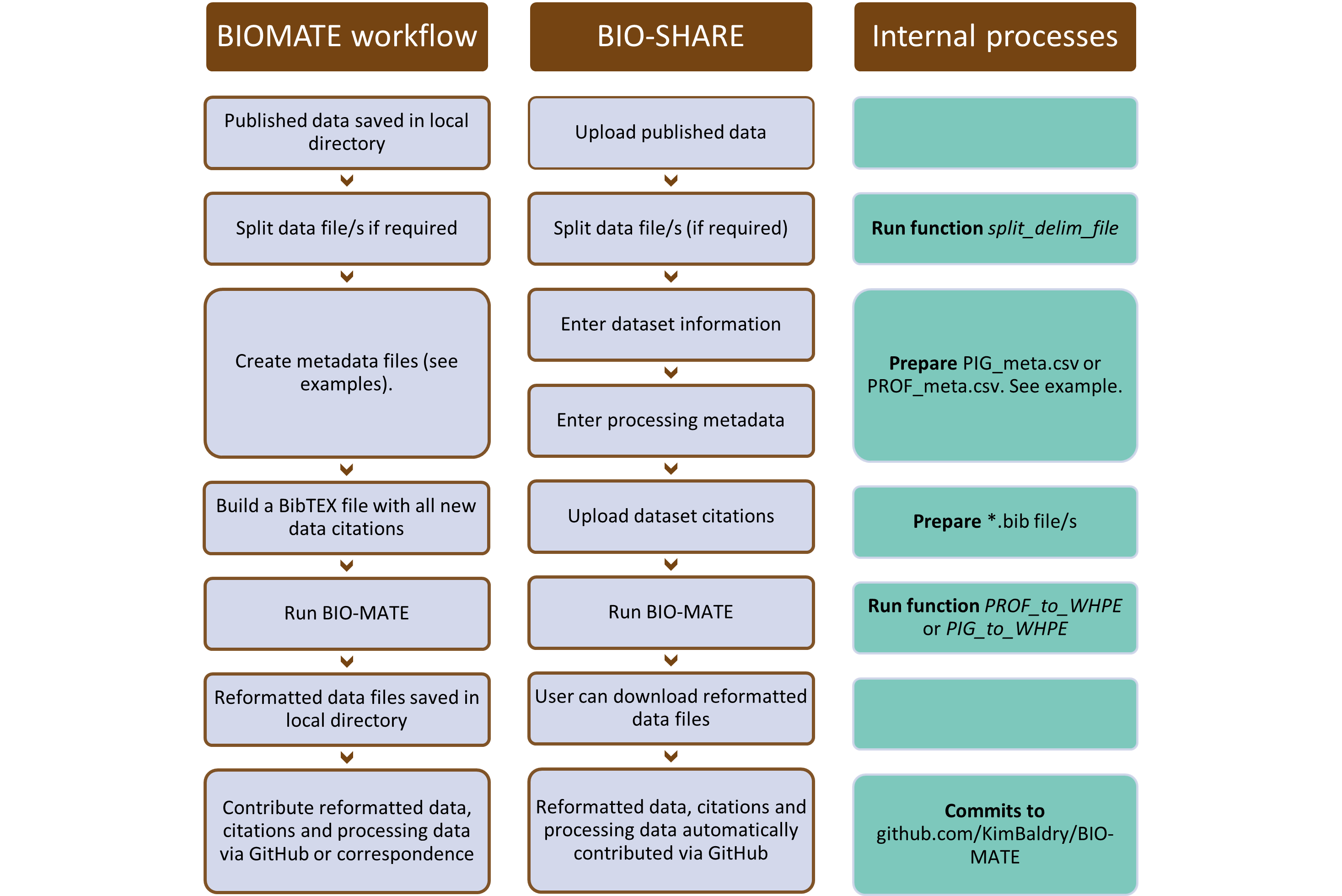


Figure 3: A schematic demonstrating the BIO-MATE workflow and how it is implemented in BIO-SHARE

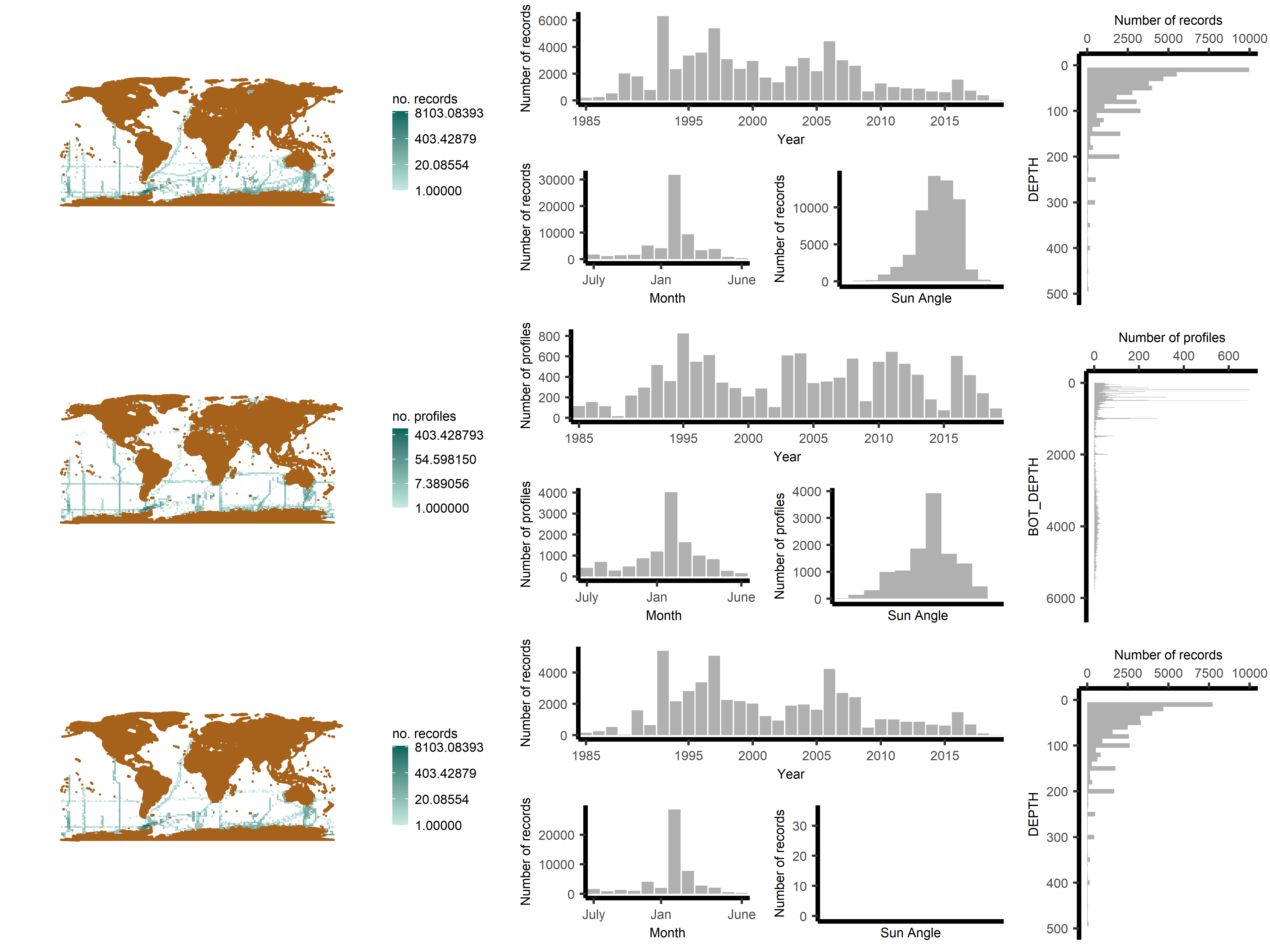


Figure 4: The spatiotemporal distribution of different data streams and bio-physical matches in the BIOMATE data compilation

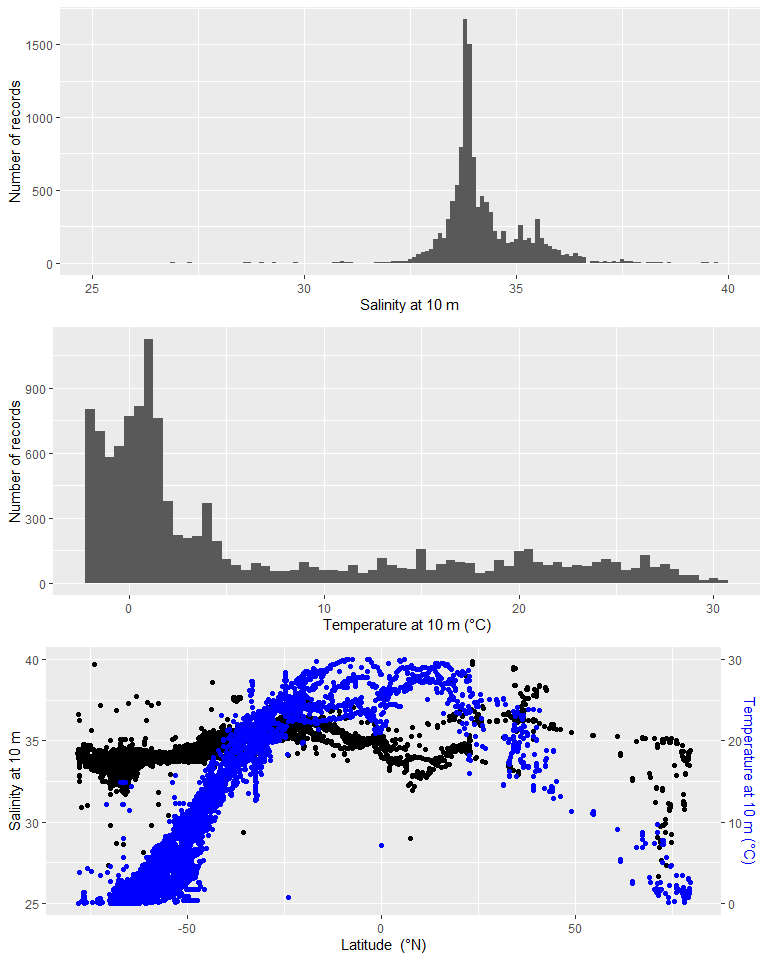


Figure 5: The distribution of temperature and salinity data measured at 10m by profiling sensors in the BIOMATE data compilation

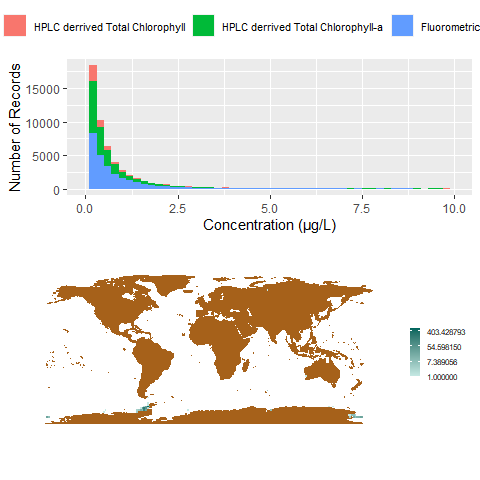


Figure 6: The (a) distribution of Chla, TCHLA and FCHLA in the BIO-MATE data compilation and (b) the location of high (>10 μg/l) Chla, TCHLA and FCHLA measurements.

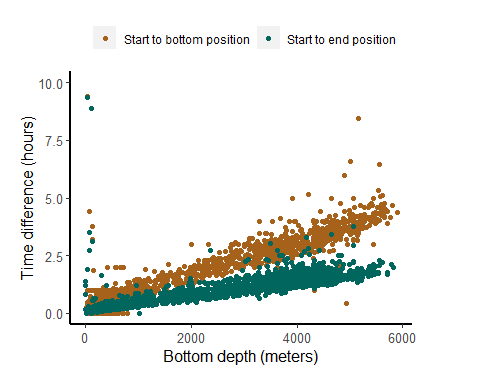
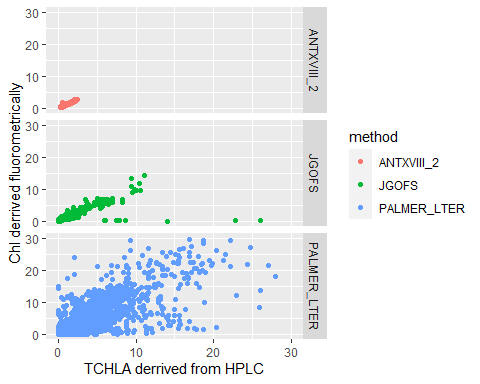


Figure 7: The time difference between the bottom and end possitions of a profiling sensor cast versus the bottom depth of the cast.

 # Tables

**Table 1:** a figure

| Processing metadata variable | Data stream/s | Description | Input Guide |
| --- | --- | --- | --- |
| **File format information** | | | |
| file\_type | all | The format of the file/s. | text delim or netcdf |
| path | all | A path to where the file/s is stored. | A pathname that is R compatible |
| extention | all | The extension of the file/s. | text delim or netcdf |
| delim | all | Only fill if rectangular text-delimited file/s. | rect |
| header\_sep | all | A separator used in headers of the file. Headers often store location data in profiling datasets and need extraction. Can be left empty. | colon, comma, dash, equals, space |
| missing\_value | all | The value or character used to indicate a missing value. | value |
| not\_detected | PIG, POC | The value or character used to indicate a variable was not detected in analysis. | value |
| **Data aquisition information** | | | |
| EXPOCODE | all | The EXPOCODE of the voyage associated with the data. | 12-digit code |
| source | all | The data repository the data files were sourced from. | The short name of the data repository used within BIO-MATE. See https://github.com/KimBaldry/BIO-MATE/product\_data/supporting\_information/BIOMATE\_SOURCES.txt. |
| PI | all | The principle investigator/s responsible for the published dataset. | Names sepparated by a dash |
| Institution | all | The institution/s who collected the data. | Names sepparated by a dash |
| contact | all | A contact for the published dataset. | E-mail address |
| citation | all | The BIO-MATE citation tag/s used to reference a BibTEX entry for the published dataset. | A BIO-MATE citation tag |
| analysis\_type | PIG, POC | The type of analysis used on water samples for the published dataset. | A code to reference an analysis type. See https://github.com/KimBaldry/BIO-MATE/product\_data/supporting\_information/BIOMATE\_METHODS.txt |
| Method | PIG, POC | The method used to analyse water samples for the published dataset. | A code to reference a method. See https://github.com/KimBaldry/BIO-MATE/product\_data/supporting\_information/BIOMATE\_METHODS.txt |
| **Location data information** | | | |
| TZ | all | The time zone for date and time information. | Time zone code |
| STNNBR | all | The name of the variable for the station number of the profiling staation. | Text. If recorded in header use header-[variable]. |
| CASTNO | all | The name of the variable for the cast number at the profiling station. | Text. If recorded in header use header-[variable]. |
| DATE | PROF | The name of the variable for the date of the profiling cast. | Text. If recorded in header use header-[variable]. |
| DATE\_analyser | PIG, POC | The name of the variable for date of observation recorded by the analyser. | Text. If recorded in header use header-[variable]. |
| DATE\_format | PROF | Format for DATE. | A format string code. See strptime in R for codes. |
| DATE\_analyser\_format | PIG, POC | Format for DATE\_analyser. | A format string code. See strptime in R for codes. |
| TIME\_s | PROF | The name of the variable for time at the start of teh profiling cast. | Text. If recorded in header use header-[variable]. |
| TIME\_b | PROF | The name of the variable for time at the bottom of the profiling cast. | Text. If recorded in header use header-[variable]. |
| TIME\_e | PROF | The name of the variable for time at the end of the profiling cast. | Text. If recorded in header use header-[variable]. |
| TIME\_analyser | PIG, POC | The name of the variable for time of observation recorded by the analyser. If recorded in header use header-[variable]. | Text |
| TIME\_format | PROF | Format for TIME. | A format string code. See strptime in R for codes. |
| TIME\_b\_format | PROF | Format for TIME\_b, if different to TIME\_format. | A format string code. See strptime in R for codes. |
| TIME\_analyser\_format | PIG, POC | Format of TIME\_analyser. | A format string code. See strptime in R for codes. |
| LATITUDE\_s | PROF | The name of the variable for latitude at the start of the profiling cast. | Text |
| LATITUDE\_b | PROF | The name of the variable for latitude at the bottom of the profiling cast. | Text |
| LATITUDE\_e | PROF | The name of the variable for latitude at the end of the profiling cast. | Text |
| LONGITUDE\_s | PROF | The name of the variable for longitude at the start of the profiling cast. | Text |
| LONGITUDE\_b | PROF | The name of the variable for longitude at the bottom of the profiling cast. | Text |
| LONGITUDE\_e | PROF | The name of the variable for longitude at the end of the profiling cast. | Text |
| LAT\_analyser | PIG, POC | The name of the variable for latitude recorded by the analyser. | Text |
| LON\_analyser | PIG, POC | The name of the variable for longitude recorded by the analyser. | Text |
| POSITION\_format | all | The format of latitude and longitude data. | A string describing the format made up of %deg (degrees), %min (minutes), %sec (seconds) and %pos (for N/S/E/W specification) |
| Sample\_ID | PIG, POC | The name of the variable containing sample identification. | Text |
| BOTTLE | PIG, POC | The name of the variable containing bottle identifications. | Text |
| Underway\_ID | PIG, POC | How underway samples are identified within the dataset. Leave blank if there are no underway values within the dataset. | [variable name]-[value] or all |
| **Profiling sensor data information** | | | |
| CTDPRS | PROF | The name of the variable for pressure collected by the profiling sensor. | Text |
| CTDPRS\_u | PROF | The units for pressure collected by the profiling sensor. | Text |
| CTDTMP | PROF | The name of the variable for temperature collected by the profiling sensor. | Text |
| CTDTMP\_u | PROF | The units for temperature collected by the profiling sensor. | Text |
| CTDSAL | PROF | The name of the variable for salinity collected by the profiling sensor. | Text |
| CTDSAL\_u | PROF | The units for salinity collected by the profiling sensor. | Text |
| CTDOXY | PROF | The name of the variable for oxygen collected by the profiling sensor. | Text |
| CTDOXY\_u | PROF | The units for oxygen collected by the profiling sensor. | Text |
| CTDFLUOR | PROF | The name of the variable for fluorescence collected by the profiling sensor. | Text |
| CTDFLUOR\_u | PROF | The units for fluorescence collected by the profiling sensor. | Text |
| CTDBEAMCP | PROF | The name of the variable for beam attenuation collected by the profiling sensor. | Text |
| CTDBEAMCP\_u | PROF | The units for beam attenuation collected by the profiling sensor. | Text |
| CTDBBP700 | PROF | The name of the variable for optical backscatter (700 nm) collected by the profiling sensor. | Text |
| CTDBBP700\_u | PROF | The units for optical backscatter (700 nm) collected by the profiling sensor. | Text |
| CTDXMISS | PROF | The name of the variable for transmittance collected by the profiling sensor. | Text |
| CTDXMISS\_u | PROF | The units for transmittance collected by the profiling sensor. | Text |
| CTDPAR | PROF | The name of the variable for photosyntheitically active radiation collected by the profiling sensor. | Text |
| CTDPAR\_u | PROF | The units for photosyntheitically active radiation collected by the profiling sensor. | Text |
| CTDNITRATE | PROF | The name of the variable for oxygen collected by the profiling sensor. | Text |
| CTDNITRATE\_u | PROF | The units for oxygen collected by the profiling sensor. | Text |
| **Pigment and POC data information** | | | |
| DEPTH | PIG,POC | The name of the variable for depth of observation recorded by the analyser. | Text |
| PIG\_u | PIG | The units for pigment measurements recorded by the analyser. | Text |
| FCHLORA | PIG | The name of the variable for fluorometrically derived chlorophyll. | Text |
| FPHEO | PIG | The name of the variable for fluorometrically derived phaeopigments. | Text |
| FPHYTIN | PIG | The name of the variable for fluorometrically derived phaeophytin. | Text |
| TCHLA | PIG | The name of the variable for HPLC derived total chlorophyll a. | Text |
| TACC | PIG | The name of the variable for HPLC derived total accessory pigments. | Text |
| DVChla | PIG | The name of the variable for HPLC derived divinyl chlorophyll a. | Text |
| Chla | PIG | The name of the variable for HPLC derived chlorophyll a. | Text |
| Chla\_ide | PIG | The name of the variable for HPLC derived chlorophyllide. | Text |
| Chla\_allom | PIG | The name of the variable for HPLC derived chlorophyll a allomers. | Text |
| Chla\_prime | PIG | The name of the variable for HPLC derived chlorophyll a prime. | Text |
| Chlb | PIG | The name of the variable for HPLC derived chlorophyll b. | Text |
| DVChlb | PIG | The name of the variable for HPLC derived divinyl chlorophyll b. | Text |
| Chlc | PIG | The name of the variable for HPLC derived chlorophyll c. | Text |
| Chlc1\_Chlc2\_Mg\_3\_8\_divinyl\_pheoporphyrin\_a5 | PIG | The name of the variable for HPLC derived chlorophyll c1 + chlorophyll c2 + Mg 3,8 divinyl pheoporphyrin a5. | Text |
| Chlc1 | PIG | The name of the variable for HPLC derived chlorophyll c1. | Text |
| Chlc1\_like | PIG | The name of the variable for HPLC derived chlorophyll c1-like. | Text |
| Chlc2 | PIG | The name of the variable for HPLC derived chlorophyll c2. | Text |
| Chlc1\_Chlc2 | PIG | The name of the variable for HPLC derived chlorophyll c1 + chlorophyll c2. | Text |
| Chlc3 | PIG | The name of the variable for HPLC derived chlorophyll c3. | Text |
| MgDVP | PIG | The name of the variable for HPLC derived Mg 2,4 divinyl pheoporphyrin a5 monomethyl ester. | Text |
| 19Hex | PIG | The name of the variable for HPLC derived 19’hexanoyloxyfucoxanthin. | Text |
| 19But | PIG | The name of the variable for HPLC derived 19’butanoyloxyfucoxanthin. | Text |
| Fucox | PIG | The name of the variable for HPLC derived fucoxanthin. | Text |
| Perid | PIG | The name of the variable for HPLC derived peridinin. | Text |
| Prasino | PIG | The name of the variable for HPLC derived prasinoxanthin. | Text |
| Allox | PIG | The name of the variable for HPLC derived alloxanthin. | Text |
| Lutein | PIG | The name of the variable for HPLC derived lutein. | Text |
| Zeax | PIG | The name of the variable for HPLC derived zeaxanthin. | Text |
| Zea\_Lut | PIG | The name of the variable for HPLC derived zeaxanthin + lutein. | Text |
| Violax | PIG | The name of the variable for HPLC derived violaxanthin. | Text |
| Alpha\_car | PIG | The name of the variable for HPLC derived alpha carotene. | Text |
| Beta\_car | PIG | The name of the variable for HPLC derived beta carotene. | Text |
| Gamma\_car | PIG | The name of the variable for HPLC derived gamma carotene. | Text |
| Epsilon\_car | PIG | The name of the variable for HPLC derived epsilon carotene. | Text |
| Alpha\_Beta\_car | PIG | The name of the variable for HPLC derived alpha + beta carotene. | Text |
| Neox | PIG | The name of the variable for HPLC derived neoxanthin. | Text |
| DD | PIG | The name of the variable for HPLC derived diadinoxanthin. | Text |
| DT | PIG | The name of the variable for HPLC derived diatoxanthin. | Text |
| Viol\_Neox | PIG | The name of the variable for HPLC derived violaxanthin + neoxanthin. | Text |
| Phaeopigments | PIG | The name of the variable for HPLC derived bulk phaeopigments. | Text |
| Phide\_a | PIG | The name of the variable for HPLC derived phaeophorbide a. | Text |
| Phytin\_a | PIG | The name of the variable for HPLC derived phaeophytin a. | Text |

**Table 2:** a figure

| group | All pigment records | Subsurface profile records (>4 depth samples above 75 m) | Surface records (<10 m) |
| --- | --- | --- | --- |
| Number of voyages | 174 | 115 | 146 |
| Number of Records | 65,834 | 46,354 | 18,025 |
| Unique samples (no replicates) | 62,020 | 43,822 | 15,916 |
| Unique samples in lat and lon (not depth) | 14,981 | 4,726 | 10,432 |
| Bio-physical matches | 50,048 | 38,458 | 12,003 |
| HPLC and fluorometry matches | 3,008 | 1,544 | 851 |
| Number of HPLC records | 26,984 | 14,934 | 7,490 |
| Number of Fluorometry records | 38,717 | 31,329 | 10,520 |
| Matches with CTDFLUOR records | 26,284 | 20,570 | 5,896 |
| Matches with CTDBBP700 records | 2,829 | 2,460 | 355 |
| Matches with CTDBEAMCP records | 8,116 | 6,718 | 1,028 |

| V1 | group |
| --- | --- |
| 127.00000 | Number of voyages |
| 11,818.00000 | Number of profiles |
| 100.00000 | Profiles with pressure records (%) |
| 99.87307 | Profiles with salinity records |
| 99.89846 | Profiles with temperature records |
| 42.43527 | Profiles with oxygen records |

**Table 3:** a figure

| Variable | Description |
| --- | --- |
| **Header information** | |
| ORIGINAL\_CTDFILE/S | The name of the original file/s |
| CTDFILE\_MOD\_DATE | The modification date of the file |
| SOURCED\_FROM | The repository where data was originally sourced from |
| DATASET\_CONTACT | Name and email of the listed dataset contact |
| DOI/s | Doi/s of original files |
| BIOMATE\_CITE\_TAGS | The BIOMATE citation tags that are associated with the data, methods and source repository |
| DATA\_CITATION/S | The full citations associted with the data |
| **Header variables** | |
| NUMBER\_HEADERS | The number of header variables |
| EXPOCODE | The EXPOCODE associated with the data |
| SHIP | The vessel on which the data was collected |
| STNNBR or EVENTNBR | The station number of the profiling station |
| CASTNO | The cast number of the profiling station |
| CTD\_IDs | An identifcation for the profiling station |
| DATE | The date of the profiling station |
| TIMEZONE | The timezone the data was collected in |
| CTD\_START\_TIME | The time at the start of the profiling station |
| CTD\_START\_LATITUDE | The latitude at the start of the profiling station |
| CTD\_START\_LONGITUDE | The longitude at the start of the profiling station |
| CTD\_BOTTOM\_TIME | The time at the bottom of the profiling station |
| CTD\_BOTTOM\_LATITUDE | The latitude at the bottom of the profiling station |
| CTD\_BOTTOM\_LONGITUDE | The longitude at the bottom of the profiling station |
| CTD\_END\_TIME | The time at the end of the profiling station |
| CTD\_END\_LATITUDE | The latitude at the end of the profiling station |
| CTD\_END\_LONGITUDE | The longitude at the end of the profiling station |
| **Ocean data variables** | |
| missing\_value | The value that corresponds to missing data within the data table |
| CTDPRS | Pressure |
| CTDTMP | Temperature |
| CTDSAL | Salinity |
| CTDDOXY | Dissolved oxygen |
| CTDFLUOR | Fluorescence |
| CTDBEAMCP | Beam attenuation |
| CTDBBP700 | Optical backscatter at 700 nm |
| CTDXMISS | Transmissometer |
| CTDPAR | Photosynthetically active radiation |
| CTDNITRATE | Nitrate |

**Table 4:** a figure

| Variable | Description |
| --- | --- |
| **Header information** | |
| ORIGINAL\_CHLFILE/S | The name of the original file/s |
| CHLFILE\_MOD\_DATE | The modification date of the file |
| SOURCED\_FROM | The repository where data was originally sourced from |
| ANALYSIS\_METHOD | The analysis method used to obtain data |
| DATASET\_CONTACT | Name and email of the listed dataset contact |
| DOI/s | Doi/s of original files |
| BIOMATE\_CITE\_TAGS | The BIOMATE citation tags that are associated with the data, methods and source repository |
| DATA\_CITATION/S | The full citations associted with the data |
| METHOD\_CITATION/S | The full citation associated with the method used to analyse the water sample for pigments |
| **Header variables** | |
| NUMBER\_HEADERS | The number of header variables |
| EXPOCODE | The EXPOCODE associated with the data |
| SHIP | The vessel on which the data was collected |
| TIMEZONE | The timezone the data was collected in |
| missing\_value | The value that corresponds to missing data within the data table |
| not\_detected | The value that corresponds to data not detected in analysis within the data table |
| **Ocean data variables** | |
| CTD\_IDs | An identifcation for a matching profiling station in the profiling sensor stream |
| DATE | The date of the profiling station |
| TIME\_s | The start time of the profiling station |
| TIME\_b | The bottom time of the profiling start date |
| TIME\_e | The end time of the profiling station |
| LATITUDE | The start latitude of the profiling station |
| LONGITUDE | The start longitude of the profiling station |
| STNNBR | The station number of the profiling station |
| CASTNO | The cast number of the profiling station |
| DATE\_analyser | The date of sampling as recorded by the analyser |
| TIME\_analyser | The time of samping as recorded by the analyser |
| LAT\_analyser | The latitude at sampling as recorded by the analyser |
| LON\_analyser | The longitude at sampling as recorded by the analyser |
| STNNBR\_analyser | The station number of the profiling station as recorded by the analyser |
| CASTNO\_analyser | The cast number of the profiling station as recorded by the analyser |
| Sample\_ID | The sample identification as recorded by the analyser |
| BOTTLE | The rosette bottle number as recorded by the analyser |
| DEPTH | The depth the sample was taken |
| FCHLORA | Fluorometrically derived chlorophyll |
| FPHEO | fluorometrically derived phaeopigments |
| FPHYTIN | fluorometrically derived phaeophytin |
| TCHLA | HPLC derived total chlorophyll a |
| TACC | HPLC derived total accessory pigments |
| DVChla | HPLC derived divinyl chlorophyll a |
| Chla | HPLC derived chlorophyll a |
| Chla\_ide | HPLC derived chlorophyllide |
| Chla\_allom | HPLC derived chlorophyll a allomers |
| Chla\_prime | PLC derived chlorophyll a prime |
| Chlb | HPLC derived chlorophyll b |
| DVChlb | HPLC derived divinyl chlorophyll b |
| Chlc | HPLC derived chlorophyll c |
| Chlc1\_Chlc2\_Mg\_3\_8\_divinyl\_pheoporphyrin\_a5 | HPLC derived chlorophyll c1 + chlorophyll c2 + Mg 3,8 divinyl pheoporphyrin a5 |
| Chlc1 | HPLC derived chlorophyll c1 |
| Chlc1\_like | HPLC derived chlorophyll c1-like |
| Chlc2 | HPLC derived chlorophyll c2 |
| Chlc1\_Chlc2 | HPLC derived chlorophyll c1 + chlorophyll c2 |
| Chlc3 | HPLC derived chlorophyll c3 |
| MgDVP | HPLC derived Mg 2,4 divinyl pheoporphyrin a5 monomethyl ester |
| 19Hex | HPLC derived 19’hexanoyloxyfucoxanthin |
| 19But | HPLC derived 19’butanoyloxyfucoxanthin |
| Fucox | HPLC derived fucoxanthin |
| Prasino | HPLC derived alloxanthin |
| Lutein | HPLC derived lutein |
| Zeax | HPLC derived zeaxanthin |
| Zea\_Lut | HPLC derived zeaxanthin + lutein |
| Violax | HPLC derived violaxanthin |
| Alpha\_car | HPLC derived alpha carotene |
| Beta\_car | HPLC derived beta carotene |
| Gamma\_car | HPLC derived gamma carotene |
| Epsilon\_car | HPLC derived epsilon carotene |
| Alpha\_Beta\_car | HPLC derived alpha + beta carotene |
| Neox | HPLC derived neoxanthin |
| DD | HPLC derived diadinoxanthin |
| DT | HPLC derived diatoxanthin |
| Viol\_Neox | HPLC derived violaxanthin + neoxanthin |
| Phaeopigments | HPLC derived bulk phaeopigments |
| Phide\_a | HPLC derived phaeophorbide a |
| Phytin\_a | HPLC derived phaeophytin a |

# References

1. Raymond W. Schmitt | Emeritus, W. H., Woods Hole Oceanographic Institution. [The ocean’s role in climate](https://doi.org/10.5670/oceanog.2018.225). *Oceanography* **issue\_volume**, (2018).

2. Ainley, D. G., Fraser, W. R., Smith, W. O., Hopkins, T. L. & Torres, J. J. [The structure of upper level pelagic food webs in the antarctic: Effect of phytoplankton distribution](https://doi.org/10.1016/0924-7963(91)90017-O). *Journal of Marine Systems* **2**, 111–122 (1991).

3. Basu, S. & Mackey, K. R. M. [Phytoplankton as key mediators of the biological carbon pump: Their responses to a changing climate](https://doi.org/10.3390/su10030869). *Sustainability* **10**, (2018).

4. Carranza, M. M. *et al.* [When mixed layers are not mixed. Storm-driven mixing and bio-optical vertical gradients in mixed layers of the southern ocean](https://doi.org/10.1029/2018JC014416). *Journal of Geophysical Research: Oceans* **123**, 7264–7289 (2018).

5. Prairie, J. C., Sutherland, K. R., Nickols, K. J. & Kaltenberg, A. M. [Biophysical interactions in the plankton: A cross-scale review](https://doi.org/10.1215/21573689-1964713). *Limnology and Oceanography: Fluids and Environments* **2**, 121–145 (2012).

6. Wihsgott, J. U. *et al.* [Observations of vertical mixing in autumn and its effect on the autumn phytoplankton bloom](https://doi.org/10.1016/j.pocean.2019.01.001). *Progress in Oceanography* **177**, 102059 (2019).

7. Brody, S. R. & Lozier, M. S. [Characterizing upper-ocean mixing and its effect on the spring phytoplankton bloom with in situ data](https://doi.org/10.1093/icesjms/fsv006). *ICES Journal of Marine Science* **72**, 1961–1970 (2015).

8. Mignot, A., D’Ortenzio, F., Taillandier, V., Cossarini, G. & Salon, S. [Quantifying observational errors in biogeochemical-argo oxygen, nitrate, and chlorophyll a concentrations](https://doi.org/10.1029/2018GL080541). *Geophysical Research Letters* **46**, 4330–4337 (2019).

9. Valente, A. *et al.* [A compilation of global bio-optical in situ data for ocean-colour satellite applications – version two](https://doi.org/10.5194/essd-11-1037-2019). *Earth System Science Data* **11**, 1037–1068 (2019).

10. Johnson, R., Strutton, P. G., Wright, S. W., McMinn, A. & Meiners, K. M. [Three improved satellite chlorophyll algorithms for the southern ocean](https://doi.org/10.1002/jgrc.20270). *Journal of Geophysical Research: Oceans* **118**, 3694–3703 (2013).

11. Sauzède, R. *et al.* [Vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: A first database for the global ocean](https://doi.org/10.5194/essd-7-261-2015). *Earth System Science Data* **7**, 261–273 (2015).

12. Roesler, C. *et al.* [Recommendations for obtaining unbiased chlorophyll estimates from in situ chlorophyll fluorometers: A global analysis of WET labs ECO sensors](https://doi.org/10.1002/lom3.10185). *Limnology and Oceanography: Methods* **15**, 572–585 (2017).

13. Verdy, A. & Mazloff, M. R. [A data assimilating model for estimating southern ocean biogeochemistry](https://doi.org/10.1002/2016JC012650). *Journal of Geophysical Research: Oceans* **122**, 6968–6988 (2017).

14. Haëntjens, N., Boss, E. & Talley, L. D. [Revisiting ocean color algorithms for chlorophyll a and particulate organic carbon in the southern ocean using biogeochemical floats](https://doi.org/10.1002/2017JC012844). *Journal of Geophysical Research: Oceans* **122**, 6583–6593 (2017).