Constraining the ocean's biological pump with in situ optical observations and supervised learning. Part 1:particle size distributions

Daniel J Clements¹, Simon Yang¹, Thomas Weber², Andrew McDonnell³, Rainer Kiko⁴, Lars Stemmann⁵, and Daniele Bianchi¹

¹Department of Atmospheric and Oceanic Sciences, University of California Los Angeles, Los Angeles, CA, USA.
²Department of Earth and Environmental Sciences, University of Rochester, Rochester, New York, USA
³College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Fairbanks, Alaska 99775-7220, USA.
⁴Sorbonne Université
⁵Sorbonne Université, CNRS, UMR 7093, Institut de la Mer de Villefranche sur mer, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-Mer, France.

November 24, 2022

Abstract

The abundance and size distribution of marine organic particles are two major factors controlling biological carbon sequestration in the ocean. These quantities are the result of complex physical-biological interactions that are difficult to observe, and their spatial and temporal patterns remain uncertain. Here, we present a novel analysis of particle size distributions (PSD) from a global compilation of in situ Underwater Vision Profiler 5 (UVP5) optical measurements. Using a machine learning algorithm, we extrapolate sparse UVP5 observations to the global ocean from well-sampled oceanographic variables. We reconstruct global maps of PSD parameters (biovolume and slope) for particles at the base of the euphotic zone. These reconstructions reveal consistent global patterns, with high chlorophyll regions generally characterized by high particle biovolume and flatter PSD slope, i.e., a high relative abundance of large vs. small particles. The resulting negative correlations between particle biovolume and slope further suggests amplified effects on sinking particle fluxes. Our approach and estimates provide a baseline for an improved understanding of particle cycles in the ocean, and pave the way to global, three-dimensional reconstructions of sinking particle fluxes from UVP5 observations.

Constraining the ocean's biological pump with *in situ* optical observations and supervised learning. Part 1: particle size distributions

D.J. Clements¹, S. Yang¹, T. Weber², A.M.P. McDonnell³, R.Kiko⁴, L.Stemmann⁴, D.Bianchi¹

¹Department of Atmospheric and Oceanic Sciences, University of California Los Angeles, Los Angeles, CA, USA.
 ²Department of Earth and Environmental Sciences, University of Rochester, Rochester, New York, USA
 ³College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Fairbanks, Alaska 99775-7220, USA.
 ⁴Sorbonne Université, CNRS, UMR 7093, Institut de la Mer de Villefranche sur mer, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-Mer, France.

Key Points:

4

5

13

14	•	We use optical observations of marine particle size distribution to reconstruct global
15		climatological particle biovolume and spectral slope.
16	•	We describe the importance of different biogeochemical variables on particle bio-
17		volume and spectral slope.
18	•	Spatial and seasonal variations of biovolume and slope have synergistic effects on
19		carbon export.

Corresponding author: D.J Clements, dclements@atmos.ucla.edu

Corresponding author: D. Bianchi, dbianchi@atmos.ucla.edu

20 Abstract

The abundance and size distribution of marine organic particles are two major factors 21 controlling biological carbon sequestration in the ocean. These quantities are the result 22 of complex physical-biological interactions that are difficult to observe, and their spa-23 tial and temporal patterns remain uncertain. Here, we present a novel analysis of par-24 ticle size distributions (PSD) from a global compilation of *in situ* Underwater Vision Pro-25 filer 5 (UVP5) optical measurements. Using a machine learning algorithm, we extrap-26 olate sparse UVP5 observations to the global ocean from well-sampled oceanographic 27 variables. We reconstruct global maps of PSD parameters (biovolume and slope) for par-28 ticles at the base of the euphotic zone. These reconstructions reveal consistent global pat-29 terns, with high chlorophyll regions generally characterized by high particle biovolume 30 and flatter PSD slope, i.e., a high relative abundance of large vs. small particles. The 31 resulting negative correlations between particle biovolume and slope further suggests am-32 plified effects on sinking particle fluxes. Our approach and estimates provide a baseline 33 for an improved understanding of particle cycles in the ocean, and pave the way to global, 34 three-dimensional reconstructions of sinking particle fluxes from UVP5 observations. 35

36 1 Introduction

The ocean absorbs CO₂ from the atmosphere, which is used by phytoplankton and other autotrophs to build their organic biomass. A fraction of this organic matter eventually sinks into the ocean interior, where much of it is remineralized back to CO₂, effectively removing carbon from the atmosphere over time scales from decades to millennia. The set of processes responsible for carbon export from the ocean's surface and sequestration into deep layers are collectively referred to as the ocean's biological pump.

This biological carbon sequestration is largely dependent on the ability of sinking 43 particles to escape shallow remineralization and reach the deep layers, the so-called par-44 ticle transfer efficiency. Large, dense organic particles tend to sink at a speed propor-45 tional to their size (Kriest, 2002). Aggregation and coagulation of particles (Alldredge 46 & Gotschalk, 1988) as well as repackaging by marine organisms (e.g., by formation of 47 fecal pellets and sinking carcasses) lead to a substantial increase in the size of organic 48 particles, and hence of their sinking velocity (Stemmann & Boss, 2012; Boyd et al., 2019). 49 Conversely, disaggregation and consumption by microorganisms and filter-feeders tend 50 to reduce the size of particles and their sinking speed. Ultimately, the abundance and 51 fate of organic matter in the surface ocean results from a delicate balance of both phys-52 ical and biogeochemical processes. The rate and effect of these processes is typically as-53 sumed to be size dependent (Burd & Jackson, 2009a; Devries et al., 2014). Thus, the abun-54 dance of particles of different sizes, i.e., the particle size distribution (PSD) is a primary 55 determinant of organic carbon export and sequestration, and retains important infor-56 mation on particle dynamics (Stemmann & Boss, 2012). 57

Importantly, this sinking of organic matter removes carbon and bioavailable ele-58 ments from the surface ocean at a rate proportional to the size of the particle (Kriest, 59 2002), eventually storing them in the interior ocean for timescales that range from decades 60 to millenia. Particle consumption in the deep ocean provides energy to deep ocean mi-61 croorganisms and food webs, while simultaneously consuming oxygen. The amount of 62 carbon removed via sinking particles thus has major implications for deep ocean ecosys-63 tems (Siegel et al., 2014), atmospheric CO₂ and climate change (Kwon et al., 2009; Palevsky 64 & Doney, 2018), and the ocean microbiome (Karl et al., 1984; Bianchi et al., 2018). All 65 these effects are influenced by the surface particle size distribution. However, quantifying the large scale abundance, distribution, and size structure of sinking organic parti-67 cles has been historically difficult. 68

Satellite-based observations allow to estimate the PSD in the surface ocean, for par ticle size ranges that typically include phytoplankton and small, slowly sinking particles

(Kostadinov et al., 2009, 2010a, 2010b). However, satellite retrievals miss larger particles that more directly contribute to particle export, and are limited to the upper few tens of meters of the ocean, thus providing little direct information on subsurface particle fluxes and transfer efficiency. Despite the limitations, satellite-based PSD estimates have proven helpful to constrain models of the ocean's biological pump (DeVries & Weber, 2017).

Recent advances in ocean optical observations enable direct determination of *in situ* 77 PSD throughout the water column (Stemmann & Boss, 2012; Boss et al., 2015; Lombard 78 et al., 2019). The Underwater Vision Profiler 5 (UVP5) is an optical particle counter that provides the *in situ* particle abundance for relatively large particles (80 μ m - 2.6 cm) in 80 a given sampled volume (Picheral et al., 2010). The UVP5 consists of a camera attached 81 to the CTD rosette, and is able to collect images at high frequency as it is lowered in 82 the water column. Vertical profiles of PSD from the UVP5 are commonly taken at up 83 to 20 images per second, with downward speeds of 1 m s⁻¹, as deep at 6 km (Picheral 84 et al., 2010). Since 2008, UVP5s have been routinely deployed on oceanographic cruises, 85 in all ocean basins.

Because UVP5 instruments observe a range of sizes that includes rapidly sinking particles, they are especially helpful for characterizing patterns and fate of sinking car-88 bon fluxes. Prior studies have utilized UVP5 observations to shed light on the ocean's 89 biological pump. For example, Guidi et al. (2008) showed that PSD observations from 90 UVP5 can be combined with sediment trap data to estimate sinking carbon fluxes. A 91 similar approach was later used to estimate regional carbon fluxes (Forest et al., 2012; 92 Guidi et al., 2016; Kiko et al., 2017), as well as regional patterns of particle transfer ef-93 ficiency and deep carbon sequestration (Guidi et al., 2015). Recently, the study by Cram 94 et al. (2018) combined UVP5 observations taken along a meridional section in the Pa-95 cific Ocean and satellite-based surface chlorophyll to reconstruct global PSD and drive 96 a model of marine particle dynamics. While these studies demonstrate the potential of 97 UVP5 observations for regional and global investigations, they are based on relatively small data sets, which limits the robustness of extrapolations to the entire ocean.

In this study, we take advantage of the rapid growth of UVP5 observations and em-100 ploy a machine learning approach to reconstruct global patterns of PSD in the upper ocean, 101 and investigate their drivers. Specifically, we train a supervised machine learning algo-102 rithm to reconstruct PSD from relatively sparse UVP5 observations and well-sampled 103 oceanographic variables. By comparing patterns in PSD with environmental drivers, we 104 further gain insight into the potential mechanisms responsible for shaping the surface 105 ocean's PSD and its variability. In a companion paper (Clements et al., 2021), these global 106 reconstructions are used to estimate global particle carbon export and investigate its re-107 gional variability and controls. 108

The rest of the paper is organized as follows. Section 2 describes the machine-learning approach used to globally extrapolate PSD globally. Section 3 presents the reconstructions of particle distributions and compares our results to previous studies, discussing the uncertainties and caveats inherent to our approach. Section 4 summarizes the main findings and discusses future directions.

114 2 Methods

Observations with a variety of optical instruments, including UVP5, reveal that the PSD of organic particles in the ocean can be well approximated by a power law over a relatively broad size range (from micrometers to centimeters) (Stemmann & Boss, 2012). Accordingly, the PSD can be described by the following equation (Stemmann & Boss, 2012): $n(s) = n_0 \cdot s^{-\beta},\tag{1}$

where s is the particle equivalent spherical diameter, or size, and n(s)ds is the num-121 ber of particles in an arbitrarily small size range [s, s+ds]. This power law approxima-122 tion depends on two parameters: the intercept n_0 (i.e., the size-independent coefficient), 123 and the slope β (the exponent for size-dependence). The intercept of the PSD represents 124 the number of particles at an arbitrary reference size, and the slope encapsulates the rel-125 ative proportion between small and large particles. For a given slope, increasing the in-126 tercept proportionally increases the total number of particles. Conversely, for a given in-127 tercept, increasing the slope (i.e., making the spectrum "steeper") increases the proportion of small particles, while decreasing the slope (i.e., making the spectrum "flatter") 129 increases the proportion of large particles. Relatively small changes in the slope can thus 130 result in dramatic changes in the size partitioning of particles and in quantities that de-131 pend on this partitioning, such as the total particle biovolume and surface area. 132

Here, we use UVP5 observations to estimate PSDs (i.e., n_0 and β) at the base of 133 the euphotic zone, by fitting Equation 1 to observed particle abundances. We then ex-134 trapolate the sparse UVP5 observations to a global grid, by training a supervised learn-135 ing algorithm to predict spatially-varying PSD parameters from well-sampled environ-136 mental predictors. We exploit the three-dimensional nature of UVP5 observations to per-137 form these calculation at a varying base of the euphotic zone, here defined by the 1% light 138 level according to Morel et al. (2007), rather than a single depth. The steps used to re-139 construct global PSD from UVP5 observations are illustrated in the workflow schematic 140 in Fig. 1, and are discussed in the following sections. 141

142

153

2.1 Reconstructions of particle size spectra from UVP5 data

We use observations from a new compilation of UVP5 measurements spanning the global ocean (Kiko et al., 2021). The data set consists of over 6700 profiles from 119 cruises, collected from 2008 to 2020 (Fig. 2). These observations provide robust particle counts for the 105 μ m - 5 mm size range at each location and depth. Under the power law assumption (Equation 1), the two parameters n_0 and β are needed to capture the PSD (Stemmann et al., 2004; Stemmann & Boss, 2012; Devries et al., 2014).

We calculate the power law slope β by fitting a linear least-squares regression through the log-transformed particle abundance and size. We then calculate the observed particle biovolume (BV) by multiplying the volume of a particle of a given size s by the observed size distribution n(s), and integrating over all size ranges:

$$BV = \int_{s_{min}}^{s_{max}} n(s) \cdot \frac{\pi}{6} \cdot s^3 \, ds. \tag{2}$$

In practice, the continuous integral is approximated by a summation over all sizebins in which the UVP5 observations are discretized.

¹⁵⁶ Under the power law assumption, the biovolume can also be expressed analytically ¹⁵⁷ as a function of the slope and intercept, by substituting Equation 1 into Equation 2:

$$BV = \int_{s_{min}}^{s_{max}} n_0 \cdot s^{-\beta} \cdot \frac{\pi}{6} \cdot s^3 \, ds = \int_{s_{min}}^{s_{max}} \frac{\pi}{6} \cdot n_0 \cdot s^{3-\beta} \, ds = \frac{\pi}{6} \cdot n_0 \cdot \left(\frac{s_{max}^{4-\beta}}{4-\beta} - \frac{s_{min}^{4-\beta}}{4-\beta}\right). \tag{3}$$

By fixing the size range, i.e., the minimum and maximum particle size that can be robustly derived from UVP5 instruments (s_{min} and s_{max} respectively), we solve Equation 3 for the intercept n_0 as a function of the PSD slope and the observed biovolume:



Figure 1. Schematic diagram illustrating the general workflow of processing UVP5 observations into a global PSD dataset. Observations are ensembled onto a normal 1 degree grid, with observation representing an average of a 20 meter vertical bin about the export horizon. PSD observations (power law slope and biovolume) are calculated for the 105 μm to 5 mm size range. The PSD slope and biovolume are globally extrapolated using a bagged Random Forest algorithm.

$$n_{0} = \frac{6 \cdot BV}{\pi} \cdot \left(\frac{s_{max}^{4-\beta}}{4-\beta} - \frac{s_{min}^{4-\beta}}{4-\beta}\right)^{-1}.$$
 (4)

We set the minimum and maximum size for this equation to the same values used to estimate the slope and biovolume from UVP5 observations. We use a minimum size $s_{min}=105 \ \mu \text{m}$ to avoid a potential slight instrument bias in the lowest size classes. We set the maximum size to $s_{max}=5 \text{ mm}$, which corresponds to the size where zooplankton start to dominate the biovolume at a variety of locations sampled by UVP5 (Forest et al., 2012; Stemmann et al., 2008; Stemmann & Boss, 2012).

162

We coarsen the temporal and spatial resolution of the UVP5 profiles by binning 169 them onto the standard monthly 1 degree-resolution grid of the World Ocean Atlas (H. Gar-170 cia et al., 2018; H. E. Garcia et al., 2019). That is, we combine multiple profiles in a given 171 grid cell and month together, thus reducing variability due to the noisy and episodic na-172 ture of particle observations. We also combine all observations within a 20 meter-thick 173 depth bin around each chosen depth horizon, to further smooth out small-scale vertical 174 variability, and to increase the significance of particle counts, especially for the largest 175 sizes. To reconstruct global PSDs, we calculate slope and biovolume at each location, 176 at the given depth horizon, using the gridded observations, and assume that these av-177 erages are representative of the climatological monthly PSD in each grid cell. 178

-5-

Although the gridding procedure reduces noise and data patchiness in many well-179 sampled regions, a significant proportion of grid cells only contains a single profile (\sim 180 45%). As a further quality check, we test the assumption that a power law distribution 181 is a good approximation for the observed PSD. For each grid cell with observations, we place an objective goodness of fit threshold to determine the robustness of the power law 183 fit. If a power law fit has a Pearson correlation coefficient R^2 of less than 0.9, we remove 184 the data point, as it likely does not closely follow a power law distribution. This qual-185 ity control step removes less than 1% of data (Supplementary Information Fig. S1). The 186 final processed UVP5 observation data set contains 2,034 gridded observations at the ex-187 port horizon, which together cover slightly less than 10% of the ocean surface. Figure 188 2 shows the spatial and temporal resolution of the final gridded data set, and an exam-189 ple of the observed PSD from UVP5 with the corresponding power law fit. 190



Figure 2. Global distribution of the UVP5 observations used in this study. (a) Number of profiles per one-degree resolution grid cell. (b) Number of months represented in each grid cell. (c) Typical particle size distribution sampled by the UVP5, in log-log space. The red dots indicate actual observations, and the black line the linear fit $(R^2 = 0.99)$.

191

2.1.1 Training and evaluating a Random Forest model

Monthly flux reconstructions require extrapolation of PSD parameters to the whole 192 ocean on monthly time scales. We use a bagged Random Forest (RF) algorithm to re-193 construct climatological PSD slope and biovolume globally, following an approach sim-194 ilar to Yang et al. (2020). A RF deploys a decision tree learning scheme to solve a re-195 gression equation iteratively, and reports the ensemble average. Using a RF, each indi-196 vidual decision tree is trained on a subset of the available data, with a subset of predic-197 tors, but the power of the method emerges when considering the ensemble average. The RF is able to learn statistical relationships between target variables (here, UVP5-derived slope and biovolume) and a series of predictors (here, environmental variables), to make 200 reconstructions that minimize the error between predicted and observed data. Because 201 a RF is highly non-linear, it runs the risk of overfitting the data, producing solutions with 202 low error, but also limited extrapolation power outside of the training data set. To mit-203

igate the risk of overfitting, the RF does not use all data points for training. Instead,
a bootstrapped sample (70%) of the data is selected for each tree in the forest. The degree of overfitting is determined by finding the error between the model and the data not
used for training, i.e., the "out-of-bag" data.

The rank of predictors is given by the out-of-bag error coupled with an internally 208 derived measure of importance, using a so-called "recursive feature elimination" approach. 209 A recursive feature elimination systematically removes the least important predictor and 210 records the out-of-bag error to describe the contribution of each predictor to the final 211 solution. When there is relatively no change in the out-of-bag error for every additional 212 predictor, these predictors are considered not important for the RF (Supplementary Fig. 213 S2). We determine statistical importance in order to establish a reduced set of predic-214 tors, reducing the risk of over-fitting while not losing predictive power. When interpret-215 ing the RF results, we apply qualitative understanding of the predictors combined with 216 the recursive feature elimination to determine if a predictor should be included in the 217 final regression or if it should be excluded. 218

219

2.1.2 Environmental Predictors

The RF algorithm relies on a set of predictors and target data at the resolution of the desired reconstruction. In our case, we use climatological monthly predictors at 1degree spatial resolution. We include a variety of predictors that are globally sampled and could be mechanistically related to particle production in the surface ocean, ranging from physical variables (e.g., temperature and salinity) to ecosystem-level quantities (e.g., primary production, euphotic zone depth). A list of all predictors is shown in Table 1.

Some of these predictors are obtained from satellite products at high spatial and 227 temporal resolution (e.g., surface chlorophyll and net primary production), and include 228 missing values caused by the presence of clouds or sea-ice. For these variables, we first average observations into monthly climatologies, then replace missing data by using a spherical interpolation algorithm (D'Errico, 2016; Yang et al., 2020). To avoid excessive 231 extrapolation in high latitude regions in wintertime, only points with at least 8 months 232 of satellite observations are used for the final reconstruction, following the approach of 233 Siegel et al. (2014). To process net primary production, we also calculate the Sverdrup 234 critical depth, where light becomes too limiting to support photosynthesis, based on cli-235 matological chlorophyll concentration and incident shortwave radiation (Siegel et al., 2002). 236 When the critical depth is exceeded, we assume that phytoplankton spent too much of their life cycle in light-limited depths, thus making net productivity negligible. Surface 238 net primary production is thus set to zero at all points where, in a given month, the mixed 239 layer depth exceeds the critical depth, before interpolating. We also include as a predic-240 tor the standard deviation of the primary production, using it as a proxy for intermit-241 tency and sub-seasonal variability. Similarly, we restrict chlorophyll and net primary pro-242 duction based on climatological sea ice cover from ERA5 reanalysis (Copernicus Climate 243 Change Service, 2017), and assume that regions with at least 30% sea ice coverage are 244 characterized by limited production. 245

We use two different depth-dependent averaging procedures to generate two-dimensional 246 predictor fields from three-dimensional variables, such as temperature. We generate a 247 "surface" predictor by taking the average of the variable over the mixed layer, and a "sub-248 surface" predictor by taking the average from the base of the mixed layer to 100 m below it. For surface-only variables (e.g., chlorophyll, net primary production) and nutri-250 ents we also include predictors that quantify the change of the variable over time, be-251 cause time variability (e.g., blooms in chlorophyll) could also be related to export flux. 252 In practice, we calculate the time derivative of each variable by taking the difference be-253 tween the month of observation and the prior month. We refer to these depth- and time-254

Table 1. Variables used to predict PSD parameters, variations (i.e., vertical or temporalchanges) and data sources. The categories are organized based on predictor type, where universalpredictors are used in every Random Forest realization.

Category	Variable	Short Name	Variations	Source
Universal				
	Topography	topo		N.G.D.C (2006)
	Temperature below MLD	$temp_deep$	Time Derivative	Locarnini et al. (2019)
	Chlorophyll	Chlorophyll_modis	Time Derivative	NASA G.S.F.C (2014)
	Oxygen	02_ml 02_deep	ML/ ML+100m Time Derivative	H. E. Garcia et al. (2019)
	Shortwave Radiation	shortwave	Time Derivative	Copernicus Climate Change Service (2017)
	Nitrate	no3_ml no3_deep	${ m ML}/{ m ML}{ m +100m}$ Time Derivative	H. Garcia et al. (2018)
	Phosphate	$po4_ml$ $po4_deep$	${ m ML}/{ m ML}{ m +100m}$ Time Derivative	H. Garcia et al. (2018)
	Salinity	salt	ML/ ML+100m	Zweng et al. (2019)
Mixed Layer				
	Mixed Layer	MLD_MIMOC	Time Derivative	Johnson et al. (2012)
	Mixed Layer	MLD_DBM	Time Derivative	de Boyer Montégut et al. (2004)
Primary Production				
	Eppley VGPM	Eppvgpm	Time Derivative	Antoine and Morel (1996)
	VGPM	vgpm	Time Derivative	Behrenfeld and Falkowski $\left(1997\right)$
	CBPM	cbpm	Time Derivative	Westberry et al. (2008)
	CAFE	cafe	Time Derivative	Silsbe et al. (2016)
NPP Standard Deviation				
	Eppley VGPM	$Eppvgpm_std$		Antoine and Morel (1996)
	VGPM	$vgpm_std$		Behrenfeld and Falkowski $\left(1997\right)$
	CBPM	$cbpm_std$		Westberry et al. (2008)
Euphotic Zone Depth				
	VGPM	zeuph_vgpm		Morel et al. (2007)
	CBPM	zeuph_vgpm		Morel et al. (2007)
Iron				
	Soluble Iron	HAM_SFE	Time Derivative	Hamilton et al. (2019)
	Labile Iron	LFE	Time Derivative	Myriokefalitakis et al. (2018)

change variables as "variations" in Table 1. We test the significance of each predictor,
including vertical and time variations, with the recursive feature elimination. Finally,
we group predictors into different categories, with variations for selected variables (Table 1). If a predictor is in the "universal" category in Table 1, it is always included in all
RF realizations. For all other categories, only one predictor is randomly chosen for each
realization, but if a predictor is chosen, all variations are included too. After processing, all predictors consist of monthly climatological two-dimensional fields.

The predictors are used to reconstruct PSD slope and intercept at the climatolog-262 ical euphotic zone depth . Each prediction is based on the ensemble average of 100 RFrealizations with variable hyper-parameters (the number of trees and their complexity). 264 with the inter-model spread representing the error. Each RF realization uses a total of 265 29 predictors randomly chosen from the categories listed in Table 1. By generating an 266 ensemble of 100 RFs for each reconstruction, with varying hyper-parameters and pre-267 dictors, we reduce biases and overfitting, making the results robust with respect to pa-268 rameter tuning and the choice of different observational products. Thus, our reconstructions are not the result of tuning the hyper-parameters, or choosing only the best predictors. We evaluate the overall robustness of the predictions by reporting goodness-of-271 fit statistics that include the correlation coefficient, the root mean square error (RMSE), 272 and the average bias, calculated by comparing predictions to *in situ* data. 273

²⁷⁴ 3 Results and Discussion

275

3.1 Particle size distribution reconstructions

Figs. 3 and 4 show the global reconstructions of PSD biovolume and slope. Our 276 reconstruction method is able to capture most of the variability of the UVP5 observa-277 tions, and robustly reproduce the gridded measurements, with global average values of 0.6 ppm for biovolume $(r^2=0.91)$ and 3.9 for slope $(r^2=0.86)$ when considering the en-279 tire data set. Observations that are not used in the training (out-of-bag) provide a more 280 stringent test for the method's robustness. As shown in Figs. 3d and 4d, these out-of-281 bag observations are also robustly predicted, with a RMSE of 2.1 ppm for biovolume ($r^2=0.74$) 282 and 0.33 for slope $(r^2=0.68)$. Relative to both the full data set and the out-of-bag ob-283 servations, our reconstructions show a negligible bias. That is, there is an overall com-284 pensation between data points where our method overestimates observations, and data 285 points where our method underestimates them.

While most observations are generally accurately reproduced, there remains a de-287 gree of uncertainty in the reconstructions, as shown by the scatter around the one-to-288 one line in Figs. 3c,d and 4c,d. Some of this remaining uncertainty could be explained 289 by the episodic nature of particle production and export, and by factors not captured by our climatological predictors. Our method operates under the assumption that the 291 input data (i.e., the UVP5 observations) consists of monthly climatological averages, rather 292 than instantaneous snapshots. By ensembling in situ UVP5 measurements into 2,034 monthly 293 data points, we reduce part of the episodic nature of these observations; however some 294 variability and patchy behavior may still exist in the gridded data. Finally, while the mean 295 bias is zero, the reconstructions show a slight underestimate of extreme values at both 296 the high and low range of the observations, i.e., our reconstructions have a slightly re-297 duced range compared to observations (Figs. 3c,d and 4c,d). This slightly reduced range in the reconstructions is typical for bagged ensemble ML methods such as the RF used here, which results in a limited ability to extrapolate data and tends to smooth out ex-300 treme values (Zhang & Lu, 2012). We discuss the consequences of this potential range 301 reduction in Section 3.5. 302



Figure 3. Observed and reconstructed particle biovolume (in parts per million, ppm) at the base of the euphotic zone. (a) Observed average biovolume. (b) Annual mean biovolume reconstructions. (c) Performance of the RF reconstruction shown as density scatter plots of predicted vs. observed biovolume (colors indicate the normalized density of observations at each point). (d) Same as (c), but using out-of-bag (OOB) predictions, i.e., predictions vs. observations withheld from training. Annotations in (b) and (c) show the coefficient of determination (r^2) , the rmse, and the global bias.

303

3.2 Global patterns in particle size distribution

Our reconstructions of the PSD for the time frame 2008 to 2020, reveal high bio-304 volume in productive regions such as high latitudes, coastal waters, and upwelling sys-305 tems, and low biovolume in the oligotrophic subtropical gyres (Fig. 3b and Supplemen-306 tary Fig S3). PSD slopes show a nearly opposite pattern, with smaller slopes (i.e., "flat-307 ter" PSD) in more productive regions, and larger slopes (i.e., "steeper" PSD) in oligotrophic 308 waters (Fig. 4b and Supplementary Fig S4), although with somewhat less pronounced 309 variations compared to biovolume. Consistent with this, we find that slope and biovol-310 ume are negatively correlated ($r^2 = 0.4, p < 0.01$ Fig. 5a,b). Spatial patterns in bio-311 volume and slope roughly follow the distribution of satellite-derived primary chlorophyll 312 and primary production estimates, suggesting that phytoplankton and photosynthesis 313 exert a strong control on total abundance of particles in any given region (Kostadinov 314 et al., 2009, 2017). Accordingly, we find a positive correlation between biovolume and 315 surface chlorophyll ($R_{observed} = 0.49, R_{reconstruct} = 0.68, p < 0.01$ Fig. 5a,b) and a 316



Figure 4. Observed and reconstructed PSD slope at the base of the euphotic zone. (a) Observed average PSD slope. (b) Annual mean PSD slope reconstructions (c) Performance of the RF reconstruction shown as density scatter plots of predicted vs. observed particulate slope (colors indicate the normalized density of observations at each point). (d) Same as (c), but using out-of-bag (OOB) predictions, i.e., predictions vs. observations withheld from training. Annotations in (b) and (c) show the coefficient of determination (r^2) , the rmse, and the global bias.

negative correlation for slope ($R_{observed} = -0.18, R_{reconstruct} = -0.37, p < 0.01$ Fig. 5c,d).

The negative correlation between particle biovolume and slope (R = -0.40, -0.64Fig. 5e,f) indicates that particle-rich regions (higher biovolume) are also characterized by an excess of large particles over small particles (i.e., flatter slope), relative to average oceanic conditions. Since large particles contribute proportionally more than smaller particles to export fluxes, given the faster sinking speed, this relationship suggests that biovolume and slope will synergistically enhance export fluxes in particle-rich regions, and depress them in particle-poor regions.

While this pattern of correlations holds true for most regions, we find few significant exceptions where the PSD slope and biovolume do not co-vary as closely as expected. For example, in the North Pacific subpolar gyre, flatter slopes are found in the open ocean (Fig. 4b), in particular close to the subpolar-subtropical transition, while the highest biovolumes are found closer to the coast and in marginal seas. Similarly, slopes in coastal upwelling systems, such as the California Current and the Arabian Sea upwelling, are



Figure 5. Relationships between PSD parameters and surface chlorophyll. (a,b) Relationship between PSD slope and chlorophyll for (a) observed and (b) predicted data. (c,d) Relationship between particle biovolume and chlorophyll for (a) observed and (b) predicted data. (e,f) Relationships between PSD slope and particle biovolume. The black line in each panel shows a linear fit between the two variables, and R is the Pearson's correlation coefficient.

not as flat as the high biovolumes would suggest. We also find relatively flatter slopesin the North Pacific subtropical gyre as compared to other oligotrophic regions.

These patterns suggest that while the partitioning between large and small par-334 ticles typically reflects the strength of primary production, as previously noted (Stemmann 335 et al., 2002, 2008), there are regions where the dynamics are more complex. Coastal up-336 welling regions are generally productive and exhibit high export (Bishop et al., 2016). 337 However, according to our reconstruction, the California Current exhibits steeper slopes 338 than expected, nearly matching the North Pacific subtropical gyre. It is possible that 339 in the coastal water, slopes are higher due to an increased number of large pytoplank-340 ton (Kostadinov et al., 2010a). Diatoms observed by the UVP5 could artificially inflate 341 the particle abundance in the smaller size ranges, resulting in a lower slope. Also, this 342 could be due to reduced surface aggregation or effective disaggregation of particles, or 343 less efficient surface remineralization, which tends to proportionally reduce small par-344 ticles faster than large ones. Conversely, relative to other oligotrophic gyres, the North 345 Pacific subtropical gyre may be characterized by somewhat larger phytoplankton cells, 346

ization, especially due to the deep euphotic zone present in the region.



Figure 6. Annual seasonal cycle of particle biovolume (blue lines, in ppm) and slope (red lines) from the Random Forest reconstructions. Each seasonal cycle is from the euphotic zone for the regions specified on the map (top).

3.3 Seasonal variability in particle size distribution

349

The seasonal dynamics of biovolume and slope confirms the general anti-correlation of these two variables, and reveals significant seasonal cycles, with maximum biovolume and minimum slope generally found in spring, and minimum biovolume and maximum slope in late fall to winter (Fig. 6). Similar to the spatial distribution, we find significant deviations from the general anti-correlation between biovolume and slope. For example, in the North Atlantic, the peak in biovolume (May) precedes the minimum in slope (July). In some of the tropical regions (e.g., in the North Pacific and North Atlantic) the anti-correlation is also less robust, with periods of several months where biovolume and
 slope increase or decrease simultaneously. As discussed above, spatial and temporal de coupling of the biovolume-slope relationship could have important consequences for the
 patterns of particle export flux.

In general, regions that show higher total biovolume and lower slopes also display higher seasonality. High latitude regions are characterized by large biovolume and flatter slopes, following the pattern of productivity for these waters. Conversely subtropical regions characterized by low biovolume also exhibit low seasonal variability. The synergistic variability between biovolume and slope suggests a reduced overall variability in carbon export in low and mid latitudes relative to high latitudes. Similarly, large biovolumes and low slopes suggest that particle fluxes would be larger in high latitudes. These hypotheses are explored further in a companion paper (Clements et al., 2021).

369

3.4 Empirical Drivers of PSD

A recursive feature elimination indicates that multiple variables are required for a robust reconstruction of PSD, as each one increases the ability of the reconstruction 371 to explain observations (Supplementary Fig. S2). Among the important features, we high-372 light chlorophyll, mixed layer depth, and oxygen, although each has a somewhat differ-373 ent importance for explaining biovolume and slope variability. Interpretation of these 374 rankings should be done with care because of the statistical nature of the RF algorithm. 375 However, while a mechanistic understanding of PSD patterns can not be directly tied 376 to these rankings, highlighted predictors can provide insights into the role of different 377 processes that may be affecting PSDs. 378

We find that biovolume at the base of the euphotic zone correlates positively and 379 significantly with chlorophyll ($R_{observed} = 0.49, R_{reconstruct} = 0.68, p < 0.01$, Fig 5a,b). 380 This is not surprising, since chlorophyll is an indicator of phytoplankton, the main source 381 of organic matter and sinking particles in the ocean (Stemmann et al., 2002). However, we find that chlorophyll is not as strong a predictor of slope, when the whole ocean is 383 considered $(R_{observed} = -0.18, R_{reconstruct} = -0.37, Fig 5c, d)$, and that additional 384 predictors are needed for robust slope reconstructions. This result reflects previous find-385 ings based on UVP5 observations along a meridional section in the Pacific Ocean (Cram 386 et al., 2018). Slope reconstructions also reveal a significant predictive power for subsur-387 face oxygen. Previous work indicates that there is a connection between oxygen and to-388 tal particle concentration (Roullier et al., 2014), whereby particle concentrations increase as oxygen decreases. Oxygen is a proxy of respiration in the water column, which in turn reflects the characteristics of both the surface community that drives export, and of the 391 subsurface community responsible for this respiration (Sarmiento & Gruber, 2006). We 392 note that the PSD slope is an emergent property that reflects the interaction of phys-303 ical and biological processes that are still poorly understood. 394

Spatial patterns in slope and biovolume share several features with estimates of phy-395 toplankton size spectra and composition from observations and models (Kostadinov et 396 al., 2009; Roy et al., 2013; Barton et al., 2013; Ward et al., 2014). Regions with higher 397 biovolume and flatter slope are dominated by larger phytoplankton, while the subtrop-398 ics, with lower biovolume and steeper slope, are dominated by smaller phytoplankton 399 (Kostadinov et al., 2009; Mouw et al., 2017). The composition and size structure of phy-400 toplankton can be linked mechanistically to the size of particles and aggregates in the 401 upper ocean (Burd & Jackson, 2009b). Large cells, for example chain-forming diatoms, can more easily aggregate to form large phytodetritus particles. More indirectly, phy-403 toplankton composition and size structure exert an important control on the size struc-404 ture of zooplankton and the upper ocean food web, thus of affecting the abundance and 405 size structure of fecal pellets and other aggregates that are the byproduct of zooplank-406 ton feeding processes (Turner, 2015). 407

Phytoplankton functional groups (e.g., Mouw et al., 2017) and abundance should 408 be considered as important controlling factors on both biovolume and slope (Guidi et 409 al., 2009; Stemmann et al., 2002), and could be used as predictors alongside other phys-410 ical and biogeochemical variables. However, methodological shortcomings and disagree-411 ment between different approaches (such as satellite based retrievals) currently limit the 412 applicability of these datasets—something that may be mitigated by future advances. 413 It is also likely that information related to phytoplankton composition and size struc-414 ture retrieved from satellite implicitly enters the RF regression via relationships with en-415 vironmental predictors such as satellite retrieved surface chlorophyll and temperature 416 (Kostadinov et al., 2017; Mouw et al., 2017). 417

418

3.5 Caveats to our approach

While the global data set of UVP5 observation enables robust global reconstruc-419 tion of PSD properties, there remain sources of uncertainty and inherent limitations that 420 could affect our estimates and call for further work. First, expanding the coverage of ob-421 servations with UVP5 and similar instruments, in particular in under-sampled regions 422 characterized by large variability, such as coastal and high latitude regions, would im-423 prove the robustness of our estimates, and shed additional light on regional particle size distribution patterns not captured by previous work. Regional correlations between en-425 vironmental properties and PSD may not be well captured by extrapolation with a RF 426 algorithm trained on data from different regions, especially when non-linear relationships 427 between variables are important. 428

Our reconstructions also rely on a two-parameter power law equation to describe the observed PSD. While our tests suggest that this assumption is globally robust, other statistical models may be more appropriate, and could result in somewhat different patterns of PSD and biovolume globally. Furthermore, we do not test how well our PSD slope translates to particles smaller or larger than the range robustly sampled by the UVP5, which may be possible by combining UVP5 observations with other optical instruments (Stemmann & Boss, 2012; Boss et al., 2015; Lombard et al., 2019).

Supervised learning methods are only as reliable as the data used for training; therefore, continued work on improving satellite reconstructions of surface chlorophyll, net primary production, and other remotely-sensed variables, in particular at high latitudes, would help improve the robustness of these methods. These remotely sensed variables also have inherent seasonal biases, which may limit the interpret-ability of the correlations observed, and have a greater inherent error compared to other features used for the reconstruction (i.e. temperature) (Bisson et al., 2020).

Some variables that are known to be mechanistically linked to particle production are not considered important by the random forest method. For example silicate, which could serve as a proxy for diatom biomass or production, did not significantly reduce the RF error when included, and thus were excluded from the final reconstructions (Supplemental Figure. S2). It is possible that our random forest method is biased to select only few of highly correlated variables, even if other features are mechanistically important (Nicodemus et al., 2010).

Lastly, different machine learning approaches are likely characterized by different 450 biases. Here, we note a slight underestimate of extreme values in reconstructed PSD prop-451 erties, which may affect the reconstructed variability in particle size spectra (Zhang & 452 Lu, 2012). Different machine learning methods (i.e. Artificial Neural Networks, Boosted 453 454 Forests, etc.) have been used to reconstruct particulate matter in the surface ocean (Liu et al., 2021). Adoption of additional machine learning algorithms in conjunction with 455 increased data coverage may eventually reduce our error. Additionally, increasing num-456 ber of measurements, more detailed analyses of particle size spectra distribution, includ-457

ing at time-series stations, and spatial clustering techniques, may allow reconstructionof interannual variability (Gregor & Gruber, 2021).

460 4 Conclusions

In this paper, we provide a new, data-constrained estimate of particle size spectra based on global UVP5 observations obtained between 2008 and 2020. It captures regional and seasonal variability in observed PSD properties, and demonstrates the ability of statistical machine learning methods to extrapolate these quantities globally. These global PSD reconstructions in turn pave the way to global reconstructions of sinking particle fluxes (Clements et al., 2021).

The statistical nature of our machine learning approach does not directly reveal 467 mechanisms behind PSD and export fluxes. However, we are able to highlight spatially 468 coherent patterns, and the seasonal variability of particle abundance and size structure. 469 Specifically, we show that the total particle biovolume and the PSD slope are character-470 ized by similar but inverse patterns, with regions of high particle biovolume generally 471 characterized by flatter slopes, i.e., relatively more abundant large particles. Similarly, the seasonal cycle of the particle slope and biovolume are inversely correlated over time 473 through most of the ocean. Importantly, because of this anti-correlation, biovolume and 474 slope variations would act synergistically on sinking particle fluxes, by enhancing them 475 in region of higher biovolume and flatter slope, and reducing them in regions of low bio-476 volume and steeper slope. We also show that biovolume and slope tend to correlate with 477 observed sea surface chlorophyll and other biogeochemical variables. Specifically, regions 478 of high chlorophyll tend to be characterized by higher particle biovolume and flatter slope, 479 suggesting an important role for primary production and phytoplankton size structure for the determination of the PSD at the lower limit of the euphotic zone. 481

UVP5 and other optical observations are not limited to the surface ocean, but are 482 generally highly resolved in the vertical direction, thus enabling fully three-dimensional 483 reconstructions of PSD. This allows a closer investigation of the processes controlling particle abundance in the water column, and makes three-dimensional reconstructions of sinking particle fluxes possible. Enhanced deployments of UVPs—also on Argo floats— 486 combined with the approaches developed in this paper could also enable to decadal or 487 even annual estimates of global PSD and particle flux through the water column. Ul-488 timately, a three-dimensional view of particle export would shed light on the ocean's abil-489 ity to sequester carbon, and inform models of change in the ocean's biological pump. 490

491 Acknowledgments

This material is based upon work supported by the U.S. National Science Foundation 492 under grants No. OCE-1635632 and OCE-1847687. D.B. acknowledges support from the 493 Alfred P. Sloan Foundation, and computational support by the Extreme Science and Engineering Discovery Environment (XSEDE) through allocation TG-OCE17001. A.M.P.M 495 acknowledges support from NSF Award No. 1654663. T.W. was supported by NSF award 496 OCE-1635414. RK acknowledges support via the BMBF funded project CUSCO, the 497 EU project TRIATLAS (European Union's Horizon 2020 programme, grant agreement 498 No 817578) and a "Make Our Planet Great Again" grant of the ANR within the "Pro-499 gramme d'Investissements d'Avenir"; reference "ANR-19-MPGA-0012". Data generated 500 by this analysis has been uploaded to BCO-DMO, DOI:10.26008/1912/bco-dmo.856942.1. 501 The individual UVP5 profiles used to generate the reconstructions can be obtained on the EcoTaxa website https://ecotaxa.obs-vlfr.fr/part/. 503

504 References

Alldredge, A. L., & Gotschalk, C. (1988). In situ settling behavior of marine snow.

506	Limnology and Oceanography, 33(3), 339–351. doi: 10.4319/lo.1988.33.3.0339
507	Antoine, D., & Morel, A. (1996, mar). Oceanic primary production: 1. Adap-
508	tation of a spectral light-photosynthesis model in view of application to
509	satellite chlorophyll observations. Global Biogeochemical Cycles, 10(1),
510	43-55. Retrieved from http://doi.wiley.com/10.1029/95GB02831 doi:
511	10.1029/95GB02831
512	Barton, A. D., Pershing, A. J., Litchman, E., Record, N. R., Edwards, K. F., Finkel,
513	Z. V., Ward, B. A. (2013). The biogeography of marine plankton traits.
514	Ecology Letters, $16(4)$, 522–534. doi: 10.1111/ele.12063
515	Behrenfeld, M. J., & Falkowski, P. G. (1997, jan). Photosynthetic rates derived
516	from satellite-based chlorophyll concentration. Limnology and Oceanography,
517	42(1), 1-20. Retrieved from https://www.google.com/search?q=Engle+et+
518	al.{\%}2C2000{\&}oq=Engle+et+al.{\%}2C2000{\&}aqs=chrome69i57
519	.11777j0j8{\&}sourceid=chrome{\&}ie=UTF-8http://doi.wiley.com/
520	10.4319/10.1997.42.1.0001 doi: 10.4319/10.1997.42.1.0001
521	Bianchi, D., Weber, T. S., Kiko, R., & Deutsch, C. (2018). Global niche of marine
522	anaerobic metabolisms expanded by particle microenvironments. Nature Geo-
523	science, 1–0. Retrieved from http://dx.dol.org/10.1038/s41561-018-0081
524	-0 (001. 10.1050/S41501-010-0001-0 Dichon I K Fong M D & Wood T I (2016) Depotie observations of high
525	uintertime earbon expert in California costal waters Biogeogeoienese 12(10)
526	wintertime carbon export in Camorina coastar waters. $Diogeosciences, 15(10),$ 3109–3120 doi: 10.5194/bg-13-3109-2016
527	Bisson K Siegel D A & DeVries T (2020) Diagnosing Mechanisms of Ocean
520	Carbon Export in a Satellite-Based Food Web Model. Frontiers in Marine
530	Science, 7, 505. Retrieved from https://www.frontiersin.org/article/
531	10.3389/fmars.2020.00505 doi: 10.3389/fmars.2020.00505
532	Boss, E., Guidi, L., Richardson, M. J., Stemmann, L., Gardner, W., Bishop, J. K.,
533	Sherrell, R. M. (2015). Optical techniques for remote and in-situ charac-
534	terization of particles pertinent to geotraces. Progress in Oceanography, 133,
535	43–54.
536	Boyd, P. W., Claustre, H., Levy, M., Siegel, D. A., & Weber, T. (2019). Multi-
537	faceted particle pumps drive carbon sequestration in the ocean. Na-
538	<i>ture</i> , 568(7752), 327–335. Retrieved from https://doi.org/10.1038/
539	s41586-019-1098-2 doi: 10.1038/s41586-019-1098-2
540	Burd, A. B., & Jackson, G. A. (2009a). Particle aggregation. Annual Review of Ma-
541	rine Science, 1(1), 65-90. Retrieved from http://www.annualreviews.org/
542	doi/10.1146/annurev.marine.010908.163904 doi: 10.1146/annurev.marine
543	Burd A B & Lackson C A (2000b) Particle aggregation Annual review of ma
544	rine science 1 65–90
545	Clements D. Vang S. Weber T. McDonnell A. Kiko B. Stemmann L. &
540	Bianchi, D. (2021). Constraining the ocean's biological pump with in situ
548	optical observations and supervised learning. part 2: Carbon flux. In Review
549	for Global Biogeochemical Cycles.
550	Copernicus Climate Change Service. (2017). Era5: Fifth generation of ecmwf
551	atmospheric reanalyses of the global climate. Copernicus Climate Change
552	Service Climate Data Store (CDS). Retrieved from https://cds.climate
553	.copernicus.eu/cdsapp#!/home (accessed: 11-13-2019)
554	Cram, J. A., Weber, T., Leung, S. W., McDonnell, A. M., Liang, J. H., & Deutsch,
555	C. (2018). The Role of Particle Size, Ballast, Temperature, and Oxygen in the
556	Sinking Flux to the Deep Sea. Global Biogeochemical Cycles, 32(5), 858–876.
557	doi: 10.1029/2017GB005710
558	de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., & Iudicone, D. (2004).
559	Mixed layer depth over the global ocean: An examination of profile data and a
560	prome-based climatology. Journal of Geophysical Research C: Oceans, $109(12)$,

561	1-20. doi: $10.1029/2004$ JC002378
562	Devries, T., Liang, J. H., & Deutsch, C. (2014). A mechanistic particle flux model
563	applied to the oceanic phosphorus cycle. <i>Biogeosciences</i> , 11(19), 5381–5398.
564	doi: 10.5194/bg-11-5381-2014
565	DeVries, T., & Weber, T. (2017). The export and fate of organic matter in
566	the ocean. New constraints from combining satellite and oceanographic
500	tracer observations Clobal Biogeochemical Cycles 31(3) 535-555 doi:
507	10 1002/2016CB005551
508	D'Ennice I (2016) Invisit name (mottab control file enchance 2012)
569	D'Errico, J. (2010). Inpaint nans (mattao central jue exchange, 2012).
570	Forest, A., Stemmann, L., Picneral, M., Burdori, L., Robert, D., Fortier, L., &
571	Babin, M. (2012). Size distribution of particles and zooplankton across
572	the shelf-basin system in southeast Beaufort Sea: Combined results from an
573	Underwater Vision Profiler and vertical net tows. Biogeosciences, $9(4)$, 1301–
574	1320. doi: $10.5194/bg-9-1301-2012$
575	Garcia, H., Weathers, K., Paver, C., Smolyar, I., Boyer, T., Locarnini, R., Rea-
576	gan, J. (2018). World Ocean Atlas 2018. Volume 4: Dissolved Inorganic Nu-
577	trients (phosphate, nitrate and nitrate+nitrite, silicate). NOAA Atlas NESDIS
578	84, 84 (July), 35.
579	Garcia, H. E., Weathers, K., Paver, C. R., Smolyar, I., Boyer, T. P., Locarnini,
580	R. A., Reagan, J. R. (2019). World Ocean Atlas 2018, Volume 3: Dissolved
581	Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation. NOAA Atlas
582	<i>NESDIS</i> , 3(83), 38 pp.
583	Gregor, L., & Gruber, N. (2021, 3). Oceansoda-ethz: A global gridded data
584	set of the surface ocean carbonate system for seasonal to decadal stud-
585	ies of ocean acidification. Earth System Science Data, 13, 777-808. doi:
586	10.5194/essd-13-777-2021
587	Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Gorsky,
588	G. (2016). Plankton networks driving carbon export in the oligotrophic ocean.
589	<i>Nature</i> , 532(7600), 465–470. Retrieved from http://dx.doi.org/10.1038/
590	nature16942 doi: 10.1038/nature16942
591	Guidi, L., Jackson, G. A., Stemmann, L., Miquel, J. C., Picheral, M., & Gorsky,
592	G. (2008). Relationship between particle size distribution and flux in the
593	mesopelagic zone. Deep-Sea Research Part I: Oceanoaraphic Research Papers.
594	55(10), 1364–1374. doi: 10.1016/j.dsr.2008.05.014
595	Guidi, L., Legendre, L., Revgondeau, G., Uitz, J., Stemmann, L., & Henson, S. A.
596	(2015 jul) A new look at ocean carbon remineralization for estimating
590	deepwater sequestration Global Biogeochemical Cucles 29(7) 1044–1059
597	Retrieved from http://doi_wilev_com/10_1002/2014GB005063doi:
590	10 1002/2014GB005063
600	Guidi L. Stemmann L. Jackson G. A. Ihanez F. Claustre H. Legendre L.
601	Gorsky G (2009) Effects of phytoplankton community on production
602	size and export of large aggregates: A world-ocean analysis Limpology and
602	Size and export of large aggregates. It world-ocean analysis. $D_{intrology}$ and $O_{cean carranhy}$ 5/(6) 1051–1063 doi: 10.4310/lo.2000.54.6.1051
603	Hamilton D S Sconzo D A Fong V Chinness I Kok I F Li I Ma
604	hamilton, D. S., Scanza, R. A., Feng, T., Guinness, J., Kok, J. F., Li, L., Ma-
605	atmospheric coluble iron and observation comparisons using the Mechanism of
606	Intermediate complexity for Modelling Iron (MIMI v1.0) Conscientific Model
607	$D_{\text{current}} = 10(0) - 2825 - 2862 - d_{\text{cir}} = 10.5104 / \text{aread} = 12.2825 - 2010$
608	Development, 12(9), 3835-3802. doi: 10.5194/gfnd-12-3855-2019
609	Johnson, G. C., Schmidtko, S., & Lyman, J. M. (2012). Relative contributions of
610	temperature and salinity to seasonal mixed layer density changes and horizon-
611	tal density gradients. Journal of Geophysical Research: Oceans, 117(4). doi:
612	10.1029/2011JC007651
613	Karl, D., Knauer, G., Martin, J., & Ward, B. (1984). Bacterial chemolithotrophy in
614	the ocean is associated with sinking particles. Nature, $309(5963)$, 54–56.
615	Kiko, R., Biastoch, A., Brandt, P., Cravatte, S., Hauss, H., Hummels, R., Stem-

616	mann, L. (2017). Biological and physical influences on marine snowfall at the
617	equator. Nature Geoscience, $10(11)$, 852–858. doi: $10.1038/NGEO3042$
618	Kiko, R., Picheral, M., Antoine, D., Babin, M., Berline, L., Biard, T., Stem-
619	mann, L. (2021). The global marine particle size distribution dataset obtained
620	with the underwater vision profiler 5 - version 1. PANGAEA.
621	Kostadinov, T. S., Cabré, A., Vedantham, H., Marinov, I., Bracher, A., Brewin,
622	R. J., Uitz, J. (2017). Inter-comparison of phytoplankton functional type
623	phenology metrics derived from ocean color algorithms and earth system mod-
624	els. Remote Sensing of Environment, 190, 102-177. Retrieved from https://
625	www.sciencedirect.com/science/article/pii/s003442571050459X doi. bttps://doi.org/10.1016/j.rso.2016.11.014
626	Kostadinov T. S. Siegel D. A. & Maritanana S. (2000) Patriaval of the particle
627	size distribution from satellite ocean color observations. <i>Lowrnal of Coonbusieal</i>
628	Research: Oceans 11/(9) 1–22 doi: 10.1029/2009JC005303
629	Kostadinov T S Siegel D A & Maritorena S (2010a) Global variability of
631	phytoplankton functional types from space: assessment via the particle size
632	distribution Biogeosciences $7(10)$ 3239–3257 Betrieved from https://
633	bg.copernicus.org/articles/7/3239/2010/ doi: 10.5194/bg-7-3239-2010
634	Kostadinov, T. S., Siegel, D. A., & Maritorena, S. (2010b). Global variability of
635	phytoplankton functional types from space: assessment via the particle size
636	distribution. Biogeosciences, 7(10), 3239–3257. Retrieved from https://
637	bg.copernicus.org/articles/7/3239/2010/ doi: 10.5194/bg-7-3239-2010
638	Kriest, I. (2002). Different parameterizations of marine snow in a 1D-model and
639	their influence on representation of marine snow, nitrogen budget and sedimen-
640	tation. $Deep$ -Sea Research Part I: Oceanographic Research Papers, $49(12)$,
641	2133–2162. doi: $10.1016/S0967-0637(02)00127-9$
642	Kwon, E. Y., Primeau, F., & Sarmiento, J. L. (2009). The impact of remineraliza-
643	tion depth on the air-sea carbon balance. Nature Geoscience, $2(9)$, 630–635.
644	Liu, H., Li, Q., Bai, Y., Yang, C., Wang, J., Zhou, Q., Wu, G. (2021). Improving
645	satellite retrieval of oceanic particulate organic carbon concentrations using
646	machine learning methods. Remote Sensing of Environment, 256 (January),
647	112316. Retrieved from https://doi.org/10.1016/j.rse.2021.112316 doi:
648	10.1010/J.rse.2021.112310
649	Locarnini, R. A., Misnonov, A. V., Baranova, O. K., Boyer, I. P., Zweng, M. M.,
650	Temperature A Michoney Technical Editor NOAA Atlas NESDIS 1(81)
651	52pp
052	Lombard F. Boss F. Waite A. M. Voot M. Uitz, I. Stemmann L
653	(2019) Globally consistent quantitative observations of planktonic ecosystems
655	Frontiers in Marine Science, 6, 196.
656	Morel, A., Huot, Y., Gentili, B., Werdell, P. J., Hooker, S. B., & Franz, B. A.
657	(2007). Examining the consistency of products derived from various ocean
658	color sensors in open ocean (Case 1) waters in the perspective of a multi-
659	sensor approach. Remote Sensing of Environment, 111(1), 69–88. doi:
660	10.1016/j.rse.2007.03.012
661	Mouw, C. B., Hardman-Mountford, N. J., Alvain, S., Bracher, A., Brewin, R. J.,
662	Bricaud, A., others (2017). A consumer's guide to satellite remote sensing
663	of multiple phytoplankton groups in the global ocean. Frontiers in Marine
664	<i>Science</i> , 4, 41.
665	Myriokefalitakis, S., Ito, A., Kanakidou, M., Nenes, A., Krol, M. C., Mahowald,
666	N. M., Duce, R. A. (2018). Reviews and syntheses: The GESAMP atmo-
667	spheric iron deposition model intercomparison study. Biogeosciences, $15(21)$,
668	0059-0084. doi: 10.5194/bg-15-6659-2018
669	NASA G.S.F.C. (2014). Modis-aqua ocean color data. NASA Goddard Space Flight
670	Center, Ocean Ecology Laboratory, Ocean Biology Processing Group. doi: dx

671	.doi.org/10.5067/AQUA/MODIS OC.2014.0
672	N.G.D.C. (2006). 2-minute gridded global relief data (etopo2) v2. National Geophys-
673	ical Data Center, NOAA. (accessed: 11-13-2019) doi: 10.7289/V5J1012Q
674	Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of
675	random forest permutation-based variable importance measures under predic-
676	tor correlation. BMC bioinformatics, 11(1), 1–13.
677	Palevsky, H. I., & Doney, S. C. (2018). How Choice of Depth Horizon Influ-
678	ences the Estimated Spatial Patterns and Global Magnitude of Ocean Car-
679	bon Export Flux. <i>Geophysical Research Letters</i> , 45(9), 4171–4179. doi:
680	10.1029/2017GL076498
681	Picheral, M., Guidi, L., Stemmann, L., Karl, D. M., Iddaoud, G., & Gorsky, G.
682	(2010). The underwater vision profiler 5: An advanced instrument for high
683	spatial resolution studies of particle size spectra and zooplankton. Limnology
684	and Oceanography: Methods, 8(SEPT), 462–473. doi: 10.4319/lom.2010.8.462
685	Roullier, F., Berline, L., Guidi, L., Durrieu De Madron, X., Picheral, M., Sciandra,
686	A., Stemmann, L. (2014). Particle size distribution and estimated carbon
687	flux across the Arabian Sea oxygen minimum zone. Biogeosciences, 11(16),
688	4541-4557. Retrieved from https://bg.copernicus.org/articles/11/4541/
689	2014/ doi: 10.5194/bg-11-4541-2014
690	Roy, S., Sathyendranath, S., Bouman, H., & Platt, T. (2013). The global distribu-
691	tion of phytoplankton size spectrum and size classes from their light-absorption
692	spectra derived from satellite data. Remote Sensing of Environment, 139,
693	185-197. Retrieved from http://dx.doi.org/10.1016/j.rse.2013.08.004
694	doi: 10.1016/j.rse.2013.08.004
695	Sarmiento, J. L., & Gruber, N. (2006). Ocean biogeochemical dynamics. Prince-
696	ton University Press. Retrieved from http://www.jstor.org/stable/
697	j.ctt3fgxqx
698	Siegel, D. A., Buesseler, K. O., Doney, S. C., Sailley, S. F., Behrenfeld, M. J., &
699	Boyd, P. W. (2014). Global assessment of ocean carbon export by combining
700	satellite observations and food-web models. Global Biogeochemical Cycles,
701	28(3), 181-196. doi: $10.1002/2013$ GB004743
702	Siegel, D. A., Doney, S. C., & Yoder, J. A. (2002). The North Atlantic spring phyto-
703	plankton bloom and Sverdrup's critical depth hypothesis. Science, $296(5568)$,
704	730-733. doi: $10.1126/science.1069174$
705	Silsbe, G. M., Behrenfeld, M. J., Halsey, K. H., Milligan, A. J., & Westberry,
706	T. K. (2016). The CAFE model: A net production model for global ocean
707	phytoplankton. Global Biogeochemical Cycles, $3\theta(12)$, 1756–1777. doi:
708	10.1002/2016GB005521
709	Stemmann, L., & Boss, E. (2012). Plankton and Particle Size and Packaging:
710	From Determining Optical Properties to Driving the Biological Pump. An-
711	nual Review of Marine Science, 4(1), 263–290. Retrieved from http://
712	www.annualreviews.org/doi/10.1146/annurev-marine-120/10-100853
713	doi: $10.1146/annurev-marine-120/10-100853$
714	Stemmann, L., Gorsky, G., Marty, JC., Picheral, M., & Miquel, JC. (2002).
715	Four-year study of large-particle vertical distribution (0-1000m) in the nw
716	Sog Research Part II. Tenical Studies in Oceanomaphy. (0(11), 2142-2162
717	Retrieved from https://www.sciencedirect.com/science/article/nii/
710	S0967064502000322 (Studies at the DVFAMED (France ICOFS) Time
730	Series Station NW M editerranean Sea) doi: https://doi.org/10.1016/
720	S0967-0645(02)00032-2
722	Stemmann L. Jackson G A & Janson D (2004) A vertical model of par-
723	ticle size distributions and fluxes in the midwater column that includes
724	biological and physical processes - Part I: Model formulation. Deen-Sea
725	Research Part I: Oceanographic Research Papers. 51(7). 865–884. doi:
2	

726	10.1016/j.dsr.2004.03.001
727	Stemmann, L., Youngbluth, M., Robert, K., Hosia, A., Picheral, M., Paterson, H.,
728	\ldots Gorsky, G. (2008). Global zoogeography of fragile macrozooplankton in the
729	upper 100-1000 m inferred from the underwater video profiler. ICES Journal
730	of Marine Science, $65(3)$, 433–442. doi: 10.1093/icesjms/fsn010
731	Turner, J. T. (2015). Zooplankton fecal pellets, marine snow, phytodetritus and the
732	ocean's biological pump. Progress in Oceanography, 130, 205–248.
733	Ward, B. A., Dutkiewicz, S., & Follows, M. J. (2014). Modelling spatial and tem-
734	poral patterns in size-structured marine plankton communities: Top-down
735	and bottom-up controls. Journal of Plankton Research, $36(1)$, $31-47$. doi:
736	10.1093/plankt/fbt097
737	Westberry, T., Behrenfeld, M. J., Siegel, D. A., & Boss, E. (2008). Carbon-based
738	primary productivity modeling with vertically resolved photoacclimation.
739	Global Biogeochemical Cycles, $22(2)$, 1–18. doi: $10.1029/2007$ GB003078
740	Yang, S., Chang, B. X., Warner, M. J., Weber, T. S., Bourbonnais, A. M., Santoro,
741	A. E., Bianchi, D. (2020). Global reconstruction reduces the uncertainty
742	of oceanic nitrous oxide emissions and reveals a vigorous seasonal cycle. <i>Pro-</i>
743	ceedings of the National Academy of Sciences of the United States of America,
744	117(22). doi: 10.1073/pnas.1921914117
745	Zhang, G., & Lu, Y. (2012). Bias-corrected random forests in regression. <i>Journal of</i>
746	Applied Statistics, 39(1), 151-160. Retrieved from https://doi.org/10.1080/
747	02664763.2011.578621 doi: 10.1080/02664763.2011.578621
748	Zweng, M. M., Reagan, J. R., Seidov, D., Boyer, T. P., Antonov, J. I., Locarnini,
749	R. A., Smolyar, I. V. (2019). World Ocean Atlas 2018, Volume 2: Salinity.

NOAA Atlas NESDIS, 2(82), 50.

Supporting Information for Constraining the ocean's biological pump with in situ optical observations and supervised learning. Part 1: particle size distributions

D.J. Clements¹, S. Yang¹, T. Weber², A.M.P. McDonnell³, R.Kiko⁴,

L.Stemmann⁴, D.Bianchi¹

¹Department of Atmospheric and Oceanic Sciences, University of California Los Angeles, Los Angeles, CA, USA.

 2 Department of Earth and Environmental Sciences, University of Rochester, Rochester, New York, USA

³College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Fairbanks, Alaska 99775-7220, USA.

⁴Sorbonne Université, CNRS, UMR 7093, Institut de la Mer de Villefranche sur mer, Laboratoire d'Océanographie de Villefranche,

Villefranche-sur-Mer, France

Contents of this file

1. Figures S1 to S4

Corresponding author: D.J. Clements (dclements@atmos.ucla.edu)

Corresponding author: D.Bianchi (dbianchi@atmos.ucla.edu)



Correlation Coefficient (r^2) of the powelaw fits through the observed particle size distribution. Each dot represents an individual profile used. Fits with $R^2 < 90$ are removed.

Figure S1. The correlation coefficient (R^2) of the linear least squares fits between the log of the particle counts and the log of their size. The slope and intercept of these fits are the powerlaw parameters to describe the PSD. r-squared less than 0.9 are removed from the dataset, before extrapolating globally.



Figure S2. The relative importance of each predictor for each predicted variable from the Euphotic zone. The y-axis shows the r-squared of a random forest, using all predictors at and to the point along the x-axis. It shows the strength of adding additional predictors. A) shows the ranking for the Biovolume, and B) for the Slope.



Figure S3. Particulate carbon biovolume at the base of the euphotic zone reconstructed from the random forest calculations, showing monthly climatologies.



Figure S4. Particulate carbon PSD slope at the base of the euphotic zone reconstructed from the random forest calculations, showing monthly climatologies.