Using machine learning and BGC-Argo floats to assess biogeochemical models and optimize observing system design

Alexandre Mignot¹, Hervé Claustre²,³, Gianpiero Cossarini⁴, Fabrizio D’Ortenzio²,³, Elodie Gutknecht¹, Julien Lamouroux¹, Paolo Lazzari⁴, Coralie Perruche¹, Stefano Salon⁴, Raphaëlle Sauzède¹, Vincent Taillandier²,³, Anna Teruzzi⁴

¹Mercator Océan International, Toulouse, France
²Laboratoire d’Océanographie de Villefranche-sur-Mer, Villefranche-sur-Mer, CNRS and Sorbonne Université, 06230 Villefranche-sur-Mer, France
³Institut de la Mer de Villefranche, CNRS and Sorbonne Université, 06230 Villefranche-sur-Mer, France
⁴National Institute of Oceanography and Applied Geophysics - OGS, Trieste, Italy

Corresponding author: Alexandre Mignot, amignot@mercator-ocean.fr

Numerical models of ocean biogeochemistry are becoming major tools to detect and predict the impact of climate change on marine resources and monitor ocean health. However, with the continuous improvement in model structure and spatial resolution, incorporation of these additional degrees of freedom into fidelity assessment has become increasingly challenging. Here, we propose a new method to inform about the model predictive skill in a concise way. The method is based on the conjoint use of a K-means clustering technique, assessment metrics and BGC-Argo observations. The K-means algorithm and the assessment metrics reduce the number of model data points to be evaluated. The metrics evaluate either the model state accuracy or the skill of the model in capturing emergent properties, such as the Deep Chlorophyll Maximum and Oxygen Minimum Zones. The use of BGC-Argo observations as the sole evaluation data set ensures the accuracy of the data as it is a homogenous data set with strict sampling methodologies and data quality control procedures. The method is applied to the global ocean biogeochemical analysis and forecasting system of the Copernicus Marine Service. The model performance is evaluated using the model efficiency statistical score that compares the model-observations misfit with the variability of the observations, and thus objectively quantifies whether the model outperforms the BGC-Argo climatology.
We show that, overall, the model surpasses the BGC-Argo climatology in predicting pH, dissolved inorganic carbon, alkalinity, oxygen, nitrate, and phosphate in the mesopelagic and the mixed layers, as well as silicate in the mesopelagic layer. However, there are still areas for improvement in reducing the model-data misfit for certain variables such as silicate, pH, and the partial pressure of CO$_2$ in the mixed layer, as well as chlorophyll-a related, Oxygen Minimum Zones-related and particulate organic carbon metrics. The method proposed here is also helpful to inform the design of the BGC-Argo network, in particular, the regions where BGC-Argo observations should be enhanced to improve the model accuracy through the assimilation of BGC-Argo data or process-oriented assessment studies. We strongly recommend increasing the number of observations in the Arctic region, while maintaining the already high-density of observations in the Southern Oceans. The model error in these regions is only slightly less than the variability observed in BGC-Argo measurements. Our study illustrates how the synergic use of modelling and BGC-Argo data can both inform about the performance of models and the design of observing systems.

### 1. Introduction

Since pre-industrial times, the ocean has taken ~26% of the total anthropogenic CO$_2$ emission (Friedlingstein et al., 2022) leading to dramatic change in the ocean’s biogeochemical (BGC) cycles, such as ocean acidification (Iida et al., 2020). Moreover, deoxygenation (Breitburg et al., 2018) and change in the biological carbon pump are now manifesting globally (Capuzzo et al., 2018; Osman et al., 2019; Roxy et al., 2016). Together with plastic pollution (Eriksen et al., 2014) and an increase in fisheries pressure (Crowder et al., 2008), major changes are therefore occurring in marine ecosystems at the global scale. In order to contextualize monitoring of ongoing changes, derive climate projections and develop better mitigation strategies, realistic numerical simulations of the oceans’ BGC state are required.

Numerical models of ocean biogeochemistry represent a prime tool to address these issues because they produce three dimensional estimates of a large number of chemical and biological variables that are dynamically consistent with the ocean circulation (Fennel et al., 2019). They can assess past and current states of the BGC ocean, produce short-term to
seasonal forecasts as well as climate projections. However, these models are far from being flawless, mostly because there are still huge knowledge gaps in the understanding of key BGC processes and, as a result, the mathematical functions that describe BGC fluxes, and ecosystems dynamics are too simplistic (Schartau et al., 2017). For instance, most models do not include a radiative component for the penetration of solar radiation in the ocean. It has been nevertheless shown that coupling such a component with a BGC model improves the representation of the dynamics of phytoplankton in the lower euphotic zone (Dutkiewicz et al., 2015; Álvarez et al., 2022). Additionally, the parameterization of the mathematical functions generally results from laboratory experiments on a few representative species and may not be suitable for extrapolation to ocean simulations that need to represent the large range of organisms present in oceanic ecosystems (Schartau et al., 2017; Ward et al., 2010).

Furthermore, the assimilation of physical data in coupled physical-BGC models that improves the physical ocean state can paradoxically degrade the simulation of the BGC state of the ocean (Fennel et al., 2019; Park et al., 2018; Gasparin et al., 2021). A rigorous assessment of BGC models is thus essential to test their predictive skills, their ability to reproduce BGC processes and estimate confidence intervals on model predictions (Doney et al., 2009; Stow et al., 2009).

However, the evaluation of BGC models is limited by the availability of data. It relies principally on a combination of different data sets from satellite (such as chlorophyll-$a$ concentration), cruises observations, permanent oceanic stations from large databases such as the World Ocean Database (e.g., Doney et al., 2009; Dutkiewicz et al., 2015; Lazzari et al., 2012, 2016; Lynch et al., 2009; Séférian et al., 2013; Stow et al., 2009). All these datasets have neither a sufficient vertical or temporal resolution, nor a synoptic view, nor provide all variables necessary to evaluate how models represent climate-relevant processes such as the air-sea CO$_2$ fluxes, the biological carbon pump, ocean acidification or deoxygenation.

In 2016, the Biogeochemical-Argo (BGC-Argo) program was launched with the goal to operate a global array of 1000 BGC-Argo floats equipped with oxygen (O$_2$), chlorophyll $a$ (Chl$a$) and nitrate (NO$_3$) concentrations, particulate backscattering (b$_{bp}$), pH and downwelling irradiance sensors (Biogeochemical-Argo Planning Group, 2016; Claustre et al., 2020).

Although the planned number of 1000 floats has not been reached yet, the BGC-Argo program has already provided a large number of quality-controlled vertical profiles of O$_2$, Chl$a$, NO$_3$, b$_{bp}$, and pH (Fig. 1). With respect to O$_2$, Chl$a$, NO$_3$, and b$_{bp}$, the North Atlantic
and the Southern Ocean are reasonably well sampled whereas pH is well sampled only in the Southern Ocean. At the regional scale, the Mediterranean Sea is also fairly well sampled by BGC-Argo floats (Salon et al., 2019; Terzić et al., 2019; D’Ortenzio et al., 2020). However, there are still large under-sampled areas like the Arctic Ocean, subtropical gyres and the subpolar North Pacific. Thanks to machine learning based methods (Bittig et al., 2018; Sauzède et al., 2017), floats equipped with O$_2$ sensors can be additionally used to derive vertical profiles of NO$_3$, phosphate (PO$_4$), silicate (Si), alkalinity (Alk), dissolved inorganic carbon (DIC), pH and pCO$_2$.

The BGC-Argo data set represents a significant improvement for the assessment of models compared to large databases such as the World Ocean Database (Boyer et al., 2013) or the Copernicus Marine Service in situ dataset (European Union-Copernicus Marine Service, 2015). Large databases are composed of data collected from various instrument types with heterogenous data sampling methodologies. Therefore, for a given variable, the accuracy numbers are not the same and change depending on the instrument type (European Union-Copernicus Marine Service, 2019). Consequently, this affects the overall accuracy over time due to the changing proportion of instrument types over the years. On the other hand, the BGC-Argo data set is a homogenous data set with strict and uniform sampling methodologies and data Quality-Control (QC) procedures. As a result, the BGC-Argo data set has a satisfactory level of accuracy, which remains stable over time (Johnson et al., 2017; Mignot et al., 2019). Moreover, the number of quality-controlled observations collected every year by the BGC-Argo fleet is now greater than any other data set (Claustre et al., 2020). Using the BGC-Argo data set as the single evaluation data set is therefore a way to ensure consistent accuracy.

The BGC-Argo floats provide multivariate observations at high vertical and temporal resolutions and for long periods of time providing nearly continuous time series of the vertical distribution of several biogeochemical variables. This is not possible with discrete, univariate vertical samplings provided by cruise cast in situ measurements or from climatological values derived from the World Ocean Atlas. All these specificities overcome the limitations of previous datasets, especially with respect to their univariate nature, as well as their limited vertical and temporal resolutions. This opens new perspectives for the evaluation of BGC models (Gutknecht et al., 2019; Salon et al., 2019; Terzić et al., 2019).
The development of BGC models, coupled with the ongoing increase in spatial and vertical resolutions, has resulted in a significant rise in the volume of model outputs. Simplification techniques are therefore required to provide decipherable information on model predictive skill. Allen et al. (2007) proposed a methodology for reducing the spatial dimensions in model assessment exercises, thereby providing concise information about the model performance. They use an unsupervised learning algorithm to classify the southern North Sea into 5 coherent BGC regions based on modelled time series of temperature, NO$_3$, PO$_4$, and Si concentrations. Then, they evaluated the predictive capabilities of the model in each BGC region (instead of each grid point), thus greatly reducing the number of points to be validated.

An additional method for reducing the dimensions of model-data comparison is the use of assessment metrics (Hipsey et al., 2020; Russell et al., 2018). In particular, metrics such as depth-averaged state variables (e.g., mixed layer averaged Chla, NO$_3$, O$_2$, etc…), mass fluxes and process rates (e.g., primary production or division rates), or emergent properties (e.g., Deep Chlorophyll Maximum (DCM), or Oxygen Minimum Zone (OMZ)) are particularly useful to reduce the number of model’s vertical layers to be compared with the observations.

The objectives of the present study are twofold. Our first aim is to propose a methodology that uses the BGC-Argo data set, an unsupervised learning algorithm and assessment metrics to simplify marine BGC model-data comparisons, and thus inform, in a concise way, about model performances. The second objective is to use this methodology to also identify ocean regions where the model-observations misfit is larger than the variability of the BGC-Argo data and thus inform the BGC-Argo observing system of regions that should be better sampled. The first step of the method consists in defining 23 assessment metrics that are used both to construct the BGC regions and then to compare the model outputs with the BGC-Argo data. Second, following the approach of Allen et al. (2007), we use an unsupervised learning algorithm, specifically a K-means clustering technique, to classify the global ocean into 8 coherent BGC regions based on the climatological modelled time series of the 23 assessments metrics. In the last step, the skill of the model in predicting the assessment metrics is evaluated in each BGC-region, using the model efficiency statistical score. Unlike other statistical metrics such the correlation coefficient, the bias or the root mean square difference, that does not quantify objectively whether the model performance is acceptable or not; the model efficiency calculates whether the model outperforms an observational climatology (Fennel et al., 2022). Finally, the method is implemented using the global ocean BGC analysis.
and forecasting system of the Copernicus Marine Service (European Union-Copernicus Marine Service, 2019).

The paper is organised as follows: section 2 presents the data sets used in the study. In section 3, we define the assessment metrics, and we detail the K-means algorithm as well as the model efficiency statistical score. In section 4, we present and discuss the results. Finally, section 5 concludes the study.

2. Data

a. BGC-Argo floats observations

The float data were downloaded from the Argo Coriolis Global Data Assembly Centre in France (ftp://ftp.ifremer.fr/argo, last accessed in January 2023). The CTD and trajectory data were quality controlled using the standard Argo protocol (Wong et al., 2015). The raw BGC signals were transformed to biogeochemical variables (i.e., O$_2$, Chl$_a$, NO$_3$, b$_{bp}$, and pH) and quality-controlled according to international BGC-Argo protocols (Johnson et al., 2018b, a; Schmechtig et al., 2015, 2018; Thierry et al., 2018; Thierry and Bittig, 2018).

In the Argo data-system, the data are available in three data modes: “Real-Time”, ”Adjusted” and ”Delayed” (Bittig et al., 2019). In the “Real-time” mode, the raw data are converted into state variables and an automatic quality-control is applied to “flag” gross outliers. In the ”Adjusted” mode, the “Real-time” data receive a calibration adjustment in an automated manner. In the “Delayed” mode, the “Adjusted” data are adjusted and validated by a scientific expert. While the “Real-Time” and “Adjusted” data are considered acceptable for operational application (data assimilation), the “Delayed” mode is designed for scientific exploitation and represent the highest quality of data with the ultimate goal, when time-series with sufficient duration will have been acquired, to possibly extract climate-related trends (Bojinski et al., 2014). However, for some variables, only a limited fraction of data is accessible in “Delayed” mode. Consequently, for each variable, we selected the highest level of data modes, where at least 80% of the data are available (see Table 1). Note that this criterion is not applied to O$_2$, where only “Delayed” mode data were selected in order to generate the pseudo-observations from CANYON-B neural network (see after). We removed data with missing location or time.
information and flagged as “Bad data” (flag = 4). Depending on the parameter and the
associated data mode, we also excluded data flagged as “potentially bad data” (flag=3) (see
Table 1). Finally, it should be noted that the status of the different modes of adjustment for bbp
is still very inhomogeneous in the global BGC-Argo database. A quality control procedure in
“Real-Time” has just been proposed to the Argo Data Management Team but is not yet
operationally implemented in the database (Dall’Olmo et al. 2022). Since there is no current
official consensus for the qualification of bbp data we decided to use for this study all data
modes but to remove the data that are flagged as “Bad data” (see details in Table 1).

Particulate Organic Carbon (POC) concentrations were derived from bbp observations. First,
three consecutive low-pass filters were applied on the vertical profiles of bbp to remove
spikes (Briggs et al., 2011): a 2-point running median followed by a 5-point running
minimum and 5-point running maximum. Then, the filtered bbp profiles were converted into
POC (mgC m⁻³) using a simplified version of the empirical POC/bbp algorithm developed by
Gali et al. (2022), i.e., for depths larger than the mixed layer depth (MLD):

\[
\frac{POC}{bbp} = c + a \cdot e^{-0.001 \cdot b \cdot (z - MLD)}, \quad (1)
\]

\[z > MLD,\]

where \( c \) is a constant deep value and, \( b \), the slope of the exponential decrease, sets to 12010
mgC m⁻³ m and -6.57, respectively, as proposed by Gali et al. (2022). The global coefficient
\( a \), is set to 37990 mgC m⁻³ m to be consistent with a relationship, developed for global
applications (i.e., \( POC = 38687.27 \cdot bbp^{0.95} \)) (European Union-Copernicus Marine Service,
2020). In the Mixed Layer (ML), \( z \) is fixed at \( z = MLD \), and the Eq. (1) simplifies to

\[
\frac{POC}{bbp} = c + a, \quad (2)
\]

\[z \leq MLD.\]

Finally, we complemented the existing BGC-Argo dataset with pseudo-observations of NO₃,
PO₄, Si, Alk, and DIC concentrations as well as pH and pCO₂ using the CANYON-B neural
network (Bittig et al., 2018). CANYON-B estimates vertical profiles of nutrients as well as
the carbonate system variables from concomitant measurements of float pressure,
temperature, salinity, and O\textsubscript{2} qualified in “Delayed” mode together with the associated
globalization and date of sampling. CANYON-B was trained and validated using the
GLODAPv2 data set (Key et al., 2015). The CANYON-B estimates of NO\textsubscript{3} and pH were
merged with measured values on the rationale that CANYON-B estimates have RMS errors
(NO\textsubscript{3} = 0.7 µmol kg\textsuperscript{-1}, pH = 0.013) (Bittig et al., 2018) that are of the same order of
magnitude as those of the BGC-Argo observations errors (NO\textsubscript{3} = 0.5 µmol kg\textsuperscript{-1}, pH = 0.07)
(Mignot et al., 2019; Johnson et al., 2017).

Finally, we verified that the RMS errors of BGC-Argo data (both measured and from
CANYON-B estimates) are lower than the RMS difference between the model and BGC-
Argo data, so that the comparison of simulated properties with the BGC-Argo data leads to a
meaningful evaluation of the model performance. We believe it is reasonable to draw
conclusions on the model uncertainty from BGC-Argo data as long as the BGC-Argo errors
are much lower than the model-observations RMS difference.

b. Global Ocean BGC analysis and forecasting system of the
Copernicus Marine Service

The global model simulation used in this study (see Appendix A.1) originates from the global
ocean hydrodynamic-biogeochemical coupled system, based on NEMO-PISCES model,
implemented and operated by Mercator Ocean for the Marine Service of the EU’s earth
observation programme Copernicus (CMEMS, 2020). The BGC component is constrained by
the assimilation of satellite Chl\textsubscript{a} concentrations, and a climatological-damping is applied to
nitrate, phosphate, oxygen, silicate - with World Ocean Atlas 2013 - to dissolved inorganic
carbon and alkalinity – with GLODAPv2 climatology (Key et al., 2015) - and to dissolved
organic carbon and iron - with a 4000-year PISCES climatological run. The BGC model is
forced in offline mode by daily averages of ocean physics, sea ice and atmospheric
conditions. The ocean physics and sea ice forcing come from the global ocean physics
analysis and forecasting system at 1/12\degree (Lellouche et al., 2018) that assimilates along-track
altimeter data, satellite Sea Surface Temperature and Sea-Ice Concentration, and \textit{in situ}
temperature and salinity vertical profiles. The BGC model has a 1/4\degree horizontal resolution, 50
vertical levels (with 22 levels in the upper 100 m, the vertical resolution is 1 m near the 
surface and decreases to 450 m resolution near the bottom).

We used daily outputs of Chla, NO3, PO4, Si, O2, pH, DIC and Alk, and weekly outputs of the 
size classes of phytoplankton, the small detrital particles and microzooplankton 
(resampled offline from weekly to daily frequency through constant interpolation) from 2009 
to 2020. Note that the method of linear resampling, while artificially increasing the number of 
data, could potentially bias the statistical results, especially in regions with poor data 
coverage. As suggested by Gali et al. (2022), the POC concentration was computed offline by 
adding together the two size classes of phytoplankton, the small detrital particles and 
microzooplankton modelled by PISCES. This particular combination of phytoplanktonic and 
non-phytoplanktonic organisms has been shown to match the small POC observed by the 
floats. The partial pressures of CO2 values were extrapolated in the mixed layer from the 
surface value estimated by the model. The Black Sea was not considered in the present 
analysis because the model solutions are of poor quality. Finally, the daily model outputs 
were collocated in time and space the closest to the BGC-Argo floats positions, and they were 
interpolated to the sampling depth of the float observations. The characteristics of the model 
are further detailed in the appendix.

3. Methods

a. Assessment metrics

In this section, we present 23 metrics used for the clustering of the ocean and for the 
assessment of the model simulation with BGC-Argo data. The metrics are associated with the 
carbonate chemistry, the biological carbon pump, and oxygen levels. Most of the metrics 
evaluate the model state accuracy through the comparison of simulated state variables with 
BGC-Argo observations depth-averaged in the mixed (hereinafter indicated with the subscript 
mixed) and mesopelagic (hereinafter indicated with the subscript meso) layers. This two-layer 
comparison between model and BGC-Argo data provides an indirect evaluation of the key 
processes and fluxes associated with the carbonate chemistry, biological carbon pump and 
oxygen levels in the mixed and mesopelagic layers. In addition, some of the metrics assess the 
skill of the model in capturing emergent properties, such as the nitracline, the DCMs and the 
OMZs. The metrics are described below and summarized in Table 2. The definition of the
metrics is the same for the model and the BGC-Argo data. The MLD is computed, following De Boyer et al. (2004), as the depth at which the change in potential density from its value at 10 m exceeded 0.03 kg m\(^{-3}\). Dall'Olmo and Mork (2014) define the mesopelagic layer as the region between the deeper of either the euphotic layer depth or the MLD, and a depth of 1000 meters. However, for ease of use, we adopt a simplified definition that considers the mesopelagic layer to be the region between the MLD and a depth of 1000 meters. To ensure the accuracy of the metrics calculation, we have checked the representation of the MLDs in the model. The model's MLDs closely match the observed data, as indicated by an overall mean square difference of approximately 30% of the total variance in the observations.

**i. Carbonate chemistry**

The uptake of ~26% anthropogenic CO\(_2\) by the global ocean (Friedlingstein et al., 2022) has altered the oceanic carbonate chemistry over the past few decades (Iida et al., 2020). Assessing how models correctly represent the oceanic carbonate chemistry is therefore critical if we aim to derive accurate climate projections on their future change. The classical variables for the study of carbonate chemistry are DIC, Alk, pH and pCO\(_2\) (Williams and Follows, 2011). These variables are assessed in the mixed (DIC\(_{\text{mixed}}\), Alk\(_{\text{mixed}}\), pH\(_{\text{mixed}}\) and pCO\(_2\)\(_{\text{mixed}}\)) and mesopelagic (DIC\(_{\text{meso}}\), Alk\(_{\text{meso}}\), pH\(_{\text{meso}}\)) layers. The partial pressure of CO\(_2\) is only assessed in the mixed layer as the evaluation of pCO\(_2\)\(_{\text{mixed}}\) plays a critical role to assess the skill of BGC models to correctly represent the air-sea CO\(_2\) flux.

**ii. Biological carbon pump**

The biological carbon pump is the transformation of nutrients and dissolved inorganic carbon into organic carbon in the upper part of the ocean through phytoplankton photosynthesis and the subsequent transfer of this organic material into the deep ocean. The functioning of this pump relies on key pools of nutrients and carbon as well as several processes that control mass fluxes between the pools. Changes in the biological carbon pump are now manifesting globally (Capuzzo et al., 2018; Osman et al., 2019; Roxy et al., 2016).
One way to indirectly evaluate the model's ability to accurately capture essential processes related to the biological carbon pump in the ocean's upper layer, such as primary production, respiration, and grazing, is to compare various ML pools [here the nutrients (NO$_3$$_{\text{mixed}}$, PO$_4$$_{\text{mixed}}$, Si$_{\text{mixed}}$, Chl$_{\text{mixed}}$ and POC$_{\text{mixed}}$)] with BGC-Argo observations. Similarly, the assessment of the mesopelagic nutrients, and POC concentration (hereinafter denoted NO$_3$$_{\text{meso}}$, PO$_4$$_{\text{meso}}$, Si$_{\text{meso}}$, and POC$_{\text{meso}}$) provides an indirect evaluation of the key mesopelagic layer processes, such as export production, respiration, etc.

In stratified systems, a DCM is formed at the base of the euphotic layer (Barbieux et al., 2019; Cullen, 2015; Letelier et al., 2004; Mignot et al., 2014, 2011). It has been suggested that the DCM plays a key role in the synthesis of organic carbon by phytoplankton (Macías et al., 2014). DCMs are therefore key features to be assessed in BGC models with respect to processes involved in the biological carbon pump such as the primary production. However, the DCM layer generally escapes detection by remote sensing. Furthermore, the DCM is also an emergent feature that develops in response to complex physical and biogeochemical interactions (Cullen, 2015). Thus, its evaluation provides critical information regarding the accuracy of the model in capturing complex patterns of key ecosystem processes. The depth and magnitude of DCM (H$_{\text{DCM}}$ and Chl$_{\text{DCM}}$) are helpful metrics for the assessment of DCM dynamics. The depth of the DCM is calculated as the depth where the maximum of Chla occurs in the profile with the criterion that H$_{\text{DCM}}$ should be deeper than the MLD. The magnitude of the DCM corresponds to the Chla value at H$_{\text{DCM}}$.

NO$_3$ is often depleted in the surface layers and is a limiting factor for phytoplankton growth in most oceanic regions. The vertical supply of NO$_3$ to the surface layers depends, among other factors, on the vertical gradient of NO$_3$ (the nitracline), and, in particular, on its depth (the nitracline depth) (Cermeno et al., 2008; Omand and Mahadevan, 2015). Therefore, the comparison of the simulated nitracline depth (H$_{\text{nit}}$) with BGC-Argo observations allows for an indirect assessment of the model performance in reproducing vertical fluxes of NO$_3$.

Following previous studies (Cermeno et al., 2008; Lavigne et al., 2013; Richardson and Bendtsen, 2019), the depth of the nitracline is identified as the first depth where NO3 is detected. A detection threshold of 1 µmol kg$^{-1}$ is used, which is an upper estimate of the accuracy of BGC-Argo NO3 data (Johnson et al., 2017; Mignot et al., 2019).

### iii. Oxygen levels
Oxygen levels in the global and coastal waters have declined over the whole water column over the past decades (Schmidtko et al., 2017) and OMZs are expanding (Stramma et al., 2008). Assessing how models correctly represent ocean oxygen levels as well as the OMZs is therefore critical to monitor their change over time. Similar to the assessment of DCMs, evaluating Oxygen Minimum Zones (OMZs) provides insight into how the model represents emergent dynamics resulting from intricate physical and biogeochemical interactions (Paulmier and Ruiz-Pino, 2009). Oxygen levels are evaluated in the mixed ($O_{2\text{mixed}}$) and mesopelagic ($O_{2\text{meso}}$) layers. OMZs are defined as oceanic regions where $O_2$ levels are lower than 20 $\mu$mol kg$^{-1}$ (Paulmier and Ruiz-Pino, 2009). OMZs are characterized by their depths ($H_{O2\text{min}}$) and their concentrations ($O_{2\text{min}}$).

**b. Bioregionalization of the global ocean**

In this study, we use the K-means clustering algorithm (Hartigan and Wong, 1979) to regionalize the ocean based on the modelled climatological monthly time series of the 23 metrics described previously. The K-means clustering algorithm is an unsupervised machine learning technique that groups similar objects together in a way that maximizes similarity between objects within a group and minimizes similarity between objects in different groups. This clustering tool has been successfully used to classify marine BGC regions in different oceanic basins based on the seasonal cycle of satellite chlorophyll (Kheireddine et al., 2021; Mayot et al., 2016; Lacour et al., 2015; D’Ortenzio and d’Alcala, 2009). The step-by-step methodology, used in this study, is described in the next section.

The first step in the analysis involved computing monthly climatological time series for the 23 metrics at each model grid cell. These time series were derived from the monthly climatological time series of state variables predicted by the model from 2009 to 2020. To account for the log-normal distribution and the wide range of values for metrics in units of Chla or POC, a log-10 transformation was applied to these metrics. Second, to take into consideration the 6-months shift in seasons between the northern and southern hemispheres, the dates for grid cells located in the Southern Hemisphere were shifted by 6 months (Bock et al., 2022). Third, to classify model grid cells based on the seasonality and amplitude of the 23 metrics, each metric was standardized by subtracting the global mean and dividing by the
global standard deviation. This ensured that each metric had a mean of 0 and a standard
deviation of 1, enabling comparison across metrics with different units. Fourth, to reduce the
dimensionality of the data set, a principal component analysis was applied to the scaled data.
Only the components that explain 99% of the variance in the data set were kept, reducing
thereby the dimensions of the data set by 85%. A K-means clustering analysis was then
performed on the resulting data set. Following Kheireddine et al. (2021), the number of
clusters was determined based on a silhouette analysis (Rousseeuw, 1987), which yielded a
value of 8 for the number of clusters.

c. Model efficiency

To quantify the model predictive skill, a model efficiency statistical score ($m_e$) was computed
for each metric and in each BGC region:

$$m_e = 1 - \frac{\sum_{i=1}^{N} (m_i - \bar{o})^2}{\sum_{i=1}^{N} (o_i - \bar{o})^2}, \quad (3)$$

where $m_i$ and $o_i$ are the model and BGC-Argo matched values, respectively, $\bar{o}$ is the BGC-
Argo climatology and $N$ is the number of matchups. Assuming that the spatial variations are
small in a given BGC-region, $\bar{o}$ represents the temporal average and $\sum_{i=1}^{N} (o_i - \bar{o})^2$ represents
the variance due to temporal fluctuations. The model efficiency tests whether the model
outperforms the BGC-Argo climatology ($0 < m_e < 1$, Fennel et al., 2022), or stated
differently, if the model-data mean square difference is lower than the observation variance,
i.e., $\frac{1}{N} \sum_{i=1}^{N} (m_i - o_i)^2 < \frac{1}{N} \sum_{i=1}^{N} (o_i - \bar{o})^2$. To ensure the robustness of $m_e$, we verified that
the number of matchups for each metric and in each BGC-region was greater than 100, then
outliers were removed using Tukey’s fences (Tukey, 1977).

4. Results and discussion

a. BGC-regions of the Global Ocean
The K-means clustering algorithm identified 8 distinct BGC-regions (Figure 2). 6 of the 8 BGC-regions correspond to well-defined spatial regions and are, thus, named accordingly, i.e., the Arctic, Equatorial (Equ.), Mediterranean Sea (Med. Sea), OMZs, Subtropical Gyres (Sub. Gyres) and Southern Oceans BGC-regions. The other two BGC-regions, i.e., the Low Nutrients Bloom (Low Nut. Bloom) and the High Nutrients Bloom (High Nut. Bloom) are located in the North Atlantic and North Pacific oceans, as well as in the lower latitudes of the Southern Oceans. These two BGC-regions correspond to ocean basins that experience a phytoplankton bloom in the springtime (Westberry et al., 2016). The main difference between these regions is that in one of them, macronutrients such as nitrate and phosphate are abundant throughout the year due to phytoplankton growth being mainly limited by iron (Williams and Follows, 2011). Finally, it should be noted that outlier grid cells were not removed from the analysis; these outliers are mainly present in grid cells close to the coast. Additionally, grid cells with bathymetry shallower than 1000 m were not included in the clustering analysis as metrics associated with mesopelagic processes cannot be calculated in these shallow grid cells.

The BGC-regions found in our study are overall coherent with the biomes estimated in Fay and McKinley (2014) (hereinafter denoted FM2014). The Arctic and Southern Oceans correspond to the FM2014 ice biome. The Sub. Gyres correspond to the FM2014 subtropical permanently stratified biome. The Equatorial BGC-region represents a larger area than the Equatorial biome in FM2014. The Low Nut. and High Nut. Bloom regions correspond to FM2014 subtropical seasonally stratified and subpolar seasonally stratified biomes, respectively. These two BGC-regions are coherent in the North Pacific and in the Southern Oceans in both studies. They differ, however, in the North Atlantic. In FM2014, the subpolar North Atlantic is divided between the subtropical seasonally stratified and subpolar seasonally stratified biomes, whereas in our study this area is only represented by one BGC-region; the Low Nut. Bloom. Finally, the Med. Sea and OMZs BGC-regions are not represented in FM2014. The main differences between our study and FM2014 are due to differences in the methodology used to identify BGC-regions. In our study, we used 23 input variables to identify BGC-regions, while in FM2014, clustering was based on only one BGC input variable (Chl$\alpha$) and three physical variables (sea surface temperature, MLD, and sea-ice fraction). Our method allows for the identification of specific BGC-regions whose function is mainly characterized by variables other than Chl$\alpha$, such as OMZs. Furthermore, our method
includes coastal areas, and identifies the Med. Sea as a BGC-region, which is not included in FM2014 because it is considered a coastal region.

b. Model performance

Figures 3-5 display the model efficiency (me) calculated for each assessment metric and BGC region. To enhance clarity, the me values are grouped by process, namely carbonate chemistry, biological carbon pump, and oxygen levels. The results are presented as bubble plots (panels b), where the size of the bubble is proportional to the me value. A bar plot (panels c) shows the median me value for a given assessment metric across all BGC regions, while another bar plot (panels a) shows the median me value for a given BGC region across all assessment metrics. Due to the limited number of assessment metrics associated with oxygen levels in most regions (i.e., 2), the mean is used instead of the median. The x and y axes in panels b are arranged in descending order based on the median me value across all assessment metrics (as shown in panels a) and the median me value across all BGC regions (as shown in panel b), respectively.

i. Carbonate chemistry

The model demonstrates improved performance in predicting certain carbonate chemistry metrics (i.e., Alk_meso, DIC_mixed, Alk_mixed, DIC_meso, and pH_meso) compared to the BGC-Argo climatology, as indicated by median me values significantly greater than 0 (Figs. 3b and c). However, the model’s ability to reproduce instantaneous variability in pH_mixed is more limited, with a me value close to 0, indicating no improvement over a simple average of observed values. Furthermore, the model underperforms the BGC-Argo climatology for pCO2_mixed across all regions. Despite these limitations, the model provides an overall better estimate of carbonate chemistry dynamics in all BGC regions compared to the BGC-Argo climatology, as evidenced by Figure 3a.

ii. Biological carbon pump

The efficiency of the model in estimating the biological carbon pump metrics varies across both metrics and regions (Fig. 4a-c). The model outperforms the BGC-Argo climatology in estimating PO4 and NO3 in the mesopelagic and mixed layer, as well as Si_meso and HNit.
However, the model’s ability to predict Si decreases significantly as one moves from the mesopelagic to the mixed layer. Additionally, the metrics associated with the first trophic level, such as Chl\textsubscript{mixed}, H\textsubscript{DCM}, Chl\textsubscript{DCM}, POC\textsubscript{mixed}, and POC\textsubscript{meso}, are systematically outperformed by the BGC-Argo climatology, with median \textit{me} values less than 0 in nearly all BGC regions (Figure 4b). Regional analysis of the median \textit{me} values (Figure 4a) shows that the model performs better than the observational mean (median \textit{me} values greater than 0) in only a few regions (i.e., the High Nut. Bloom, the Low Nut. Bloom, the Med. Sea, and the OMZs) indicating that the model performs relatively well in these regions, but may not be as accurate in the other regions.

iii. Oxygen levels

The model provides better estimates of mixed and mesopelagic O\textsubscript{2} concentrations in most BGC regions compared to the BGC-Argo climatology, as evidenced by consistently positive \textit{me} values in Figure 5b. However, the BGC-Argo climatology provides a better representation of the magnitude of O\textsubscript{2min} compared to the model, while the model performs better than the climatology in predicting H\textsubscript{O2min}, but only in the OMZs BGC-region. These results suggest that while the model performs well in estimating mixed and mesopelagic O\textsubscript{2} concentrations in most BGC regions, it doesn’t accurately capture the depth and magnitude of OMZs.

iv. Discussion

The model outperforms the BGC-Argo climatology for DIC, Alk, NO\textsubscript{3}, PO\textsubscript{4}, in the mesopelagic layer and mixed layers and Si in the mesopelagic layer. We attribute this good performance to the effective application of climatological damping. As described in the Appendix, the climatological damping mitigates the effects of physical data assimilation in the offline coupled hydrodynamic-biogeochemical system, which can lead to unrealistic drift of various biogeochemical variables. Specifically, we used the World Ocean Atlas 2013 (Garcia et al., 2013, 2014) for NO\textsubscript{3}, PO\textsubscript{4}, O\textsubscript{2}, and Si, and the Global Ocean Data Analysis Project version 2 (GLODAPv2) climatology (Key et al., 2015) for DIC and Alk. However, our analysis revealed that the model's performance in estimating Si in the mixed layer is significantly degraded comparing to the mesopelagic layer, indicating the presence of additional factors affecting the model's ability to accurately estimate this variable. Further
investigation is required to identify these factors and improve the model's performance in estimating Si in the mixed layer.

For the three Chla-related metrics, the model performs worse than the BGC-Argo climatology. This is unexpected, as the model incorporates a reduced-order Kalman filter (Lellouche et al., 2013) that assimilates daily L4 remotely sensed surface Chla, providing a mixed-layer correction to the modeled Chla (see Appendix). We verified that the assimilation of satellite Chla improves the model's ability to predict Chla, as the model-BGC-Argo data misfit is lower compared to a simulation without assimilation (not shown). Furthermore, the model-satellite misfit was also found to be lower than the variability of the satellite data (European Union-Copernicus Marine Service, 2019). These results suggest that discrepancies between the assimilated satellite Chla product and the BGC-Argo data may be responsible for the observed model-BGC-Argo data misfit. Therefore, we suggest that future studies investigate the consistency between ocean colour products and BGC-Argo Chla products on a global scale, as these two products are expected to be assimilated together in future operational BGC systems (Ford, 2021).

Overall, the model also performs worse or no better than the BGC-Argo climatology in predicting POC concentrations, the magnitude and depth of OMZs, pH_{\text{mixed}} and pCO_{2\text{mixed}}. The poor performance of PISCES-based simulations relative to BGC-Argo POC observations has been extensively studied in Gali et al. (2022). They pointed out that the large model-data misfit could be the result of an imperfect BGC-Argo POC-b_{bp} conversion factor, unsuitable model parameters associated with POC dynamics and missing processes in the model structure. Similarly, the poor model skill in capturing the OMZs dynamics has also already been documented in several studies (Busecke et al., 2022; Schmidt et al., 2021; Cabré et al., 2015). All these studies suggested that improving the ocean circulation in physical models may be the most important factor to improve the accuracy of OMZs model predictions. Finally, the negative model efficiencies for pH_{\text{mixed}} and pCO_{2\text{mixed}} can be attributed to the fact that these variables are driven by DIC, Alk, temperature, and salinity. Therefore, even small uncertainties in the model estimates of DIC, Alk (as shown in Figure 3b), temperature, and salinity (Lellouche et al., 2018) can result in poor model performance in capturing the variability of pH and pCO₂. This highlights the importance of accurately modelling these four variables to improve model estimates of pH and pCO₂.
c. Recommendation for the design of the BGC-Argo observing system

Observing System Simulation Experiments (OSSE) have been the primary tool to inform about the design of the BGC-Argo observing system (Ford, 2021; Biogeochemical-Argo Planning Group, 2016). OSSEs typically comprise a realistic “nature run”, which represents “the truth” from which synthetic observations are sampled. The synthetic observations represent the observing system to be designed. To test its impact on improving model’s predictive skill, the synthetic observations are then assimilated in an “assimilative run”. The accuracy of the “assimilative run” is then evaluated against the “nature run”. Here, we use the real BGC-Argo observations to inform about the design of the BGC-Argo network. More specifically, our aim is to inform about the regions where the model errors are greater than the variability of the BGC-Argo data, and consequently where BGC-Argo observations should be enhanced to improve the model accuracy through BGC-Argo data assimilation or process-oriented assessment studies.

For a given BGC-region, we compute a single multivariate score corresponding to the median of the 23 me associated with each assessment metric (Fig. 6). This is consistent with the fact that the BGC-Argo floats, that are now deployed, observe the 5 variables used to derive the assessment metrics, i.e., $O_2$, Chl$\alpha$, NO$_3$, $b_{bp}$ and pH. In the Arctic and in the Southern Ocean BGC-regions (typically North of 60°N and South of 60°S), the median me is barely greater than 0, suggesting that in these regions, the model performs no better than a simple mean of the observed values. In these two regions, the model is not well constrained by the assimilation of remotely sensed Chl$\alpha$ because satellite observations of ocean colour are not possible for most of the year due to ubiquitous cloud cover. In addition, the lack of in situ observations makes the climatological forcing less efficient in these regions. Together, these factors are likely to lead to poorer model performance compared to other regions. Consequently, we strongly recommend enhancing the Arctic region where BGC-Argo observations are scarce (Fig.1), and where the winter-spring months are particularly undersampled (not shown). We also recommend maintaining the already-high-density of BGC-Argo observations in the Southern Ocean. These observations are critical to better constrain
the model in these two regions where the constraint of models by assimilation of satellite observations is not possible for most of the year.

5. Conclusion

In this study, we propose a method based on the global data set of BGC-Argo observations, a K-means clustering algorithm and 23 assessment metrics to simplify model-data comparison and inform on Copernicus Marine Service forecasting system predictive skill and the design of the BGC-Argo observing system. The K-means algorithm identified 8 BGC-regions in the model simulation that are consistent with Fay and McKinley (2014) study. Within each BGC-region and for each assessment metric, we compute a model efficiency statistical score that quantifies whether the model outperforms the BGC-Argo climatology by comparing the model-BGC-Argo data mean square difference with the observation variance.

Overall, the model surpasses the BGC-Argo climatology in predicting pH, DIC, Alk, O$_2$, NO$_3$ and PO$_4$ in the mesopelagic and the mixed layers, as well as Si in the mesopelagic layer. For the other metrics, whose model predictions are outperformed by the BGC-Argo climatology, we provide suggestions to reduce the model-data misfit and thus to increase the model efficiency. Regarding the estimation of Si in the mixed layer, we suggest the presence of additional factors that may affect the model's ability to accurately estimate this variable. Further investigation is necessary to identify these factors and improve the model's performance in this regard. For Chla-related metrics, we recommend checking the consistency between ocean colour products and BGC-Argo Chla products at the global scale as it may explain part of the misfit between the model, that assimilates satellite Chla, and BGC-Argo observations. The discrepancies between modelled and observed POC and OMZs have been already investigated in previous studies. It has been suggested that improving the BGC-Argo POC-b$_{bp}$ conversion factor, tuning the model parameters and implementing missing processes in the model structure could decrease the model-data inconsistencies associated with POC dynamics. Similarly, improving the ocean circulation in physical models should improve the accuracy of OMZ model predictions. Finally, pH$_{mixed}$ and pCO$_2$$_{mixed}$ should be better modelled if the uncertainties associated with DIC, Alk, temperature and salinity in the mixed layer are reduced.
The proposed method can also be used to optimize the design of the BGC-Argo network. In particular, the regions where BGC-Argo observations should be enhanced to reduce the model-data misfit through the assimilation of BGC-Argo data or process-oriented assessment studies. We strongly recommend enhancing the observation density in the Arctic region and maintaining the already high density of observations in the Southern Oceans. These are two regions where the model error is barely less than the variability of BGC-Argo observations, and where it is not possible to use satellite observations to constrain the models through assimilation most of the year.
### Table 1. Data mode and QC flags of the BGC-Argo observations used in this study.

In the Argo data-system, the data are available in three data modes: "Real-Time", "Adjusted" and "Delayed". See section 2a for a brief description of each data mode. The flags “3” and “4” refer to “potentially bad data” and “bad data”, respectively. See also Bittig et al. (2019), for a more detailed description of Argo data modes and flags.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Data mode</th>
<th>Data mode of associated pressure, temperature and salinity profiles</th>
<th>QC flags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chla</td>
<td>Adjusted and Delayed</td>
<td>Real time, Adjusted and Delayed</td>
<td>• Real time (P,T,S): All flags except 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Adjusted or Delayed: All flags except 3 and 4</td>
</tr>
<tr>
<td>O₂</td>
<td>Delayed</td>
<td>Delayed</td>
<td>• All flags except 3 and 4</td>
</tr>
<tr>
<td>NO₃</td>
<td>Adjusted and Delayed</td>
<td>Real time, Adjusted and Delayed</td>
<td>• Real time (P,T,S): All flags except 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Adjusted or Delayed: All flags except 3 and 4</td>
</tr>
<tr>
<td>pH</td>
<td>Adjusted and Delayed</td>
<td>Real time, Adjusted and Delayed</td>
<td>• Real time (P,T,S): All flags except 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Adjusted or Delayed: All flags except 3 and 4</td>
</tr>
<tr>
<td>b&lt;sub&gt;b&lt;/sub&gt;</td>
<td>Adjusted and Delayed</td>
<td>Real time, Adjusted and Delayed</td>
<td>• Real time (P,T,S): All flags except 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Adjusted or Delayed (P,T,S): All flags except 3 and 4</td>
</tr>
</tbody>
</table>
• Adjusted or Delayed (bp):
  All flags except 4
**Table 2.** Assessment metrics used to assess the model simulation with BGC-Argo data. For each metric, the level of assessment, as described in Hipsey et al. (2020) is also indicated.

<table>
<thead>
<tr>
<th>Process</th>
<th>Metric</th>
<th>Definition</th>
<th>units</th>
<th>Assessment level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbonate chemistry</td>
<td>pCO₂ mixed</td>
<td>Depth-averaged pCO₂ in the mixed layer</td>
<td>µatm</td>
<td>State variable</td>
</tr>
<tr>
<td></td>
<td>DIC mixed</td>
<td>Depth-averaged DIC in the mixed layer</td>
<td>µmol kg⁻¹</td>
<td>State variable</td>
</tr>
<tr>
<td></td>
<td>Alk mixed</td>
<td>Depth-averaged Alk in the mixed layer</td>
<td>µmol kg⁻¹</td>
<td>State variable</td>
</tr>
<tr>
<td></td>
<td>DIC meso</td>
<td>Depth-averaged DIC in the mesopelagic layer</td>
<td>µmol kg⁻¹</td>
<td>State variable</td>
</tr>
<tr>
<td></td>
<td>Alk meso</td>
<td>Depth-averaged Alk in the mesopelagic layer</td>
<td>µmol kg⁻¹</td>
<td>State variable</td>
</tr>
<tr>
<td></td>
<td>pH mixed</td>
<td>Depth-averaged pH in the mixed layer</td>
<td>total</td>
<td>State variable</td>
</tr>
<tr>
<td></td>
<td>pH meso</td>
<td>Depth-averaged pH in the mesopelagic layer</td>
<td>total</td>
<td>State variable</td>
</tr>
<tr>
<td>Biological carbon pump</td>
<td>Chl mixed</td>
<td>Depth-averaged Chlα in the mixed layer</td>
<td>mg m⁻³</td>
<td>State variable</td>
</tr>
<tr>
<td></td>
<td>NO₃ mixed</td>
<td>Depth-averaged NO₃ in the mixed layer</td>
<td>µmol kg⁻¹</td>
<td>State variable</td>
</tr>
<tr>
<td></td>
<td>PO₄ mixed</td>
<td>Depth-averaged PO₄ in the mixed layer</td>
<td>µmol kg⁻¹</td>
<td>State variable</td>
</tr>
<tr>
<td></td>
<td>Si mixed</td>
<td>Depth-averaged Si in the mixed layer</td>
<td>µmol kg⁻¹</td>
<td>State variable</td>
</tr>
<tr>
<td>State/Property</td>
<td>Description</td>
<td>Unit</td>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>-------------</td>
<td>------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>NO$<em>3</em>{\text{meso}}$</td>
<td>Depth-averaged NO$_3$ in the mesopelagic layer</td>
<td>µmol kg$^{-1}$</td>
<td>State variable</td>
<td></td>
</tr>
<tr>
<td>PO$<em>4</em>{\text{meso}}$</td>
<td>Depth-averaged PO$_4$ in the mesopelagic layer</td>
<td>µmol kg$^{-1}$</td>
<td>State variable</td>
<td></td>
</tr>
<tr>
<td>Si$_{\text{meso}}$</td>
<td>Depth-averaged Si in the mesopelagic layer</td>
<td>µmol kg$^{-1}$</td>
<td>State variable</td>
<td></td>
</tr>
<tr>
<td>POC$_{\text{mixed}}$</td>
<td>Depth-averaged POC in the mixed layer</td>
<td>mg m$^{-3}$</td>
<td>State variable</td>
<td></td>
</tr>
<tr>
<td>POC$_{\text{meso}}$</td>
<td>Depth-averaged POC in the mesopelagic layer</td>
<td>mg m$^{-3}$</td>
<td>State variable</td>
<td></td>
</tr>
<tr>
<td>Chl$_{\text{DCM}}$</td>
<td>Magnitude of DCM</td>
<td>mg m$^{-3}$</td>
<td>Emergent property</td>
<td></td>
</tr>
<tr>
<td>H$_{\text{DCM}}$</td>
<td>Depth of DCM</td>
<td>m</td>
<td>Emergent property</td>
<td></td>
</tr>
<tr>
<td>H$_{\text{nit}}$</td>
<td>Depth of nitracline</td>
<td>m</td>
<td>Emergent property</td>
<td></td>
</tr>
<tr>
<td>Oxygen levels</td>
<td>O$<em>2</em>{\text{mixed}}$</td>
<td>Depth-averaged O$_2$ in the mixed layer</td>
<td>µmol kg$^{-1}$</td>
<td>State variable</td>
</tr>
<tr>
<td>O$<em>2</em>{\text{meso}}$</td>
<td>Depth-averaged O$_2$ in the mesopelagic layer</td>
<td>µmol kg$^{-1}$</td>
<td>State variable</td>
<td></td>
</tr>
<tr>
<td>O$_{2\text{min}}$</td>
<td>Value of O$_2$ minimum</td>
<td>µmol kg$^{-1}$</td>
<td>Emergent property</td>
<td></td>
</tr>
<tr>
<td>H$_{O2\text{min}}$</td>
<td>Depth of O$_2$ minimum</td>
<td>m</td>
<td>Emergent property</td>
<td></td>
</tr>
</tbody>
</table>
Figures

Figure 1. Spatial and temporal coverage of BGC-Argo quality-controlled pH, NO₃, Chla, O₂ and bpop profiles. (a) Number of quality-controlled profiles for the entire period per 4°x4° bin. (b) Number of quality-controlled profiles per year. Note that this study only uses data from 2009 to 2020 to evaluate model performance.
Figure 2. Spatial distribution of the 8 BGC-regions obtained with a K-means clustering method applied to a dataset of modelled climatological monthly time series of the 23 assessment metrics.
Figure 3. Bubble plot of model efficiency statistical score ($m_e$) as a function of BGC-regions [Arctic, Equatorial (Equ.), High Nutrients Bloom (High Nut. Bloom), Low Nutrients Bloom (Low Nut. Bloom), Mediterranean Sea (Med. Sea), Oxygen Minimum Zones (OMZs), Southern Oceans, Subtropical Gyres (Sub. Gyres)] and assessment metrics associated with the carbonate chemistry [depth-averaged pCO$_2$ in the mixed layer (pCO$_2$ mixed), depth-averaged DIC in the mixed layer (DIC$_{\text{mixed}}$), depth-averaged Alk in the mixed layer (Alk$_{\text{mixed}}$), depth-averaged DIC in the mesopelagic layer (DIC$_{\text{meso}}$), depth-averaged Alk in the mesopelagic layer (Alk$_{\text{meso}}$), depth-averaged pH in the mixed layer (pH$_{\text{mixed}}$), and depth-averaged pH in the mesopelagic layer (pH$_{\text{meso}}$)] (b). The size of a bubble is proportional to the value of $m_e$. For a given assessment metric, the median values of $m_e$ over all BGC regions are represented as a bar plot (c). Similarly, for a given BGC region, the median values of $m_e$ across all assessment metrics are represented as a bar plot (a). In (b), The x and y axes are arranged in descending order of the median value of $m_e$ over all assessment metrics (panels a) and the median value of $m_e$ over all BGC regions, respectively. The blue and orange colours correspond to a positive and negative $m_e$. 
Figure 4. Same as Figure 3 but for assessment metrics associated with the biological carbon pump [depth-averaged Chlα in the mixed layer (Chl_{mixed}), depth-averaged NO₃ in the mixed layer (NO₃_{mixed}), depth-averaged PO₄ in the mixed layer (PO₄_{mixed}), depth-averaged Si in the mixed layer (Si_{mixed}), depth-averaged NO₃ in the mesopelagic layer (NO₃_{meso}), depth-averaged PO₄ in the mesopelagic layer (PO₄_{meso}), depth-averaged Si in the mesopelagic layer (Si_{meso}), depth-averaged POC in the mixed layer (POC_{mixed}), depth-averaged POC in the mesopelagic layer (POC_{meso}), magnitude of DCM (Chl_{DCM}), depth of DCM (H_{DCM}), and depth of nitracline (H_{nit})].
Figure 5. Same as Figure 3 but for assessment metrics associated with the oxygen levels [depth-averaged O$_2$ in the mixed layer (O$_2$$_{\text{mixed}}$), depth-averaged O$_2$ in the mesopelagic layer (O$_2$$_{\text{meso}}$), value of O$_2$ minimum (O$_2$$_{\text{min}}$), and depth of O$_2$ minimum (H$_{O2\text{min}}$)]. Note that in (a), the bar plot represents the mean value of $m_c$ over all assessment metrics.
Figure 6. Median of the 23 $m_e$ associated with each assessment metric by BGC-region.
Appendix

A.1 The CMEMS global hydrodynamic-biogeochemical model

The model used in this study features the offline coupled NEMO–PISCES model, with a 1/4°
horizontal resolution 50 vertical levels (with 22 levels in the upper 100 m, the vertical
resolution is 1m near the surface and decreases to 450m resolution near the bottom) and daily
temporal resolution, covering the period from 2009 to 2017.

The biogeochemical model PISCES v2 (Aumont et al., 2015) is a model of intermediate
complexity designed for global ocean applications, and is part of NEMO modelling platform.
It features 24 prognostic variables and includes five nutrients that limit phytoplankton growth
(nitrate, ammonium, phosphate, silicate and iron) and four living compartments: two
phytoplankton size classes (nanophytoplankton and diatoms, resp. small and large) and two
zooplankton size classes (microzooplankton and mesozooplankton, resp. small and large); the
bacterial pool is not explicitly modelled. PISCES distinguishes three non-living detrital pools
for organic carbon, particles of calcium carbonate and biogenic silicate. Additionally, the
model simulates the carbonate system and dissolved oxygen. PISCES has been successfully
used in a variety of biogeochemical studies, both at regional and global scale (Bopp et al.,
2005; Gehlen et al., 2006, 2007; Gutknecht et al., 2019; Lefèvre et al., 2019; Schneider et al.,
2008; Séférian et al., 2013; Steinacher et al., 2010; Tagliabue et al., 2010).

The dynamical component is the latest Mercator Ocean global 1/12° high-resolution ocean
model system, extensively described and validated in Lellouche et al. (2013, 2018). This
system provides daily and 1/4°-coarsened fields of horizontal and vertical current velocities,
vertical eddy diffusivity, mixed layer depth, sea ice fraction, potential temperature, salinity,
sea surface height, surface wind speed, freshwater fluxes and net surface solar shortwave
irradiance that drive the transport of biogeochemical tracers. This system also features a
reduced-order Kalman filter based on the Singular Evolutive Extended Kalman filter (SEEK)
formulation introduced by Pham et al. (1998), that assimilates, on a 7-day assimilation cycle,
along-track altimeter data, satellite Sea Surface Temperature and Sea-Ice Concentration from
OSTIA, and in situ temperature and salinity vertical profiles from the CORA 4.2 in situ database.

In addition, the biogeochemical component of the coupled system also embeds a reduced order Kalman filter (similar to the above mentioned) that operationally assimilates daily L4 remotely sensed surface chlorophyll (European Union-Copernicus Marine Service, 2022). Thanks to a multivariate formulation of model error covariances, the system is able to provide a 3D correction to the nanophytoplankton, diatoms and nitrates model concentrations, from the surface chlorophyll data provided by satellite observations.

In parallel, a climatological-damping is applied to nitrate, phosphate, oxygen, silicate - with World Ocean Atlas 2013 - to dissolved inorganic carbon and alkalinity – with GLODAPv2 climatology (Key et al., 2015) - and to dissolved organic carbon and iron - with a 4000-year PISCES climatological run. This relaxation is set to mitigate the impact of the physical data assimilation in the offline coupled hydrodynamic-biogeochemical system, leading significant rises of nutrients in the Equatorial Belt area, and resulting in an unrealistic drift of various biogeochemical variables e.g. chlorophyll, nitrate, phosphate (Fennel et al., 2019; Park et al., 2018). The time-scale associated with this climatological damping is set to 1 year and allows a smooth constraint that has been shown to be efficient to reduce the model drift.
Data availability. The BGC model data can be downloaded from the Copernicus Marine Environmental Monitoring Service (https://resources.marine.copernicus.eu/?option=com_csw&view=details&product_id=GLOBAL_ANALYSIS_FORECAST_BIO_001_028). The BGC-Argo data were downloaded from the Argo Global Data Assembly Centre in France (ftp://ftp.ifremer.fr/argo/).

Authors Contribution: AM, GC, FD, SS and VT originated the study. AM, HC, FD, RS and VT designated the study. AM and RS process the BGC-Argo floats data. AM analysed the data. AM wrote the first draft of the manuscript. HC, GC, FD, EG, PL, CP, SS, RS, VT and AT contributed to the subsequent drafts. All authors read and approved the final draft.

Competing Interests: The authors declare no competing financial interests.

Materials and correspondence: Correspondence and request for material should be addressed to mignot@mercator-ocean.fr

Acknowledgements: This study has been conducted using the Copernicus Marine Service products. The BGC-Argo data were collected and made freely available by the International Argo program and the national programs that contribute to it (https://www.argo.jcommops.org). The Argo program is part of the Global Ocean Observing System. Part of this work was performed within the framework of the BIOOPTIMOD and MASSIMILI CMEMS Service Evolution Projects. This paper represents a contribution to the following research projects: NAOS (funded by the Agence Nationale de la Recherche in the framework of the French “Equipement d’avenir” program, grant ANR J11R107-F), remOcean (funded by the European Research Council, grant 246777), and the French Bio-Argo program (BGC-Argo France; funded by CNES-TOSCA, LEFE-GMMC).
References


Boyer, T. P., Antonov, J. I., Baranova, O. K., Garcia, H. E., Johnson, D. R., Mishonov, A. V.,
2013.

Breitburg, D., Levin, L. A., Oschlies, A., Grégoire, M., Chavez, F. P., Conley, D. J., Garçon,
V., Gilbert, D., Gutiérrez, D., Isensee, K., Jacinto, G. S., Limburg, K. E., Montes, I., Naqvi, S.
W. A., Pitcher, G. C., Rabalais, N. N., Roman, M. R., Rose, K. A., Seibel, B. A., Telszewski,
M., Yasuhara, M., and Zhang, J.: Declining oxygen in the global ocean and coastal waters,

Briggs, N., Perry, M. J., Cetinić, I., Lee, C., D’Asaro, E., Gray, A. M., and Rehm, E.: High-
resolution observations of aggregate flux during a sub-polar North Atlantic spring bloom,

Pacific Ocean Oxygen Minimum Zone and Its Core in a Warming World, AGU Adv., 3,

Cabré, A., Marinov, I., Bernardello, R., and Bianchi, D.: Oxygen minimum zones in the
tropical Pacific across CMIP5 models: mean state differences and climate change trends,

McQuatters-Gollop, A., Silva, T., van Leeuwen, S. M., and Engelhard, G. H.: A decline in
primary production in the North Sea over 25 years, associated with reductions in zooplankton
abundance and fish stock recruitment, Glob. Change Biol., 24, e352–e364,

Cermeno, P., Dutkiewicz, S., Harris, R. P., Follows, M., Schofield, O., and Falkowski, P. G.: The role of nutricline depth in regulating the ocean carbon cycle, Proc. Natl. Acad. Sci., 105,

Claustre, H., Johnson, K. S., and Takeshita, Y.: Observing the Global Ocean with

Crowder, L. B., Hazen, E. L., Aivissar, N., Bjorkland, R., Latanich, C., and Ogburn, M. B.: The Impacts of Fisheries on Marine Ecosystems and the Transition to Ecosystem-Based

Cullen, J. J.: Subsurface Chlorophyll Maximum Layers: Enduring Enigma or Mystery
135111, 2015.

Doney, S. C., Lima, I., Moore, J. K., Lindsay, K., Behrenfeld, M. J., Westberry, T. K.,
ocean ecosystem-biogeochemistry models against field and remote sensing data, J. Mar. Syst.,


