A compilation of global bio-optical in situ data for ocean-colour satellite applications – version two
Valente, André; Sathyendranath, Shubha; Brotas, Vanda; Groom, Steve; Grant, Michael; Taberner, Malcolm; Antoine, David; Arnone, Robert; Balch, William M.; Barker, Kathryn; Barlow, Ray; Bélanger, Simon; Berthon, Jean-françois; Beiktepe, ükrü; Borsheim, Yngve; Bracher, Astrid; Brando, Vittorio; Canuti, Elisabetta; Chavez, Francisco; Cianca, Andrés

Published in:
Earth System Science Data
Publication date:
2019
Publisher rights:
Full Open Access paper
The re-use license for this item is:
CC BY
The Document Version you have downloaded here is:
Publisher's PDF, also known as Version of record

The final published version is available direct from the publisher website at:
10.5194/essd-11-1037-2019

Link to author version on UHI Research Database

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the UHI Research Database are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights:
1) Users may download and print one copy of any publication from the UHI Research Database for the purpose of private study or research.
2) You may not further distribute the material or use it for any profit-making activity or commercial gain
3) You may freely distribute the URL identifying the publication in the UHI Research Database

Take down policy
If you believe that this document breaches copyright please contact us at RO@uhi.ac.uk providing details; we will remove access to the work immediately and investigate your claim.

Download date: 16. Mar. 2020
A compilation of global bio-optical in situ data for ocean-colour satellite applications – version two

André Valente\textsuperscript{1}, Shubha Sathyendranath\textsuperscript{2}, Vanda Brotas\textsuperscript{1,2}, Steve Groom\textsuperscript{2}, Michael Grant\textsuperscript{2,3}, Malcolm Taberner\textsuperscript{3}, David Antoine\textsuperscript{4,5}, Robert Arnone\textsuperscript{6}, William M. Balch\textsuperscript{7}, Kathryn Barker\textsuperscript{8,9,10}, Ray Barlow\textsuperscript{11}, Simon Bélanger\textsuperscript{12}, Jean-François Berthon\textsuperscript{13}, Şükri Beşiktepe\textsuperscript{14}, Yngve Borsheim\textsuperscript{15}, Astrid Bracher\textsuperscript{16,17}, Vittorio Brando\textsuperscript{18}, Elisabetta Canuti\textsuperscript{13}, Francisco Chavez\textsuperscript{19}, Andrés Cianca\textsuperscript{20}, Hervé Claustre\textsuperscript{4}, Lesley Clementson\textsuperscript{9}, Richard Crout\textsuperscript{21}, Robert Frouin\textsuperscript{22}, Carlos García-Soto\textsuperscript{23,24}, Stuart W. Gibb\textsuperscript{25}, Richard Gould\textsuperscript{21}, Stanford B. Hooker\textsuperscript{26}, Mati Kahrui\textsuperscript{22}, Milton Kampel\textsuperscript{27}, Holger Klein\textsuperscript{28}, Susanne Kratzer\textsuperscript{29}, Raphael Kudela\textsuperscript{30}, Jesús Ledesma\textsuperscript{31}, Hubert Loisel\textsuperscript{32}, Patricia Matrai\textsuperscript{7}, David McKee\textsuperscript{33}, Brian G. Mitchell\textsuperscript{22}, Tiffany Moisan\textsuperscript{34,1}, Frank Muller-Karger\textsuperscript{35}, Leonie O’Dowd\textsuperscript{36}, Michael Ondrusek\textsuperscript{37}, Trevor Platt\textsuperscript{2}, Alex J. Poulton\textsuperscript{38}, Michel Repecaud\textsuperscript{39}, Thomas Schroeder\textsuperscript{9}, Timothy Smyth\textsuperscript{2}, Denise Smythe-Wright\textsuperscript{40}, Heidi M. Sosik\textsuperscript{41}, Michael Twardowski\textsuperscript{42}, Vincenzo Vellucci\textsuperscript{2}, Kenneth Voss\textsuperscript{43}, Jeremy Werdell\textsuperscript{26}, Marcel Wernand\textsuperscript{44,7}, Simon Wright\textsuperscript{45}, and Giuseppe Zibordi\textsuperscript{13}

\textsuperscript{1}MARE – Marine and Environmental Sciences Centre, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisbon, Portugal
\textsuperscript{2}Plymouth Marine Laboratory, Plymouth, PL1 3DH, UK
\textsuperscript{3}EUMETSAT, Eumetsat-Allee 1, 64295 Darmstadt, Germany
\textsuperscript{4}Sorbonne Université, CNRS, Laboratoire d’Océanographie de Villefranche, LOV, 06230 Villefranche-sur-Mer, France
\textsuperscript{5}Remote Sensing and Satellite Research Group, School of Earth and Planetary Sciences, Curtin University, Perth, WA 6845, Australia
\textsuperscript{6}University of Southern Mississippi, Stennis Space Center, MS, USA
\textsuperscript{7}Bigelow Laboratory for Ocean Sciences, 60 Bigelow Dr., East Boothbay, ME 04544, USA
\textsuperscript{8}ARGANS Ltd, Plymouth, UK
\textsuperscript{9}CSIRO Oceans and Atmosphere, Perth, Western Australia, Australia
\textsuperscript{10}Australian Research Data Commons, Caulfield East, Australia
\textsuperscript{11}Bayworld Centre for Research and Education, Cape Town, South Africa
\textsuperscript{12}Université du Québec à Rimouski, Rimouski, Quebec, Canada
\textsuperscript{13}European Commission, Joint Research Centre, Ispra, Italy
\textsuperscript{14}Dokuz Eylül University, Institute of Marine Science and Technology, Izmir, Turkey
\textsuperscript{15}Institute of Marine Research, Bergen, Norway
\textsuperscript{16}Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany
\textsuperscript{17}Institute of Environmental Physics, University Bremen, Bremen, Germany
\textsuperscript{18}CNR – ISMAR, Rome, Italy
\textsuperscript{19}Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA
\textsuperscript{20}PLOCAN – Oceanic Platform of the Canary Islands, Carretera de Taliarte, 35214 Telde, Gran Canaria, Spain
\textsuperscript{21}Naval Research Laboratory, Stennis Space Center, MS, USA
\textsuperscript{22}Scripps Institution of Oceanography, University of California San Diego, CA, USA
\textsuperscript{23}Spanish Institute of Oceanography (IEO), Corazón de María 8, 28002 Madrid, Spain
\textsuperscript{24}Plentziako Itsas Estazioa/Euskal Herriko Unibertsitatea (PIE/EHU), Areatza z/g, 48620 Plentzia, Spain
\textsuperscript{25}Environmental Research Institute, North Highland College, University of the Highlands and Islands, Thurso, Scotland, UK
\textsuperscript{26}NASA Goddard Space Flight Center, Greenbelt, MD, USA

Published by Copernicus Publications.
Abstract. A global compilation of in situ data is useful to evaluate the quality of ocean-colour satellite data records. Here we describe the data compiled for the validation of the ocean-colour products from the ESA Ocean Colour Climate Change Initiative (OC-CCI). The data were acquired from several sources (including, inter alia, MOBY, BOUSSOLE, AERONET-OC, SeaBASS, NOMAD, MERMAID, AMT, ICES, HOT and GeP&CO) and span the period from 1997 to 2018. Observations of the following variables were compiled: spectral remote-sensing reflectances, concentrations of chlorophyll $a$, spectral inherent optical properties, spectral diffuse attenuation coefficients and total suspended matter. The data were from multi-project archives acquired via open internet services or from individual projects, acquired directly from data providers. Methodologies were implemented for homogenization, quality control and merging of all data. No changes were made to the original data, other than averaging of observations that were close in time and space, elimination of some points after quality control and conversion to a standard format. The final result is a merged table designed for validation of satellite-derived ocean-colour products and available in text format. Metadata of each in situ measurement (original source, cruise or experiment, principal investigator) was propagated throughout the work and made available in the final table. By making the metadata available, provenance is better documented, and it is also possible to analyse each set of data separately. This paper also describes the changes that were made to the compilation in relation to the previous version (Valente et al., 2016). The compiled data are available at https://doi.org/10.1594/PANGAEA.898188 (Valente et al., 2019).
1 Introduction

Currently, there are several sets of in situ bio-optical data, worldwide, suitable for validation of ocean-colour satellite data. Whereas some are managed by the data producers, others are in international repositories with contributions from multiple scientists. Many have rigid quality controls and are built specifically for ocean-colour validation. The use of only any one of these datasets would limit the number of data in validation exercises. It is, therefore, vital to acquire and merge all these datasets into a single unified dataset to maximize the number of matchups available for validation, their distribution in time and space, and, consequently, to reduce uncertainties in the validation exercise. However, merging several datasets together can be a complicated task. First it is necessary to acquire and harmonize all datasets into a single standard format. Second, during the merging, duplicates between datasets have to be identified and removed. Third, the metadata should be propagated throughout the process and made available in the final merged product. Ideally, the compiled dataset would be made available as a simple text table, to facilitate ease of access and manipulation. In this work such unification of multiple datasets is presented. This was done for the validation of the ocean-colour products from the ESA Ocean Colour Climate Change Initiative (OC-CCI), but with the intent to serve the broader user community as well.

A merged dataset is not without drawbacks: it is likely to be large and so not always easy to manipulate; because the merging is done on pre-existing, processed databases, it is not possible to have full control of the whole processing chain; the dataset would be a compilation of observations collected by several investigators using different instruments, sampling methods and protocols, which might eventually have been modified by the processing routines used by the repositories or archives. To minimize these potential drawbacks, we have, for the most part, incorporated only datasets that have emerged from the long-term efforts of the ocean-colour and biological oceanographical communities to provide scientists with high-quality in situ data, and we implemented additional quality checks on the data to enhance confidence in the quality of the merged product. Nevertheless, it is still recognized that different and unpredictable uncertainties may affect data from the diverse sources as a result of the application of a variety of field/laboratory instruments, methods and data reduction schemes.

In Sect. 2 the methodologies used to harmonize and integrate all data, as well as a description of individual datasets acquired, are provided. In Sect. 3 the geographic distribution and other characteristics of the final merged dataset are shown. Section 4 provides an overview of the data.

2 Data and methods

2.1 Preprocessing and merging

The compiled global set of bio-optical in situ data described in this work has an emphasis, though not exclusive, on open-ocean data. It comprises the following variables: remote-sensing reflectance (rrs), chlorophyll a concentration (chl), algal pigment absorption coefficient (aph), detrital and coloured dissolved organic matter absorption coefficient (adg), particle backscattering coefficient (bbp), diffuse attenuation coefficient for downward irradiance (kd) and total suspended matter (tsm). The variables rrs, aph, adg, bbp and kd are spectrally dependent, and this dependence is, hereafter, implied. The data were compiled from 27 sources (MOBY, BOUSSOLE, AERONET-OC, SeaBASS, NOMAD, MERMAID, AMT, ICES, HOT, GeP&CO, AWI, ARCSSPP, BARENTSSEA, BATS, BIOCHEM, BODC, CALCOFI, CCELTER, CIMT, COASTCOLOUR, ESTOC, IMOS, MAREDAT, PALMER, SEADATANET, TPSS and TARA); each one described in Sect. 2.2. The data sources in this work should also be viewed as groups of data that were acquired from a specific source, standardized with a specific method and later merged into the compilation. The compiled in situ observations have a global distribution and cover the period 1997 to 2018. The listed variables, with the exception of total suspended matter, were chosen as they are the operational satellite ocean-colour products of the ESA OC-CCI project, which currently focuses on the merging of four ocean-colour satellite sensors: the Medium Resolution Imaging Spectrometer (MERIS) of ESA, the Moderate Resolution Imaging Spectroradiometer (MODIS) of NASA, the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) of NASA, and the Visible Infrared Imaging Radiometer Suite (VIIRS) of NASA and the National Oceanic and Atmospheric Administration (NOAA) to create a time series of satellite data.

This is the second version of the compilation of global bio-optical in situ data described by Valente et al. (2016). A track-change file of the manuscript of the first version can be found in the Supplement. The new version has more data and a higher temporal and spatial coverage. The increases in the number of observations are mainly for chla, rrs and aph. In comparison with Valente et al. (2016), the observations of chla and aph have doubled in number and provide a better spatial coverage, especially in the Southern and Arctic Ocean. The rrs values also increased in number, but not as much in spatial coverage, because most of the new observations came from fixed locations.

The present second version is a compilation of data from sources used in the first version (MOBY, BOUSSOLE, AERONET-OC, SeaBASS, NOMAD, MERMAID, AMT, ICES, HOT and GeP&CO) plus data from additional sources (AWI, ARCSSPP, BARENTSSEA, BATS, BIOCHEM, BODC, CALCOFI, CCELTER, CIMT, COASTCOLOUR, ESTOC, IMOS, MAREDAT, PALMER, SEA-
DATANET, TPSS and TARA). The main differences from the first version were (1) some of the data sources used in the first version were updated (MOBY, AERONET, SeaBASS and HOT), (2) new data sources were added, (3) a new variable was compiled (total suspended matter), (4) the format of the database was modified and (5) two new flags were added.

Concerning the change in format, in Valente et al. (2016) the compilation was provided as one unique two-dimensional table. Now, given its increased size (136,250 rows and 1286 columns compared with 80,524 rows and 267 columns previously), the table has been broken into three smaller tables that relate to each other via one unique key identifying each row. One additional table is also provided to help with data manipulation. Despite this change, the compilation should still be viewed conceptually as one unique table, and as such, it is still described in that way. In the present version, two flags were added: flag_time and flag_chl_method. The first is because in the present version three data sources were used (ESTOC, MAREDAT and TPSS) where information on time (hour of the day) was not available. The time for these observations was set to 12:00:00 UTC and the observations were flagged with “1” in the column flag_time. A second flag was necessary, because in two data sources (ARCSSPP and SEADATANET) there was uncertainty on whether the compiled chlorophyll concentrations were measured using fluorometric, spectrophotometric or high-performance liquid chromatography (HPLC) methods. The compiled chlorophyll observations from these two data sources were flagged with “1” in the column flag_chl_method and were marked as chla_fluor.

Remote-sensing reflectance (rrs) is a primary ocean-colour product defined as \( \text{rrs} = \frac{L_w}{E_s} \), where \( L_w \) is the upward water-leaving radiance and \( E_s \) is the total downward irradiance at sea level. Another quantity that is often required is the normalized water-leaving radiance (\( nL_w \)) (Gordon and Clark, 1981), which is related to remote-sensing reflectance via \( \text{rrs} = nL_w/F_o \), where \( F_o \) is the top-of-the-atmosphere solar irradiance. If not directly available, remote-sensing reflectance was calculated through the equations described above, depending on the format of the original data. The original data were acquired in an advanced form (e.g. time-averaged, extrapolated to surface) from nine data sources designed for ocean-colour validation and applications (MOBY, BOUSSOLE, AERONET-OC, SeaBASS, NOMAD, MERMAID, COASTCOLOUR, TARA, AWI), therefore only requiring the conversion to a common format. In the processing made by the space agencies, the quantity \( \text{rrs} \) is normalized to a single Sun-viewing geometry (Sun at zenith and nadir viewing) taking in account the bidirectional effects as described in Morel and Gentili (1996) and Morel et al. (2002). Thus, for consistency with satellite rrs product, the latter normalization was applied to the in situ rrs.

Chlorophyll \( a \) concentration is the conventional measure for phytoplankton biomass and one of the most widely used satellite ocean-colour products (IOCCG, 2008). To validate satellite-derived chlorophyll \( a \) concentration, two different variables were compiled: one of these represents chlorophyll \( a \) measurements made through fluorometric or spectrophotometric methods, referred to hereafter as chla_fluor and the other is the chlorophyll concentration derived from HPLC measurements, referred to hereafter as chla_hplc. The chlorophyll data were compiled from the following 25 data sources: BOUSSOLE, SeaBASS, NOMAD, MERMAID, AMT, Ices, HOT, GeP&CO, AWI, ARCSSPP, BAR- ENTSSEE, BATS, BIOCHEM, BODC, CALCOFI, CCEL- TER, CIMT, COASTCOLOUR, ESTOC, IMOS, MARE- DAT, PALMER, SEADATANET, TPSS and TARA. One requirement for chla_fluor measurements was that they were made using in vitro methods (i.e. based on extractions of chlorophyll \( a \)). Although this severely decreased the number of observations, since in situ fluorometry (e.g. fluorometers mounted on CTDs) is widely available in oceanographic databases, it was decided to exclude such data because of potential problems with the calibration of in situ fluorometer data. The variable chla_hplc was calculated by summing all reported chlorophyll \( a \) derivatives, including divinyl chlorophyll \( a \), epimers, allomers and chlorophyllide \( a \). The two chlorophyll variables are retained separately in the database to facilitate their use. HPLC measurements could be considered of higher quality, but fluorometric measurements are more numerous. Thus one option for users is to use chla_fluor only when there are no chla_hplc measurements available. To be consistent with satellite-derived chlorophyll values, which are derived from the light emerging from the upper layer of the ocean, all chlorophyll observations in the top 10 m (replicates at the same depth, or measurements at multiple depths) were averaged if the coefficient of variation among observations was less than 50 %, otherwise they were discarded. The averages were then assigned to the surface. The depth of 10 m was chosen as a compromise between clear oligotrophic and turbid eutrophic waters. Other methods, such as chlorophyll depth averages using local attenuation conditions (Morel and Maritorena, 2001), require observations at multiple depths, which, given our decision to use only in vitro measurements, would have reduced considerably the final number of observations.

With regard to the inherent optical properties (aph, adg, bbp), if not already calculated and provided in the contributed datasets, they were computed from related variables that were available: particle absorption (ap), detrital absorption (ad), coloured dissolved organic matter (CDOM) absorption (ag) and total backscattering (bb). The following equations were used: \( \text{adg} = \text{ad} + \text{ag} \), \( \text{ap} = \text{aph} + \text{ad} \), and \( \text{bb} = \text{bbp} + \text{bbw} \). For the latter equation, the variable \( \text{bbw} \) was computed using \( \text{bbw} = \text{bw}/2 \), where \( \text{bw} \) is the scattering coefficient of seawater derived from Zhang et al. (2009). The diffuse attenuation coefficient for downward irradiance (kd) did not require any conversion and was compiled as originally acquired. Observations of inherent optical properties (surface values) and the diffuse attenuation coefficient...
for downward irradiance were acquired in total from six data sources designed for ocean-colour validation and applications (SeaBASS, NOMAD, MERMAID, AWI, COASTCOLOUR, TPSS), thus already subject to the processing routines of these datasets. Concerning total suspended matter, these data were compiled as originally available from MERMAID and COASTCOLOUR.

The merged dataset was compiled from 27 sets of in situ data, which were obtained individually either from archives that incorporate data from multiple contributors (SeaBASS, NOMAD, MERMAID, ICES, ARCSSPP, BIOCHEM, BODC, COASTCOLOUR, MAREDAT, SEADATANET) or from particular contributors, measurement programmes or projects (MOBY, BOUSSOLE, AERONET-OC, HOT, GeP&CO, AMT, AWI, BARENTSSEA, BATS, CALCOFI, CCELTER, CIMT, ESTOC, IMOS, PALMER, TPSS, TARA) and were subsequently homogenized and merged. Data contributors are listed in Table 2. There were methodological differences between datasets. Therefore, after acquisition, and prior to any merging, each set of data was preprocessed for quality control and converted to a common format. During this process, data were discarded if they had (1) unrealistic or missing date and geographic coordinate fields, (2) poor quality (e.g. original flags) or method of observation that did not meet the criteria for the dataset (e.g. in situ fluorescence for chlorophyll concentration), and (3) spuriously high or low data. For the last item, the following limits were imposed: [0.001–100] mg m\(^{-3}\) for chla_fluor and chla_hplc; [0–0.15] \(\text{sr}^{-1}\) for rr; [0.0001–10] m\(^{-1}\) for aph, adg and bbp; [0–1000] g m\(^{-3}\) for tsm; and [aw(\(\lambda\)) – 10] m\(^{-1}\) for kd, where aw is the pure water absorption coefficients derived from Pope and Fry (1997). Also during this stage, three metadata strings were attributed to each observation: dataset, subdataset and contributor. The dataset contains the name of the original set of data and can only be one of the following: aoc, boussole, mermaid, moby, nomad, seabass, hot, ices, amt, gepco, arcsspp, awi, barentssea, bats, biochem, bode, calcofi, cc, ccelter, cimt, estoc, imos, palmer, sdatanet, tpss or tara. The subdataset starts with the dataset identifier and is followed by additional information about the data, as <dataset>_<cruise/station/site> (e.g. seabass_car81). The contributor contains the name of the data contributor. An effort was made to homogenize the names of data contributors from the different sets of data. These three metadata are the link to trace each observation to its origin and were propagated throughout the processing. Finally, this processing stage ended with each set of data being scanned for replicate variable data and replicate station data, which when found were averaged if the coefficient of variation was less than 50%, otherwise they were discarded. Replicates were defined as multiple observations of the same variable, with the same date, time, latitude and longitude. For the latter case, a search window of 5 min in time and 200 m in distance was given to account for station drift. A small number of observations that were identified as replicates had different subdataset identifiers (i.e. different cruise names). These observations were considered suspicious if the values were different and discarded. If the values were the same, one of the observations was retained. This possibly originated from the same group of data being contributed to an archive by two different data contributors.

Once a set of data was homogenized, its data were integrated into a unique table. This final merging focused on the removal of duplicates between the sets of data. Although some duplicates are known (e.g. MOBY, BOUSSOLE, AERONET-OC and NOMAD data are found in SeaBASS and MERMAID), others are unknown (e.g. how many data of GeP&CO, ICES, AMT and HOT are within NOMAD, SeaBASS and MERMAID). Therefore, duplicates were identified using the metadata (dataset and subdataset) when possible and temporal–spatial matches, as an additional precaution. For temporal–spatial matches, several thresholds were used, but typically 5 min and 200 m were taken to be sufficient to identify most duplicated data, which reflected small differences in time, latitude and longitude, between the different sets of data. Larger thresholds were used in some cases as a precautionary procedure. This was the case when searching for NOMAD data in other datasets, because NOMAD includes a few cases where merging of radiometric and pigment data was done with large spatial–temporal thresholds (Werdell and Bailey, 2005). A large temporal threshold was also used when integrating observations from the three data sources that did not have time available (ESTOC, MAREDAT and TPSS). In regard to all data, if duplicates were found, data from the NOMAD dataset were selected first, followed by data from individual projects or contributors (MOBY, BOUSSOLE, AERONET-OC, AMT, HOT, GeP&CO, AWI, BARENTSSEA, BATS, CALCOFI, CCELTER, CIMT, ESTOC, IMOS, PALMER, TPSS and TARA), and finally for the remaining datasets (SeaBASS, MERMAID, ICES, ARCSSPP, BIOCHEM, BODC, COASTCOLOUR, MAREDAT and SEADATANET). This procedure was chosen to preserve the NOMAD dataset as a whole, since it is widely used in ocean-colour validation. It should be noted that, by this procedure, data from individual projects or contributors may be listed under NOMAD (e.g. some PALMER data are found in NOMAD with metadata string nomad_palmer_iter). After giving priority to NOMAD, the priority was generally given to data from individual projects or contributors, but due to an incremental approach, where only new data are added to previous versions of the compilation, some data from individual projects or contributors (BATS, CALCOFI, CIMT, PALMER and TPSS) added in later stages may be found under other data sources. This occurs mainly for BATS and CALCOFI, which have their earlier chlorophyll data in SeaBASS with metadata strings seabass_bats* and seabass_cal*, and also CIMT, which has some of its data under COASTCOLOUR. After all data from
Table 1. The standard variables, nomenclatures and units in the final table.

<table>
<thead>
<tr>
<th>Variable/column</th>
<th>Description and units</th>
</tr>
</thead>
<tbody>
<tr>
<td>idx</td>
<td>Unique key identifying each row</td>
</tr>
<tr>
<td>time</td>
<td>GMT, &lt;YYYY-MM-DD&gt;T<a href="">HH:MM:SS</a>Z</td>
</tr>
<tr>
<td>lat</td>
<td>Decimal degree, −90 : 90, south negative</td>
</tr>
<tr>
<td>long</td>
<td>Decimal degree, −180 : 180, west negative</td>
</tr>
<tr>
<td>depth_water</td>
<td>Sampling depth (m) – all assigned to zero</td>
</tr>
<tr>
<td>chla_hplc</td>
<td>Total chlorophyll a concentration determined from the HPLC method (mg m(^{-3}))</td>
</tr>
<tr>
<td>chla_fluor</td>
<td>Chlorophyll a concentration determined from fluorometric or spectrophotometric methods (mg m(^{-3}))</td>
</tr>
<tr>
<td>rrs_(&lt;band&gt;)</td>
<td>Remote-sensing reflectance (sr(^{-1}))</td>
</tr>
<tr>
<td>aph_(&lt;band&gt;)</td>
<td>Algal pigment absorption coefficient (m(^{-1}))</td>
</tr>
<tr>
<td>adg_(&lt;band&gt;)</td>
<td>Detrital plus CDOM absorption coefficient (m(^{-1}))</td>
</tr>
<tr>
<td>bbp_(&lt;band&gt;)</td>
<td>Particle backscattering coefficient (m(^{-1}))</td>
</tr>
<tr>
<td>kd_(&lt;band&gt;)</td>
<td>Diffuse attenuation coefficient for downward irradiance (m(^{-1}))</td>
</tr>
<tr>
<td>tsm</td>
<td>Total suspended matter (g m(^{-3}))</td>
</tr>
<tr>
<td>etopo1</td>
<td>Water depth from ETOPO1 (m)</td>
</tr>
<tr>
<td>chla_hplc_dataset</td>
<td>Metadata string for chla_hplc</td>
</tr>
<tr>
<td>chla_hplc_subdataset</td>
<td>Metadata string for chla_hplc</td>
</tr>
<tr>
<td>chla_hplc_contributor</td>
<td>Metadata string for chla_hplc</td>
</tr>
<tr>
<td>chla_fluor_dataset</td>
<td>Metadata string for chla_fluor</td>
</tr>
<tr>
<td>chla_fluor_subdataset</td>
<td>Metadata string for chla_fluor</td>
</tr>
<tr>
<td>chla_fluor_contributor</td>
<td>Metadata string for chla_fluor</td>
</tr>
<tr>
<td>rrs_dataset</td>
<td>Metadata string for rrs</td>
</tr>
<tr>
<td>rrs_subdataset</td>
<td>Metadata string for rrs</td>
</tr>
<tr>
<td>rrs_contributor</td>
<td>Metadata string for rrs</td>
</tr>
<tr>
<td>aph_dataset</td>
<td>Metadata string for aph</td>
</tr>
<tr>
<td>aph_subdataset</td>
<td>Metadata string for aph</td>
</tr>
<tr>
<td>aph_contributor</td>
<td>Metadata string for aph</td>
</tr>
<tr>
<td>adg_dataset</td>
<td>Metadata string for adg</td>
</tr>
<tr>
<td>adg_subdataset</td>
<td>Metadata string for adg</td>
</tr>
<tr>
<td>adg_contributor</td>
<td>Metadata string for adg</td>
</tr>
<tr>
<td>bbp_dataset</td>
<td>Metadata string for bbp</td>
</tr>
<tr>
<td>bbp_subdataset</td>
<td>Metadata string for bbp</td>
</tr>
<tr>
<td>bbp_contributor</td>
<td>Metadata string for bbp</td>
</tr>
<tr>
<td>kd_dataset</td>
<td>Metadata string for kd</td>
</tr>
<tr>
<td>kd_subdataset</td>
<td>Metadata string for kd</td>
</tr>
<tr>
<td>kd_contributor</td>
<td>Metadata string for kd</td>
</tr>
<tr>
<td>tsm_dataset</td>
<td>Metadata string for tsm</td>
</tr>
<tr>
<td>tsm_subdataset</td>
<td>Metadata string for tsm</td>
</tr>
<tr>
<td>tsm_contributor</td>
<td>Metadata string for tsm</td>
</tr>
<tr>
<td>flag_time</td>
<td>“1” if observation without time (set to 12:00:00 UTC)</td>
</tr>
<tr>
<td>flag_chl_method</td>
<td>“1” if observation as unknown chlorophyll method</td>
</tr>
</tbody>
</table>

a given source were free of duplicates, they were merged consecutively by variable in the final table. During this process, we also searched for rows (stations) that were separated from each other by time differences less than 5 min and horizontal spatial differences of less than 200 m. When such rows were found, the observations in those rows were merged into a single row. The compiled merged data were compared with the original sets to certify that no errors occurred during the merging. As a final step, a water column (station) depth was recorded for each observation, which was the closest water column depth from the ETOPO1 global relief model (National Geophysical Data Center ETOPO1; Amante and Eakins, 2009). For observations where the closest water depth was above sea level (e.g. data collected very near the coast), it was given the value of zero.

Data processing thus included two major steps: preprocessing and merging. The first step was related to each set of contributing datasets in particular and aimed to identify problems and convert the data of interest to a standard format. The second step dealt with the integration of data into one unique file and included the elimination of duplicated
Table 2. Original sets of data and data contributors in the final table.

<table>
<thead>
<tr>
<th>Data source</th>
<th>Description</th>
<th>Data contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marine Optical Buoy (MOBY)</td>
<td>Daily observations of remote-sensing reflectance, measured by a fixed mooring system, located west of the Hawaiian island of Lanai. Data compiled between 1997 and 2012. Data were obtained from the MOBY website. Compiled standard variable: rrs.</td>
<td>Paul DiGiacomo, Kenneth Voss</td>
</tr>
<tr>
<td>Bouée pour l’acquisition d’une Série Optique à Long Terme (BOUSSOLE)</td>
<td>High-frequency (15 min) observations of remote-sensing reflectance, from a fixed mooring system, located in the western Mediterranean Sea. Measurements of chlorophyll a concentration are also available at the mooring locations. Remote-sensing reflectance and chlorophyll a data were compiled between 2003–2012 and 2001–2012, respectively. Data were provided by David Antoine. Compiled standard variables: rrs and chla_hplc.</td>
<td>David Antoine, Vicenzo Vellucci</td>
</tr>
<tr>
<td>AErosol RObotic NETwork-Ocean Color (AERONET-OC)</td>
<td>Daily observations of remote-sensing reflectance, measured by modified sun photometers. Data compiled between 2002 and 2012. Sites included Abu Al Bukhoosh (~ 25° N, ~ 53° E), COVE_SEAPRISM (~ 36° N, ~ 75° W), Gloria (~ 44° N, ~ 29° E), Gustav_Dalen_Tower (~ 58° N, ~ 17° E), Helsinki Lighthouse (~ 59° N, ~ 24° E), LISCO (~ 40° N, ~ 73° W), Lucinda (~ 18° S, ~ 146° E), MVCO (~ 41° N, ~ 70° W), Palgrunden (~ 58° N, ~ 13° E), Venice (~ 45° N, ~ 12° E) and WaveCIS_Site_CSI_6 (~ 28° N, ~ 90° W). Data were obtained from the AERONET-OC website. Compiled standard variable: rrs.</td>
<td>Robert Arnone (WaveCIS), Sam Ahmed (LISCO), Vittorio Brando (Lucinda), Dick Crout (WaveCIS), Hui Feng (MVCO), Alex Gilerson (LISCO), Rick Gould (WaveCIS), Brent Holben (COVE_SEAPRISM), Susanne Kratzer (Palgrunden), Thomas Schroeder (Lucinda), Heidi M. Sosik (MVCO), Giuseppe Zibordi (Abu Al Bukhoosh, Gloria, Gustav Dalen Tower, Helsinki Lighthouse and Venice)</td>
</tr>
<tr>
<td>SeaWiFS Bio-optical Archive and Storage System (SeaBASS)</td>
<td>Global archive of in situ marine data from multiple contributors. Bio-optical global data between 1997 and 2012 were extracted from the SeaBASS website. Pigment data were mostly extracted using the Pigment search tool, which provides data directly from the archives. Radiometric data were extracted using the Validation tool, which only provides in situ data with matchups for ocean-colour sensors. Compiled standard variables: rrs, chla_hplc, chl_fluor, aph, adg, bbp and kd.</td>
<td>Robert Arnone, Kevin Arrigo, William Balch, Ray Barlow, Mike Behrenfeld, Sukru Besiktepe, Emmanuel Boss, Chris Brown, Douglas Capone, Ken Carder, Francisco Chavez, Alex Chekaluk, Jay-Chung Chen, Dennis Clark, Herve Claustre, Lesley Clementson, Jorge Corredor, Glenn Cota, Yves Dandonneau, Heidi Dierssen, David Eslinger, Piotr Flatau, Robert Frouin, Carlos Garcia, Alex Gilerson, Joaquim Goes, Gwo-Ching Gong, Adriana Gonzalez-Silvera, Rick Gould, Larry Harding, Jon Hare, Stan B. Hooker, Chuanmin Hu, Milton Kampel, Sung-Ho Kang, Gary Kirkpatrick, Oleg Kopelevich, Sam Laney, Pierre Larouche, Jesus Ledesma, Zhongping Lee, Ricardo Letelier, Marlon Lewis, Steven Lohrenz, Mary Luz Canon, Antonio Mannino, John Marra, Chuck McClain, Christophe Menkes, Mark Miller, Greg Mitchell, Ru Morrison, James Mueller, Frank Muller-Karger, Ruben Negri, James Nelson, Norman Nelson, Mary Jane Perry, David Phinney, John Porter, Collin Roesler, David Siegel, Mike Sieracki, Jeffrey Smart, Raymond Smith, Heidi Sosik, James Spinhirne, Dariusz Stramski, Rick Stumpf, Ajit Subramaniam, Chuck Trees, Michael Twardowski, Kenneth Voss, Marcel Wernand, Ronald Zaneveld, Eric Zettler, Giuseppe Zibordi, Richard Zimmerman</td>
</tr>
</tbody>
</table>
Table 2. Continued.

<table>
<thead>
<tr>
<th>Data source</th>
<th>Description</th>
<th>Data contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERIS Match-up In situ Database (MERMAID)</td>
<td>Global database of in situ bio-optical data matched with concurrent MERIS Level 2 satellite ocean-colour products. The Extract matchup tool to acquire data was used. Data were compiled between 2002 and 2012. Access has been granted through a signed service level agreement. Compiled standard variables: rrs, chla_hplc, chl_fluor, aph, adg, bbp and kd.</td>
<td>Simon Belanger, Jean-Francois Berthon, Vanda Brotas, Elisabetta Canuti, Pierre Yves Deschamps, Annelies Hommersom, Mati Kahru, Holger Klein, Hubert Loisel, David McKee, Greg Mitchell, Michael Ondrusek, Michel Repecaud, David Siegel, Gavin Tilstone, Giuseppe Zibordi</td>
</tr>
<tr>
<td>Atlantic Meridional Transect (AMT)</td>
<td>Multidisciplinary programme that makes biological, chemical and physical oceanographic measurements during an annual voyage between the United Kingdom and destinations in the South Atlantic. It has compiled observations of chlorophyll a concentration between 1997 (AMT5) and 2005 (AMT17). Data were provided by the British Oceanographic Data Centre (BODC). Compiled standard variables: chla_hplc and chl_fluor.</td>
<td>Ray Barlow, Stuart Gibb, Victoria Hill, Patrick Holligan, Gerald Moore, Leonie O'Dowd, Alex Poulton, Emilio Suarez</td>
</tr>
<tr>
<td>International Council for the Exploration of the Sea (ICES)</td>
<td>Database of several collections of data related to the marine environment. It has compiled observations of chlorophyll a concentration in the northern European seas, between 1997 and 2012. Data were provided by the ICES database on the marine environment (2014, Copenhagen, Denmark). Compiled standard variables: chla_hplc and chl_fluor.</td>
<td>Not available</td>
</tr>
<tr>
<td>Hawaii Ocean Time-series (HOT)</td>
<td>Multidisciplinary programme that makes repeated biological, chemical and physical oceanographic observations near Oahu, Hawaii. Measurements of chlorophyll a concentration between 1997 and 2012 were extracted from the project website. Compiled standard variables: chla_hplc and chl_fluor.</td>
<td>Bob Bidigare, Matthew Church, Ricardo Letelier, Jasmine Nahorniak</td>
</tr>
<tr>
<td>Geochemistry, Phyttoplankton, and Color of the Ocean (GeP&amp;CO)</td>
<td>Programme of in situ data collection aboard a merchant ship from France to New Caledonia, between 1999 and 2002. Measurements of chlorophyll a concentration were obtained from the project website. Compiled standard variables: chla_hplc and chl_fluor.</td>
<td>Yves Dandonneau</td>
</tr>
<tr>
<td>ARCSSPP</td>
<td>Arctic System Science Primary Production database. Available from the NODC FTP site. Compiled standard variable: chla_fluor.</td>
<td>Patricia Matrai</td>
</tr>
<tr>
<td>Data source</td>
<td>Description</td>
<td>Data contributors</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>BARENTSSEA</td>
<td>Data collection from cruises of the Institute of Marine Research (Norway) mainly around the Barents Sea. Provided by Knut Yngve Børtsheim. Compiled standard variable: chla_fluor.</td>
<td>Knut Yngve Børtsheim</td>
</tr>
<tr>
<td>BATS</td>
<td>Data collection from the Bermuda Atlantic Time-series Study. Available from BATS website. Compiled standard variables: chla_fluor and chla_hplc.</td>
<td>Not available</td>
</tr>
<tr>
<td>BODC</td>
<td>British Oceanographic Data Centre. Mainly European seas. Provided by BODC. Compiled standard variables: chla_fluor and chla_hplc.</td>
<td>Not available</td>
</tr>
<tr>
<td>COASTCOLOUR</td>
<td>Quality-controlled compilation of bio-optical data in several coastal sites. Available from PANGAEA. Compiled standard variables: chla_fluor, chla_hplc, rrs, aph, adg, bbp and tsm.</td>
<td>Not available</td>
</tr>
</tbody>
</table>
Table 2. Continued.

<table>
<thead>
<tr>
<th>Data source</th>
<th>Description</th>
<th>Data contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPSS</td>
<td>Compilation of bio-optical data predominantly from the northwestern Atlantic but also from the Indian Ocean, South Pacific and central Atlantic. Provided by Trevor Platt and Shubha Sathyendranath. Compiled standard variables: chla_hplc, chla_fluor and aph.</td>
<td>Trevor Platt, Shubha Sathyendranath.</td>
</tr>
<tr>
<td>TARA</td>
<td>Data collection from the TARA global transects. Provided by Emmanuel Boss. All data available in SeaBASS. Compiled standard variables: chla_hplc and rrs.</td>
<td>Emmanuel Boss</td>
</tr>
</tbody>
</table>

data between the individual sets of data. In the next subsections a brief overview of each original set of data is provided.

2.2 Preprocessing of each set of data

2.2.1 Marine Optical Buoy (MOBY)

MOBY is a fixed mooring system operated by the National Oceanic and Atmospheric Administration (NOAA) that provides a continuous time series of water-leaving radiance and surface irradiance in the visible region of the spectra since 1997. The site is located a few kilometres west of the Hawaiian island of Lanai where the water depth is about 1200 m. Since its deployment, MOBY measurements have been the primary basis for the on-orbit vicarious calibrations of the SeaWiFS and MODIS ocean-colour sensors. A full description of the MOBY system and processing is provided in Clark et al. (2003). Data are freely available for scientific use at the MOBY Gold directory. The products of interest are the Scientific Time Series files, which refer to MOBY data averaged over sensor-specific wavelengths and particular hours of the day (around 20:00–23:00 UTC). For this work, the satellite band-average products for SeaWiFS, MODIS AQUA, MERIS, VIIRS and the Ocean and Land Colour Instrument (OLCI) were compiled from the R2017 Reprocessing. The in-band-average subproduct was used, and to maintain the highest quality only data determined from the upper two arms (Lw1) and flagged as good quality were acquired. Data from the MOBY203 deployment were discarded due to the absence of surface irradiance data. The compiled variable was the remote-sensing reflectance, rrs, which was computed from the original water-leaving radiance (Lw) and surface irradiance (Es). The water-leaving radiances were corrected for the bidirectional nature of the light field (Morel and Gentili, 1996; Morel et al., 2002) using the same lookup table and method as that used in the SeaWiFS Data Analysis System (SeaDAS) processing code. The MOBY data were reprocessed in 2017 (MOBY R2017 Reprocessing) to include various improvements in the calibration of the instrument and post-processing, which include (1) a new method to extrapolate the upwelling radiance attenuation coefficient to the surface, (2) an increase in arm depth by 0.234 m and (3) a single pixel shift in the data for the red spectrograph collected at a bin factor of 384. Only the last two changes were included in present compilation. As mentioned before, the MOBY data compiled in this work are sensor-specific. Therefore, attention is necessary to use the correct MOBY data when validating a particular sensor. The way MOBY data are stored in the final merged table is consistent with the original wavelengths; however, these wavelengths can differ from what is sometimes expected to be the central wavelength of a given band and sensor. Irrespective of the wavelength where MOBY data are stored in the final table, for validation of bands 1–6 of SeaWiFS, MOBY data stored in the final merged table at 412, 443, 490, 510, 555 and 670 nm, respectively, should be used. For validation of bands 1–6 of MODIS AQUA, MOBY data stored in the final merged table at 416, 442, 489, 530, 547 and 665 nm, respectively, should be used. For validation of bands 1–7 of MERIS, MOBY data stored in the final merged table at 410.5, 440.4, 487.8, 507.7, 557.6, 617.5 and 662.4 nm, respectively, are the appropriate data. For validation of bands 2–8 of OLCI, MOBY data stored in the final merged table at 410.5, 440.4, 487.8, 507.7, 557.6, 617.5 and 662.4 nm, respectively, are the appropriate data.
data. Finally, for validation of bands 1–5 of VIIRS, MOBY data stored in the final merged table at 412.9, 444.5, 481.2, 556.3 and 674.6 nm, respectively, are the appropriate data.

2.2.2 BOUée pour l’acquisiiSîion de Séries Optiques à Long termE (BOUSSOLE)

The BOUSSOLE project started in 2001 with the objective of establishing a time series of bio-optical properties in oceanic waters to support the calibration and validation of ocean-colour satellite sensors (Antoine et al., 2006). The project consists of a monthly cruise programme and a permanent optical mooring (Antoine et al., 2008). The mooring collects radiometry and inherent optical properties (IOPs) in continuous mode every 15 min at two depths (4 and 9 m nominally). The monthly cruises are devoted to the mooring servicing, to the collection of vertical profiles of radiometry and IOPs, and to water sampling at 11 depths from the surface down to 200 m, for subsequent analyses including phytoplankton pigments, particulate absorption, CDOM absorption and suspended particulate matter load. The BOUSSOLE mooring is in the western Mediterranean Sea at a water depth of 2400 m. All pigment (2001–2012) and radiometric (2003–2012) data were provided by the principal investigator. The compiled variables were rrs and chla_hplc. Observations of the diffuse attenuation coefficient (kd) were not included in the present compilation, as they were under internal quality revision at the time of data acquisition. Remote-sensing reflectance was computed from the original fully normalized water-leaving radiance (nLw_ex), which is the normalized water-leaving radiance (nLw, previously described), with a correction for the bidirectional nature of the light field (Morel and Gentili, 1996; Morel et al., 2002). The solar irradiance (Fo) was computed from two available variables in the original set of data: the normalized water-leaving radiance (nLw) and the remote-sensing reflectance (rrs), using the equation Fo = nLw/rrs. Only radiometric observations that meet the following criteria were used: (1) tilt of the buoy was less than 10°, (2) the buoy was not lowered by more than 2 m as compared to its nominal water line (to ensure the Es reference sensor is above water and exempt from sea spray) and (3) the solar irradiance was within 10% of its theoretical clear-sky value (determined from Gregg and Carder, 1990). The latter criterion was used to select clear skies only. An additional quality control was to remove observations that were 50% higher or lower than the daily average. This removed a small number of spikes in the time series. The final quality control step was to remove days where the standard deviation was more than half of the daily average. This was meant to identify days with high variability. Very few days (N = 2) were removed with this test. These quality control criteria were applied per wavelength, which resulted in some observations with an incomplete spectrum.

2.2.3 AERosol RObotic NETwork-Ocean Color (AERONET-OC)

AERONET-OC is a component of AERONET, including sites where sun photometers operate with a modified measurement protocol leading to the determination of the fully normalized water-leaving radiance (Zibordi et al., 2006, 2009). As a result of collaboration between the Joint Research Centre (JRC) and NASA, this component has been specifically developed for the validation of ocean-colour radiometric products. The strength of AERONET-OC is “the production of standardized measurements that are performed at different sites with identical measuring systems and protocols, calibrated using a single reference source and method, and processed with the same codes” (Zibordi et al., 2006, 2009). All high-quality data (Level-2) were acquired from the project website for 11 sites: Abu Al Bukhoosh (~ 25° N, ~ 53° E), COVE_SEAPRISM (~ 36° N, ~ 75° W), Gloria (~ 44° N, ~ 29° E), Gustav_Dalen_Tower (~ 58° N, ~ 17° E), Helsinki Lighthouse (~ 59° N, ~ 24° E), LISCO (~ 40° N, ~ 73° W), Lucinda (~ 18° S, ~ 146° E), MVCO (~ 41° N, ~ 70° W), Palgrunden (~ 58° N, ~ 13° E; Philipson et al., 2016), Venice (~ 45° N, ~ 12° E) and WaveCIS_Site_CSI_6 (~ 28° N, ~ 90° W). The compiled variable was rrs. Remote-sensing reflectance was computed from the original fully normalized water-leaving radiance (see Sect. 2.2.2 for definition). The solar irradiance (Fo), which is not part of the AERONET-OC data, was computed from the Thullier et al. (2003) solar spectrum irradiance, by averaging Fo over a wavelength-centred 10 nm window. Data were compiled for the exact wavelengths of each record, which can change over time for a given site depending on the specific instrument deployed. In comparison with the previous compilation of AERONET-OC data from the Lucinda site, a calibration correction was applied by NASA affecting instrument SN-520. All radiometric data from this instrument provided by NASA prior to October 2018 were underestimated by approximately a factor of 2 due to incorrect application of instrument gains during the processing.

2.2.4 SeaWiFS Bio-optical Archive and Storage System (SeaBASS)

SeaBASS is one of the largest archives of in situ marine bio-optical data (Werdell et al., 2003). It is maintained by NASA’s Ocean Biology Processing Group (OBPG) and includes measurements of optical properties, phytoplankton pigment concentrations, and other related oceanographic and atmospheric data. The SeaBASS database consists of in situ data from multiple contributors, collected using a variety of measurement instruments with consistent, community-vetted protocols from several marine platforms such as fixed buoys, handheld radiometers and profiling instruments. Quality control of the received data includes a rigorous series of pro-
tocols that range from file format verification to inspection of the geophysical data values (Werdell et al., 2003). Radiometric data were acquired through the Validation search tool, which provided in situ data with matchups for particular ocean-colour sensors (Bailey and Werdell, 2006). The criterion in the search query was defined to have the minimal flag conditions in the satellite data, to retrieve a greater number of matchups and, therefore, in situ data. Regarding phytoplankton pigment data, the majority were acquired through the Pigment search tool, which provided pigment data directly from the archives. As was stated in the SeaBASS website, the Pigment search tool was originally designed to return only in vitro fluorometric measurements, which is consistent with our approach, but over time chlorophyll $a$ measurements made using other methods (e.g. in situ fluorometry) were included in the retrieved pigment data. In the pigment data used in this work, a large number of in situ fluorometric measurements from continuous underway instruments were identified and discarded. These data were initially identified from cruises with more than 50 observations per day and then rechecked in the SeaBASS website to confirm whether indeed they were continuous underway measurements. A total of 120,412 such measurements were identified and discarded. Given the large volume of this group of data, it is possible that some chlorophyll $a$ observations from in situ methods may have escaped the scrutiny and persisted into the final merged dataset. The Pigment search tool was recently discontinued, and, instead, the File search tool can be used, which was also used here to acquire chlorophyll observations for more recent years. The compiled variables from SeaBASS data were $rr$, $chl_{hplc}$, $chl_{fluor}$, $aph$, $bbp$ and $kd$. No conversion was necessary since all variables were acquired in the desired format.

### 2.2.5 NASA bio-Optical Marine Algorithm Dataset (NOMAD)

NOMAD is a publicly available dataset compiled by the NASA OBPG at the Goddard Space Flight Center. It is a high-quality global dataset of coincident radiometric and phytoplankton pigment observations for use in ocean-colour algorithm development and satellite-data product-validation activities (Werdell and Bailey, 2005). The source of the bio-optical data is the SeaBASS archive; therefore, many dependencies exist between these two datasets, which were addressed during the merging. The current version (version 2.0 ALPHA, 2008) includes data from 1991 to 2007 and an additional set of observations of inherent optical properties. The current version was used in this work, but with an additional set of columns of remote-sensing reflectance corrected for the bidirectional effects (Morel and Gentili, 1996; Morel et al., 2002). This additional set of columns was provided directly by the NOMAD creators. The compiled variables were $rr$, $chl_{hplc}$, $chl_{fluor}$, $aph$, $bbp$ and $kd$. Conversion was necessary only for $aph$, $adg$ and $bbp$ and followed the procedures described in Sect. 2.1. For the calculation of $bbp$ the variable $bb$ was used with a smooth fitting to remove noise. A portion of the NOMAD data were optically weighted (for methods see Werdell and Bailey, 2005). These data are not consistent with the protocols chosen in this work, but these observations were retained since NOMAD is a widely used dataset in ocean-colour validation.

### 2.2.6 MERIS Match-up In situ Database (MERMAID)

MERMAID provides in situ bio-optical data matched with concurrent and comparable MERIS Level 2 satellite ocean-colour products (Barker, 2013a, b). The MERMAID in situ database consists of data from multiple contributors, measured using a variety of instruments and protocols from several marine platforms such as fixed buoys, handheld radiometers and profiling instruments. Comprehensive quality control and protocols are used by MERMAID to integrate all the data into a common and comparable format (Barker, 2013a, b). Access to MERMAID data is limited to the MERIS Validation Team, the MERIS Quality Working Group and to the in situ data contributors. For this work, access has been granted to the MERMAID database through a signed service level agreement. The MERMAID data includes subsets of several datasets used in this compilation (MOBY, AERONET-OC, BOUSSOLE, NOMAD). These observations were removed from the MERMAID dataset to avoid duplication (as discussed in Sect. 2.1). The compiled variables were $rr$, $chl_{hplc}$, $chl_{fluor}$, $aph$, $bbp$, $kd$ and $tsm$. Remote-sensing reflectance was calculated by dividing the original fully normalized water-leaving reflectance ($R_{w_ex}$), which is the water-leaving reflectance ($R_w = \pi L_w/E_s$), with a correction for the bidirectional nature of the light field (Morel and Gentili, 1996; Morel et al., 2002), by $\pi$. Conversion was also necessary for $aph$, $adg$ and $bbp$ and followed the procedures described in Sect. 2.1.

### 2.2.7 Hawaii Ocean Time-series (HOT)

HOT programme provides repeated comprehensive observations of the hydrography, chemistry and biology of the water column at a station located 100 km north of Oahu, Hawaii, since October 1988 (Karl and Michaels, 1996). This site is representative of the North Pacific subtropical gyre. Cruises are made approximately once a month to the deep-water station ALOHA (A Long-Term Oligotrophic Habitat Assessment; 22°45’N, 158°00’W). Pigment data ($chl_{hplc}$ and $chl_{fluor}$) were extracted directly from the project website. Radiometric measurements from the HOT project are also available, but observations of $rr$ and $kd$ from the HOT project were acquired in this work as part of the SeaBASS dataset.
2.2.8 Geochemy, Phytoplankton, and Color of the Ocean (GeP&CO)

GeP&CO is part of the French PROOF programme and aims to describe and understand the variability of phytoplankton populations, as well as to assess its consequences on the geochemistry of the oceans (Dandonneau and Niang, 2007). It is based on the quarterly travels of the merchant ship Contship London from France to New Caledonia in the Pacific. A scientific observer sailed on each trip and operated the sampling for surface water, filtration, various measurements and checking at several times of each day. The experiment started in October 1999 and finished in July 2002. Pigment data were extracted from the project website. Additional pigment data obtained during the OISO-4 cruise in the southern Indian Ocean on board R/V Marion Dufresne (January–February 2000) were added. The samples were measured by Yves Dandonneau following the method used in the GeP&CO project. The compiled variable was chla_hplc and chla_fluor.

2.2.9 Atlantic Meridional Transect (AMT)

AMT is a multidisciplinary programme, which undertakes biological, chemical and physical oceanographic research during an annual voyage between the UK and destinations in the South Atlantic (Robinson et al., 2006). The programme was established in 1995 and since then has completed 28 research cruises. Pigment data between 1997 (AMT5) and 2005 (AMT17) were provided by the British Oceanographic Data Centre (BODC) following a specific request for discrete observations of chlorophyll \( a \) concentration since 1997. The AMT data were isolated by searching for the string AMT in the cruise columns, and the respective principal investigators were then searched for individually in a separated metadata file. Data not flagged with highest quality or without method of measurement were not used. For any interest in the original data, BODC is the point of contact, which ensures that if there are any updates, the most recent data are supplied. The compiled variables are chla_hplc and chla_fluor.

2.2.10 International Council for the Exploration of the Sea (ICES)

ICES is a network of more than 4000 scientists from almost 300 institutes, with 1600 scientists participating in activities annually. The ICES Data Centre manages a number of large dataset collections related to the marine environment covering the northeastern Atlantic, Baltic Sea, Greenland Sea and Norwegian Sea. The majority of data originate from national institutes that are part of the ICES network of member countries. Data were provided (on 28 April 2014) from the ICES database on the marine environment (Copenhagen, Denmark) following a specific request. The ICES data were made available under the ICES data policy, and if there is any conflict between this and the policy adopted by the users, then the ICES policy applies. The compiled variables were chla_hplc and chla_fluor.

2.2.11 Arctic System Science Primary Production (ARCSSPP)

The ARCSSPP database is a synthesis of observations between 1954 and 2006 from the Arctic Ocean and northern seas (Matrai et al., 2013). The observations were acquired from data repositories, publications or provided by individual investigators. The database includes quality-controlled observations of productivity and chlorophyll \( a \), photosynthetically available radiation and hydrographic parameters. This collection of data was acquired at http://www.nodc.noaa.gov/cgi-bin/OAS/prd/accession/download/63065 (last access: 10 July 2019). For the present work, only observations of chlorophyll \( a \) concentration with known time zones were used. The compiled chlorophyll observations were from discrete samples, but the exact method (either chla_fluor or chla_hplc) was not available for all observations. Thus, the ARCSSPP chlorophyll observations were marked as chla_fluor, although some might have been from HPLC measurements, and were flagged with “1” in the column flag_chla_method. The compiled variable was chla_fluor.

2.2.12 Data provided by Astrid Bracher, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI)

In this work, the AWI data source refers to the group of observations that were provided to the OC-CCI project by Astrid Bracher. These are bio-optical observations collected during several cruises in the Atlantic and Pacific Ocean. All data were available through the PANGAEA repository. Observations of concentration of chlorophyll \( a \) as well as 1 nm spectrally resolved remote-sensing reflectances and the algal pigment absorption coefficient were considered. The methods for these observations are described by Taylor et al. (2011a). For chlorophyll, data from the following cruises were used: ANT-XXIV/1, ANT-XXIV/4, ANT-XXVI/4 and MSM18/3 (https://doi.org/10.1594/PANGAEA.847820; Bracher et al., 2015); SO202/2 (https://doi.org/10.1594/PANGAEA.820607; Zindler et al., 2013); ANT-XXVII/2 (https://doi.org/10.1594/PANGAEA.848590; Bracher, 2015); ANT-XXVIII/3 and SO218 (https://doi.org/10.1594/PANGAEA.819099; Taylor et al., 2011b); and ANT-XXVIII/3 and SO218 (https://doi.org/10.1594/PANGAEA.848591; Soppa et al., 2014). Concerning remote-sensing reflectances, the observations taken during cruises ANT-XXIV/4 and ANT-XXVI/4 (https://doi.org/10.1594/PANGAEA.847820; Bracher et al., 2015) and cruise ANT-XXV/1 (https://doi.org/10.1594/PANGAEA.819099; Taylor et al., 2011b) were gathered. The remote-sensing reflectances...
were corrected for the bidirectional nature of the light field (Morel and Gentili, 1996; Morel et al., 2002). The absorption coefficients were taken during cruise SO202/2 (https://doi.org/10.1594/PANGAEA.820607; Zindler et al., 2013), cruise ANT_XXV/1 (https://doi.org/10.1594/PANGAEA.819099; Taylor et al., 2011b), and cruises ANT-XXV/3 and ANT-XXVIII/3 (https://doi.org/10.1594/PANGAEA.819617; Soppa et al., 2013). The compiled variables were chla_hplc, rrs and aph.

2.2.13 Bermuda Atlantic Time-series Study (BATS)

BATS is a long-term study by the Bermuda Institute of Ocean Sciences based on regular cruises in the western Atlantic Ocean (Sargasso Sea) since 1988. The cruises at the BATS site (∼31°40′ N, 64°10′ W) sample ocean temperature and salinity but are focused on biogeochemical variables such as nutrients, dissolved inorganic carbon, oxygen, HPLC of pigments, primary production and sediment trap flux. In this work all the phytoplankton pigment data available from the BATS website (http://bats.bios.edu/bats-data/, last access: 10 July 2019) were considered, which also included regional and transect cruises not specific to the nominal BATS site. The compiled variables were chla_hplc and chla_fluor.

2.2.14 Data provided by Knut Yngve Børsheim (BARENTSSEA)

The BARENTSSEA data source refers to a group of observations that were provided to OC-CCI project by Knut Yngve Børsheim. This collection was developed using data from the archives of the Institute of Marine Research (Norway). It comprises observations of temperature, salinity and chlorophyll a routinely collected by cruises, mainly in the North Sea, the Norwegian Sea and the Barents Sea between 1997 and 2013. The chlorophyll a concentration was measured by filtering and extraction using Turner fluorometers. The compiled variable was chla_fluor.

2.2.15 The Fisheries and Oceans Canada database for biological and chemical data (BIOCHEM)

BioChem is an archive of marine biological and chemical data maintained by Fisheries and Oceans Canada (DFO, 2018; Devine et al., 2014). The available observations are from department research initiatives and collected in areas of Canadian interest. Available parameters include pH, nutrients, chlorophyll, dissolved oxygen and other plankton data (species and biomass). Chlorophyll measurements from in vitro fluorometric methods were extracted (from http://www.dfo-mpo.gc.ca/science/data-donnees/biochem/index-eng.html, last access: 10 July 2019) with close guidance by the BioChem help desk, confirming quality and methods. The used data span from 1997 to 2014 and were mainly from the Gulf of St Lawrence (western North Atlantic). The compiled variable was chla_fluor.

2.2.16 British Oceanographic Data Centre (BODC)

BODC is the designated marine science data centre for the United Kingdom. The data used in this work derive from a specific request for discrete observations of chlorophyll a concentration since 1997. Initially, this request was used to compile AMT data (see Sect. 2.2.9). The remaining data comprising observations of chlorophyll a concentration from fluorometric and HPLC methods, mostly sampled in the North Atlantic, were analysed and added (the dataset string for this data source is bdoc). Data not flagged with highest quality or without method of measurement were discarded. The compiled variables were chla_hplc and chla_fluor.

2.2.17 California Cooperative Oceanic Fisheries Investigations (CALCOFI)

CalCOFI is a partnership of the California Department of Fish and Wildlife, National Oceanic and Atmospheric Administration Fisheries Service, and Scripps Institution of Oceanography. CalCOFI has conducted quarterly cruises off southern and central California since 1949. Data collected in the upper 500 m include temperature, salinity, oxygen, nutrients, chlorophyll, primary productivity, plankton biodiversity and biomass. For this work, only observations of chlorophyll a concentration derived from fluorometric methods flagged with highest quality were used. Data were acquired from the file CalCOFI_Database_194903-201701_csv_20Sept2017.zip available at http://www.calcofi.org (last access: 10 July 2019) The compiled variable was chla_fluor.

2.2.18 California Current Ecosystem Long-Term Ecological Research (CCELTER)

CCELTER investigates the California Current coastal pelagic ecosystem, with a focus on long-term forcing. The CCELTER data include primary and derived measurements from both Process and CalCOFI-augmented cruises, as well as other time series. CCELTER data include variables from the physical environment, biogeochemistry and biological populations/communities. For this work chlorophyll observations measured from discrete bottle samples from CCELTER Process cruises determined by extraction and bench fluorometry (https://doi.org/10.6073/pasta/7feb632dabb30f0e79683017721a83c7; Goericke, 2017) were compiled. The compiled variable was chla_fluor.
2.2.19 Center for Integrated Marine Technologies (CIMT)

CIMT was a non-operational programme where marine scientists from different disciplines and institutions combine their efforts on observations directed towards understanding the central California upwelling system. The CIMT archived data include coastal ocean observations from satellites, shipboard data, moorings and large marine animal movements. For this work, pigment data from discrete bottle samples taken during CIMT monthly cruises were used. Data were acquired from the project website (https://cimt.ucsc.edu/data_portal.htm, last access: 10 July 2019). The compiled variable was chla_fluor.

2.2.20 CoastColour Round Robin (COASTCOLOUR)

COASTCOLOUR datasets were designed to evaluate the performance of ocean-colour satellite algorithms in the retrieval of water quality parameters in coastal waters (Nechad et al., 2015a). Three types of COASTCOLOUR datasets are available: (1) a matchup dataset where in situ bio-optical observations are available simultaneously with a cloud-free MERIS product, (2) an in situ reflectance dataset where an in situ reflectance is available simultaneously with an in situ measurement of chlorophyll \(a\) concentration and/or total suspended matter, and (3) a simulated dataset where reflectances were generated by a radiative transfer model. This work used the matchup dataset, which includes most of the in situ measurements and is available at https://doi.org/10.1594/PANGAEA.841950 (Nechad et al., 2015b). The matchup dataset provides optical, biogeochemical and physical data collections at 17 sites across the globe. From this dataset, observations of reflectance, chlorophyll \(a\), total suspended matter and IOPs were compiled. The remote-sensing reflectances were corrected for the bidirectional nature of the light field (Morel and Gentili, 1996; Morel et al., 2002). The compiled variables were rrs, chla_hplc, chla_fluor, aph, adg, bbp and tsm.

2.2.21 European Station for Time series in the Ocean, Canary Islands (ESTOC)

ESTOC is an open-ocean monitoring site located in the eastern North Atlantic subtropical gyre. ESTOC was initiated in 1991 with particle flux measurements and in 1994 began standard observations of the water column, in addition to the deployment of a current meter mooring. The core parameters measured at ESTOC include salinity, temperature, current speed, nutrients, chlorophyll, inorganic carbon, particulate organic carbon and nitrogen, and sinking particle flux (Neuer et al., 2007). For this work measurements of chlorophyll \(a\) concentration from monthly cruises from 1994 to 2011 were used. These data were provided to CCI following a specific request. The time of day was unavailable and was set to 12:00:00 UTC. These observations were flagged with “1” in the column flag_time. The compiled variable was chla_fluor.

2.2.22 Integrated Marine Observing System (IMOS)

IMOS is a national collaborative research infrastructure supported by Australian Government. Since 2006, IMOS has operated a wide range of observing equipment throughout the coastal and open ocean around Australia, making all data openly available to the scientific community and other stakeholders and users. In this work, the IMOS dataset refers only to a data collection entitled IMOS National Reference Station (NRS) – Phytoplankton HPLC Pigment Composition Analysis, which was acquired from the Australian Ocean Data Network Portal (https://portal.aodn.org.au, last access: 10 July 2019). This dataset comprises phytoplankton pigment composition measured by HPLC collected as part of the IMOS National Mooring Network – National Reference Station field sampling. Pigment sampling was conducted on a monthly basis with small vessels at nine sites. The IMOS also hosts the Satellite Remote Sensing Bio-optical Database, which comprises phytoplankton pigment composition measured by HPLC collected as part of a suite of bio-optical parameters from samples collected from research voyages in Australian waters; however, for this work, the observations from the IMOS Bio-optical Database were acquired as a subset of the SeaBASS dataset. The compiled variable was chla_hplc.
2.2.23 MARine Ecosystem DATa (MAREDAT)

The MAREDAT database is a global assemblage of pigments measured by HPLC (Peloquin et al., 2013a) from the combination of 136 independent field datasets, solicited from investigators and databases. The database provides high-quality measurements of taxonomic pigments including chlorophyll $a$ and $b$, 19'-butanoyloxyfucoxanthin, 19'-hexanoyloxyfucoxanthin, alloxanthin, divinyl chlorophyll $a$, fucoxanthin, lutein, peridinin, prasinoxanthin, violaxanthin and zeaxanthin. The database is available through PANGAEA (https://doi.org/10.1594/PANGAEA.793246; Peloquin et al., 2013b). For this work only measurements of total chlorophyll $a$ flagged with high quality were used. The time of day was unavailable and was set to 12:00:00 UTC. These observations were flagged with “1” in the column flag_time. The compiled variable was chla_hplc.

2.2.24 Palmer station Long-Term Ecological Research (PALMER)

PALMER is a monitoring station located in western Antarctic Peninsula. The Palmer station investigates the marine ecology of the Southern Ocean with a focus on the pelagic marine ecosystem, including sea ice habitats, regional oceanography and nesting sites of seabird predators. The PALMER data include measurements of meteorological, oceanographic, sea ice, predators, nutrients and biogeochemistry, pigments, primary production, zooplankton and microbe parameters. This work used the measurements of chlorophyll analysed by HPLC and fluorometry taken at the Palmer station (https://doi.org/10.6073/pasta/0624c7d161d3b5486d7ba06c2e50ee21; Schofield et al., 2018a; and https://doi.org/10.6073/pasta/dea95430a6ad844ceaa023ee1ced650d3; Schofield et al., 2018b) and from the annual cruises off the coast of the western Antarctic Peninsula (https://doi.org/10.6073/pasta/4d583713667a0f52b9d2937a26d0d82e; Schofield et al., 2018c; and https://doi.org/10.6073/pasta/c479b92d42ace1ce37f9a977e214952; Schofield et al., 2017). The compiled variables were chla_hplc and chla_fluor.

2.2.25 SeaDataNet archive (SEADATANET)

SeaDataNet is a Pan-European infrastructure for ocean and marine data management. It aims to develop a standardized
2.2.26 Data provided by Trevor Platt and Shubha Sathyendranath (TPSS)

In this work, the TPSS data source refers to a group of observations that were provided to this compilation by Trevor Platt and Shubha Sathyendranath. This is a collection of bio-optical in situ data collected during cruises predominantly in the northwestern Atlantic but also from the Indian Ocean, South Pacific and central Atlantic (see Sathyendranath et al., 2009, for additional details regarding the cruises). It comprises measurements of phytoplankton pigments and algal pigment absorption coefficients. The time of day was unavailable and was set to 12:00:00 UTC. These observations were flagged with “1” in the column flag_chla_method. The compiled variables were chla_hplc, chla_fluor and aph.

2.2.27 Bio-optical data from Tara expeditions (TARA)

The Tara expeditions consist of several cruises around the world, some with durations of several years, designed to study and understand the distribution of planktonic organisms in the world ocean. The discrete observations of remote-sensing reflectance and chlorophyll a concentration from HPLC measurements taken during the Tara Oceans (2009–2013) and Mediterranean (2014) expeditions were considered in this work. These data were provided to the ESA OCCI project by Emmanuel Boss and were available in the SeaBASS archive. The remote-sensing reflectances were corrected for the bidirectional nature of the light field (Morel and Gentili, 1996; Morel et al., 2002). The compiled variables were chla_hplc and rrs.

3 Results

In this work several sets of bio-optical in situ data were acquired, homogenized and merged into a single table. The table comprises in situ observations between 1997 and 2018, with a global distribution, and includes the following variables: remote-sensing reflectance (rrs), chlorophyll a concentration (chla), algal pigment absorption coefficient (aph), detrital and coloured dissolved organic matter absorption (adg), particle backscattering coefficient (bbp), diffuse attenuation coefficient for downward irradiance (kd) and total suspended matter (tsm). All observations in the table were processed in such a way that they can be compared directly with satellite-derived ocean-colour data. The table consists of 136,250 rows and 1286 columns. Each row represents a unique station in space and time, separated from the rest by at least 5 min and 200 m. For each observation in a given station, there are three metadata strings: dataset, subdataset and contributor. The columns of the table take the form described in Table 1. The data contributors are indicated in Table 2. Regarding spectral variables, all original wavelengths were preserved, which requires a large number of unique wavelengths to be maintained in the database. No band shifting was performed (though some archived data in some data sources may have been merged with nearby wavelengths) and no minimum number of wavelengths per observation was imposed. This allows further manipulation of the table for different purposes. In the following paragraphs, the table is analysed and the final group of observations is described for each contributing dataset; however, the numbers reported here do not reflect the original numbers in each dataset, since duplicates across contributing datasets were removed (e.g. NOMAD and others were removed from MERMAID).

Observations of remote-sensing reflectance are available at 611 unique wavelengths (i.e. columns), between 404.7 and 1022.1 nm (Fig. 1). In total there are 59,781 observations (i.e. rows) with remote-sensing reflectance in the table. The total number of observations are partitioned per contributing datasets as follows: AERONET-OC (31,574), BOUS-
Figure 5. Global distribution of remote-sensing reflectance per dataset in the final table. The data sources are identified with different colours. Points show locations where at least one observation is available. Crosses show sites from which time series data of remote-sensing reflectance are available.

Figure 6. Comparison of coincident observations of chlorophyll $a$ concentration derived with different methods (chla_fluor and chla_hplc). The data were transformed prior to regression analysis to account for their log-normal distribution.

Figure 7. Number of observations per chlorophyll $a$ concentration acquired with different methods (chla_fluor and chla_hplc).

SOLE (17 364), MOBY (5466), NOMAD (3326), MERMAID (885), SeaBASS (698), AWI (54), COASTCOLOUR (307) and TARA (107). Data from AERONET-OC, BOUSSOLE and MOBY correspond to continuous time series, and, hence, the higher number of observations. Data distribution at 44X and 55X nm is provided in Fig. 2a and b, respectively. Data were first searched for at 445 and 555 nm and then with a search window up to 8 nm to include also data at 547 nm. Median values at 44X nm range from 0.003 m$^{-1}$ (AERONET-OC) and 0.009 m$^{-1}$ (MOBY), whereas at 55X nm the median values lie between 0.001 m$^{-1}$ (AWI) and 0.007 m$^{-1}$ (COASTCOLOUR). The observations are unevenly distributed between each month of the year in both hemispheres, with a higher coverage in summer months (Fig. 3). There are fewer data in the Southern Hemisphere than in the Northern Hemisphere (Fig. 3). For additional analysis, rrs band ratios were plotted against each other (490 : 555 versus 412 : 443, Fig. 4). Most points are within the boundaries of the NOMAD dataset, but some scattered points were found. These points were retained in the table to allow further manipulation with different quality control criteria. Complementary analysis of remote-sensing reflectance data is made when other variables are concurrently avail-
Figure 8. Global distribution of chlorophyll a concentration per interval of the observed value. All chlorophyll data were considered, but for a given station, HPLC data were selected if available.

Figure 9. Global distribution of chlorophyll a concentration per dataset in the final table. All chlorophyll data were considered, but for a given station, HPLC data were selected if available. Crosses show sites from where data of chlorophyll are available in a specific geographic location.

able and discussed below (see Figs. 11 and 16). The geographic distribution of remote-sensing reflectance observations (Fig. 5) shows a higher number of observations in some coastal regions, such as those of North America and northern Europe. The central regions of the ocean show a lower number of observations, with the Atlantic Ocean having the highest density in relation to the other oceans. The best geographic coverage is provided by the NOMAD database. Data from SeaBASS are fewer in number but are still important. Data from MERMAID are mainly located along the coasts of Europe, North America and the central region of the North Atlantic Ocean. The observations from COASTCOLOUR are concentrated in 17 coastal sites around the world, while AWI data are available for the Atlantic, Pacific and Southern Ocean. TARA data are spread across several regions, with the highest data density in the Mediterranean Sea.
For chlorophyll $a$ concentration, two types of observations were compiled, one measured by fluorometric or spectrophotometric methods (chla_fluor) and the other measured by HPLC methods (chla_hplc). A comparison of both measurements (Fig. 6), when available at the same station, shows good agreement (Trees et al., 1985). As stated before, the analysis was done on the final merged table; thus no data were filtered and the good relation can be explained in part by the quality control implemented by the data providers and curators of repositories such as NOMAD and SeaBASS (Werdell and Bailey, 2005). The total number of rows with concurrent chla_fluor and chla_hplc is 5344, with contributions from SeaBASS (39%), TPSS (18%), NOMAD (13%), PALMER (9%), BATS (6%), COASTCOLOUR (5%), MERMAID (4%), HOT (4%), and AMT+GeP&CO+BODC+CCELTER+CALCOFI (2%). The chla_fluor observations are available in 61,525 stations (rows), with values ranging from 0.001 to 100 mg m$^{-3}$ (Fig. 7). They are from NOMAD (2350), SeaBASS (18,122), MERMAID (3711), ICES (5421), HOT (702),...
Figure 11. A remote-sensing reflectance maximum band ratio (as defined in text) ([443, 490, 510] / 555 or [443, 490, 510] / 560 if 555 not available) as a function of chlorophyll a concentration. All chlorophyll data were considered, but for a given station, HPLC data were selected if available. Data within 2 nm of the wavelengths were used. For reference, the solid and dotted lines show the NASA OC4 and OC4E v6 standard algorithms, respectively (https://oceancolor.gsfc.nasa.gov/atbd/chlor_a/, last access: 10 July 2019). The total number of points was 3814, of which 79 % were from NOMAD.

A. Valente et al.: A compilation of global bio-optical in situ data – version two

The spatial distribution of the chlorophyll values in the subtropical gyres and higher values in coastal and upwelling areas. There is a close correspondence between the spatial patterns of the average and standard deviation maps (Fig. 10b and c), which may be an indicator of the data quality. Coincident observations of chlorophyll a concentration and remote-sensing reflectance are available at 3814 stations. These observations are mostly from NOMAD (79 %), MERMAID (9 %), COASTCOLOUR (6 %), and SeaBASS (5 %). The maximum of three selected band ratios of remote-sensing reflectance is plotted against chlorophyll a concentration (Fig. 11). The chla values used are the combined HPLC and fluorometric chlorophyll a, and for the rrs, the closest spectral observation within 2 nm was used. The maximum band ratios were calculated as the maximum of [rrs(443) / rrs(555), rrs(490) / rrs(555), rrs(555) / rrs(490)].
Figure 12. The distribution of (a) $\text{aph}$ at 44 nm, (b) $\text{aph}$ at 55 nm, (c) $\text{adg}$ at 44 nm, (d) $\text{adg}$ at 55 nm, (e) $\text{bbp}$ at 44 nm, (f) $\text{bbp}$ at 55 nm, (g) $\text{kd}$ at 44 nm, and (h) $\text{kd}$ at 55 nm. Data were first searched for at 445 and 555 nm and then with a search window up to 8 nm to include data at 547 nm. The graphical convention is identical to Fig. 2.

The inherent optical properties ($\text{aph}$, $\text{adg}$ and $\text{bbp}$) are available at 550 unique wavelengths between 300 and 850 nm. There is a total of 3293, 1654 and 792 observations, for $\text{aph}$, $\text{adg}$ and $\text{bbp}$, respectively. For $\text{aph}$ the total number of observations is distributed among NOMAD (1190), TPSS (966), COASTCOLOUR (593), AWI (458), SeaBASS (14) and MERMAID (72). For $\text{adg}$ the contributions are as follows: NOMAD (1079), COASTCOLOUR (531), SeaBASS (11) and MERMAID (33). The $\text{bbp}$ observations come from NOMAD (371), COASTCOLOUR (154), SeaBASS (32) and MERMAID (235). The data distribution of $\text{aph}$, $\text{adg}$ and $\text{bbp}$ at 44 nm and 55 nm for each dataset is provided in Fig. 12a–f. Median values of $\text{aph}$, $\text{adg}$ and $\text{bbp}$ at 44 and 55 nm for each dataset are summarized in Table 3. For additional analysis, the following band ratios for the absorption coefficients were calculated: $\text{aph}(490)/\text{aph}(443)$, $\text{aph}(412)/\text{aph}(443)$, $\text{adg}(443)/\text{adg}(490)$ and $\text{adg}(412)/\text{adg}(443)$. Data within 2 nm of the wavelengths were used to maximize the number of points. The distribution of the ratios is shown in Fig. 13. Several observations were found to be outside the thresholds used in the IOCCG report 5 (2006) for quality control (see dotted vertical black lines in Fig. 13). These points are highlighted here for information but retained in the database, as these were mostly from NOMAD and there was an inter-
est to preserve this dataset as a whole. Also, not discarding these data allows further manipulation with different quality control criteria. On the annual scale, the observations of the inherent optical properties are strongly underrepresented in the Southern Hemisphere where there is a complete absence of data in several months of the year (Fig. 3). Overall, the geographic coverage for observations of aph, adg and bbp (Fig. 14) is poor, with most open ocean regions not being sampled, except for the Atlantic Ocean. Small clusters of data are located in particular coastal regions.

Finally, for the diffuse attenuation coefficient for downward irradiance (kd) there are 25 unique wavelengths between 405 and 709 nm. There is a total of 2454 observations from NOMAD (2266), SeaBASS (118) and MERMAID (70). Data distribution of kd at 44 nm and 55 nm for each dataset is shown in Fig. 12g and h. No kd data at these wavelengths were available for the SeaBASS dataset (only at 490 nm). Median values of kd at 44 nm span between 0.08 m$^{-1}$ (NOMAD) and 0.1 m$^{-1}$ (MERMAID), whereas at 55 nm the kd values are approximately 0.1 m$^{-1}$ (NOMAD and MERMAID). NOMAD provides the best geographical coverage (Fig. 15), with a higher coverage in the Atlantic, compared with other oceans. With the exception of the coastal regions of North America and the Sea of Japan, most coastal regions are not sampled. In the Northern Hemisphere, kd is distributed roughly evenly across all months of the year, but in the Southern Hemisphere there are few data points during the austral winter and none at all in September (Fig. 3). For total suspended matter (tsm) there is a total of 1546 observations divided between COASTCOLOUR (1199) and MERMAID (347). The observations of tsm are available in a greater number in the Northern Hemisphere (Fig. 3) and are distributed across several coastal regions around Europe, the Mediterranean Sea, the South China Sea, Indonesia and Australia (Fig. 15).

Although most of the stations with concurrent variables are from the NOMAD dataset, for completeness, an examination of bio-optical relationships is provided (Fig. 16). The relation between aph at 443 nm and chlorophyll $a$ (Fig. 16 a) agrees with Bricaud et al. (2004). A total of 2953 points exist with these two variables available (34 % from NOMAD, 32 % from TPSS, 11 % from AWI, 11% from COASTCOLOUR, and the remaining 12 % from MERMAID and SeaBASS). The relation between the sum of aph and adg at 443 nm and rrs at 443 nm (Fig. 16 b) shows a similar dispersion, with the exception of some scattered points, to an equivalent analysis on the IOCCG report 5 (see their Fig. 2.3). Again, the scattered data were retained in the final table to preserve the NOMAD dataset. A total of 1112 points exist for which these three variables are available (97 % from NOMAD). The relation between the ratio rrs(490) / rrs(555) and kd(490) (Fig. 16c) shows a good agreement with the NASA KD2S standard algorithm (https://oceancolor.gsfc.nasa.gov/atbd/kd_490/). A total of 2280 points exist for which these three variables are available (93 % from NOMAD). The relation between the ratio rrs(490) / rrs(555) and bbp at 555 nm (Fig. 16 c) shows a good agreement with the relation suggested by Tiwari and Shanmugam (2013). A total of 365 points exist for which these three variables are available (89 % from NOMAD).

4 Data availability

Information about the data availability can be found in Appendix B.

5 Conclusions

In this work, a compilation of bio-optical in situ data is presented, resulting from the acquisition, homogenization and integration of several sets of data obtained from different sources. The compiled data have a global coverage and span the period from 1997 to 2018. Minimal changes were made to the original data, other than the ones occurring from conversion to standard format and quality control. In situ measurements of the following variables were compiled: remote-sensing reflectance, chlorophyll $a$ concentration, algal pigment absorption coefficient, detrital and coloured dissolved organic matter absorption coefficient, particle backscattering coefficient, diffuse attenuation coefficient for downward irradiance and total suspended matter.

The final set of data consists of a substantial number of in situ observations, available in a simple text table and processed in a way that could be used directly for the evaluation
Table 3. Summary of median values for aph, adg and bbp at 44\(X\) and 55\(X\) nm for each dataset (as shown in Fig. 12a–f). Data were first searched for at 445 and 555 nm and then with a search window up to 8 nm to include data at 547 nm.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Median aph 44 nm</th>
<th>Median aph 55 nm</th>
<th>Median adg 44 nm</th>
<th>Median adg 55 nm</th>
<th>Median bbp 44 nm</th>
<th>Median bbp 55 nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeaBASS</td>
<td>0.0549</td>
<td>0.0074</td>
<td>0.0711</td>
<td>0.0222</td>
<td>0.0035</td>
<td>0.0025</td>
</tr>
<tr>
<td>MERMAID</td>
<td>0.0282</td>
<td>0.0052</td>
<td>0.1149</td>
<td>0.0286</td>
<td>0.0080</td>
<td>0.0052</td>
</tr>
<tr>
<td>NOMAD</td>
<td>0.0353</td>
<td>0.0046</td>
<td>0.0515</td>
<td>0.0112</td>
<td>0.0030</td>
<td>0.0022</td>
</tr>
<tr>
<td>COASTCOLOUR</td>
<td>0.0665</td>
<td>0.0096</td>
<td>0.1259</td>
<td>0.0175</td>
<td>0.0047</td>
<td>0.0037</td>
</tr>
<tr>
<td>AWI</td>
<td>0.0208</td>
<td>0.0032</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TPSS</td>
<td>0.0454</td>
<td>0.0071</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Figure 14. Global distribution of observations of inherent optical properties (algal pigment absorption coefficient aph, detrital plus CDOM absorption coefficient adg, and particle backscattering coefficient bbp) in the final table.

Figure 15. Global distribution of diffuse attenuation coefficient for downward irradiance (kd) and total suspended matter (tsm) per dataset in the final table. The tsm and kd points from MERMAID overlap each other in the western Black Sea (~40° N, 30° E) and the Arctic (~70° N, 120° W).
Figure 16. Examples of bio-optical relationships in the final merged table: (a) $\text{aph}(443)$ versus chlorophyll $\alpha$. The total number of points (2953) is divided between AWI (334), COASTCOLOUR (335), MERMAID (214), NOMAD (991), SeaBASS (124) and TPSS (955). For reference the solid line shows the regression from Bricaud et al. (2004). (b) $[\text{aph}(443) + \text{adg}(443)]$ versus rrs(443). The total number of points (1112) is divided between MERMAID (33) and NOMAD (1079). (c) $[\text{rrs}(490)/\text{rrs}(555)]$ versus $\text{kd}(490)$. The total number of points (2280) is divided between MERMAID (62), NOMAD (2117) and SeaBASS (101). For reference the solid line shows the NASA KD2S standard algorithm (https://oceancolor.gsfc.nasa.gov/atbd/kd_490/, last access: 10 July 2019). (d) $[\text{rrs}(490)/\text{rrs}(555)]$ versus bbp(555). The total number of points (365) is divided between MERMAID (33), NOMAD (324), and COASTCOLOUR+SeaBASS (4). For reference the solid line shows the relation proposed by Tiwari and Shanmugam (2013). A search window of 2 nm was used for panels (a) and (b), and a search window of 5 nm was used for panels (c) and (d) to include data at 560 nm when not available at 555 nm.
purposes, including the validation of retrievals from recent space-borne sensors such as Landsat 8 and Sentinel-2 and 3. It may also be useful in the preparation of future sensors like NASA PACE. The objective of publishing the compilation is to make it easily accessible to the broader community.

Note on former version

A former version of this article was published on 3 June 2016 and is available at https://doi.org/10.5194/essd-8-235-2016.
Appendix A: Notation

ad Detrital absorption coefficient (m$^{-1}$)
adg Detrital plus CDOM absorption coefficient (m$^{-1}$)
AERONET-OC AERosol RObotic NETwork-Ocean Color
ag CDOM absorption coefficient (m$^{-1}$)
AMT Atlantic Meridional Transect
ap Particle absorption coefficient (m$^{-1}$)
aph Algal pigment absorption coefficient (m$^{-1}$)
ARCSSPP Arctic System Science Primary Production
AWI Data collection from Astrid Bracher
aw Pure water absorption coefficient (m$^{-1}$)
BARENTSSEA Data collection from Knut Yngve Børsheim
BATS Bermuda Atlantic Time-series Study
bb Total backscattering coefficient (m$^{-1}$)
bbp Particle backscattering coefficient (m$^{-1}$)
bbw Backscattering coefficient of seawater (m$^{-1}$)
BIOCHEM The Fisheries and Oceans Canada database for biological and chemical data
BODC British Oceanographic Data Centre
BOUSSOLE Bouée pour l’acquisition d’une Série Optique à Long Terme
CALCOFI California Cooperative Oceanic Fisheries Investigations
CCELTER California Current Ecosystem Long Term Ecological Research
CDOM Coloured Dissolved Organic Matter
chla Chlorophyll a concentration (mg m$^{-3}$)
chla_fluor Chlorophyll a concentration determined from fluorometric or spectrophotometric methods (mg m$^{-3}$)
chla_hplc Total chlorophyll a concentration determined from the HPLC method (mg m$^{-3}$)
CIMT Center for Integrated Marine Technology
COASTCOLOUR Compilation of data in several coastal sites
Es Surface irradiance (or above-water downwelling irradiance) (mW cm$^{-2}$ μm$^{-1}$)
ESA European Space Agency
ESTOC Estación Europea de Series Temporales del Oceano
Fo Top-of-the-atmosphere solar irradiance (mW cm$^{-2}$ μm$^{-1}$)
GeP&CO Geochemistry, Phytoplankton, and Color of the Ocean
HOT Hawaii Ocean Time-series
HPLC High-performance liquid chromatography
ICES International Council for the Exploration of the Sea
IMOS Integrated Marine Observing System
kd Diffuse attenuation coefficient for downward irradiance (m$^{-1}$)
Lw Water-leaving radiance (or above-water upwelling radiance) (mW cm$^{-2}$ μm$^{-1}$ sr$^{-1}$)
MAREDAT Compilation of data in several coastal sites
MERIS Medium Resolution Imaging Spectrometer
MERMAID MERIS Match-up In situ Database
MOBY Marine Optical Buoy
MODIS Moderate Resolution Imaging Spectroradiometer
NASA National Aeronautics and Space Administration
nLw Normalized water-leaving radiance (mW cm$^{-2}$ μm$^{-1}$ sr$^{-1}$)
nLw_ex nLw with a correction for bidirectional effects (mW cm$^{-2}$ μm$^{-1}$ sr$^{-1}$)
NOMAD NASA bio-Optical Marine Algorithm Dataset
OC-CCI Ocean Colour Climate Change Initiative
OLCI Ocean and Land Colour Instrument
PALMER Palmer station Long-Term Ecological Research
rrs Remote-sensing reflectance (sr$^{-1}$)
Rw Irradiance reflectance (dimensionless)
SeaBASS SeaWiFS Bio-optical Archive and Storage System
SEADATANET Archive of in situ marine data
SeaWiFS Sea-viewing Wide Field-of-view Sensor
TARA Data collection from global transects
TPSS Data collection from Trevor Platt and Shubha Sathyendranath
VIIRS Visible Infrared Imaging Radiometer Suite
Appendix B: Data availability

The compiled data are available at https://doi.org/10.1594/PANGAEA.898188 (Valente et al., 2019). The database is composed of three main tables: table insitudb_chla.csv with the observations of chla_fluor and chla_hplc, table insitudb_rrs.csv with observations of rrs and table insitudb_iopskdtsm.csv with remaining observations (aph, adg, bbp, kd and tsm). The rows within the three tables relate to each other via a unique key (column idx). The three tables can be viewed conceptually as one table with all data. To help with data manipulation, six auxiliary tables derived from the previous three main tables are provided. The table insitudb_metadata.csv contains all available metadata and helps, for example, to find rows (i.e. idx) with multiple variables (e.g. rrs and chla_fluor). The table auxiliary_table_contributors.csv contains the number of observations per data contributor, variable and dataset. The remaining four tables (insitudb_rrs_satbands2.csv, insitudb_rrs_satbands6.csv, insitudb_iopskdtsm_satbands2.csv and insitudb_iopskdtsm_satbands6.csv) contain the spectral data of the main tables (i.e. insitudb_rrs.csv and insitudb_iopskdtsm.csv) aggregated within ±2 and ±6 nm, respectively, of SeaWiFS, MODIS AQUA, MERIS, VIIRS and OLCI sensor bands. The tables are generated by assigning, in each row of the main tables (i.e. insitudb_rrs.csv and insitudb_iopskdtsm.csv), the closest spectral observation within 2 nm (or 6 nm) of a sensor band. The centre wavelengths of each band and sensor used in the generation of the files are the following: SeaWiFS bands 1–8 were centred at [412, 443, 490, 510, 555, 670, 765, 865] nm, respectively; MODIS-AQUA bands 1–9 were centred at [412, 443, 488, 531, 547, 667, 678, 748, 869] nm, respectively; MERIS bands 1–13 were centred at [412, 442, 490, 510, 560, 620, 665, 681, 709, 753, 779, 865, 885] nm, respectively; VIIRS bands 1–5 were centred at [410, 443, 486, 551, 671] nm, respectively; OLCI bands 1–7 were centred at [412, 442, 490, 510, 560, 620, 665] nm. An exception to this procedure was made to confirm that the correct MOBY data are stored in the files (see Sect. 2.2.1. for discussion on how MOBY wavelengths are stored in the main file). Finally, a readme file is provided to help the user. Table B1 shows how the compiled data look. The example of a query for available chlorophyll data from subdataset seabass_car81 is given.

Table B1. Example of how the compiled data look. The result if the compilation is queried for the chlorophyll data from subdataset seabass_car81 is given.

<table>
<thead>
<tr>
<th>idx</th>
<th>time</th>
<th>lat</th>
<th>long</th>
<th>chla_fluor</th>
<th>chla_fluor_dataset</th>
<th>chla_fluor_subdataset</th>
<th>chla_fluor_contributor</th>
</tr>
</thead>
<tbody>
<tr>
<td>30266</td>
<td>2002-08-06T09:02:00Z</td>
<td>10.5</td>
<td>-64.67</td>
<td>0.185</td>
<td>seabass</td>
<td>seabass_car81</td>
<td>Frank Muller-Karger</td>
</tr>
</tbody>
</table>
Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/essd-11-1037-2019-supplement.

Author contributions. AV compiled the database, carried out the integration and quality checking, and drafted the manuscript. The first six authors are part of the ESA OC-CCI team and contributed to the design of the compilation and to the quality checking, as well as contributing data. The remaining authors are listed alphabetically and are data contributors (see their respective dataset in Table 2) or individuals responsible for the development of a particular dataset (e.g. JW for NOMAD and KB for MERMAID). All data contributors (listed in Table 2) were contacted for authorization of data publishing and offered co-authorship. In the case of the ICES dataset the permission for publishing was given by the ICES team. All the authors have critically reviewed the manuscript. MW and TM passed away before submission. We regard their approval of this work as implicit.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This paper is a contribution to the ESA OC-CCI project. This work is also a contribution to project PEstOE/MAR/UI0199/2014. We would like to thank the efforts of the teams responsible for collection of the data in the field and of the teams responsible for processing and storing the data in archives, without which this work would not be possible. We thank Tâmoghna Acharyya and Robert Brewin at Plymouth Marine Laboratory for their initial contribution to this work. We thank the NOAA (US) for making available the MOBY data and Yong Sung Kim for the help in questions about MOBY data. BOUSSOLE is supported and funded by the European Space Agency (ESA), the Centre National de la Recherche Scientifique (CNRS), the Institut National des Sciences de l’Univers (INSU), the Sorbonne Université (SU) and the Institut de la Mer de Villefranche (IMEV). We thank ACRI-ST, ARGANS and ESA for access to the MERMAID database (http://hermes.acri.fr/mermaid, last access: 10 July 2019). We thank Annelies Hommersom, Pierre Yves Deschamps, Gavin Tilstone and David Siegel for allowing the use of MERMAID data for which they are principal investigators. We thank the British Oceanographic Data Centre (BODC) for access to AMT data and in particular to Polly Hadziabdic and Rob Thomas for their help in questions about the AMT dataset. We thank Victoria Hill, Patrick Holligan, Gerald Moore and Emilio Suarez for the use of AMT data for which they are principal investigators. We thank Sam Ahmed, Hui Feng, Alex Gilerson and Brent Holben for allowing the use of the AERONET-OC data for which they are principal investigators. We thank also the AERONET staff and site support people. The Australian Integrated Marine Observing System (IMOS) and CSIRO are acknowledged for funding the Lucinda AERONET-OC site. We thank Bob Bidigare, Matthew Church, Ricardo Letelier and Jasmine Nahorniak for making the HOT data available, as well as the National Science Foundation for support of the HOT research (grant OCE 09-26766). We thank Yves Dandonneau for allowing the use of GeP&CO data. We thank the ICES database on the marine environment (Copenhagen, Denmark, 2014) for allowing the use of their archived data and Marilynn Sørensen for the help with questions about the ICES dataset. We thank all ICES contributors for their data. We thank Eric Zettler and Sea Education Association. The CARIACO Ocean Time-Series Program also provided significant decade-long bio-optical information used in this study. These data were obtained from NOMAD and SeaBASS. We thank NASA, SeaBASS and the Ocean Biology Processing Group (OBPG) for access to SeaBASS and NOMAD data. We thank NASA for project funding for data collection. We thank Chris Proctor from SeaBASS for his valuable and prompt help in a variety of questions. We are deeply thankful to the data contributors of NOMAD and SeaBASS: Kevin Arrigo, Mike Behrenfeld, Emmanuel Boss, Chris Brown, Mary Luz Canon, Douglas Capone, Ken Carder, Alex Chekalyuk, Jay-Chung Chen, Dennis Clark, Jorge Corredor, Glenn Cota, Yves Dandonneau, Heidi Dierssen, David Esslinger, Pietr Flatau, Alex Gilerson, Joaquim Goes, Gwo-Ching Gong, Adriana Gonzalez-Silvera, Larry Harding, Jon Hare, Chunmin Hu, Sung-Ho Kang, Gary Kirkpatrick, Oleg Kopelевич, Sam Laney, Pierre Larouche, Zhongping Lee, Ricardo Letelier, Marlon Lewis, Steven Lohrenz, Antonio Mannino, John Marra, Chuck McClain, Christophe Menkes, Mark Miller, Ru Morison, James Mueller, Ruben Negri, James Nelson, Norman Nelson, Mary Jane Perry, David Phinney, John Porter, Collin Roesler, David Siegel, Mike Sieracki, Jeffrey Smart, Raymond Smith, James Spinohirne, Dariusz Stramski, Rick Stumpf, Ajit Subramaniam, Chuck Trees, Ronald Zaneveld, Eric Zettler and Richard Zimmerman. For the BIOCHEM data we thank the Fisheries and Oceans Canada and the following data contributors: Diane Archambault, Hughes Benoit, Esther Bonneau, Eugene Colbourne, Alain Gagne, Yves Gagnon, Tom Hurlbut, Catherine Johnson, Pierre Joly, Maurice Levassuer, Jean-Francois Lussier, Sonia Michaud, Patrick Ouellet, Jacques Plourde, Stephane Plourde, Luc Savoie, Michael Scarratt, Philippe Schwab, Michael Sturr and Francois Villeneuve. We also thank Laure Devine for the help in processing the BIOCHEM dataset. We thank Ralph Goericker for allowing the use of the CalCOFI and CCELETER data. CalCOFI research is supported by contributions from the participating agencies: the California State Department of Fish and Wildlife, NOAA, National Marine Fisheries Service, Southwest Fisheries Science Center, and the University of California, Integrative Oceanography Division at the Scripps Institution of Oceanography, UCSD. The authors would like to thank the Oceanic Platform of the Canary Islands (PLOCAN) and its staff for making freely available the use of this ESTOC dataset. We thank the following MAREDAT data providers: Robert Bidigare, Denise Cummings, Giacomino DiTullio, Chris Gallienne, Ralf Goericker, Patrick Holligan, David Karl, Michael Landry, Michael Lomas, Michael Lucas, Jean-Claude Marty, Walker Smith, Rick Stumpf, Emilio Suarez, Koji Suzuki, Maria Vernet and Simon Wright. We thank Oscar Schofield, Raymond Smith and Maria Vernet for allowing the use of the PALMER data. Data from the Palmer LTER data repository were supported by the Office of Polar Programs, NSF grants OPP-9011927, OPP-9632763 and OPP-0217282. We thank the SeaDataNet-Pan-European infrastructure for ocean and marine data management (http://www.seadatnet.org, last access: 10 July 2019). We thank Emmanuel Boss for the TARA data. Funding for the collection and processing of the TARA dataset was provided by the NASA Ocean Biology and Biogeochemistry programme un-
der grants NNX11A14G, NNX09AU43G, NNX13AE58G and NNX15AC08G to the University of Maine. Vanda Brotas received a sabbatical grant from FCT SFRH/BSAB/142981/201. We would like to honour the memory of Marcel Wernand and Tiffany Moisan, authors who contributed to the first version.

Financial support. This research has been supported by the ESA Climate Change Initiative – Ocean Colour project (ref: AO-1/6207/09/I-LG).

Review statement. This paper was edited by David Carlson and reviewed by two anonymous referees.

References


Schofield, O., Vernet, M., and Prezelin, B.: Photosynthetic pigments of water column samples and analyzed with High Performance Liquid Chromatography (HPLC), collected aboard Palmer LTER annual cruises off the coast of the Western Antarctica Peninsula, 1991–2016. Environmental Data Initiative, https://doi.org/10.6073/pasta/4d5837113667a0f52b92d9237a26dd0d82e, 2018c.


