Comparing satellite and BGC-Argo chlorophyll estimation: a phenological study

Alberto Baudena¹, Vincent Taillandier², Nicolas Mayot¹, Alexandre Mignot³, Wilhem Riom¹, and Fabrizio D’Ortenzio⁴

¹Laboratoire d’Océanographie de Villefranche sur Mer
²Sorbonne Université, CNRS, Laboratoire d’Océanographie de Villefranche
³Mercator Ocean International
⁴Sorbonne Universités, UPMC Univ. Paris 06, and CNRS UMR 7093, LOV, Observatoire océanologique

June 10, 2024

Abstract

Ocean primary production is a key process that regulates marine ecosystems and the global climate, but its estimation is still affected by multiple uncertainties. Typically, the chlorophyll-a concentration (CHL) is used to characterise this process, as it is considered a proxy of phytoplankton biomass. To date, the most common observing systems for studying CHL are ocean colour satellites and BGC-Argo floats. Those are complementary systems: satellite observations provide global coverage but are limited to the ocean surface, while BGC-Argo floats provide punctual observations along the whole water column. Comparison of these two observing systems has been performed only at regional or single-float scales, while at global scale this results in large uncertainties due to the relatively low and irregular BGC-Argo coverage. Here, we propose a different method, by comparing satellite and BGC-Argo climatological annual time series within seven different bioregions, each characterised by a homogeneous phytoplankton phenology, allowing us to smooth the uncertainties. By comparing the mean values, the amplitudes, and the shapes of the two time series, we are able to identify regions (a) where they agree (58-61% of the ocean surface area); (b) where the BGC-Argo float network should be extended (generally regions with less than 5 profiles each 100x100 km² square); (c) where the discrepancy is likely due to satellite or (d) BGC-Argo performance. Use of either BGC-Argo and satellite data in regions b—d should be carried carefully and we provide, for each region, suggestions on which system could be affected by the largest uncertainties.
Comparing satellite and BGC-Argo chlorophyll estimation: a phenological study

Alberto Baudena1,2,*, Vincent Taillandier1, Nicolas Mayot1, Alexandre Mignot2, Wilhem Riom1, and Fabrizio D’Ortenzio1

1Sorbonne Université, CNRS, Laboratoire d’Océanographie de Villefranche, UMR 7093 LOV, Villefranche-sur-Mer, France
2Consiglio Nazionale Delle Ricerche- Istituto di Scienze Marine (CNR-ISMAR), Lerici (SP), Italy
3Mercator Ocean International, 31400 Toulouse, France
*To whom correspondence should be addressed; E-mail: alberto.baudena@gmail.com.

Abstract

Ocean primary production is a key process that regulates marine ecosystems and the global climate, but its estimation is still affected by multiple uncertainties. Typically, the chlorophyll-a concentration (CHL) is used to characterise this process, as it is considered a proxy of phytoplankton biomass. To date, the most common observing systems for studying CHL are ocean colour satellites and BGC-Argo floats. Those are complementary systems: satellite observations provide global coverage but are limited to the ocean surface, while BGC-Argo floats provide punctual observations along the whole water column. Comparison of these two observing systems has been performed only at regional or single-float scales, while at global scale this results in large uncertainties due to the relatively low and irregular BGC-Argo coverage. Here, we propose a different method, by comparing satellite and BGC-Argo climatological annual time series within seven different bioregions, each characterised by a homogeneous phytoplankton phenology, allowing us to smooth the uncertainties. By comparing the mean values, the amplitudes, and the shapes of the two time series, we are able to identify regions (a) where they agree (58-61% of the ocean surface area); (b) where the BGC-Argo float network should be extended (generally regions with less than 5 profiles each 100x100 km² square); (c) where the discrepancy is likely due to satellite or (d) BGC-Argo performance. Use of either BGC-Argo and satellite data in regions b–d should be carried carefully and we provide, for each region, suggestions on which system could be affected by the largest uncertainties.

Plain language summary

The capacity of oceans to produce photosynthetic activity and hence participating in global primary productivity is mainly assessed through the measurement of chlorophyll (CHL). This pigment, responsible for photosynthetic process, is mainly measured by two observing systems to date: satellite and biogeochemical
Argo drifters. While these systems are complementary (satellite have broadly coverage limited at the surface, BGC-Argo drifters punctual observations along the whole water column), comparisons of their output only exist at regional or single-drifter levels. Here, rather than comparing them punctually (which introduce extremely large noise levels), we group the observations according to 7 bioregions, smoothing the uncertainties. Each bioregion is characterised by a homogeneous phenology, i.e. a typical CHL time series along the year. By comparing the time series observed by satellite with the one obtained from BGC-Argo floats, we show that these two observing systems agree in most of the global ocean surface area (58-61%). We identify some of the reasons explaining the discrepancy found in the other areas, in some cases due to BGC-Argo drifters observing capacities, in others due to satellite ones. Finally, we show where are the regions where the BGC-Argo sampling density should be increased.

1 Introduction.

Ocean primary production is considered to contribute to 50% of the global net primary production (Field et al., 1998). One of the main sources of error in this estimation is the extreme spatial and temporal variability of the global ocean distribution of phytoplankton. This is required to evaluate and predict the consequent primary production and is hardly characterised and quantified.

To date, mapping the global ocean phytoplankton distribution is mainly assessed through the systematic observation of the chlorophyll-a concentration (CHL), which is the main algal pigment responsible for the photosynthetic process in phytoplankton cells. CHL is generally considered a good, although not perfect, proxy of phytoplankton biomass concentration in the ocean (see for example discussion in Huot et al. 2007). Therefore, estimates of ocean primary productivity strongly relies on the occurrence of CHL measurements.

CHL is estimated by several methods. The most efficient in terms of spatial and temporal resolution is the satellite ocean colour. Since 1997, systematic and gap free databases of remote sensing CHL, at kilometric resolution, have been obtained thanks to an uninterrupted series of ocean colour missions: SeaWiFS, MODIS, MERIS, VIIRS, and OLCI (McClain et al., 2022). These observations are massively used to identify global or regional patterns in the CHL distribution. In addition, they are used to characterise specific phytoplankton dynamics as well as seasonal, frontal, or mesoscale blooms (see a recent review of the ocean colour application in Sathyendranath et al. 2023). Ocean colour CHL is also the primary data source for validating phytoplankton concentration simulated by global ocean biogeochemistry models (Gregg, 2008; Pradhan et al., 2019) and, more recently, it has been successfully assimilated in operational models (e.g. in the global Copernicus marine service biogeochemical forecasting system: Mignot et al. 2023). However, ocean colour satellites are unable to sense the ocean surface in presence of clouds or strong winds, resulting in large portions of unobserved ocean areas (Stock et al., 2020). Products generated by using observations obtained by different satellite sensors, could partially palliate this problem by increasing the spatial and temporal coverage of CHL satellite maps, although they could introduce additional errors (Garnesson et al. 2019).

Furthermore, although uncertainty of global satellite CHL retrieval is estimated
to range between 16.5% and 30% (Hu et al., 2012) regional characteristics could strongly degrade space derived estimations (e.g. Chen et al. 2021).

Another diffused method to estimate CHL is through fluorescence measurements: the excitation of a water parcel containing algal pigments with blue/green light generates an emission of light in the red, which is proportional to CHL. Fluorimeters (i.e. sensors measuring fluorescence) are routinely used to determine in situ CHL, in particular from ship, providing then a method to obtain depth profiles of CHL (see for example Petit et al. 2022 and references therein). Fluorimeters are usually calibrated by high performance liquid chromatography (HPLC) or spectrofluorometers. To date, fluorescence-based CHL estimations are undoubtedly the most large source of in situ CHL observations. More recently, the number of available fluorescence-based CHL profiles has dramatically increased thanks to the massive deployment of Argo profiling floats equipped with fluorometers (BGC-Argo, Claustre et al. 2020). A BGC-Argo float is a robotic drifter, equipped with physical and biogeochemical sensors, able to autonomously sample the first 2 kms of the water column and to transmit in real-time the acquired data. To date, the BGC-Argo floats have acquired more than 200k CHL profiles ((Stoer et al., 2023)) at locations and seasons generally undersampled by traditional research vessels. However, the calibration and validation of CHL data from BGC-Argo floats is not straightforward (see Roesler et al. 2017; Schmechtig et al. 2023; Lavigne et al. 2012), mainly because the HPLC data obtained simultaneously to BGC-Argo profile are mainly limited at the float deployment or at the recovery, (Johnson et al., 2017; Taillandier et al., 2018; D’Ortenzio et al., 2020). In addition, the BGC-Argo network has a limited spatial coverage, and is still under development (Group, 2016).

Nonetheless, by providing CHL at depth, under the clouds and at high latitudes (i.e. where and when the ocean colour satellites cannot provide data), autonomously and in real time, the BGC-Argo network is the ideal observing system to complement ocean colour satellites. The combined use of the two observing systems offers extraordinary opportunities to characterise the spatio-temporal distribution of CHL at global scale. Several studies have focused on the combination of the two data sets to validate and assimilate ocean biogeochemistry models (e.g. Mignot et al. 2023; Cossarini et al. 2019; Teruzzi et al. 2021). Other studies have compared CHL fields obtained by few floats missions (i.e. Chiswell et al. 2022; Boss et al. 2008; D’Ortenzio et al. 2021) or focusing on regional or sub-basin scale (i.e. D’Ortenzio et al. 2020; Haenjens et al. 2017; Wojtasiewicz et al. 2018). To date, however, very few studies have compared the two data systems with a strict matchup approach at global scale as generally required for satellite calibration/validation exercise (Werdell and Bailey, 2005).

Independently to the intrinsic uncertainties related to the two observing systems, a matchup analysis performed at global scale is not trivial. The high demanding criteria required to perform calibration and validation exercises for ocean colour (as described in Werdell and Bailey 2005) are often too severe to obtain a sufficient number of matchups. These criteria are then weakened, for example, by enlarging the time window (from ±3 hours to ±12 hours, as in Wojtasiewicz et al. 2018). In addition, the irregular distribution of the BGC-Argo profiles, with regions of strong density and others completely unobserved by floats, could generate bias in the matchup analysis, pushing toward regional approaches (as in D’Ortenzio et al. 2021).

In this context, we propose an alternative method to compare these two datasets.
Rather than comparing BGC-Argo and satellite CHL at the locations where BGC-Argo profiles are obtained, we compare their climatological annual time series within defined regions. CHL time series represent a pivotal information to characterise the phenology of phytoplankton, i.e., the identification of key moments in the phytoplankton annual cycle, such as bloom onset or decay. Studying phytoplankton phenology is considered crucial within the framework of climate change: a modification in the timing of the main phytoplankton dynamics, induced by alterations in its environment, could have significant repercussions on the entire marine food web (Ji et al., 2010). In this sense, by using phenology as a comparison metric, we explore how the two datasets perform in giving a global picture of the CHL distribution and dynamics, and how they could simultaneously be used to compensate for their intrinsic limitations (see also Gittings et al. 2019).

Comparing time series obtained from the two observing systems, however, is not completely trivial. Although for each location (i.e. a pixel) a CHL time series could be generated from a satellite data set, this is not the case for BGC-Argo floats, which could change positions at each time step. In such a sense, satellite time series could be considered as an Eulerian observation, when BGC-Argo times series are close to a Lagrangian data frame. To overcome this issue, we will use here a method to aggregate multiple individual (i.e. satellite and BGC-Argo) time series, within large ocean regions. For this, we use the concept of “bioregion”, which could be defined as an oceanic region characterised by homogeneous biogeochemical characteristics. In our case, the bioregions are defined as oceanic regions sharing similar phenological characteristics and, more precisely, temporal series with similar shapes. This approach has already been applied to ocean colour time series and used to identify phenological patterns in the Mediterranean (D’Ortenzio and Ribera d’Alcalà, 2009; Mayot et al., 2016), in the Red Sea (Kheireddine et al., 2021), in the Labrador Sea (Marchese et al., 2019), and in the Southern Ocean (Ardyna et al., 2017). It was also used to compare two different ocean colour data sets obtained at 20 years of interval (D’Ortenzio et al., 2012). Finally, and more importantly for the present study, maps of satellite-derived phenological bioregions have been used as masks to aggregate sparse in situ data and to generate CHL time series comparable to those obtained from satellite (Lacour et al., 2015; Mayot et al., 2017; D’Ortenzio et al., 2021).

Here, we applied this approach with the specific aim of comparing the capability of ocean colour satellites and BGC-Argo observing systems to evaluate CHL seasonality at global scale. The advantage of this approach is that it allows us (i) to identify ocean regions that share the same phytoplankton phenology in a robust way; (ii) to smooth the uncertainties, by regrouping and averaging satellite and BGC-Argo data within each cluster (iii) to conduct the comparison in phenologically homogeneous regions, avoiding potential biases.

2 Materials and Methods

This paper is based on the analysis of temporal series of CHL computed in a climatological way. The key concept is that, for a given location, we build a time series which reveals how the chlorophyll changes, on average, along a typical year. This section is organised as follows.
In Subsecs. 2.1–2.5, we expose the calculation of time series of satellite CHL covering the global ocean (referred to as a climatology), and how they are clustered using the phenological procedure of D’Ortenzio and Ribera d’Alcalà (2009).
In Subsec. 2.6, we describe the processing of BGC-Argo data and the calculation of BGC-Argo-derived seasonal CHL time series specific to each cluster.
In Subsec. 2.7, we expose how satellite and BGC-Argo time series are statistically compared within each cluster.

2.1 Satellite data
The surface CHL satellite data set used is the OCEANCOLOUR_GLO_BGC_L3_MY_009_103 product which is provided by Copernicus Marine Environment Monitoring Service (CMEMS, http://marine.copernicus.eu/) platform (product version cmems_obs-oce_glo_bgc-plankton_my_l3-multi-4km_P1D). This L3 product has 4 km spatial resolution, is provided daily from 1998 to present day and merges observations from SeaWiFS, MODIS, MERIS, VIIRS-SNPP & JPSS1, OLCI-S3A & S3B missions. Consequently, the accuracy and coverage of the satellite CHL used here are dependent on the number and quality of the in orbit sensors (increasing in the last years when the number of simultaneous sensors was higher).

2.2 Climatology calculation
The clusterization method used requires, for each satellite pixel, a complete time series (i.e. without temporal gaps). Observations missing due e.g. to clouds or algorithmic errors are solved by interpolation (Subsec. 2.2.1). However, polar regions require a specific processing, mainly because they cannot be observed during winter months. This reduces the number of observations available and results in relatively long temporal gaps, which need to be interpolated, introducing then a high level of noise. Although noise is not an intrinsic problem, the application of the partition method to such time series prevents a rapid convergence of the clusterization and the identification of an optimal number of clusters, resulting in a numerical failure of the method.
To overcome this issue, a possibility would be to exclude polar regions from our analyses. However, a large number of BGC-Argo profiles (used to compare with satellite data) are in polar regions. In addition, poles are crucial locations regulating the global primary productivity and being importantly affected by climate change (Post et al., 2019). First, we calculate a climatology (referred to as reference climatology) with very strict criteria (please see below the details). This allows us to obtain less noisy time series, and permits the partition method to converge and identify an optimal number of clusters (and reference time series). In this climatology, locations without enough data coverage are excluded, and only the last 5 years of data (2018-2022), those with larger accuracy are used. Other criteria are applied in order to reduce the noise in the time series (see below the details). The reference climatology covers the 57°S-57°N band and does not include coastal regions as they are affected by larger CHL uncertainties (Van Oostende et al., 2018; Zeng et al., 2023; McCluskey et al., 2022).
To extend our coverage to polar zones and coastal regions, the second step consists into the calculation of a second climatology, which we refer to as extended climatology. This is based on less strict criteria and on the use of the entirety
of the chlorophyll observations over the last 25 years. Given the larger noise
and uncertainty, the partition method fails to converge and identify an optimal
number of clusters when directly applied to the extended climatology. For this
reason, we use the reference time series, identified with the reference climatology,
to partition the extended climatology. We associate, to each time series of the
extended climatology, one of the reference time series, based on their similarity
(Euclidean distance, details below). In this way, we are able to also include
the polar and coastal regions in our partition. We then verify the resemblance
of these two partitions (one obtained from the reference climatology, the other
from the extended one) to assure they are consistent (Supplementary Table S.1
and Supplementary Fig. S.4).
Details about the calculation of these climatologies are reported in the following
subsections.

2.2.1 Reference climatology

A first climatology, named reference climatology, is calculated using all daily
CHL fields between 2018 and 2022. The use of this time window is due to
the fact that the quality of CHL data is better compared to previous periods,
following the number and quality of in orbit sensors. We use a procedure
already used in previous works (D’Ortenzio and Ribera d’Alcalà, 2009; Marchese
et al., 2019; Mayot et al., 2016; Kheireddine et al., 2021; Ardyna et al., 2017;
D’Ortenzio et al., 2012), readapting it to study the global ocean. We proceed
as follows.

1. We consider the chlorophyll field of the 1st January 2018. Then, we ex-
clude all pixels within 1° (~111 km) of the coast. This is because uncer-
tainty of CHL data in coastal regions is generally larger than in the open
ocean (Van Oostende et al., 2018; Zeng et al., 2023; McCluskey et al.,
2022). We also exclude pixels in the Black and Caspian Seas, and in the
Hudson Bay, as they are enclosed or quasi-enclosed basins.

2. In order to smooth the uncertainties, we reduce the spatial resolution of
the chlorophyll field of the 1st January 2018 from the nominal resolution
(0.04°) to 1°. For this purpose, we build a grid of 1° resolution covering the
global ocean. For each grid cell, we take the median of all the chlorophyll
values within the grid cell (~625 values per cell). If, in that grid cell, less
than α=25% of the values are available (due e.g. to cloud coverage), we
flag that cell as not available (NA).

3. The temporal resolution of the time series is eight days. Thus, step 1
and 2 are repeated for all the days between the 1st and the 8th January
2018. Hence, for a given grid cell, we obtain 8 chlorophyll values. We
consider the median of these values as representative of the chlorophyll
concentration in that grid cell for the 1-8 January 2018 period. Values
flagged as NA are excluded from the median. If all eight values in a grid
cell over an eight-day period are marked as NA, the CHL in that grid cell
(for the 1-8 January 2018 period) is marked as NA. This is repeated for
all the grid cells.

4. Step 3 is repeated for all years between 2018 and 2022, obtaining 5 values
for each grid cell. We consider the median of these 5 values as representa-
Table 1: List of parameters (rows) used for the calculation of each climatology (columns).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Reference climatology</th>
<th>Extended climatology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time period</td>
<td>2018—2022</td>
<td>1998—2022</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>50%</td>
<td>25%</td>
</tr>
<tr>
<td>$\beta$</td>
<td>60%</td>
<td>20%</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>75%</td>
<td>50%</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>4 weeks</td>
<td>14 weeks</td>
</tr>
</tbody>
</table>

5. For each grid cell, if the CHL climatological time series has less than $\gamma=25\%$ of the available values, we exclude that grid cell from the analysis. Similarly, if the CHL climatological time series shows temporal gaps of more than $\epsilon=4$ weeks, we discard that grid cell. Otherwise, we remove any existing temporal gaps using a monotonic splice interpolation (pchip_interpolate function in the python package scipy). This type of interpolation is chosen because it performs a non-linear interpolation and, at the same time, does not create artificial peaks in the time series.

6. Each climatological CHL time series is smoothed using a running mean over 40 days (i.e. 5 time steps).

7. We shift the climatological CHL time series located in the Southern hemisphere of 6 months, so that Austral and Boreal summers (winter) occur in June-August (December-February)

8. Finally, we normalise each climatological CHL time series by dividing it by its annual climatological maximum value. Thus, each climatological CHL time series has a maximum value equal to 1. We refer to these as normalised time series and we indicate the normalised chlorophyll as $nCHL$. We distinguish them from the non normalised time series that have real CHL concentrations, and which are used for the comparison with real CHL values measured by BGC-Argo floats, as detailed in the subsections below. To further distinguish them from the BGC-Argo time series, we refer to them also as satellite time series.

2.2.2 Extended climatology

A second climatology, named extended climatology, is calculated with the same exact procedure but with three differences. First, we include pixels close to
the shore. Second, we adopt different parameters (Table 1) which reduce the number of values flagged as \( NA \), allowing us to extend the global coverage of the climatology to higher latitudes. Third, we use CHL data between 1998 and 2022.

### 2.3 Partition method for the reference climatology

The time series of the reference climatology are partitioned using the \( k \)-means method (Hartigan and Wong, 1979; Ahmed et al., 2020). This method is chosen as it has already been successfully applied to a variety of case studies including clustering of temporal series of satellite CHL (D’Ortenzio and Ribera d’Alcalà, 2009; Marchese et al., 2019; Magot et al., 2016; Kheireddine et al., 2021; Ardyna et al., 2017; D’Ortenzio et al., 2012). The \( k \)-means is an unsupervised machine learning technique which groups together time series so that the distance among them is minimum, while maximising the distance between different groups. As distance, the Euclidean distance is used. Given two time series of \( n \) time steps \( t_1 = \{a_1, a_2, ..., a_n\} \), and \( t_2 = \{b_1, b_2, ..., b_n\} \), the Euclidean distance between \( t_1 \) and \( t_2 \) is:

\[
\sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}
\]

The \( k \)-means algorithm needs one parameter, namely the number of clusters \( k \) in which the time series should be regrouped. At the end of \( k \)-means process, each time series of the reference climatology is assigned to one of the \( k \) clusters, meaning that each pixel of the climatology is assigned to the obtained cluster. This gives a geographical distribution of the memberships which we call hereafter bioregion following D’Ortenzio and Ribera d’Alcalà (2009).

In summary, for each of the \( k \) clusters obtained from the \( k \)-means partition, we obtain a spatial distribution (that we assume and call bioregion sensu D’Ortenzio and Ribera d’Alcalà 2009) and a reference time series (see below Subsec. 2.5 for definition).

### 2.4 Identification of the optimal number of clusters

Different tests (Elbow, Silhouette, Calinski-Harabasz) are used to determine the optimal number of clusters of the reference climatology (D’Ortenzio and Ribera d’Alcalà 2009; Supplementary Material S.2 and S.2.1). We identify the number of optimal clusters applying these techniques to the reference climatology (rather than to the extended one) due to the higher accuracy of this climatology. Indeed, the stricter criteria used to calculate it (compared to the criteria used for the other two climatologies, Table 1) make its time series less noisy and more robust.

The results of the tests vary between 5 and 9, although all (except Calinski-Harabasz which indicates 3 clusters and was discarded as its partition resulted too broad for the purposes of this paper, Supplementary Fig. S.5) converged on 7 and 8. We finally decided on 7 after visual inspection of the time series, which suggests that two of the 8 times series, although considered statistically...
different, show very close seasonal traits (see Supplementary Material S.2.1 and Supplementary Fig. S.8).

We also stress that this number is consistent (i) with the study of D’Ortenzio and Ribera d’Alcalà (2009, 7 clusters), Mignot et al. (2023, 8 clusters), which was based on a k-means partition as well, and with the global study of Bock et al. (2022, 6 clusters); (ii) when changing the resolution of the grid over which the climatology is calculated (Supplementary Fig. S.9); (iii) with the result obtained without shifting time series in the Southern hemisphere of 6 months (Supplementary Fig. S.10).

2.5 Partition method for the extended climatology

Here we describe the methodology used to partition the extended climatology. We do not use the k-means method (used with the reference climatology) as it does not allow us to identify an optimal number of clusters when directly applied to the extended climatology. This is because of the presence, close to the shore and at high latitudes, of time series with limited data quality or gaps. The method used to partition the extended climatology employs the 7 reference time series together with their relative thresholds of acceptance. These are calculated as follows. First, we consider all the time series belonging to bioregion 1. The average of these time series is considered as representative of the reference time series of bioregion 1. We then calculate, for all the time series belonging to bioregion 1, the Euclidean distance with the reference time series, obtaining values. The threshold of acceptance of bioregion 1 is defined as the maximum of the distances. This is repeated for all the bioregions, resulting in 7 reference time series and corresponding thresholds of acceptance.

For each time series of the extended climatology we calculate the Euclidean distance with the 7 reference time series, obtaining values. The minimum of these values indicates the final classification of the time series of the extended climatology, providing that this minimum is lower than the corresponding threshold of acceptance. Otherwise, no bioregion is assigned (only less than 0.3% of the time series of the extended climatology; Supplementary Table S.1).

Overall, the percentage of surface covered by each bioregion does not change consistently (Supplementary Table S.1). Then, since the extended climatology covers a domain which is larger than the one covered by the reference climatology, we only consider the grid cells in which both climatologies are defined (i.e. in which both time series are not flagged as NA). We find that 72.9% of the pixels of the extended climatology have been assigned to the same bioregion of the reference climatology (Supplementary Fig. S.4a and c). The difference is explained by the use of a different number of years to calculate the two climatologies (2018-2022 period for the reference climatology, 1998-2022 period for the extended climatology; Supplementary Material S.3 and Supplementary Fig. S.4a and b), and is not due to the partition method. In addition, we calculate the Euclidean distance between the time series of the extended climatology and the reference time series of the bioregion it should belong to (which can differ from the cluster it has been assigned to). We find that 97.3% of the time series of the extended climatology are within the threshold of acceptance of the corresponding reference time series. For each bioregion obtained using the extended climatology, we calculated its mean time series : they are all significantly similar to the corresponding reference time series (Supplementary Fig. S.6). When the
silhouette value for each bioregion of the extended climatology is calculated, positive values are obtained for almost all the time series (Supplementary Fig. S.14). Overall, these results indicate that the method used to partition the extended climatology does not affect the pattern of the bioregions significantly, and corroborate our choice of extending our analyses to the coastal and polar regions.

2.6 BGC-Argo data

2.6.1 BGC-Argo data processing

BGC-Argo data were collected and made freely available by the International Argo Program and the national programs that contribute to it (https://argo.ucsd.edu, https://www.ocean-ops.org; Argo (2020)). The Argo Program is part of the Global Ocean Observing System. BGC-Argo are autonomous drifters which sample the water column and measure state variables including temperature, pressure, and salinity. In addition, different sensors allow the drifter to measure biogeochemical variables such as CHL, oxygen, particulate backscattering (b\textsubscript{bp}), nitrates, pH, and downwelling irradiance. BGC-Argo data are subjected to several quality control procedures (Schmechtig et al., 2023, 2015; Thierry et al., 2022; Johnson et al., 2023). First, raw data are converted to “real time” state variables and outliers are identified and flagged. Second, automated algorithms calibrate real time data, providing “adjusted” data. Finally, adjusted data are verified and validated by an expert, providing “delayed” mode data. As the latter data have only a limited coverage, we considered data both in “real time”, “adjusted”, and “delayed” mode, following the methodological approach of Mignot et al. (2023). However, we exclude from each profile all the values classified as “probably bad data”, “bad data”, or “missing value” (~30% of the profiles were integrally excluded in this way).

We only select BGC-Argo floats which measured CHL, excluding CHL values larger than 40 mg/m\textsuperscript{3} from that profile, obtaining a total of 84186 profiles across the global ocean. The use of “real time” data allowed us to augment the spatial and temporal coverage, although the quality of the CHL evaluation will be certainly impacted. In any case, these data only represent ~11% of the profiles for CHL (Stoer et al., 2023).

For each BGC-Argo profile, we calculate the CHL averaged in the first optical depth, which is usually considered as representative of the CHL measured by satellite observations (Morel and Berthon, 1989). The first optical depth was calculated as in Morel and Berthon (1989).

2.6.2 BGC-Argo time series

In order to compare CHL observations from BGC-Argo with the CHL time series obtained from satellite data, we define a BGC-Argo time series as follows:

- For each bioregion, we select all the BGC-Argo profiles which are collected within that bioregion.
- For each 8-day time step (e.g. 1-8 January, or 9-16 January, etc.), we select all the BGC-Argo profiles measured in correspondence of that time
step and we consider their CHL averaged in the first optical depth. We calculate the median of these values and we repeat this for each time step, generating a final time series of 46 values.

- A running average over a 40 days windows (i.e. 5 time steps), is finally carried out.

Furthermore, we calculate other diagnostics:

- For each time step (e.g. 1-8 January, or 9-16 January, etc.), we count the number of BGC-Argo profiles available, obtaining a time series of 46 values. We then consider the median, minimum, and maximum of this time series (referred to as “median #BGCprof ts”, “min #BGCprof ts”, “max #BGCprof ts” respectively, in Supplementary Table S.3)

- For all the \(N_{GC}\) grid cell of a bioregion, we count the number of BGC-Argo profiles within it, obtaining \(N_{GC}\) values. We then consider the mean, median, and maximum value of this \(N_{GC}\) values (referred to as “# BGC prof/cell (mean)”, “# BGC prof/cell (median)”, “# BGC prof/cell (max)” respectively, in Table 3 and Supplementary Table S.3)

- For each bioregion, we calculate (i) the spatial area, expressed both in terms of pixels, and squares of 100 km size; (ii) the percentage of surface covered by the bioregion, expressed as the ratio between the number of pixels composing it and the total number of oceanic pixels multiplied by 100; (iii) the percentage of BGC-Argo profiles which are in that cluster compared to all the BGC-Argo profiles available (“Percentage of all BGC profiles” in Table 3).

2.7 Comparison of BGC-Argo and satellite time series

The similarity between the mean time series of the cluster and the BGC-Argo time series is tested in two ways:

- We verify that both time series are normally distributed, then we perform a t-test (Supplementary Material S.4; Student 1908)

- We calculate the Euclidean distance between the two time series and we compare it with the threshold of acceptance of the reference cluster. If the Euclidean distance is lower, then we consider the BGC time series as significantly similar to the mean time series of the cluster. This second test is applied only to compare normalised time series. In this case, before applying the test, the BGC time series are normalised by dividing them by their maximum.

3 Results

3.1 Partition of the global ocean

The global ocean is partitioned in \(k=7\) bioregions (Fig. 1). The number of bioregions is chosen using multiple statistical tests (Methods and Supplementary Material S.2 and S.2.1). The pattern of the partition is robust with respect
Figure 1: Regionalisation of the global ocean in $k=7$ bioregions (upper panel) obtained from the 1998-2022 extended climatology. In the lower panels, each plot represents the mean normalised nCHL time series of a specific bioregion (i.e. average of all the time series belonging to that bioregion, thick solid lines) along with their uncertainty (thin solid lines, standard deviation) identified by a different colour (right-hand legend in the upper panel).
Table 2: Number of BGC-Argo profiles in the 7 bioregions along with their mean satellite and BGC-Argo CHL values (and their respective standard deviation), and results of the comparison (t-test, Methods).

<table>
<thead>
<tr>
<th>Bioregion</th>
<th>BGC profiles</th>
<th>Satellite chl, mean [mg/m³]</th>
<th>Satellite chl, std [mg/m³]</th>
<th>BGC-Argo chl, mean [mg/m³]</th>
<th>BGC-Argo chl, std [mg/m³]</th>
<th>t test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13889</td>
<td>0.15</td>
<td>0.28</td>
<td>0.23</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>15550</td>
<td>0.12</td>
<td>0.25</td>
<td>0.13</td>
<td>0.15</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>15758</td>
<td>0.13</td>
<td>0.17</td>
<td>0.16</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5636</td>
<td>0.27</td>
<td>0.33</td>
<td>0.37</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>13187</td>
<td>0.31</td>
<td>0.44</td>
<td>0.32</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>9487</td>
<td>0.46</td>
<td>0.50</td>
<td>0.52</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>15994</td>
<td>0.47</td>
<td>0.50</td>
<td>0.65</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

3.2 Global comparison with BGC-Argo observations

Figure 2 shows the positions of the available BGC-Argo profiles. The CHL time series obtained from the two data sets are plotted separately for each bioregion in the lower panels. Note that satellite data are re-transformed from nCHL to CHL, implying a larger variability (standard deviation 2.7 times larger than in the normalised case). Table 2 also summarises statistics of the different bioregions for the two data sets.

The spatial distribution of BGC-Argo profiles is not homogenous, resulting in a
Figure 2: Comparison of CHL time series from satellite and BGC-Argo. The CHL time series are non-normalised (i.e. real values). The upper panel shows the partition of the global ocean in $k=7$ bioregions as reported in Fig. 1 with the black dots indicating the position of the 84186 BGC profiles used in the present study (Methods). In the lower panels, each plot represents (i) the mean CHL time series of the bioregion from satellite data as coloured solid lines (thicker coloured solid line: average time series; thinner coloured solid line: uncertainty, expressed as standard deviation; colours in the right-hand legend of the upper panel); and (ii) the BGC-Argo time series (black dots: median of chlorophyll of that time step; black solid line: 40-days running average of the black dots, Methods).
number of BGC-Argo profiles per bioregion highly variable (from 5636 profiles in bioregion 4 to 15994 profiles in bioregion 7). Overall, satellite and BGC-Argo mean CHL values are close except for bioregion 1, which exhibits BGC-Argo CHL almost double that of satellite CHL. The shapes of the satellite and BGC-Argo time series are similar for all the bioregions considered (see also the comparison of normalised time series in Supplementary Fig. S.7), with six out of the seven bioregions successfully passing the t-test. Some differences are observed in bioregion 5, which shows a CHL increase observed by BGC-Argo time series in summer but not in satellite data. In bioregion 6, the bloom initiation detected in satellite and BGC-Argo data temporally matches, while the annual maximum is delayed by about one month in BGC-Argo time series. In bioregion 7, the mid-summer increase in CHL is smaller in the satellite time series than in the BGC-Argo data.

3.3 Analysis of variability of satellite time series

In general, the CHL time series have large variabilities (the standard deviation of the time series represents, on average, the 53% of their mean values against 20% for the normalised case; Table 2 and thin solid coloured lines in Fig. 1 and Fig. 2). To deepen the comparison between the satellite and BGC-Argo time series, we proceed to an additional analysis. First, we hypothesised that the high seasonal variability (a large large standard deviation of the time series), is the result of grouping temporal series which are geographically distant. For example, bioregion 5 is observed in Equatorial and polar regions, with seasonal mean values potentially different. Indeed, our partition method, which aims to identify different phenological regimes, is based on the use of nCHL time series. This can result in bioregions that, while sharing the same shape of the nCHL, are located in regions known to have mean values of CHL largely different.

To solve this point, each bioregion is split in a sub-bioregion at high latitudes (HL) by considering only time series obtained in pixels located poleward than 38° latitude, and in a sub-bioregion at low latitudes (LL; points located around the equator lower than 38° latitude). In that way, bioregion 1 is split into sub-bioregion 1 LL and sub-bioregion 1 HL; the same for bioregion 2, 3, etc. Another source of the variability, likely artifactual, derives from the irregular distribution of BGC-Argo profiles, which do not cover the global ocean homogeneously, potentially inducing biases in the satellite BGC-Argo comparison. Therefore, to calculate the mean satellite time series for each sub-bioregion, we consider only pixels located less than 1° from a BGC-Argo profile.

The results confirm that the agreement between satellite and BGC-Argo time series is good, but not valid everywhere (Fig. 3 and Table 3). In particular, 8 out of 14 sub-bioregions show a significant agreement between satellite and BGC-Argo time series (sub-bioregions 1, 3, 4, and 5 at low latitudes, and sub-bioregions 2, 3, 4, and 7 at high latitudes), while 6 do not (sub-bioregions 2, 6, and 7 LL, and sub-bioregions 1, 5, and 6 HL).

Subsequently, we calculate different properties for each LL and HL sub-bioregion (Table 3 and Subsec. 2.6.2). In general, satellite and BGC-Argo time series are significantly different in sub-bioregions where the BGC-Argo sampling density is low (less than 5 BGC-Argo profiles per grid cell; Supplementary Fig. 5.2), while
Figure 3: Bioregions split into high and low latitude sub-bioregions. The upper panel shows the partition of the global ocean in $k=7$ bioregions, as reported in Fig. 1. The white lines show the 38°N and S latitude used to split the bioregions. The 7 lower panels on the left (first two columns) show the low latitude sub-bioregions, while the 7 lower panels on the right (third and fourth columns) show the high latitude sub-bioregions. Each plot represents (i) the mean CHL time series of the sub-bioregion non-normalised calculated considering only grid cells located less than 1° from a BGC-Argo profile, as coloured solid lines (colours in the right-hand legend of the upper panel). (ii) The BGC-Argo time series (grass green dots: median of chlorophyll of that time step; grass green solid line: 40-days running average of the black dots, Methods). The title of each panel indicates the number of the bioregion, the percentage of cells belonging to it (in brackets), and whether the comparison between the satellite and BGC-Argo time series is significant or not (t-test=“yes”, Methods).
they agree where the BGC-Argo sampling density is large (6 BGC-Argo profiles per cell). This is particularly visible in sub-bioregion 6 LL, where a potentially erroneous double peak in the BGC time series can be explained by one of the lowest density of BGC profiles there (5.0). On the other hand, the remarkable agreement found in sub-bioregion 3 HL can be due to the large BGC-Argo sampling density there (34.5, the highest).

When focusing on the sub-bioregions where BGC-Argo and satellite time series do not match, different categories of discrepancy can be qualitatively identified (Table 3). (i) In some sub-bioregions, a strong peak is present in the BGC-Argo time series while, in the satellite time series, the amplitude is much less pronounced. Furthermore, on average, BGC-Argo time series shows values which are almost two-fold larger than satellite time series. It is the case of sub-bioregions 1, 5 HL and, notably, also of the portion of bioregion 1 in the Southern Pacific (Supplementary Material S.6 and Supplementary Fig. S.1). A similar discrepancy is found in sub-bioregion 6 HL with the difference that, there, the satellite time series shows a more visible peak. We define sub-bioregions with these characteristics as belonging to the category “similar shapes, different amplitudes and mean values” (SS, DA&MV). (ii) Conversely, in some regions, BGC-Argo time series show constantly lower values than satellite ones and similar shapes and amplitudes (sub-bioregion 2 LL). We define this category as “similar shapes and amplitudes, different mean values” (SS&A, DMV). Finally, (iii) in some sub-bioregions, the low number of BGC-Argo profiles is not sufficient to produce a reasonable BGC-Argo time series. It is the case of sub-bioregions 6 and 7 LL in which only 0.17% and 1.73% of all BGC profiles were carried, respectively. We define this category as “low BGC-Argo sampling” (LBGCS).

4 Discussion

The obtained regionalisation of the global ocean (Fig. 1) compares well with the regionalisation proposed by Longhurst (1995) and D’Ortenzio and Ribera d’Alcalà (2009, Supplementary Material S.5). In particular, bioregion 1 corresponds to model number 4 of Longhurst (Tropics), while bioregion 2 closely matches Tropical bioregion of D’Ortenzio and Ribera d’Alcalà (2009). Bioregion 3 corresponds to model 3 of Longhurst (Subtropical winter nutrient-limited), while bioregion 4 is similar to the Bloom bioregion of D’Ortenzio and Ribera d’Alcalà (2009). Bioregion 5 shares characteristics of both model 1 (Polar irradiance-limited) and 2 (midlatitude nutrient-limited spring production peak) of Longhurst. Finally, both bioregions 6 and 7 correspond to model number 1 of Longhurst (1995), with bioregion 6 characterised by a stronger decline in CHL. The pattern of our regionalisation compares also well with the one obtained by Fay and McKinley (Supplementary Material S.5), in particular concerning bioregions 2, 3, and 4. Differences are present in the Southern Ocean where the biomes found by Fay and McKinley are organised in latitudinal bands, while the disposition of our bioregions is less regular. A good qualitative agreement is also found between our partition and the one by Bock et al. (2022, Supplementary Material S.5). Larger differences are found with the regionalisation of Mignot et al. (2023), possibly due to the very large number of parameters used in that
<table>
<thead>
<tr>
<th>Bioregion</th>
<th>Satellite chl, mean [mg/m³]</th>
<th>Satellite chl, amplitude [mg/m³]</th>
<th>BGC-Argo chl, mean [mg/m³]</th>
<th>BGC-Argo chl, amplitude [mg/m³]</th>
<th>Surface [10⁶*100 km²]</th>
<th># BGC prof/cell (mean)</th>
<th>Percentage cells covered [%]</th>
<th>Percentage of all BGC prof [%]</th>
<th>t test</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1LL</td>
<td>0.11</td>
<td>0.01</td>
<td>0.09</td>
<td>0.03</td>
<td>6293</td>
<td>6.3</td>
<td>15.64</td>
<td>10.22</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>1HL</td>
<td>0.15</td>
<td>0.02</td>
<td>0.39</td>
<td>0.34</td>
<td>1850</td>
<td>3.6</td>
<td>6.77</td>
<td>5.33</td>
<td>no</td>
<td>SS, DA&amp;MV</td>
</tr>
<tr>
<td>2LL</td>
<td>0.09</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>7148</td>
<td>7.2</td>
<td>18.5</td>
<td>13.41</td>
<td>no</td>
<td>SS&amp;A, DMV</td>
</tr>
<tr>
<td>2HL</td>
<td>0.26</td>
<td>0.15</td>
<td>0.25</td>
<td>0.19</td>
<td>471</td>
<td>14.4</td>
<td>1.61</td>
<td>3.71</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>3LL</td>
<td>0.11</td>
<td>0.12</td>
<td>0.09</td>
<td>0.20</td>
<td>6820</td>
<td>7.6</td>
<td>17.95</td>
<td>14.32</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>3HL</td>
<td>0.18</td>
<td>0.21</td>
<td>0.19</td>
<td>0.34</td>
<td>201</td>
<td>34.5</td>
<td>0.64</td>
<td>2.87</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>4LL</td>
<td>0.16</td>
<td>0.20</td>
<td>0.22</td>
<td>0.45</td>
<td>1338</td>
<td>3.8</td>
<td>3.69</td>
<td>2.51</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>4HL</td>
<td>0.21</td>
<td>0.27</td>
<td>0.29</td>
<td>0.48</td>
<td>909</td>
<td>5.8</td>
<td>3.55</td>
<td>3.69</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>5LL</td>
<td>0.20</td>
<td>0.13</td>
<td>0.10</td>
<td>0.10</td>
<td>1108</td>
<td>5.7</td>
<td>2.77</td>
<td>1.58</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>5HL</td>
<td>0.19</td>
<td>0.09</td>
<td>0.45</td>
<td>0.60</td>
<td>2977</td>
<td>4.2</td>
<td>11.23</td>
<td>12.91</td>
<td>no</td>
<td>SS, DA&amp;MV</td>
</tr>
<tr>
<td>6LL</td>
<td>0.55</td>
<td>0.77</td>
<td>0.30</td>
<td>0.64</td>
<td>148</td>
<td>5.0</td>
<td>0.39</td>
<td>0.17</td>
<td>no</td>
<td>LBGCS</td>
</tr>
<tr>
<td>6HL</td>
<td>0.32</td>
<td>0.63</td>
<td>0.53</td>
<td>1.21</td>
<td>1177</td>
<td>11.2</td>
<td>5.04</td>
<td>10.48</td>
<td>no</td>
<td>M</td>
</tr>
<tr>
<td>7LL</td>
<td>0.36</td>
<td>0.51</td>
<td>0.22</td>
<td>0.37</td>
<td>816</td>
<td>6.4</td>
<td>2.02</td>
<td>1.73</td>
<td>no</td>
<td>LBGCS</td>
</tr>
<tr>
<td>7HL</td>
<td>0.30</td>
<td>0.38</td>
<td>0.49</td>
<td>1.04</td>
<td>2378</td>
<td>6.9</td>
<td>10.19</td>
<td>17.08</td>
<td>yes</td>
<td>M</td>
</tr>
</tbody>
</table>

Table 3: Properties of the bioregions split in high and low latitudes sub-bioregions, results of the comparison between the BGC-Argo and satellite time series (t test), and classification of the sub-bioregions in four categories: (i) Similar Shape, Different Amplitude and Mean Value (SS, DA&MV); (ii) Similar Shape and Amplitude, Different Mean Value (SS&A, DMV); (iii) Low BGC-Argo sampling (LBGCS); (iv) Match (M). Sub-bioregion 6HL is classified as Match due to the good qualitative agreement between BGC-Argo and satellite chlorophyll measurements (Supplementary Table S.3).
Figure 4: Summary of satellite and BGC-Argo time series agreement. Regions where both time series match (M) are shown in green. Regions where the density of BGC-Argo profiles is too low to permit comparison with satellite time series (LBGCS) are shown in yellow. Regions where satellite and BGC-Argo time series present similar shapes and amplitudes, but different mean values, are shown in orange (SS&A, DMV). Regions where satellite and BGC-Argo time series share similar shapes but different amplitudes and mean values are shown in red (SS, DA&MV). Black dots show the sub-bioregion 6 HL, for which we found a good qualitative agreement between BGC-Argo and satellite chlorophyll measurements. However, attention must be paid when comparing BGC-Argo and satellite chlorophyll data in that region.
work and to the use of model rather than satellite outputs.

The good qualitative matchup of our findings with previous works is important especially in the light of the fact that we base our regionalisation uniquely on the shape of the CHL time series, and not on a larger number of assumptions and variables. This highlights how CHL phenological cycles are the expression of multiple marine processes. More importantly, previous works were able to match the shape of the satellite and BGC-Argo time series, but they obtained very different mean CHL values (D’Ortenzio et al., 2020). Here, to our knowledge, we show for the first time an agreement not only from the point of view of the seasonal shape, but also with respect to the mean CHL values. Furthermore, the partition is robust with respect to changes in the size of the grid cell used to calculate the climatologies and to the use of the shift of 6 months applied to the Southern hemisphere time series. The agreement between BGC-Argo and satellite CHL is partially explained by a large variability associated with the satellite time series. To investigate the processes underpinning this variability, we have split bioregions in high and low latitude sub-bioregions and considered only pixels located in proximity (less than 1°) of a BGC-Argo profile to calculate mean satellite time series. We find agreement in 8 out of 14 sub-bioregions, corresponding to 58% of our domain (reported in green in Fig. 4). In general, we find that for a BGC-Argo sampling density larger than 5 profiles per grid cell, the likelihood of an agreement between satellite and BGC-Argo time series is larger. This implicitly suggests a way to determine whether the BGC-Argo sampling density in a given bioregion is, to date, good enough. An exception to this scenario is provided by bioregion 4 LL, which shows agreement of satellite and BGC-Argo time series despite having only 3 BGC-Argo profiles per grid cell. Conversely, further BGC-Argo float deployments are needed in bioregion 2 LL, despite a relatively large BGC-Argo sampling density there (≈6 BGC profiles per cell).

We qualitatively identify three types of discrepancy between BGC-Argo and satellite time series.

- Similar shapes, different amplitudes and mean values: These regions are characterised by a prominent CHL enhancement observed by BGC-Argo during summer, which is observed by satellite as well but with a lower amplitude. In addition, the CHL values observed by BGC-Argo floats are consistently larger than the ones observed by satellite across the year. These sub-bioregions are found at high latitudes (sub-bioregions 1, 5, 6 HL, and also bioregion 1 considered exclusively in the Southern Pacific, Supplementary Material S.6 and Supplementary Fig. S.1) and present a relatively large number of BGC-Argo profiles (5%, 13%, and 11% of all profiles, respectively). Thus, the mismatch is unlikely due to insufficient BGC-Argo sampling (in addition, in sub-bioregion 6 HL the density of BGC-Argo profiles is among the highest: 10 profiles per grid cell). We suggest that this mismatch could rather be due to well-known problems in the conversion of fluorescence measured by the floats into CHL Behrenfeld et al. (2006, 2009); Bock et al. (2022). Notably, this discrepancy has already been reported in the region of sub-bioregion 1 HL, in particular in the Southern Pacific (Mignot et al. 2023; Bock et al. 2022, Supplementary Fig. S.1). Another possibility is the scarce number of years for which
observations are available in satellite CHL images at high latitudes (Supplementary Fig. S.14); hence, the climatological satellite time series may have missed important productivity events, which may explain the low value of the satellite-derived CHL during the annual maximum in summer. Hence, we stress that attention must be used when analysing CHL in these regions, particularly in the Southern Ocean during Austral summer. For these reasons, we report them in red in Fig. 4. When considering all the time series composing sub-bioregion 6 HL (and not only those closer than 1° to a BGC-Argo profile), satellite and BGC-Argo time series agree (Supplementary Table S.3). For this reason, Sub-bioregion 6 HL is shown in green (with black dots), and leads to an agreement between satellite and BGC-Argo time series in 61% of our domain.

- Similar shapes and amplitudes, different mean values: In these regions (sub-bioregion 2 LL), mainly located in a band around 25°N and 25°S, BGC-Argo and satellite time series show a similar phenology, both in terms of shape (t test significant, Supplementary Table S.3) and amplitude of the time series (Table 3). However, the chlorophyll reported by BGC-Argo floats is constantly lower than the one measured by satellite. This discrepancy could be due to the overestimation of CHL by satellite at low latitudes as reported in previous studies (Omta et al., 2009; Clow et al., 2024). Conversely, as the density and number of BGC-Argo profiles are among the largest there, the BGC-Argo coverage seems not to explain this difference. Hence, these regions are reported in orange in Fig. 4.

- Insufficient BGC-Argo sampling: In sub-bioregions 6 and 7 LL, mainly located in the Arabian Sea and close to Amazon delta, the low number of BGC-Argo profiles did not allow us to calculate a robust BGC-Argo time series. This likely explains the mismatch with satellite measurements. These results suggest the need for further BGC-Argo deployments in these regions. Hence, sub-bioregions 6 and 7 LL are reported in yellow in Fig. 4.

5 Conclusions

Our results point to the fact that the agreement between CHL observed by satellite and BGC-Argo floats is valid in the majority of the global ocean (∼58-61% of the surface area), but not everywhere. Our method provides a novel way to determine regions where satellite and BGC-Argo observations agree, and regions in which they do not, based on the phytoplankton phenology described with annual cycles of CHL. The reasons for this mismatch are likely due to poor satellite performance (in particular in sub-bioregion 2 LL) or to problems in the relationship used to convert fluorescence to CHL in BGC-Argo floats (sub-bioregions 1, 5, and 6 HL). Our methods also allow us to identify regions lacking BGC-Argo data (sub-bioregions 6 and 7 LL). More generally, we suggest larger BGC-Argo float deployments in sub-bioregions showing discrepancy between satellite and BGC-Argo time series. In conclusion, we stress that attention must be paid when coupling together satellite and BGC-Argo CHL observations in these regions (e.g. to constrain biogeochemical models or for marine spatial planning),
as they cover ~40% of the global ocean. The framework used in the present study can be applied in the comparison of other biogeochemical variables.

Acknowledgements

AB, VT, and FDO received support from the CNES/TOSCA project SEASONs. This study has been conducted using E.U. Copernicus Marine Service Information (product OCEANCOLOUR_GLO_BGC_L3_MY_009_103; https://doi.org/10.48670/moi-00280). The authors would like to thank the International Argo and BGC-Argo Program, as well as the NASA and Mercator Ocean agencies. The authors are grateful to Inés Jambou for her help in the design of Figs 1–3 and the different tables, as well as Elodie Gutknecht, Coralie Perruche, and Giulio Tirabassi for their useful advise.

Data availability

All the data necessary to reproduce all the plots shown in the present paper will be published in zenodo upon publication.
Supplementary Material
The methods used to calculate the optimal number of clusters are:

- **Elbow method** The elbow method (Sugar, 1998; Marutho et al., 2018) is based on the calculation of the total within-clusters sum of squares. In other words, for a given cluster we calculate the average of the Euclidean distances between each time series belonging to the cluster and the mean time series of that cluster. This is repeated for all the clusters, resulting in $k$ values (one for each cluster), which are summed together. This provides the total within-clusters sum of squares for the partition in $k$ clusters, which is considered representative of the total variance of that partition. The elbow method consists in identifying the value $k$ for which the variance stops decreasing, i.e. the addition of a supplemental cluster does not bring any novel information.

- **Silhouette method** The silhouette method (Rousseeuw, 1987) measures the similarity of the time series within the same cluster and compares it with the separation between clusters. The silhouette value varies between -1 and 1 and it indicates, for each time series, whether it was appropriately assigned to the corresponding cluster (silhouette value tending to 1) or not (silhouette values tending to -1). We use the silhouette value averaged over all the time series, and we calculated this metric for $k$ between 2 and 25, similarly to Kheireddine et al. (2021).

- **Calinski-Harabasz method** This method (Caliński and Harabasz, 1974) identifies the optimal number of clusters which maximises the cohesion of the time series within the same cluster, while at the same time maximising the separation of different clusters. This method relies upon distance from the mean time series of each cluster and the mean global time series.

### S.2.1 Detailed identification of the optimal number of clusters

The elbow method indicates an optimal number of clusters between 5 and 9. The silhouette analysis and the Calinski-Harabasz index indicate an optimal number of clusters of 3 (Supplementary Fig. S.12 and S.13). However, this solution was discarded because it separates the global ocean in large, well-known regions (an Equatorial-Tropical band, a temperate band, and a polar band, Supplementary Fig. S.5). In fact, in the present study, in order to identify agreement and differences between BGC-Argo and satellite chlorophyll observations, we focus on smaller regions. Furthermore, the silhouette analysis has a local maximum at 6 and 7 clusters. This validity of this value is confirmed by a supplementary analysis. When passing from 7 to 8 clusters, the first cluster is split in two, while the other clusters remain unchanged (Supplementary Fig. S.8). The two mean time series obtained from the split of the first cluster are consistently similar to the mean time series of the first cluster. Conversely, when passing from 5 to 6 clusters (or from 6 to 7), a novel time series is introduced, increasing the variance explained by the partition.
S.3 Third climatology (extended climatology with a short time frame)

The third climatology (referred also as “extended climatology with a short time frame”) is calculated with the same exact procedure and parameters used for the extended climatology. The only difference is that we use L3 chlorophyll fields of 2018-2022 time period only (and not the 1998-2022 time period, Table S.2).

This climatology is partitioned using the reference time series of the reference climatology (analogously to what has been done to partition the extended climatology, Section 2.5). The same tests used to compare the partitions obtained from the reference and extended climatologies are applied. The percentage of surface covered by each bioregion does not change consistently when considering the partition with the reference or with the third climatology (Supplementary Table S.1). Furthermore, 93.3% of the time series of the third climatology are assigned to the same bioregion than the reference climatology (Supplementary Fig. S.4a and c). The improvement compared to the extended climatology (for which only 72.9% of the time series were correctly assigned) can be explained by the fact that the reference and third climatology are calculated using data from 25 years (1998-2022 period) compared to the extended climatology (based on 5 years only: 2018-2022 period). Hence, this difference is due to temporal changes and not the partition method. All in all, 97.6% of the time series of the third climatology are within the threshold of acceptance of the corresponding reference time series. These findings indicate that the 7 main time series of the third climatology are consistently similar to the ones of the reference and extended climatologies, and corroborate our choice of extending our analyses to the coastal and polar regions.

S.4 T-test detailed description

After verifying that both time series (BGC and satellite ones) are normally distributed, we perform a t-test (Student (1908)) using the respective mean and standard deviations.

\[
t = \frac{|x_{BGC_i} - x_{Sat_i}|}{\sqrt{\frac{\sigma^2_{BGC_i}}{n_{BGC_i}} + \frac{\sigma^2_{Sat_i}}{n_{Sat_i}}}}
\]

where \(x_{BGC_i}\) is the average of the BGC-Argo time series in the \(i\)-th cluster, \(x_{Sat_i}\) is the average of the satellite time series in the \(i\)-th cluster, \(\sigma^2_{BGC_i}\) and \(\sigma^2_{Sat_i}\) are the variances of the two time series, and \(n_{BGC_i}\) and \(n_{Sat_i}\) are the number of elements in each time series. The results of this test is compared with the reference value at 5% of significance obtained using \(n_{BGC_i} + n_{Sat_i} - 2\) degrees of freedom. If smaller than the reference value, the time series are considered statistically similar.
S.5 Comparison of our partition with previous works

• Comparison of time series of Figure 1 with previous works.
  - The 7 mean time series identified compare well with the phenological classification of Longhurst (1995) and D’Ortenzio et al. (2012). Bioregion 1 corresponds to model number 4 of Longhurst (1995, Tropics), while bioregion 2 shares characteristics of both model 4 and 3 (Subtropical winter nutrient-limited) and closely matches D’Ortenzio et al. (2012, Tropical bioregion). Bioregion 3 corresponds to model 3, while bioregion 4 is similar to Bloom bioregion of D’Ortenzio et al. (2012). Bioregion 4 is also similar to model 2 of Longhurst (1995) but is characterised by a sharper spring peak. Bioregion 5 is similar to both model 1 (Polar irradiance-limited) and 2 (midlatitude nutrient-limited spring production peak) of Longhurst (1995). Finally, both bioregions 6 and 7 correspond to model number 1 of Longhurst (1995), with bioregion 6 which is characterised by a sharper chl decrease.

• Comparison of partition pattern of Figure 1 with previous works
  - Longhurst (1995): the latitudinal structure of our partition aligns qualitatively well with the distribution of the Longhurst provinces (1995, 1998), in particular in the central Pacific and Atlantic Oceans. In addition, bioregion 3 in the North Pacific closely resembles the North Pacific Polar Front province. Other province correctly identified by our partition are the Western Australian and Indonesian Coast, the Tasman Sea, the Northwest Arabian Sea upwelling, and the Benguela current coast. Conversely, our partition in the Southern Ocean is more scattered compared to the Subantarctic water ring and the Antarctic provinces.
  - Fay and McKinley (2014): The partition we obtained compares relatively well with the one obtained by Fay and McKinley (2014). Bioregion 2 and 3 together match remarkably well with their subtropical permanently stratified biome. Bioregion 4 is similar to their subtropical seasonally stratified biome, while bioregion 7 to the subpolar seasonally stratified biome, especially in the North Atlantic. Bioregion 1 is similar to their equatorial biome, but in our case it extends over a larger band and is also present in the Southern Pacific. In the Southern Ocean, the biomes found by Fay and McKinley (2014) are organised in latitudinal bands, while the disposition of our bioregions is less regular.
  - Bock et al. (2022): the high-chlorophyll bloom bioregion matches well with our bioregion 7, which is characterised by a strong seasonal bloom as well. The points classified by Bock et al. (2022) as Equatorial cluster belongs to our bioregion 1, which shows typical Equatorial dynamics. In the Indian Ocean, our partition correctly separates the Equatorial cluster (our bioregion 1) from the Arabian cluster (our bioregions 2 and 6).
Table S.1: Percentage of time series belonging to each of the 7 bioregions (columns) for the reference climatology (first row), the extended (second row) and third climatology (third row). The last column indicates the percentage of time series of the second and third climatology which were not assigned to any bioregion as their Euclidean distance was larger than the corresponding threshold of acceptance (Methods).

<table>
<thead>
<tr>
<th>Bioregion 1 % surface</th>
<th>Bioregion 2 % surface</th>
<th>Bioregion 3 % surface</th>
<th>Bioregion 4 % surface</th>
<th>Bioregion 5 % surface</th>
<th>Bioregion 6 % surface</th>
<th>Bioregion 7 % surface</th>
<th>Not assigned [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference climatology</td>
<td>22.92</td>
<td>23.22</td>
<td>20.64</td>
<td>8.33</td>
<td>9.9</td>
<td>6.08</td>
<td>8.92</td>
</tr>
<tr>
<td>2018-2022 climatology</td>
<td>22.4</td>
<td>20.11</td>
<td>18.59</td>
<td>7.24</td>
<td>14</td>
<td>5.43</td>
<td>12.21</td>
</tr>
<tr>
<td>1998-2022 climatology</td>
<td>18.78</td>
<td>19.91</td>
<td>17.42</td>
<td>7.96</td>
<td>11.21</td>
<td>10.27</td>
<td>14.9</td>
</tr>
</tbody>
</table>

S.6 Equatorial and Southern Pacific case study.

Here, we focus on cluster 3 in the Pacific Ocean, as this cluster has a consistent presence both at low and high latitudes there, and a significant agreement between satellite and BGC time series only at low latitudes, but not at high latitudes in the Equatorial Pacific (EP), but not in the Southern Pacific (SP). Therefore, we consider bioregion 1 in the Equatorial Pacific (bioregion 1 EP) as the grid cells located in the Pacific Ocean whose latitude is comprised between 21°S and 12°N, and bioregion 1 in the Southern Pacific (bioregion 1 SP) as grid cells whose latitude is comprised between 33°S and 58°S. For each of these two subclusters, we calculate the mean time series of chlorophyll measured by satellite and by the BGC-Argo fleet.

In bioregion 1 EP, chlorophyll concentration is rather steady and low across the year (Fig. S.1a and b). In general, there is a good agreement between satellite and BGC observations, though the former are in general lower than BGC observations (0.1 mg/m³ against 0.2 mg/m³).

In bioregion 1 SP, BGC chlorophyll observations are 3 to 4 times larger than satellite-based measurements (0.3-0.4 mg/m³ against 0.1 mg/m³). Bioregion 1 SP is characterised by a deep ML during winter months that, during spring, becomes shallower and is associated with a seasonal DCM. Subsequently, the ML deepens again, making the DCM disappear, and triggering an increase of integrated chlorophyll in the mixed layer (Fig. S.1c).
Table S.2: List of parameters (rows) used for the calculation of each climatology (columns) as in Table S.2, but including the third climatology as well (named extended climatology, short time frame)
Table S.3: Properties of the bioregions split in high and low latitudes sub-bioregions as reported in Table 3. Here, we report also properties concerning the number of BGC Argo profiles composing each time series, and the results of the comparison between the BGC-Argo and satellite time series using the normalised values.
Figure S.1: Satellite and BGC-Argo time series in the Equatorial Pacific portion of bioregion 1 (first two panels) and the Southern Pacific portion of bioregion 1 (third and fourth columns). Panel a and c: NCHL time series. Panel b and d: non normalised CHL time series (i.e. showing real chlorophyll values). In each panel, the thick dark-blue solid lines indicate the mean chlorophyll time series observed by satellite, with associated uncertainty (standard deviation, thin coloured line), while the green dots and the solid line the mean BGC-Argo time series.
Figure S.2: First plot: boxplot of number of cells composing bioregions in which satellite and BGC time series are significantly different (left box) and in which are significantly similar (right box), listed in Table 3. The other plots show the mean and maximum number of BGC profiles per cell, the percentage of all BGC profile within a given bioregion, the number of BGC profiles for a given time step (median, minimum, and maximum), and the number of BGC profiles per 100×100 km surface square.
Figure S.3: $k$-means partition of the reference climatology with reference time series. The reference climatology is calculated with stricter criteria than the other two climatologies (Table 1 and Supplementary Table S.2), and does not include coastal or polar regions (Methods). The upper panel shows the 7 bioregions (right-hand legend) obtained with the $k$-means partition. Each of the lower panels shows the within-bioregion average of the normalised time series (also referred to as the reference time series) as thick coloured lines. The thin coloured lines show the associated uncertainty (standard deviation).
Figure S.4: Top left panel: results of $k$-means partition applied to the reference climatology (as reported in Supplementary Figure S.3). Top right panel: results of partition obtained using the 7 reference time series on the extended climatology, but only for the pixels where the reference climatology is defined. The second climatology includes coastal and polar regions, and is based on 2018-2022 data (Methods). In this panel, 93.3% of the grid cells are assigned to the same bioregion than the reference (upper left panel). Lower left panel: results of partition obtained using the 7 reference time series on the third climatology, but only for the grid cells for which the reference climatology is defined. The third climatology includes coastal and polar regions, and is based on 1998-2022 data (Methods). In this panel, 72.9% of the pixels are assigned to the same bioregion than the reference (upper left panel). In addition, 97.3% of the time series are within the threshold of acceptance of the corresponding reference time series. Lower right panel: same as lower left panel, with the difference that all the grid points are reported (hence, also coastal and polar regions). Please note that this panel is identical to the one reported in Fig. 1.
Figure S.5: Partition of the reference climatology with $k=3$ bioregions (value indicated by the silhouette and Calinski-Harabasz indexes, Methods and Supplementary Fig. S.12 and S.13) and using the $k$-means method. The lower panels show the within-bioregion average of the normalised time series (thicker coloured solid line: mean value; thinner coloured solid line, uncertainty: standard deviation; colours indicated in the right-hand legend in the upper panel). The global ocean results partitioned in three large, well-known bioregions: an equatorial-tropical band, a temperate band, and a polar bioregion.
Figure S.6: Comparison between reference time series and average time series obtained from partition of the extended climatology. The upper panel shows the partition obtained using the 7 reference time series on the extended climatology, as reported in Fig. 1. Each of the lower panels shows a reference time series as black dots (large dots: average value; small dots: standard deviation). In addition, we report the within-bioregion average of the normalised time series of the extended climatology (thicker coloured solid line: mean value; thinner coloured solid line, uncertainty: standard deviation; colours indicated in the right-hand legend in the upper panel). For each panel, the similarity between the two time series was tested (i) with a t test and (ii) by calculating the Euclidean distance between the two time series, and by comparing whether it was lower than the corresponding threshold of acceptance of the reference time series. All the tie series resulted significantly similar to the corresponding reference time series.
Figure S.7: Comparison between BGC and satellite normalised time series. The upper panel shows the partition obtained using the 7 reference time series on the extended climatology with, superposed, the position of the BGC-Argo profiles (black dots), as reported in Fig. 2. Each of the lower panels shows the within-bioregion average of the normalised time series of the extended climatology (thicker coloured solid line: mean value; thinner coloured solid line, uncertainty: standard deviation; colours indicated in the right-hand legend in the upper panel). In each panel, the black dots represent the median value measured by the BGC-Argo floats for each 8-day time step, while the black solid line represents the moving average of the black dots (over 40 days, i.e. 5 time steps). Both the black dots and the black solid line values were normalised by dividing them by the maximum value of the moving average time series. Finally, the similarity between the two time series (the coloured line obtained from satellite data and the black line obtained from BGC-Argo floats) was tested (i) with a t test and (ii) by calculating the Euclidean distance between the two time series, and by comparing whether it was lower than the corresponding threshold of acceptance of the reference time series. The results of the tests are reported in the title of each panel.
Figure S.8: Identification of the optimal number of clusters. Each column shows the mean time series (within-cluster average) obtained using different $k$ values (from $k=6$ to $k=9$) to partition the reference climatology. Time series are sorted so that similar time series appear on the same column/row. When passing from $k=6$ to $k=7$ clusters, a novel time series appears (second row, second column). However, when passing from $k=7$ to $k=8$ clusters, the time series in the third row (cluster 1 in partition with $k=7$) is split into two very similar time series (clusters 1 and 8 in partition with $k=8$). The same occurs when passing from $k=8$ to $k=9$ clusters (this time in the second row, cluster 7 in partition with $k=8$ becomes clusters 1 and 7 in partition with $k=9$). Hence, we identify an optimal number of clusters $k=7$. 

<table>
<thead>
<tr>
<th>Bio region</th>
<th>Partition with $k=6$ bioregions</th>
<th>Partition with $k=7$ bioregions</th>
<th>Partition with $k=8$ bioregions</th>
<th>Partition with $k=9$ bioregions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1.png" alt="Graph 1" /></td>
<td><img src="image2.png" alt="Graph 2" /></td>
<td><img src="image3.png" alt="Graph 3" /></td>
<td><img src="image4.png" alt="Graph 4" /></td>
</tr>
</tbody>
</table>
Figure S.9: Sensitivity of the results with respect to the spatial resolution of the climatology. Left panels: partition of the reference climatology in \( k=7 \) bioregions and corresponding reference time series, as reported in Supplementary Fig. S.3. Here, the size of a grid cell is 1°. Right panels show the exact same results obtained this time with a climatology of grid cell size 0.50°. The lower panel shows the exact same results with a climatology of grid cell size 0.25° (only the map is reported).
Figure S.10: Sensitivity of the results when not shifting time series in the Southern hemisphere of 6 months. The upper panel shows the partition of the global ocean in $k=7$ bioregions, as reported in the upper panel of Fig. 1. The lower panel shows the partition obtained when not shifting time series in the Southern hemisphere of 6 months. In that case, we identify an optimal number of bioregions $k=9$. Even if the optimal number of bioregions among the two panels is different, however, their spatial distribution is almost identical. The presence of 2 additional bioregions in the lower panel can be explained by the fact that some time series in the Southern hemisphere contain the same information than time series in the Northern hemisphere, only shifted by 6 months. Thus, in the lower panel, the algorithm is forced to find two further time series. This result also corroborates our choice about shifting time series in the Southern hemisphere of 6 months.
Figure S.11: Elbow test: total within-clusters sum of squares ($y$-axis) when varying $k$ (the number of cluster used to partition the reference climatology, $x$-axis).
Figure S.12: Silhouette analysis results: silhouette value (y-axis) when varying $k$ (x-axis)
Figure S.13: Calinski-Harabasz results: Calinski-Harabasz value (y-axis) when varying $k$ (x-axis). The blue dashed vertical line shows the maximum of the Calinski-Harabasz value.

Figure S.14: The colour of each pixel shows the percentage of available satellite CHL data used in the calculation of the extended climatology. This was calculated as the ratio of the number of CHL time series not flagged as NA and the total number of years considered (25) and then multiplied by 100.
Figure S.14: (Includes panels on this page and previous one). Silhouette value for each bioregion of the extended climatology. The right panels show, for each bioregion, the histogram of the silhouette values of the time series belonging to that bioregion. The colour of the histogram is dark if the silhouette value is larger than the median of all the silhouette values of that bioregion, light if it is negative, and intermediate if it is between 0 and the median. The left panel shows the distribution of the points belonging to that bioregion, coloured according to the histogram colour bar.
References


Student, The probable error of a mean, Biometrika, 6, 1–25, 1908.

Sugar, C. A., Techniques for clustering and classification with applications to medical problems, Ph.D. thesis, United States – California, 1998, copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works.


Comparing satellite and BGC-Argo chlorophyll estimation: a phenological study

Alberto Baudena1,2,∗, Vincent Taillandier1, Nicolas Mayot1, Alexandre Mignot2, Wilhem Riom1, and Fabrizio D’Ortenzio1

1Sorbonne Université, CNRS, Laboratoire d’Océanographie de Villefranche, UMR 7093 LOV, Villefranche-sur-Mer, France
2Consiglio Nazionale Delle Ricerche- Istituto di Scienze Marine (CNR-ISMAR), Lerici (SP), Italy
3Mercator Ocean International, 31400 Toulouse, France
∗To whom correspondence should be addressed;
E-mail: alberto.baudena@gmail.com.

Abstract

Ocean primary production is a key process that regulates marine ecosystems and the global climate, but its estimation is still affected by multiple uncertainties. Typically, the chlorophyll-a concentration (CHL) is used to characterise this process, as it is considered a proxy of phytoplankton biomass. To date, the most common observing systems for studying CHL are ocean colour satellites and BGC-Argo floats. Those are complementary systems: satellite observations provide global coverage but are limited to the ocean surface, while BGC-Argo floats provide punctual observations along the whole water column. Comparison of these two observing systems has been performed only at regional or single-float scales, while at global scale this results in large uncertainties due to the relatively low and irregular BGC-Argo coverage. Here, we propose a different method, by comparing satellite and BGC-Argo climatological annual time series within seven different bioregions, each characterised by a homogeneous phytoplankton phenology, allowing us to smooth the uncertainties. By comparing the mean values, the amplitudes, and the shapes of the two time series, we are able to identify regions (a) where they agree (58-61% of the ocean surface area); (b) where the BGC-Argo float network should be extended (generally regions with less than 5 profiles each 100x100 km² square); (c) where the discrepancy is likely due to satellite or (d) BGC-Argo performance. Use of either BGC-Argo and satellite data in regions b–d should be carried carefully and we provide, for each region, suggestions on which system could be affected by the largest uncertainties.

Plain language summary

The capacity of oceans to produce photosynthetic activity and hence participating in global primary productivity is mainly assessed through the measurement of chlorophyll (CHL). This pigment, responsible for photosynthetic process, is mainly measured by two observing systems to date: satellite and biogeochemical
Argo drifters. While these systems are complementary (satellite have broadly coverage limited at the surface, BGC-Argo drifters punctual observations along the whole water column), comparisons of their output only exist at regional or single-drifter levels. Here, rather then comparing them punctually (which introduce extremely large noise levels), we group the observations according to 7 bioregions, smoothing the uncertainties. Each bioregion is characterised by a homogeneous phenology, i.e. a typical CHL time series along the year. By comparing the time series observed by satellite with the one obtained from BGC-Argo floats, we show that these two observing systems agree in most of the global ocean surface area (58-61%). We identify some of the reasons explaining the discrepancy found in the other areas, in some cases due to BGC-Argo drifters observing capacities, in others due to satellite ones. Finally, we show where are the regions where the BGC-Argo sampling density should be increased.

1 Introduction.

Ocean primary production is considered to contribute to 50% of the global net primary production (Field et al., 1998). One of the main sources of error in this estimation is the extreme spatial and temporal variability of the global ocean distribution of phytoplankton. This is required to evaluate and predict the consequent primary production and is hardly characterised and quantified.

To date, mapping the global ocean phytoplankton distribution is mainly assessed through the systematic observation of the chlorophyll-a concentration (CHL), which is the main algal pigment responsible for the photosynthetic process in phytoplankton cells. CHL is generally considered a good, although not perfect, proxy of phytoplankton biomass concentration in the ocean (see for example discussion in Huot et al. 2007). Therefore, estimates of ocean primary productivity strongly rely on the occurrence of CHL measurements.

CHL is estimated by several methods. The most efficient in terms of spatial and temporal resolution is the satellite ocean colour. Since 1997, systematic and gap free databases of remote sensing CHL, at kilometric resolution, have been obtained thanks to an uninterrupted series of ocean colour missions: SeaWiFS, MODIS, MERIS, VIIRS, and OLCI (McClain et al., 2022). These observations are massively used to identify global or regional patterns in the CHL distribution. In addition, they are used to characterise specific phytoplankton dynamics as well as seasonal, frontal, or mesoscale blooms (see a recent review of the ocean colour application in Sathyendranath et al. 2023). Ocean colour CHL is also the primary data source for validating phytoplankton concentration simulated by global ocean biogeochemistry models (Gregg, 2008; Pradhan et al., 2019) and, more recently, it has been successfully assimilated in operational models (e.g. in the global Copernicus marine service biogeochemical forecasting system: Mignot et al. 2023). However, ocean colour satellites are unable to sense the ocean surface in presence of clouds or strong winds, resulting in large portions of unobserved ocean areas (Stock et al., 2020). Products generated by using observations obtained by different satellite sensors, could partially palliate this problem by increasing the spatial and temporal coverage of CHL satellite maps, although they could introduce additional errors (Garnesson et al. 2019). Furthermore, although uncertainty of global satellite CHL retrieval is estimated
to range between 16.5% and 30% (Hu et al., 2012) regional characteristics could strongly degrade space derived estimations (e.g. Chen et al. 2021).

Another diffused method to estimate CHL is through fluorescence measurements: the excitation of a water parcel containing algal pigments with blue/green light generates an emission of light in the red, which is proportional to CHL. Fluorimeters (i.e. sensors measuring fluorescence) are routinely used to determine in situ CHL, in particular from ship, providing then a method to obtain depth profiles of CHL (see for example Petit et al. 2022 and references therein). Fluorimeters are usually calibrated by high performance liquid chromatography (HPLC) or spectrofluorometers. To date, fluorescence-based CHL estimations are undoubtedly the most large source of in situ CHL observations. More recently, the number of available fluorescence-based CHL profiles has dramatically increased thanks to the massive deployment of Argo profiling floats equipped with fluorometers (BGC-Argo, Claustre et al. 2020). A BGC-Argo float is a robotic drifter, equipped with physical and biogeochemical sensors, able to autonomously sample the first 2 kms of the water column and to transmit in real-time the acquired data. To date, the BGC-Argo floats have acquired more than 200k CHL profiles ((Stoer et al., 2023)) at locations and seasons generally undersampled by traditional research vessels. However, the calibration and validation of CHL data from BGC-Argo floats is not straightforward (see Roesler et al. 2017; Schmechtig et al. 2023; Lavigne et al. 2012), mainly because the HPLC data obtained simultaneously to BGC-Argo profile are mainly limited at the float deployment or at the recovery, (Johnson et al., 2017; Taillandier et al., 2018; D’Ortenzio et al., 2020). In addition, the BGC-Argo network has a limited spatial coverage, and is still under development (Group, 2016).

Nonetheless, by providing CHL at depth, under the clouds and at high latitudes (i.e. where and when the ocean colour satellites cannot provide data), autonomously and in real time, the BGC-Argo network is the ideal observing system to complement ocean colour satellites. The combined use of the two observing systems offers extraordinary opportunities to characterise the spatio-temporal distribution of CHL at global scale. Several studies have focused on the combination of the two data sets to validate and assimilate ocean biogeochemistry models (e.g. Mignot et al. 2023; Cossarini et al. 2019; Teruzzi et al. 2021). Other studies have compared CHL fields obtained by few floats missions (i.e. Chiswell et al. 2022; Boss et al. 2008; D’Ortenzio et al. 2021) or focusing on regional or sub-basin scale (i.e. D’Ortenzio et al. 2020; Häńjtjens et al. 2017; Wojtasiewicz et al. 2018). To date, however, very few studies have compared the two data systems with a strict matchup approach at global scale as generally required for satellite calibration/validation exercise (Werdell and Bailey, 2005). Independently to the intrinsic uncertainties related to the two observing systems, a matchup analysis performed at global scale is not trivial. The high demanding criteria required to perform calibration and validation exercises for ocean colour (as described in Werdell and Bailey 2005) are often too severe to obtain a sufficient number of matchups. These criteria are then weakened, for example, by enlarging the time window (from ±3 hours to ±12 hours, as in Wojtasiewicz et al. 2018). In addition, the irregular distribution of the BGC-Argo profiles, with regions of strong density and others completely unobserved by floats, could generate bias in the matchup analysis, pushing toward regional approaches (as in D’Ortenzio et al. 2021).

In this context, we propose an alternative method to compare these two datasets.
Rather than comparing BGC-Argo and satellite CHL at the locations where BGC-Argo profiles are obtained, we compare their climatological annual time series within defined regions. CHL time series represent a pivotal information to characterise the phenology of phytoplankton, i.e., the identification of key moments in the phytoplankton annual cycle, such as bloom onset or decay. Studying phytoplankton phenology is considered crucial within the framework of climate change: a modification in the timing of the main phytoplankton dynamics, induced by alterations in its environment, could have significant repercussions on the entire marine food web (Ji et al., 2010). In this sense, by using phenology as a comparison metric, we explore how the two datasets perform in giving a global picture of the CHL distribution and dynamics, and how they could simultaneously be used to compensate for their intrinsic limitations (see also Gittings et al. 2019).

Comparing time series obtained from the two observing systems, however, is not completely trivial. Although for each location (i.e. a pixel) a CHL time series could be generated from a satellite data set, this is not the case for BGC-Argo floats, which could change positions at each time step. In such a sense, satellite time series could be considered as an Eulerian observation, when BGC-Argo times series are close to a Lagrangian data frame. To overcome this issue, we will use here a method to aggregate multiple individual (i.e. satellite and BGC-Argo) time series, within large ocean regions. For this, we use the concept of “bioregion”, which could be defined as an oceanic region characterised by homogeneous biogeochemical characteristics. In our case, the bioregions are defined as oceanic regions sharing similar phenological characteristics and, more precisely, temporal series with similar shapes. This approach has already been applied to ocean colour time series and used to identify phenological patterns in the Mediterranean (D’Ortenzio and Ribera d’Alcalà, 2009; Mayot et al., 2016), in the Red Sea (Kheireddine et al., 2021), in the Labrador Sea (Marchese et al., 2019), and in the Southern Ocean (Ardyna et al., 2017). It was also used to compare two different ocean colour data sets obtained at 20 years of interval (D’Ortenzio et al., 2012). Finally, and more importantly for the present study, maps of satellite-derived phenological bioregions have been used as masks to aggregate sparse in situ data and to generate CHL time series comparable to those obtained from satellite (Lacour et al., 2015; Mayot et al., 2017; D’Ortenzio et al., 2021).

Here, we applied this approach with the specific aim of comparing the capability of ocean colour satellites and BGC-Argo observing systems to evaluate CHL seasonality at global scale. The advantage of this approach is that it allows us (i) to identify ocean regions that share the same phytoplankton phenology in a robust way; (ii) to smooth the uncertainties, by regrouping and averaging satellite and BGC-Argo data within each cluster (iii) to conduct the comparison in phenologically homogeneous regions, avoiding potential biases.

2 Materials and Methods

This paper is based on the analysis of temporal series of CHL computed in a climatological way. The key concept is that, for a given location, we build a time series which reveals how the chlorophyll changes, on average, along a typical year. This section is organised as follows.
In Subsecs. 2.1–2.5, we expose the calculation of time series of satellite CHL covering the global ocean (referred to as a climatology), and how they are clustered using the phenological procedure of D’Ortenzio and Ribera d’Alcalà (2009).

In Subsec. 2.6, we describe the processing of BGC-Argo data and the calculation of BGC-Argo-derived seasonal CHL time series specific to each cluster.

In Subsec. 2.7, we expose how satellite and BGC-Argo time series are statistically compared within each cluster.

### 2.1 Satellite data

The surface CHL satellite data set used is the OCEANCOLOUR\_GLO\_BGC\_L3\_MY\_009\_103 product which is provided by Copernicus Marine Environment Monitoring Service (CMEMS, http://marine.copernicus.eu/) platform (product version cmems_obs-oc_glo_bgc-plankton_my_l3-multi-4km_P1D). This L3 product has 4 km spatial resolution, is provided daily from 1998 to present day and merges observations from SeaWiFS, MODIS, MERIS, VIIRS-SNPP & JPSS1, OLCI-S3A & S3B missions. Consequently, the accuracy and coverage of the satellite CHL used here are dependent on the number and quality of the in orbit sensors (increasing in the last years when the number of simultaneous sensors was higher).

### 2.2 Climatology calculation

The clusterization method used requires, for each satellite pixel, a complete time series (i.e. without temporal gaps). Observations missing due e.g. to clouds or algorithmic errors are solved by interpolation (Subsec. 2.2.1). However, polar regions require a specific processing, mainly because they cannot be observed during winter months. This reduces the number of observations available and results in relatively long temporal gaps, which need to be interpolated, introducing then a high level of noise. Although noise is not an intrinsic problem, the application of the partition method to such time series prevents a rapid convergence of the clusterization and the identification of an optimal number of clusters, resulting in a numerical failure of the method.

To overcome this issue, a possibility would be to exclude polar regions from our analyses. However, a large number of BGC-Argo profiles (used to compare with satellite data) are in polar regions. In addition, poles are crucial locations regulating the global primary productivity and being importantly affected by climate change (Post et al., 2019).

First, we calculate a climatology (referred to as reference climatology) with very strict criteria (please see below the details). This allows us to obtain less noisy time series, and permits the partition method to converge and identify an optimal number of clusters (and reference time series). In this climatology, locations without enough data coverage are excluded, and only the last 5 years of data (2018-2022), those with larger accuracy are used. Other criteria are applied in order to reduce the noise in the time series (see below the details). The reference climatology covers the 57°S-57°N band and does not include coastal regions as they are affected by larger CHL uncertainties (Van Oostende et al., 2018; Zeng et al., 2023; McCluskey et al., 2022).

To extend our coverage to polar zones and coastal regions, the second step consists into the calculation of a second climatology, which we refer to as extended climatology. This is based on less strict criteria and on the use of the entirety.
of the chlorophyll observations over the last 25 years. Given the larger noise
and uncertainty, the partition method fails to converge and identify an optimal
number of clusters when directly applied to the extended climatology. For this
reason, we use the reference time series, identified with the reference climatology,
to partition the extended climatology. We associate, to each time series of the
extended climatology, one of the reference time series, based on their similarity
(Euclidean distance, details below). In this way, we are able to also include
the polar and coastal regions in our partition. We then verify the resemblance
of these two partitions (one obtained from the reference climatology, the other
from the extended one) to assure they are consistent (Supplementary Table S.1
and Supplementary Fig. S.4).
Details about the calculation of these climatologies are reported in the following
subsections.

2.2.1 Reference climatology
A first climatology, named reference climatology, is calculated using all daily
CHL fields between 2018 and 2022. The use of this time window is due to
the fact that the quality of CHL data is better compared to previous periods,
following the number and quality of in orbit sensors. We use a procedure
already used in previous works (D’Ortenzio and Riberà d’Alcalà, 2009; Marchese
et al., 2019; Mayot et al., 2016; Kheireddine et al., 2021; Ardyna et al., 2017;
D’Ortenzio et al., 2012), readapting it to study the global ocean. We proceed
as follows.

1. We consider the chlorophyll field of the 1st January 2018. Then, we ex-
clude all pixels within 1° (~111 km) of the coast. This is because uncer-
tainty of CHL data in coastal regions is generally larger than in the open
ocean (Van Oostende et al., 2018; Zeng et al., 2023; McCluskey et al.,
2022). We also exclude pixels in the Black and Caspian Seas, and in the
Hudson Bay, as they are enclosed or quasi-enclosed basins.

2. In order to smooth the uncertainties, we reduce the spatial resolution of
the chlorophyll field of the 1st January 2018 from the nominal resolution
(0.04°) to 1°. For this purpose, we build a grid of 1° resolution covering the
global ocean. For each grid cell, we take the median of all the chlorophyll
values within the grid cell (~625 values per cell). If, in that grid cell, less
than α=25% of the values are available (due e.g. to cloud coverage), we
flag that cell as not available (NA).

3. The temporal resolution of the time series is eight days. Thus, step 1
and 2 are repeated for all the days between the 1st and the 8th January
2018. Hence, for a given grid cell, we obtain 8 chlorophyll values. We
consider the median of these values as representative of the chlorophyll
concentration in that grid cell for the 1-8 January 2018 period. Values
flagged as NA are excluded from the median. If all eight values in a grid
cell over an eight-day period are marked as NA, the CHL in that grid cell
(for the 1-8 January 2018 period) is marked as NA. This is repeated for
all the grid cells.

4. Step 3 is repeated for all years between 2018 and 2022, obtaining 5 values
for each grid cell. We consider the median of these 5 values as representa-
2.2.1 Reference climatology

A first climatology, named reference climatology, is calculated with the same procedure as the reference climatology, but with three differences. First, we include pixels close to

tive of the mean chlorophyll for the 1-8 January 2018-2022 time step. If, for a given grid cell, more than \( \beta = 60\% \) of the 2018-2022 values are flagged as NA, we flag that cell to NA. In this way, we obtain the first step (1-8 January) of the 2018-2022 climatological time series (see Supplementary Fig. S.14 for a map of the percentage of values available).

Step 4 is repeated for each of the time steps composing a climatological year (each time step is composed of 8 days: 1-8 January; 9-16 January, etc.)

5. For each grid cell, if the CHL climatological time series has less than \( \gamma = 25\% \) of the available values, we exclude that grid cell from the analysis. Similarly, if the CHL climatological time series shows temporal gaps of more than \( \epsilon = 4 \) weeks, we discard that grid cell. Otherwise, we remove any existing temporal gaps using a monotonic splice interpolation (pchip_interpolate function in the python package scipy). This type of interpolation is chosen because it performs a non-linear interpolation and, at the same time, does not create artificial peaks in the time series.

6. Each climatological CHL time series is smoothed using a running mean over 40 days (i.e. 5 time steps).

7. We shift the climatological CHL time series located in the Southern hemisphere of 6 months, so that Austral and Boreal summers (winter) occur in June-August (December-February)

8. Finally, we normalise each climatological CHL time series by dividing it by its annual climatological maximum value. Thus, each climatological CHL time series has a maximum value equal to 1. We refer to these as normalised time series and we indicate the normalised chlorophyll as \( nCHL \). We distinguish them from the non normalised time series that have real CHL concentrations, and which are used for the comparison with real CHL values measured by BGC-Argo floats, as detailed in the subsections below. To further distinguish them from the BGC-Argo time series, we refer to them also as satellite time series.

2.2.2 Extended climatology

A second climatology, named extended climatology, is calculated with the same exact procedure but with three differences. First, we include pixels close to

<table>
<thead>
<tr>
<th>Time period</th>
<th>Reference climatology</th>
<th>Extended climatology</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018–2022</td>
<td>50 %</td>
<td>25 %</td>
</tr>
<tr>
<td>1998–2022</td>
<td>60 %</td>
<td>20 %</td>
</tr>
<tr>
<td>1998–2022</td>
<td>75 %</td>
<td>50 %</td>
</tr>
<tr>
<td>ε</td>
<td>4 weeks</td>
<td>14 weeks</td>
</tr>
</tbody>
</table>

Table 1: List of parameters (rows) used for the calculation of each climatology (columns).
the shore. Second, we adopt different parameters (Table 1) which reduce the number of values flagged as NA, allowing us to extend the global coverage of the climatology to higher latitudes. Third, we use CHL data between 1998 and 2022.

### 2.3 Partition method for the reference climatology

The time series of the reference climatology are partitioned using the $k$-means method (Hartigan and Wong, 1979; Ahmed et al., 2020). This method is chosen as it has already been successfully applied to a variety of case studies including clustering of temporal series of satellite CHL (D’Ortenzio and Ribera d’Alcalà, 2009; Marchese et al., 2019; Magot et al., 2016; Kheireddine et al., 2021; Ardyna et al., 2017; D’Ortenzio et al., 2012). The $k$-means is an unsupervised machine learning technique which groups together time series so that the distance among them is minimum, while maximising the distance between different groups. As distance, the Euclidean distance is used. Given two time series of $n$ time steps $t_1 = \{a_1, a_2, ..., a_n\}$, and $t_2 = \{b_1, b_2, ..., b_n\}$, the Euclidean distance between $t_1$ and $t_2$ is:

$$\sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$  \hspace{1cm} (1)

The $k$-means algorithm needs one parameter, namely the number of clusters $k$ in which the time series should be regrouped.

At the end of $k$-means process, each time series of the reference climatology is assigned to one of the $k$ clusters, meaning that each pixel of the climatology is assigned to the obtained cluster. This gives a geographical distribution of the memberships which we call hereafter *bioregion* following D’Ortenzio and Ribera d’Alcalà (2009).

In summary, for each of the $k$ clusters obtained from the $k$-means partition, we obtain a spatial distribution (that we assume and call *bioregion sensu* D’Ortenzio and Ribera d’Alcalà 2009) and a reference time series (see below Subsec. 2.5 for definition).

### 2.4 Identification of the optimal number of clusters

Different tests (Elbow, Silhouette, Calinski-Harabasz) are used to determine the optimal number of clusters of the reference climatology (D’Ortenzio and Ribera d’Alcalà 2009; Supplementary Material S.2 and S.2.1). We identify the number of optimal clusters applying these techniques to the reference climatology (rather than to the extended one) due to the higher accuracy of this climatology. Indeed, the stricter criteria used to calculate it (compared to the criteria used for the other two climatologies, Table 1) make its time series less noisy and more robust.

The results of the tests vary between 5 and 9, although all (except Calinski-Harabasz which indicates 3 clusters and was discarded as its partition resulted too broad for the purposes of this paper, Supplementary Fig. S.5) converged on 7 and 8. We finally decided on 7 after visual inspection of the time series, which suggests that two of the 8 times series, although considered statistically
different, show very close seasonal traits (see Supplementary Material S.2.1 and Supplementary Fig. S.8). We also stress that this number is consistent (i) with the study of D’Ortenzio and Ribera d’Alcalà (2009, 7 clusters), Mignot et al. (2023, 8 clusters), which was based on a \( k \)-means partition as well, and with the global study of Bock et al. (2022, 6 clusters); (ii) when changing the resolution of the grid over which the climatology is calculated (Supplementary Fig. S.9); (iii) with the result obtained without shifting time series in the Southern hemisphere of 6 months (Supplementary Fig. S.10).

2.5 Partition method for the extended climatology

Here we describe the methodology used to partition the extended climatology. We do not use the \( k \)-means method (used with the reference climatology) as it does not allow us to identify an optimal number of clusters when directly applied to the extended climatology. This is because of the presence, close to the shore and at high latitudes, of time series with limited data quality or gaps. The method used to partition the extended climatology employs the 7 reference time series together with their relative thresholds of acceptance. These are calculated as follows. First, we consider all the \( n_1 \) time series belonging to bioregion 1. The average of these time series is considered as representative of the reference time series of bioregion 1. We then calculate, for all the \( n_1 \) time series belonging to bioregion 1, the Euclidean distance with the reference time series, obtaining \( n_1 \) values. The threshold of acceptance of bioregion 1 is defined as the maximum of the \( n_1 \) distances. This is repeated for all the 7 bioregions, resulting in 7 reference time series and corresponding thresholds of acceptance. For each time series of the extended climatology we calculate the Euclidean distance with the 7 reference time series, obtaining 7 values. The minimum of these values indicates the final classification of the time series of the extended climatology; providing that this minimum is lower than the corresponding threshold of acceptance. Otherwise, no bioregion is assigned (only less than 0.3% of the time series of the extended climatology; Supplementary Table S.1).

Overall, the percentage of surface covered by each bioregion does not change consistently (Supplementary Table S.1). Then, since the extended climatology covers a domain which is larger than the one covered by the reference climatology, we only consider the grid cells in which both climatologies are defined (i.e. in which both time series are not flagged as NA). We find that 72.9% of the pixels of the extended climatology have been assigned to the same bioregion of the reference climatology (Supplementary Fig. S.4a and c). The difference is explained by the use of a different number of years to calculate the two climatologies (2018-2022 period for the reference climatology, 1998-2022 period for the extended climatology; Supplementary Material S.3 and Supplementary Fig. S.4a and b), and is not due to the partition method. In addition, we calculate the Euclidean distance between the time series of the extended climatology and the reference time series of the bioregion it should belong to (which can differ from the cluster it has been assigned to). We find that 97.3% of the time series of the extended climatology are within the threshold of acceptance of the corresponding reference time series. For each bioregion obtained using the extended climatology, we calculated its mean time series: they are all significantly similar to the corresponding reference time series (Supplementary Fig. S.6). When the
silhouette value for each bioregion of the extended climatology is calculated, positive values are obtained for almost all the time series (Supplementary Fig. S.14). Overall, these results indicate that the method used to partition the extended climatology does not affect the pattern of the bioregions significantly, and corroborate our choice of extending our analyses to the coastal and polar regions.

2.6 BGC-Argo data

2.6.1 BGC-Argo data processing

BGC-Argo data were collected and made freely available by the International Argo Program and the national programs that contribute to it (https://argo.ucsd.edu, https://www.ocean-ops.org; Argo (2020)). The Argo Program is part of the Global Ocean Observing System. BGC-Argo are autonomous drifters which sample the water column and measure state variables including temperature, pressure, and salinity. In addition, different sensors allow the drifter to measure biogeochemical variables such as CHL, oxygen, particulate backscattering (b_p), nitrates, pH, and downwelling irradiance. BGC-Argo data are subjected to several quality control procedures (Schmechtig et al., 2023, 2015; Thierry et al., 2022; Johnson et al., 2023). First, raw data are converted to “real time” state variables and outliers are identified and flagged. Second, automated algorithms calibrate real time data, providing “adjusted” data. Finally, adjusted data are verified and validated by an expert, providing “delayed” mode data. As the latter data have only a limited coverage, we considered data both in “real time”, “adjusted”, and “delayed” mode, following the methodological approach of Mignot et al. (2023). However, we exclude from each profile all the values classified as “probably bad data”, “bad data”, or “missing value” (∼30% of the profiles were integrally excluded in this way).

We only select BGC-Argo floats which measured CHL, excluding CHL values larger than 40 mg/m^3 from that profile, obtaining a total of 84186 profiles across the global ocean. The use of “real time” data allowed us to augment the spatial and temporal coverage, although the quality of the CHL evaluation will be certainly impacted. In any case, these data only represent ∼11% of the profiles for CHL (Stoer et al., 2023).

For each BGC-Argo profile, we calculate the CHL averaged in the first optical depth, which is usually considered as representative of the CHL measured by satellite observations (Morel and Berthon, 1989). The first optical depth was calculated as in Morel and Berthon (1989).

2.6.2 BGC-Argo time series

In order to compare CHL observations from BGC-Argo with the CHL time series obtained from satellite data, we define a BGC-Argo time series as follows:

- For each bioregion, we select all the BGC-Argo profiles which are collected within that bioregion.
- For each 8-day time step (e.g. 1-8 January, or 9-16 January, etc.), we select all the BGC-Argo profiles measured in correspondence of that time step.
step and we consider their CHL averaged in the first optical depth. We calculate the median of these values and we repeat this for each time step, generating a final time series of 46 values.

- A running average over a 40 days windows (i.e. 5 time steps), is finally carried out.

Furthermore, we calculate other diagnostics:

- For each time step (e.g. 1-8 January, or 9-16 January, etc.), we count the number of BGC-Argo profiles available, obtaining a time series of 46 values. We then consider the median, minimum, and maximum of this time series (referred to as “median #BGCprof ts”, “min #BGCprof ts”, “max #BGCprof ts” respectively, in Supplementary Table S.3)

- For all the \( N_{GC} \) grid cell of a bioregion, we count the number of BGC-Argo profiles within it, obtaining \( N_{GC} \) values. We then consider the mean, median, and maximum value of this \( N_{GC} \) values (referred to as “# BGC prof/cell (mean)”, “# BGC prof/cell (median)”, “# BGC prof/cell (max)” respectively, in Table 3 and Supplementary Table S.3)

- For each bioregion, we calculate (i) the spatial area, expressed both in terms of pixels, and squares of 100 km size; (ii) the percentage of surface covered by the bioregion, expressed as the ratio between the number of pixels composing it and the total number of oceanic pixels multiplied by 100; (iii) the percentage of BGC-Argo profiles which are in that cluster compared to all the BGC-Argo profiles available (“Percentage of all BGC profiles” in Table 3).

2.7 Comparison of BGC-Argo and satellite time series

The similarity between the mean time series of the cluster and the BGC-Argo time series is tested in two ways:

- We verify that both time series are normally distributed, then we perform a t-test (Supplementary Material S.4; Student 1908)

- We calculate the Euclidean distance between the two time series and we compare it with the threshold of acceptance of the reference cluster. If the Euclidean distance is lower, then we consider the BGC time series as significantly similar to the mean time series of the cluster. This second test is applied only to compare normalised time series. In this case, before applying the test, the BGC time series are normalised by dividing them by their maximum.

3 Results

3.1 Partition of the global ocean

The global ocean is partitioned in \( k=7 \) bioregions (Fig. 1). The number of bioregions is chosen using multiple statistical tests (Methods and Supplementary Material S.2 and S.2.1). The pattern of the partition is robust with respect
Figure 1: Regionalisation of the global ocean in $k=7$ bioregions (upper panel) obtained from the 1998-2022 extended climatology. In the lower panels, each plot represents the mean normalised nCHL time series of a specific bioregion (i.e. average of all the time series belonging to that bioregion, thick solid lines) along with their uncertainty (thin solid lines, standard deviation) identified by a different colour (right-hand legend in the upper panel).
Table 2: Number of BGC-Argo profiles in the 7 bioregions along with their mean satellite and BGC-Argo CHL values (and their respective standard deviation), and results of the comparison (t-test, Methods).

<table>
<thead>
<tr>
<th>Bioregion</th>
<th># BGC profiles</th>
<th>Satellite chl, mean [mg/m³]</th>
<th>Satellite chl, std [mg/m³]</th>
<th>BGC-Argo chl, mean [mg/m³]</th>
<th>BGC-Argo chl, std [mg/m³]</th>
<th>t test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13889</td>
<td>0.15</td>
<td>0.34</td>
<td>0.28</td>
<td>0.23</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>15550</td>
<td>0.12</td>
<td>0.25</td>
<td>0.13</td>
<td>0.15</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>15738</td>
<td>0.13</td>
<td>0.20</td>
<td>0.17</td>
<td>0.16</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>5636</td>
<td>0.27</td>
<td>0.35</td>
<td>0.33</td>
<td>0.37</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>13197</td>
<td>0.31</td>
<td>0.70</td>
<td>0.44</td>
<td>0.32</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>9487</td>
<td>0.46</td>
<td>0.61</td>
<td>0.50</td>
<td>0.52</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>15994</td>
<td>0.47</td>
<td>0.82</td>
<td>0.50</td>
<td>0.65</td>
<td>yes</td>
</tr>
</tbody>
</table>

3.2 Global comparison with BGC-Argo observations

Figure 2 shows the positions of the available BGC-Argo profiles. The CHL time series obtained from the two data sets are plotted separately for each bioregion in the lower panels. Note that satellite data are re-transformed from nCHL to CHL, implying a larger variability (standard deviation 2.7 times larger than in the normalised case). Table 2 also summarises statistics of the different bioregions for the two data sets.

The spatial distribution of BGC-Argo profiles is not homogenous, resulting in a
Figure 2: Comparison of CHL time series from satellite and BGC-Argo. The CHL time series are non-normalised (i.e. real values). The upper panel shows the partition of the global ocean in $k=7$ bioregions as reported in Fig. 1 with the black dots indicating the position of the 84186 BGC profiles used in the present study (Methods). In the lower panels, each plot represents (i) the mean CHL time series of the bioregion from satellite data as coloured solid lines (thicker coloured solid line: average time series; thinner coloured solid line: uncertainty, expressed as standard deviation; colours in the right-hand legend of the upper panel); and (ii) the BGC-Argo time series (black dots: median of chlorophyll of that time step; black solid line: 40-days running average of the black dots, Methods).
number of BGC-Argo profiles per bioregion highly variable (from 5636 profiles in bioregion 4 to 15994 profiles in bioregion 7). Overall, satellite and BGC-Argo mean CHL values are close except for bioregion 1, which exhibits BGC-Argo CHL almost double that of satellite CHL.

The shapes of the satellite and BGC-Argo time series are similar for all the bioregions considered (see also the comparison of normalised time series in Supplementary Fig. S.7), with six out of the seven bioregions successfully passing the t-test. Some differences are observed in bioregion 5, which shows a CHL increase observed by BGC-Argo time series in summer but not in satellite data. In bioregion 6, the bloom initiation detected in satellite and BGC-Argo data temporally matches, while the annual maximum is delayed by about one month in BGC-Argo time series. In bioregion 7, the mid-summer increase in CHL is smaller in the satellite time series than in the BGC-Argo data.

3.3 Analysis of variability of satellite time series

In general, the CHL time series have large variabilities (the standard deviation of the time series represents, on average, the 53% of their mean values against 20% for the normalised case; Table 2 and thin solid coloured lines in Fig. 1 and Fig. 2). To deepen the comparison between the satellite and BGC-Argo time series, we proceed to an additional analysis.

First, we hypothesised that the high seasonal variability (a large standard deviation of the time series), is the result of grouping temporal series which are geographically distant. For example, bioregion 5 is observed in Equatorial and polar regions, with seasonal mean values potentially different. Indeed, our partition method, which aims to identify different phenological regimes, is based on the use of nCHL time series. This can result in bioregions that, while sharing the same shape of the nCHL, are located in regions known to have mean values of CHL largely different.

To solve this point, each bioregion is split in a sub-bioregion at high latitudes (HL) by considering only time series obtained in pixels located poleward than 38° latitude, and in a sub-bioregion at low latitudes (LL; points located around the equator lower than 38° latitude). In that way, bioregion 1 is split into sub-bioregion 1 LL and sub-bioregion 1 HL; the same for bioregion 2, 3, etc.

Another source of the variability, likely artifactual, derives from the irregular distribution of BGC-Argo profiles, which do not cover the global ocean homogeneously, potentially inducing biases in the satellite BGC-Argo comparison.

Therefore, to calculate the mean satellite time series for each sub-bioregion, we consider only pixels located less than 1° from a BGC-Argo profile.

The results confirm that the agreement between satellite and BGC-Argo time series is good, but not valid everywhere (Fig. 3 and Table 3). In particular, 8 out of 14 sub-bioregions show a significant agreement between satellite and BGC-Argo time series (sub-bioregions 1, 3, 4, and 5 at low latitudes, and sub-bioregions 2, 3, 4, and 7 at high latitudes), while 6 do not (sub-bioregions 2, 6, and 7 LL, and sub-bioregions 1, 5, and 6 HL).

Subsequently, we calculate different properties for each LL and HL sub-bioregion (Table 3 and Subsec. 2.6.2). In general, satellite and BGC-Argo time series are significantly different in sub-bioregions where the BGC-Argo sampling density is low (less than 5 BGC-Argo profiles per grid cell; Supplementary Fig. S.2), while
Figure 3: Bioregions split into high and low latitude sub-bioregions. The upper panel shows the partition of the global ocean in $k=7$ bioregions, as reported in Fig. 1. The white lines show the $38^\circ\mathrm{N}$ and $S$ latitude used to split the bioregions. The 7 lower panels on the left (first two columns) show the low latitude sub-bioregions, while the 7 lower panels on the right (third and fourth columns) show the high latitude sub-bioregions. Each plot represents (i) the mean CHL time series of the sub-bioregion non-normalised calculated considering only grid cells located less than 1° from a BGC-Argo profile, as coloured solid lines (colours in the right-hand legend of the upper panel). (ii) The BGC-Argo time series (grass green dots: median of chlorophyll of that time step; grass green solid line: 40-days running average of the black dots, Methods). The title of each panel indicates the number of the bioregion, the percentage of cells belonging to it (in brackets), and whether the comparison between the satellite and BGC-Argo time series is significant or not (t-test=“yes”, Methods).
they agree where the BGC-Argo sampling density is large (6 BGC-Argo profiles per cell). This is particularly visible in sub-bioregion 6 LL, where a potentially erroneous double peak in the BGC time series can be explained by one of the lowest density of BGC profiles there (5.0). On the other hand, the remarkable agreement found in sub-bioregion 3 HL can be due to the large BGC-Argo sampling density there (34.5, the highest).

When focusing on the sub-bioregions where BGC-Argo and satellite time series do not match, different categories of discrepancy can be qualitatively identified (Table 3). (i) In some sub-bioregions, a strong peak is present in the BGC-Argo time series while, in the satellite time series, the amplitude is much less pronounced. Furthermore, on average, BGC-Argo time series shows values which are almost two-fold larger than satellite time series. It is the case of sub-bioregions 1, 5 HL and, notably, also of the portion of bioregion 1 in the Southern Pacific (Supplementary Material S.6 and Supplementary Fig. S.1). A similar discrepancy is found in sub-bioregion 6 HL with the difference that, there, the satellite time series shows a more visible peak. We define sub-bioregions with these characteristics as belonging to the category “similar shapes, different amplitudes and mean values” (SS, DA&MV). (ii) Conversely, in some regions, BGC-Argo time series show constantly lower values than satellite ones and similar shapes and amplitudes (sub-bioregion 2 LL). We define this category as “similar shapes and amplitudes, different mean values” (SS&A, DMV). Finally, (iii) in some sub-bioregions, the low number of BGC-Argo profiles is not sufficient to produce a reasonable BGC-Argo time series. It is the case of sub-bioregions 6 and 7 LL in which only 0.17% and 1.73% of all BGC profiles were carried, respectively. We define this category as “low BGC-Argo sampling” (LBGCS).

4 Discussion

The obtained regionalisation of the global ocean (Fig. 1) compares well with the regionalisation proposed by Longhurst (1995) and D’Ortenzio and Ribera d’Alcalà (2009, Supplementary Material S.5). In particular, bioregion 1 corresponds to model number 4 of Longhurst (Tropics), while bioregion 2 closely matches Tropical bioregion of D’Ortenzio and Ribera d’Alcalà (2009). Bioregion 3 corresponds to model 3 of Longhurst (Subtropical winter nutrient-limited), while bioregion 4 is similar to the Bloom bioregion of D’Ortenzio and Ribera d’Alcalà (2009). Bioregion 5 shares characteristics of both model 1 (Polar irradiance-limited) and 2 (midlatitude nutrient-limited spring production peak) of Longhurst. Finally, both bioregions 6 and 7 correspond to model number 1 of Longhurst (1995), with bioregion 6 characterised by a stronger decline in CHL. The pattern of our regionalisation compares also well with the one obtained by Fay and McKinley (Supplementary Material S.5), in particular concerning bioregions 2, 3, and 4. Differences are present in the Southern Ocean where the biomes found by Fay and McKinley are organised in latitudinal bands, while the disposition of our bioregions is less regular. A good qualitative agreement is also found between our partition and the one by Bock et al. (2022, Supplementary Material S.5). Larger differences are found with the regionalisation of Mignot et al. (2023), possibly due to the very large number of parameters used in that
<table>
<thead>
<tr>
<th>Bioregion</th>
<th>Satellite chl, mean [mg/m³]</th>
<th>Satellite chl, amplitude [mg/m³]</th>
<th>BGC-Argo chl, mean [mg/m³]</th>
<th>BGC-Argo chl, amplitude [mg/m³]</th>
<th>Surface [100*100 km²]</th>
<th># BGC prof/cell (mean)</th>
<th>Percentage cells covered [%]</th>
<th>Percentage of all BGC prof [%]</th>
<th>t test</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1LL</td>
<td>0.11</td>
<td>0.01</td>
<td>0.09</td>
<td>0.03</td>
<td>6293</td>
<td>6.3</td>
<td>15.64</td>
<td>10.22</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>1HL</td>
<td>0.15</td>
<td>0.02</td>
<td>0.39</td>
<td>0.34</td>
<td>1830</td>
<td>3.6</td>
<td>6.77</td>
<td>5.33</td>
<td>no</td>
<td>SS, DA&amp;MV</td>
</tr>
<tr>
<td>2LL</td>
<td>0.09</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>7148</td>
<td>7.2</td>
<td>18.5</td>
<td>13.41</td>
<td>no</td>
<td>SS&amp;A, DMV</td>
</tr>
<tr>
<td>2HL</td>
<td>0.26</td>
<td>0.15</td>
<td>0.25</td>
<td>0.19</td>
<td>471</td>
<td>14.4</td>
<td>1.61</td>
<td>3.71</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>3LL</td>
<td>0.11</td>
<td>0.12</td>
<td>0.09</td>
<td>0.20</td>
<td>6620</td>
<td>7.6</td>
<td>17.95</td>
<td>14.32</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>3HL</td>
<td>0.18</td>
<td>0.21</td>
<td>0.19</td>
<td>0.34</td>
<td>201</td>
<td>34.5</td>
<td>0.64</td>
<td>2.87</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>4LL</td>
<td>0.16</td>
<td>0.20</td>
<td>0.22</td>
<td>0.45</td>
<td>1338</td>
<td>3.8</td>
<td>3.69</td>
<td>2.51</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>4HL</td>
<td>0.21</td>
<td>0.27</td>
<td>0.29</td>
<td>0.48</td>
<td>999</td>
<td>5.8</td>
<td>3.55</td>
<td>3.69</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>5LL</td>
<td>0.20</td>
<td>0.13</td>
<td>0.10</td>
<td>0.10</td>
<td>1108</td>
<td>5.7</td>
<td>2.77</td>
<td>1.58</td>
<td>yes</td>
<td>M</td>
</tr>
<tr>
<td>5HL</td>
<td>0.19</td>
<td>0.09</td>
<td>0.45</td>
<td>0.60</td>
<td>2977</td>
<td>4.2</td>
<td>11.23</td>
<td>12.91</td>
<td>no</td>
<td>SS, DA&amp;MV</td>
</tr>
<tr>
<td>6LL</td>
<td>0.55</td>
<td>0.77</td>
<td>0.30</td>
<td>0.64</td>
<td>148</td>
<td>5.0</td>
<td>0.39</td>
<td>0.17</td>
<td>no</td>
<td>LBGCS</td>
</tr>
<tr>
<td>6HL</td>
<td>0.32</td>
<td>0.63</td>
<td>0.53</td>
<td>1.21</td>
<td>1177</td>
<td>11.2</td>
<td>5.04</td>
<td>10.48</td>
<td>no</td>
<td>M*</td>
</tr>
<tr>
<td>7LL</td>
<td>0.36</td>
<td>0.51</td>
<td>0.22</td>
<td>0.37</td>
<td>816</td>
<td>6.4</td>
<td>2.02</td>
<td>1.73</td>
<td>no</td>
<td>LBGCS</td>
</tr>
<tr>
<td>7HL</td>
<td>0.30</td>
<td>0.38</td>
<td>0.49</td>
<td>1.04</td>
<td>2378</td>
<td>6.9</td>
<td>10.19</td>
<td>17.08</td>
<td>yes</td>
<td>M</td>
</tr>
</tbody>
</table>

Table 3: Properties of the bioregions split in high and low latitudes sub-bioregions, results of the comparison between the BGC-Argo and satellite time series (t test), and classification of the sub-bioregions in four categories: (i) Similar Shape, Different Amplitude and Mean Value (SS, DA&MV); (ii) Similar Shape and Amplitude, Different Mean Value (SS&A, DMV); (iii) Low BGC-Argo sampling (LBGCS); (iv) Match (M). Sub-bioregion 6HL is classified as Match due to the good qualitative agreement between BGC-Argo and satellite chlorophyll measurements (Supplementary Table S.3).
Figure 4: Summary of satellite and BGC-Argo time series agreement. Regions where both time series match (M) are shown in green. Regions where the density of BGC-Argo profiles is too low to permit comparison with satellite time series (LBGCS) are shown in yellow. Regions where satellite and BGC-Argo time series present similar shapes and amplitudes, but different mean values, are shown in orange (SS&A, DMV). Regions where satellite and BGC-Argo time series share similar shapes but different amplitudes and mean values are shown in red (SS, DA&MV). Black dots show the sub-bioregion 6 HL, for which we found a good qualitative agreement between BGC-Argo and satellite chlorophyll measurements. However, attention must be paid when comparing BGC-Argo and satellite chlorophyll data in that region.
work and to the use of model rather than satellite outputs.

The good qualitative matchup of our findings with previous works is important especially in the light of the fact that we base our regionalisation uniquely on the shape of the CHL time series, and not on a larger number of assumptions and variables. This highlights how CHL phenological cycles are the expression of multiple marine processes. More importantly, previous works were able to match the shape of the satellite and BGC-Argo time series, but they obtained very different mean CHL values (D'Ortenzio et al., 2020). Here, to our knowledge, we show for the first time an agreement not only from the point of view of the seasonal shape, but also with respect to the mean CHL values. Furthermore, the partition is robust with respect to changes in the size of the grid cell used to calculate the climatologies and to the use of the shift of 6 months applied to the Southern hemisphere time series. The agreement between BGC-Argo and satellite CHL is partially explained by a large variability associated with the satellite time series. To investigate the processes underpinning this variability, we have split bioregions in high and low latitude sub-bioregions and considered only pixels located in proximity (less than 1°) of a BGC-Argo profile to calculate mean satellite time series. We find agreement in 8 out of 14 sub-bioregions, corresponding to 58% of our domain (reported in green in Fig. 4). In general, we find that for a BGC-Argo sampling density larger than 5 profiles per grid cell, the likelihood of an agreement between satellite and BGC-Argo time series is larger. This implicitly suggests a way to determine whether the BGC-Argo sampling density in a given bioregion is, to date, good enough. An exception to this scenario is provided by bioregion 4 LL, which shows agreement of satellite and BGC-Argo time series despite having only 3 BGC-Argo profiles per grid cell. Conversely, further BGC-Argo float deployments are needed in bioregion 2 LL, despite a relatively large BGC-Argo sampling density there (∼6 BGC profiles per cell).

We qualitatively identify three types of discrepancy between BGC-Argo and satellite time series.

- **Similar shapes, different amplitudes and mean values**: These regions are characterised by a prominent CHL enhancement observed by BGC-Argo during summer, which is observed by satellite as well but with a lower amplitude. In addition, the CHL values observed by BGC-Argo floats are consistently larger than the ones observed by satellite across the year. These sub-bioregions are found at high latitudes (sub-bioregions 1, 5, 6 HL, and also bioregion 1 considered exclusively in the Southern Pacific, Supplementary Material S.6 and Supplementary Fig. S.1) and present a relatively large number of BGC-Argo profiles (5%, 13%, and 11% of all profiles, respectively). Thus, the mismatch is unlikely due to insufficient BGC-Argo sampling (in addition, in sub-bioregion 6 HL the density of BGC-Argo profiles is among the highest: 10 profiles per grid cell). We suggest that this mismatch could rather be due to well-known problems in the conversion of fluorescence measured by the floats into CHL Behrenfeld et al. (2006, 2009); Bock et al. (2022). Notably, this discrepancy has already been reported in the region of sub-bioregion 1 HL, in particular in the Southern Pacific (Mignot et al. 2023; Bock et al. 2022, Supplementary Fig. S.1). Another possibility is the scarce number of years for which
observations are available in satellite CHL images at high latitudes (Supplementary Fig. S.14); hence, the climatological satellite time series may have missed important productivity events, which may explain the low value of the satellite-derived CHL during the annual maximum in summer. Hence, we stress that attention must be used when analysing CHL in these regions, particularly in the Southern Ocean during Austral summer.

For these reasons, we report them in red in Fig. 4. When considering all the time series composing sub-bioregion 6 HL (and not only those closer than 1° to a BGC-Argo profile), satellite and BGC-Argo time series agree (Supplementary Table S.3). For this reason, Sub-bioregion 6 HL is shown in green (with black dots), and leads to an agreement between satellite and BGC-Argo time series in 61% of our domain.

• **Similar shapes and amplitudes, different mean values**: In these regions (sub-bioregion 2 LL), mainly located in a band around 25°N and 25°S, BGC-Argo and satellite time series show a similar phenology, both in terms of shape (t test significant, Supplementary Table S.3) and amplitude of the time series (Table 3). However, the chlorophyll reported by BGC-Argo floats is constantly lower than the one measured by satellite. This discrepancy could be due to the overestimation of CHL by satellite at low latitudes as reported in previous studies (Omta et al., 2009; Clow et al., 2024). Conversely, as the density and number of BGC-Argo profiles are among the largest there, the BGC-Argo coverage seems not to explain this difference. Hence, these regions are reported in orange in Fig. 4.

• **Insufficient BGC-Argo sampling**: In sub-bioregions 6 and 7 LL, mainly located in the Arabian Sea and close to Amazon delta, the low number of BGC-Argo profiles did not allow us to calculate a robust BGC-Argo time series. This likely explains the mismatch with satellite measurements. These results suggest the need for further BGC-Argo deployments in these regions. Hence, sub-bioregions 6 and 7 LL are reported in yellow in Fig. 4.

5 Conclusions

Our results point to the fact that the agreement between CHL observed by satellite and BGC-Argo floats is valid in the majority of the global ocean (~58-61% of the surface area), but not everywhere. Our method provides a novel way to determine regions where satellite and BGC-Argo observations agree, and regions in which they do not, based on the phytoplankton phenology described with annual cycles of CHL. The reasons for this mismatch are likely due to poor satellite performance (in particular in sub-bioregion 2 LL) or to problems in the relationship used to convert fluorescence to CHL in BGC-Argo floats (sub-bioregions 1, 5, and 6 HL). Our methods also allow us to identify regions lacking BGC-Argo data (sub-bioregions 6 and 7 LL). More generally, we suggest larger BGC-Argo float deployments in sub-bioregions showing discrepancy between satellite and BGC-Argo time series. In conclusion, we stress that attention must be paid when coupling together satellite and BGC-Argo CHL observations in these regions (e.g. to constrain biogeochemical models or for marine spatial planning),
as they cover ∼40% of the global ocean. The framework used in the present study can be applied in the comparison of other biogeochemical variables.

Acknowledgements

AB, VT, and FDO received support from the CNES/TOSCA project SEASON. This study has been conducted using E.U. Copernicus Marine Service Information (product OCEANCOLOUR_GLO_BGC_L3_MY_009_103; https://doi.org/10.48670/moi-00280). The authors would like to thank the International Argo and BGC-Argo Program, as well as the NASA and Mercator Ocean agencies. The authors are grateful to Inés Jambou for her help in the design of Figs 1–3 and the different tables, as well as Elodie Gutknecht, Coralie Perruche, and Giulio Tirabassi for their useful advise.

Data availability

All the data necessary to reproduce all the plots shown in the present paper will be published in zenodo upon publication.
S.2 Methods to identify the optimal number of clusters

The methods used to calculate the optimal number of clusters are:

- **Elbow method** The elbow method (Sugar, 1998; Marutho et al., 2018) is based on the calculation of the total within-clusters sum of squares. In other words, for a given cluster we calculate the average of the Euclidean distances between each time series belonging to the cluster and the mean time series of that cluster. This is repeated for all the clusters, resulting in \( k \) values (one for each cluster), which are summed together. This provides the total within-clusters sum of squares for the partition in \( k \) clusters, which is considered representative of the total variance of that partition. The elbow method consists in identifying the value \( k \) for which the variance stops decreasing, i.e. the addition of a supplemental cluster does not bring any novel information.

- **Silhouette method** The silhouette method (Rousseeuw, 1987) measures the similarity of the time series within the same cluster and compares it with the separation between clusters. The silhouette value varies between -1 and 1 and it indicates, for each time series, whether it was appropriately assigned to the corresponding cluster (silhouette value tending to 1) or not (silhouette values tending to -1). We use the silhouette value averaged over all the time series, and we calculated this metric for \( k \) between 2 and 25, similary to Kheireddine et al. (2021).

- **Calinski-Harabasz method** This method (Caliński and Harabasz, 1974) identifies the optimal number of clusters which maximises the cohesion of the time series within the same cluster, while at the same time maximising the separation of different clusters. This method relies upon distance from the mean time series of each cluster and the mean global time series.

S.2.1 Detailed identification of the optimal number of clusters

The elbow method indicates an optimal number of clusters between 5 and 9. The silhouette analysis and the Calinski-Harabasz index indicate an optimal number of clusters of 3 (Supplementary Fig. S.12 and S.13). However, this solution was discarded because it separates the global ocean in large, well-known regions (an Equatorial-Tropical band, a temperate band, and a polar band, Supplementary Fig. S.5). In fact, in the present study, in order to identify agreement and differences between BGC-Argo and satellite chlorophyll observations, we focus on smaller regions. Furthermore, the silhouette analysis has a local maximum at 6 and 7 clusters. This validity of this value is confirmed by a supplementary analysis. When passing from 7 to 8 clusters, the first cluster is split in two, while the other clusters remain unchanged (Supplementary Fig. S.8). The two mean time series obtained from the split of the first cluster are consistently similar to the mean time series of the first cluster. Conversely, when passing from 5 to 6 clusters (or from 6 to 7), a novel time series is introduced, increasing the variance explained by the partition.
S.3 Third climatology (extended climatology with a short time frame)

The third climatology (referred also as “extended climatology with a short time frame”) is calculated with the same exact procedure and parameters used for the extended climatology. The only difference is that we use L3 chlorophyll fields of 2018-2022 time period only (and not the 1998-2022 time period, Table S.2).

This climatology is partitioned using the reference time series of the reference climatology (analogously to what has been done to partition the extended climatology, Section 2.5). The same tests used to compare the partitions obtained from the reference and extended climatologies are applied. The percentage of surface covered by each bioregion does not change consistently when considering the partition with the reference or with the third climatology (Supplementary Table S.1). Furthermore, 93.3% of the time series of the third climatology are assigned to the same bioregion than the reference climatology (Supplementary Fig. S.4a and c). The improvement compared to the extended climatology (for which only 72.9% of the time series were correctly assigned) can be explained by the fact that the reference and third climatology are calculated using data from 25 years (1998-2022 period) compared to the extended climatology (based on 5 years only: 2018-2022 period). Hence, this difference is due to temporal changes and not the partition method. All in all, 97.6% of the time series of the third climatology are within the threshold of acceptance of the corresponding reference time series. These findings indicate that the 7 main time series of the third climatology are consistently similar to the ones of the reference and extended climatologies, and corroborate our choice of extending our analyses to the coastal and polar regions.

S.4 T test detailed description

After verifying that both time series (BGC and satellite ones) are normally distributed, we perform a t-test (Student (1908)) using the respective mean and standard deviations.

\[ t = \frac{|x_{BGC_i} - x_{Sat_i}|}{\sqrt{\frac{\sigma^2_{BGC_i}}{n_{BGC_i}} + \frac{\sigma^2_{Sat_i}}{n_{Sat_i}}}} \]  

where \( x_{BGC_i} \) is the average of the BGC-Argo time series in the \( i \)-th cluster, \( x_{Sat_i} \) is the average of the satellite time series in the \( i \)-th cluster, \( \sigma^2_{BGC_i} \) and \( \sigma^2_{Sat_i} \) are the variances of the two time series, and \( n_{BGC_i} \) and \( n_{Sat_i} \) are the number of elements in each time series. The results of this test is compared with the reference value at 5% of significance obtained using \( n_{BGC_i}+n_{Sat_i}-2 \) degrees of freedom. If smaller than the reference value, the time series are considered statistically similar.
S.5 Comparison of our partition with previous works

• Comparison of time series of Figure 1 with previous works.
  
  – The 7 mean time series identified compare well with the phenological classification of Longhurst (1995) and D’Ortenzio et al. (2012). Bioregion 1 corresponds to model number 4 of Longhurst (1995, Tropics), while bioregion 2 shares characteristics of both model 4 and 3 (Subtropical winter nutrient-limited) and closely matches D’Ortenzio et al. (2012, Tropical bioregion). Bioregion 3 corresponds to model 3, while bioregion 4 is similar to Bloom bioregion of D’Ortenzio et al. (2012). Bioregion 4 is also similar to model 2 of Longhurst (1995) but is characterised by a sharper spring peak. Bioregion 5 is similar to both model 1 (Polar irradiance-limited) and 2 (midlatitude nutrient-limited spring production peak) of Longhurst (1995). Finally, both bioregions 6 and 7 correspond to model number 1 of Longhurst (1995), with bioregion 6 which is characterised by a sharper chl decrease.

• Comparison of partition pattern of Figure 1 with previous works
  
  – Longhurst (1995): the latitudinal structure of our partition aligns qualitatively well with the distribution of the Longhurst provinces (1995, 1998), in particular in the central Pacific and Atlantic Oceans. In addition, bioregion 3 in the North Pacific closely resembles the North Pacific Polar Front province. Other province correctly identified by our partition are the Western Australian and Indonesian Coast, the Tasman Sea, the Northwest Arabian Sea upwelling, and the Benguela current coast. Conversely, our partition in the Southern Ocean is more scattered compared to the Subantarctic water ring and the Antarctic provinces.

  – Fay and McKinley (2014): The partition we obtained compares relatively well with the one obtained by Fay and McKinley (2014). Bioregion 2 and 3 together match remarkably well with their subtropical permanently stratified biome. Bioregion 4 is similar to their subtropical seasonally stratified biome, while bioregion 7 to the subpolar seasonally stratified biome, especially in the North Atlantic. Bioregion 1 is similar to their equatorial biome, but in our case it extends over a larger band and is also present in the Southern Pacific. In the Southern Ocean, the biomes found by Fay and McKinley (2014) are organised in latitudinal bands, while the disposition of our bioregions is less regular.

  – Bock et al. (2022): the high-chlorophyll bloom bioregion matches well with our bioregion 7, which is characterised by a strong seasonal bloom as well. The points classified by Bock et al. (2022) as Equatorial cluster belongs to our bioregion 1, which shows typical Equatorial dynamics. In the Indian Ocean, our partition correctly separates the Equatorial cluster (our bioregion 1) from the Arabian cluster (our bioregions 2 and 6).
Bioregion 1, % surface
Bioregion 2, % surface
Bioregion 3, % surface
Bioregion 4, % surface
Bioregion 5, % surface
Bioregion 6, % surface
Bioregion 7, % surface
Not assigned [%]

Table S.1: Percentage of time series belonging to each of the 7 bioregions (columns) for the reference climatology (first row), the extended (second row) and third climatology (third row). The last column indicates the percentage of time series of the second and third climatology which were not assigned to any bioregion as their Euclidean distance was larger than the corresponding threshold of acceptance (Methods).

S.6 Equatorial and Southern Pacific case study.

Here, we focus on cluster 3 in the Pacific Ocean, as this cluster has a consistent presence both at low and high latitudes there, and a significant agreement between satellite and BGC time series only at low latitudes, but not at high latitudes in the Equatorial Pacific (EP), but not in the Southern Pacific (SP).

Therefore, we consider bioregion 1 in the Equatorial Pacific (bioregion 1 EP) as the grid cells located in the Pacific Ocean whose latitude is comprised between 21°S and 12°N, and bioregion 1 in the Southern Pacific (bioregion 1 SP) as grid cells whose latitude is comprised between 33°S and 58°S.

For each of these two subclusters, we calculate the mean time series of chlorophyll measured by satellite and by the BGC-Argo fleet.

In bioregion 1 EP, chlorophyll concentration is rather steady and low across the year (Fig. S.1a and b). In general, there is a good agreement between satellite and BGC observations, though the former are in general lower than BGC observations (0.1 mg/m³ against 0.2 mg/m³).

In bioregion 1 SP, BGC chlorophyll observations are 3 to 4 times larger than satellite-based measurements (0.3-0.4 mg/m³ against 0.1 mg/m³). Bioregion 1 SP is characterised by a deep ML during winter months that, during spring, becomes shallower and is associated with a seasonal DCM. Subsequently, the ML deepens again, making the DCM disappear, and triggering an increase of integrated chlorophyll in the mixed layer (Fig. S.1c).
Table S.2: List of parameters (rows) used for the calculation of each climatology (columns) as in Table S.2, but including the third climatology as well (named *extended climatology, short time frame*).

<table>
<thead>
<tr>
<th>Time period</th>
<th>Reference climatology</th>
<th>Extended climatology</th>
<th>Extended climatology, short time frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018—2022</td>
<td>50 %</td>
<td>25 %</td>
<td>25 %</td>
</tr>
<tr>
<td>1998—2022</td>
<td>50 %</td>
<td>25 %</td>
<td>20 %</td>
</tr>
<tr>
<td>2018—2022</td>
<td>75 %</td>
<td>50 %</td>
<td>50 %</td>
</tr>
<tr>
<td>ε</td>
<td>4 weeks</td>
<td>14 weeks</td>
<td>14 weeks</td>
</tr>
</tbody>
</table>
Table S.3: Properties of the bioregions split in high and low latitudes sub-bioregions as reported in Table 3. Here, we report also properties concerning the number of BGC Argo profiles composing each time series, and the results of the comparison between the BGC-Argo and satellite time series using the normalised values.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Satellite chl, mean [mg/m³]</th>
<th>Satellite chl, amplitude [mg/m³]</th>
<th>BGC-Argo chl, mean [mg/m³]</th>
<th>BGC-Argo chl, amplitude [mg/m³]</th>
<th>Surface [100*100 km²]</th>
<th># BGC prof/cell (mean)</th>
<th>Buffer, # BGC prof/cell (median)</th>
<th># BGC prof/cell (max)</th>
<th>Percentage of all BGC prof [%]</th>
<th>Median n° BGC profiles per time step</th>
<th>Min n° BGC profiles per time step</th>
<th>Max n° BGC profiles per time step</th>
<th>T test, normalised time series</th>
<th>Euclidean test, normalised time series</th>
</tr>
</thead>
<tbody>
<tr>
<td>1UL</td>
<td>0.11</td>
<td>0.01</td>
<td>0.09</td>
<td>0.03</td>
<td>5254</td>
<td>1.8</td>
<td>6.3</td>
<td>yes</td>
<td>yes</td>
<td>3</td>
<td>94</td>
<td>15.64</td>
<td>10.22</td>
<td>205</td>
</tr>
<tr>
<td>1HL</td>
<td>0.15</td>
<td>0.02</td>
<td>0.39</td>
<td>0.34</td>
<td>2273</td>
<td>2.2</td>
<td>3.6</td>
<td>no</td>
<td>no</td>
<td>2</td>
<td>122</td>
<td>6.77</td>
<td>5.33</td>
<td>100</td>
</tr>
<tr>
<td>2UL</td>
<td>0.08</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>8216</td>
<td>2.0</td>
<td>7.2</td>
<td>no</td>
<td>no</td>
<td>3</td>
<td>139</td>
<td>1.65</td>
<td>13.41</td>
<td>259</td>
</tr>
<tr>
<td>2HL</td>
<td>0.26</td>
<td>0.15</td>
<td>0.25</td>
<td>0.19</td>
<td>542</td>
<td>6.4</td>
<td>14.4</td>
<td>yes</td>
<td>yes</td>
<td>4</td>
<td>314</td>
<td>1.61</td>
<td>3.71</td>
<td>74</td>
</tr>
<tr>
<td>3UL</td>
<td>0.11</td>
<td>0.12</td>
<td>0.09</td>
<td>0.20</td>
<td>8032</td>
<td>2.2</td>
<td>7.6</td>
<td>yes</td>
<td>yes</td>
<td>3</td>
<td>314</td>
<td>17.95</td>
<td>14.32</td>
<td>296</td>
</tr>
<tr>
<td>3HL</td>
<td>0.18</td>
<td>0.21</td>
<td>0.34</td>
<td>0.26</td>
<td>215</td>
<td>12.4</td>
<td>34.5</td>
<td>yes</td>
<td>yes</td>
<td>10</td>
<td>290</td>
<td>0.64</td>
<td>2.97</td>
<td>51</td>
</tr>
<tr>
<td>4UL</td>
<td>0.16</td>
<td>0.20</td>
<td>0.22</td>
<td>0.20</td>
<td>1241</td>
<td>1.9</td>
<td>3.8</td>
<td>yes</td>
<td>yes</td>
<td>2</td>
<td>70</td>
<td>3.69</td>
<td>2.51</td>
<td>48</td>
</tr>
<tr>
<td>4HL</td>
<td>0.21</td>
<td>0.07</td>
<td>0.39</td>
<td>0.48</td>
<td>1193</td>
<td>2.6</td>
<td>5.8</td>
<td>yes</td>
<td>yes</td>
<td>2</td>
<td>235</td>
<td>3.55</td>
<td>3.69</td>
<td>73</td>
</tr>
<tr>
<td>5UL</td>
<td>0.20</td>
<td>0.13</td>
<td>0.10</td>
<td>0.10</td>
<td>901</td>
<td>1.6</td>
<td>5.7</td>
<td>no</td>
<td>yes</td>
<td>3</td>
<td>107</td>
<td>2.77</td>
<td>1.98</td>
<td>30</td>
</tr>
<tr>
<td>5HL</td>
<td>0.19</td>
<td>0.09</td>
<td>0.45</td>
<td>0.60</td>
<td>2977</td>
<td>3.2</td>
<td>4.2</td>
<td>no</td>
<td>no</td>
<td>2</td>
<td>108</td>
<td>11.23</td>
<td>12.91</td>
<td>268</td>
</tr>
<tr>
<td>6UL</td>
<td>0.55</td>
<td>0.77</td>
<td>0.30</td>
<td>0.64</td>
<td>148</td>
<td>1.2</td>
<td>5.0</td>
<td>no</td>
<td>no</td>
<td>1</td>
<td>42</td>
<td>0.29</td>
<td>0.17</td>
<td>3</td>
</tr>
<tr>
<td>6HL</td>
<td>0.32</td>
<td>0.63</td>
<td>0.53</td>
<td>1.21</td>
<td>1183</td>
<td>5.7</td>
<td>11.2</td>
<td>yes</td>
<td>no</td>
<td>3</td>
<td>788</td>
<td>5.04</td>
<td>10.48</td>
<td>217</td>
</tr>
<tr>
<td>7UL</td>
<td>0.36</td>
<td>0.51</td>
<td>0.02</td>
<td>0.37</td>
<td>690</td>
<td>2.4</td>
<td>6.4</td>
<td>no</td>
<td>no</td>
<td>3</td>
<td>79</td>
<td>2.02</td>
<td>1.73</td>
<td>32</td>
</tr>
<tr>
<td>7HL</td>
<td>0.30</td>
<td>0.38</td>
<td>0.49</td>
<td>1.04</td>
<td>2078</td>
<td>4.6</td>
<td>6.9</td>
<td>yes</td>
<td>yes</td>
<td>2</td>
<td>504</td>
<td>10.19</td>
<td>17.09</td>
<td>326</td>
</tr>
</tbody>
</table>
Figure S.1: Satellite and BGC-Argo time series in the Equatorial Pacific portion of bioregion 1 (first two panels) and the Southern Pacific portion of bioregion 1 (third and fourth columns). Panel a and c: NCHL time series. Panel b and d: *non normalised* CHL time series (i.e. showing real chlorophyll values). In each panel, the thick dark-blue solid lines indicate the mean chlorophyll time series observed by satellite, with associated uncertainty (standard deviation, thin coloured line), while the green dots and the solid line the mean BGC-Argo time series.
Figure S.2: First plot: boxplot of number of cells composing bioregions in which satellite and BGC time series are significantly different (left box) and in which are significantly similar (right box), listed in Table 3. The other plots show the mean and maximum number of BGC profiles per cell, the percentage of all BGC profile within a given bioregion, the number of BGC profiles for a given time step (median, minimum, and maximum), and the number of BGC profiles per 100×100 km surface square.
Figure S.3: *k*-means partition of the reference climatology with reference time series. The reference climatology is calculated with stricter criteria than the other two climatologies (Table 1 and Supplementary Table S.2), and does not include coastal or polar regions (Methods). The upper panel shows the 7 bioregions (right-hand legend) obtained with the *k*-means partition. Each of the lower panels shows the within-bioregion average of the normalised time series (also referred to as the reference time series) as thick coloured lines. The thin coloured lines show the associated uncertainty (standard deviation).
Figure S.4: Top left panel: results of k-means partition applied to the reference climatology (as reported in Supplementary Figure S.3). Top right panel: results of partition obtained using the 7 reference time series on the extended climatology, but only for the pixels where the reference climatology is defined. The second climatology includes coastal and polar regions, and is based on 2018-2022 data (Methods). In this panel, 93.3% of the grid cells are assigned to the same bioregion than the reference (upper left panel). Lower left panel: results of partition obtained using the 7 reference time series on the third climatology, but only for the grid cells for which the reference climatology is defined. The third climatology includes coastal and polar regions, and is based on 1998-2022 data (Methods). In this panel, 72.9% of the pixels are assigned to the same bioregion than the reference (upper left panel). In addition, 97.3% of the time series are within the threshold of acceptance of the corresponding reference time series. Lower right panel: same as lower left panel, with the difference that all the grid points are reported (hence, also coastal and polar regions). Please note that this panel is identical to the one reported in Fig. 1.
Figure S.5: Partition of the reference climatology with $k=3$ bioregions (value indicated by the silhouette and Calinski-Harabasz indexes, Methods and Supplementary Fig. S.12 and S.13) and using the $k$-means method. The lower panels show the within-bioregion average of the normalised time series (thicker coloured solid line: mean value; thinner coloured solid line, uncertainty: standard deviation; colours indicated in the right-hand legend in the upper panel). The global ocean results partitioned in three large, well-known bioregions: an equatorial-tropical band, a temperate band, and a polar bioregion.
Figure S.6: Comparison between reference time series and average time series obtained from partition of the extended climatology. The upper panel shows the partition obtained using the 7 reference time series on the extended climatology, as reported in Fig. 1. Each of the lower panels shows a reference time series as black dots (large dots: average value; small dots: standard deviation). In addition, we report the within-bioregion average of the normalised time series of the extended climatology (thicker coloured solid line: mean value; thinner coloured solid line, uncertainty: standard deviation; colours indicated in the right-hand legend in the upper panel). For each panel, the similarity between the two time series was tested (i) with a t test and (ii) by calculating the Euclidean distance between the two time series, and by comparing whether it was lower than the corresponding threshold of acceptance of the reference time series. All the tie series resulted significantly similar to the corresponding reference time series.
Figure S.7: Comparison between BGC and satellite normalised time series. The upper panel shows the partition obtained using the 7 reference time series on the extended climatology with, superposed, the position of the BGC-Argo profiles (black dots), as reported in Fig. 2. Each of the lower panels shows the within-bioregion average of the normalised time series of the extended climatology (thicker coloured solid line: mean value; thinner coloured solid line, uncertainty: standard deviation; colours indicated in the right-hand legend in the upper panel). In each panel, the black dots represent the median value measured by the BGC-Argo floats for each 8-day time step, while the black solid line represents the moving average of the black dots (over 40 days, i.e. 5 time steps). Both the black dots and the black solid line values were normalised by dividing them by the maximum value of the moving average time series. Finally, the similarity between the two time series (the coloured line obtained from satellite data and the black line obtained from BGC-Argo floats) was tested (i) with a t test and (ii) by calculating the Euclidean distance between the two time series, and by comparing whether it was lower than the corresponding threshold of acceptance of the reference time series. The results of the tests are reported in the title of each panel.
Figure S.8: Identification of the optimal number of clusters. Each column shows the mean time series (within-cluster average) obtained using different $k$ values (from $k=6$ to $k=9$) to partition the reference climatology. Time series are sorted so that similar time series appear on the same row. When passing from $k=6$ to $k=7$ clusters, a novel time series, not present in the time series obtained with $k=6$, appears (seventh row, second column). However, when passing from $k=7$ to $k=8$ clusters, the time series in the third row (cluster 1 in partition with $k=7$) is split in two very similar time series (clusters 1 and 8 in partition with $k=8$). The same occurs when passing from $k=8$ to $k=9$ clusters (this time in the second row: cluster 7 in partition with $k=8$ becomes clusters 1 and 7 in partition with $k=9$). Hence, we identify an optimal number of clusters $k=7$. 
Figure S.9: Sensitivity of the results with respect to the spatial resolution of the climatology. Left panels: partition of the reference climatology in $k=7$ bioregions and corresponding reference time series, as reported in Supplementary Fig. S.3. Here, the size of a grid cell is 1°. Right panels show the exact same results obtained this time with a climatology of grid cell size 0.50°. The lower panel shows the exact same results with a climatology of grid cell size 0.25° (only the map is reported).
Figure S.10: Sensitivity of the results when not shifting time series in the Southern hemisphere of 6 months. The upper panel shows the partition of the global ocean in $k=7$ bioregions, as reported in the upper panel of Fig. 1. The lower panel shows the partition obtained when not shifting time series in the Southern hemisphere of 6 months. In that case, we identify an optimal number of bioregions $k=9$. Even if the optimal number of bioregions among the two panels is different, however, their spatial distribution is almost identical. The presence of 2 additional bioregions in the lower panel can be explained by the fact that some time series in the Southern hemisphere contain the same information than time series in the Northern hemisphere, only shifted by 6 months. Thus, in the lower panel, the algorithm is forced to find two further time series. This result also corroborates our choice about shifting time series in the Southern hemisphere of 6 months.
Figure S.11: Elbow test: total within-clusters sum of squares (y-axis) when varying $k$ (the number of clusters used to partition the reference climatology, x-axis).
Figure S.12: Silhouette analysis results: silhouette value (y-axis) when varying \( k \) (x-axis)
Figure S.13: Calinski-Harabasz results: Calinski-Harabasz value (y-axis) when varying $k$ (x-axis). The blue dashed vertical line shows the maximum of the Calinski-Harabasz value.

Figure S.14: The colour of each pixel shows the percentage of available satellite CHL data used in the calculation of the extended climatology. This was calculated as the ratio of the number of CHL time series not flagged as NA and the total number of years considered (25) and then multiplied by 100.
Figure S.14: (Includes panels on this page and previous one). Silhouette value for each bioregion of the extended climatology. The right panels show, for each bioregion, the histogram of the silhouette values of the time series belonging to that bioregion. The colour of the histogram is dark if the silhouette value is larger than the median of all the silhouette values of that bioregion, light if it is negative, and intermediate if it is between 0 and the median. The left panel shows the distribution of the points belonging to that bioregion, coloured according to the histogram colour bar.
References


Student, The probable error of a mean, Biometrika, 6, 1–25, 1908.

Sugar, C. A., Techniques for clustering and classification with applications to medical problems, Ph.D. thesis, United States – California, 1998, copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works.


