



HAL
open science

Global observation of plankton communities from space

Hiroto Kaneko, Hisashi Endo, Nicolas Henry, Cédric Berney, Frédéric Mahé,
Julie Poulain, Karine Labadie, Odette Beluche, Roy El Hourany, Samuel
Chaffron, et al.

► To cite this version:

Hiroto Kaneko, Hisashi Endo, Nicolas Henry, Cédric Berney, Frédéric Mahé, et al.. Global observation of plankton communities from space. 2022. hal-03855690

HAL Id: hal-03855690

<https://hal.science/hal-03855690>

Preprint submitted on 16 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 **Global observation of plankton communities**

2 **from space**

3 Hiroto Kaneko¹, Hisashi Endo¹, Nicolas Henry^{2,3}, Cédric Berney^{2,4}, Frédéric Mahé^{5,6}, Julie
4 Poulain⁷, Karine Labadie⁸, Odette Beluche⁸, Roy El Hourany⁹, *Tara* Oceans Coordinators[†],
5 Samuel Chaffron^{3,10}, Patrick Wincker⁷, Ryosuke Nakamura¹¹, Lee Karp-Boss¹², Emmanuel
6 Boss¹², Chris Bowler¹³, Colomban de Vargas^{2,4}, Kentaro Tomii^{14,*}, Hiroyuki Ogata^{1,*}

7

8 ¹Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan

9 ²CNRS, Sorbonne Université, FR2424, ABiMS, Station Biologique de Roscoff, 29680

10 Roscoff, France

11 ³Research Federation for the study of Global Ocean Systems Ecology and Evolution,

12 FR2022/Tara GOSEE, 75016 Paris, France

13 ⁴Sorbonne Université, CNRS, Station Biologique de Roscoff, UMR7144, ECOMAP, 29680

14 Roscoff, France.

15 ⁵UMR PHIM, CIRAD, Montpellier, France

16 ⁶PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD,

17 Montpellier, France

18 ⁷Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry,

19 Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France

20 ⁸Genoscope, Institut François Jacob, Commissariat à l'Energie Atomique (CEA), Université

21 Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France

22 ⁹Univ. Littoral Côte d'Opale, Univ. Lille, CNRS, IRD, UMR 8187, LOG, Laboratoire

23 d'Océanologie et de Géosciences, F 62930 Wimereux, France

24 ¹⁰Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes,
25 France

26 ¹¹Digital Architecture Research Center, National Institute of Advanced Industrial Science and
27 Technology (AIST), Tokyo, Japan

28 ¹²School of Marine Sciences, University of Maine, Orono, Maine 04469, USA.

29 ¹³Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'École
30 Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005
31 Paris, France.

32 ¹⁴Artificial Intelligence Research Center, National Institute of Advanced Industrial Science
33 and Technology (AIST), Tokyo, Japan

34

35 * Corresponding Authors:

36 Kentaro Tomii (Email address: k-tomii@aist.go.jp)

37 Hiroyuki Ogata (Email address: ogata@kuicr.kyoto-u.ac.jp)

38

39 †The *Tara* Oceans Coordinators are listed in Author Contributions Section.

40

41

42 **Abstract**

43 Satellite remote sensing from space is a powerful way to monitor the global dynamics of
44 marine plankton. Previous research has focused on developing models to predict the size or
45 taxonomic groups of phytoplankton. Here we present an approach to identify representative
46 communities from a global plankton network that included both zooplankton and
47 phytoplankton and using global satellite observations to predict their biogeography. Six
48 representative plankton communities were identified from a global co-occurrence network
49 inferred using a novel rDNA 18S V4 planetary-scale eukaryotic metabarcoding dataset.
50 Machine learning techniques were then applied to train a model that predicted these
51 representative communities from satellite data. The model showed an overall 67% accuracy
52 in the prediction of the representative communities. The prediction based on 17 satellite-
53 derived parameters showed better performance than based only on temperature and/or the
54 concentration of chlorophyll *a*. The trained model allowed to predict the global
55 spatiotemporal distribution of communities over 19-years. Our model exhibited strong
56 seasonal changes in the community compositions in the subarctic-subtropical boundary
57 regions, which were consistent with previous field observations. This network-oriented
58 approach can easily be extended to more comprehensive models including prokaryotes as
59 well as viruses.

60

61 **Introduction**

62 Monitoring the global dynamics of marine plankton is essential to understand the function of
63 marine microbial ecosystem and its interaction and evolution with climate change. It can also
64 facilitate the discovery of new plankton species. However, it is impossible to obtain global
65 plankton samples at high spatial and temporal density using research ships alone due to the
66 extent of the ocean. Regular and global remote sensing using satellites can potentially be used

67 to solve this problem. Because plankton pigments absorb light, the spectrum of light reflected
68 from the ocean surface that is observed by satellites (remote sensing reflectance) has a
69 specific relationship with the plankton composition. Environmental parameters such as sea
70 surface temperature (SST) are also related with the plankton composition [1].

71 Several models for predicting plankton community satellite-derived data have been
72 developed over the past decades [2, 3]. Most of them focused on phytoplankton because these
73 species contain pigments, such as chlorophylls, carotenoids, and phycobilins, to capture light
74 energy for photosynthesis [4]. The abundances of three size classes, micro-phytoplankton (>
75 20 μm), nano-phytoplankton (2–20 μm) and pico-phytoplankton (0.2–2 μm), can be
76 predicted with simple models only integrating the concentration of chlorophyll *a* ([Chl]),
77 which is the core of the photosynthetic unit [5–7]. More advanced models have also been
78 developed to predict size classes using remote sensing reflectance [8–11]. The abundance of
79 taxonomic groups of phytoplankton is another target of predictive models. The abundance of
80 diatoms, prymnesiophytes (haptophytes), green algae and *Prochlorococcus* can be predicted
81 using [Chl] [5]. Another model named PhytoDOAS uses remote sensing reflectance data at
82 high spectral resolution for predicting the abundance of coccolithophores, dinoflagellates,
83 cyanobacteria, and diatoms [12, 13]. Some models have been developed to predict the
84 dominant size or taxonomic groups rather than the abundance [14–16]. One of this type of
85 models named PHYSAT can predict communities dominated by diatoms, haptophytes,
86 *Prochlorococcus* and *Synechococcus* defined by pigment concentration ratio [14, 15].
87 Another model has been developed to predict the distribution of the biogeochemical
88 provinces [17]. However, zooplankton abundances have remained difficult to predict from
89 satellite-derived data as they do not perform photosynthesis and tend to be transparent [18].
90 Nevertheless, some copepods harbor the carotenoid astaxanthin and their swarms can be
91 observed from space [19].

92 These previous methods present a limitation regarding the numbers of defined
93 plankton groups, as most of them are based on empirical relationships between pigments and
94 light absorption. Even though these methods allow to provide a synoptic view on the
95 spatiotemporal extent of the main groups of phytoplankton, they are lacking taxonomic
96 resolution and do not reproduce the complexity of a planktonic community. In order to tackle
97 this point, here, we present a machine learning trained model for global satellite observations
98 of the representative communities as captured by a global ocean plankton network. Its targets
99 are mixed auto-, hetero- and mixotrophic protist communities delineated from rDNA 18S V4
100 metabarcoding data at a high taxonomic resolution. Our approach is a network-oriented one,
101 which was inspired by the Bayesian network model used to predict the metabarcoding-based
102 bacterial composition in the English Channel [20]. There are two difficulties in predicting the
103 species composition directly from satellite-derived data. The first difficulty is the substantial
104 number of response variables as compared to predictor variables. There are hundreds of
105 thousands of species represented in the metabarcoding dataset but only nine bands of visible
106 light acquired by multispectral sensors are available as predictor variables. The second
107 difficulty is the small number of samples. In this study, we used the largest available
108 compilation of eukaryotic metabarcoding data, complemented with a novel sequence data
109 from the *Tara Oceans* expeditions, but only a few hundred samples were available for
110 analysis after appropriate filtration. By focusing on ecological networks, these two
111 difficulties were alleviated by reducing the number of variables (dimensionality) in the
112 metabarcoding data. Ecological networks tend to be structured, and are non-randomly
113 assembled [21]. Indeed, a previous study showed through an unsupervised approach for
114 community delineation that the global plankton network is self-organized by marine biomes
115 [22]. We took advantage of this property of plankton networks to reduce dimensionality and
116 convert the problem into a multiclass prediction.

117

118 **Materials and Methods**

119

120 **Satellite data**

121 Ocean color data acquired by the Moderate Resolution Imaging Spectroradiometer (MODIS)
122 on board the Aqua and the Terra satellites were used in this study. Level-3 data, mapped to a
123 5° (ca. 9 km on the Equator) square monthly grid, were downloaded from the Ocean Color
124 Web operated by NASA (<https://oceancolor.gsfc.nasa.gov/>). The data included 17 parameters
125 consisting of remote sensing reflectance ($R_{rs}(\lambda)$) from 10 visible light wavelengths (412, 443,
126 469, 488, 531, 547, 555, 645, 667, and 678 nm), six environmental parameters derived from
127 R_{rs} ([Chl], diffuse attenuation coefficient for downwelling irradiance at 490 nm, particulate
128 organic/inorganic carbon concentration, photosynthetically available radiation, and
129 normalized fluorescence line height), and SST based on infrared measurements. The data
130 were acquired from January 2003 to December 2021. To reduce the number of missing
131 values, the data from both satellites were used. In case the values from both satellites were
132 available for a grid, averaged values were used because they were well correlated (Fig. S1).

133

134 **Two-dimensional (2-D) projection of satellite-derived parameters**

135 To capture the range of all possible satellite-derived parameter values, a 2-D projection was
136 performed by randomly selecting grid cells. Twenty thousand grid cells were randomly
137 selected from all the 5° square grids. After removing grid cells on land or in coastal regions
138 and those with missing data, 8,419 grid cells remained (Fig. S2). A sampling month was
139 randomly selected from 120 months (January 2009 to December 2018) for each grid cell. The
140 satellite-derived parameters for these randomly selected grid cells and months were
141 standardized by subtracting the mean and scaling to unit variance. Finally, the 8,419 points

142 with the 17 parameters were projected onto a 2-D map by Uniform Manifold Approximation
143 and Projection (UMAP) using the Python package `umap-learn` [23].

144

145 **Metabarcoding data**

146 Amplicon sequence data (837 127 965 million reads) targeting 18S V4 regions from 1 011
147 samples (1 191 datasets) collected through the *Tara* Oceans expeditions were generated and
148 registered under the EMBL/EBI-ENA EukBank project. Raw sequencing data were
149 downloaded from the EMBL/EBI-ENA EukBank umbrella project in their native format.
150 When applicable, reads were merged and trimmed (`vsearch` [24], `cutadapt` [25]) to cover the
151 18S V4 region, as defined by the primers TAREuk454FWD1 and TAREukREV3, resulting in
152 347 327 830 unique sequences, representing 1 672 099 024 reads. After clustering (`swarm`
153 [26], chimera detection (`uchime` [27]), quality-based filtering, and post-treatments based on
154 occurrence patterns (`swarm`, `lulu` [28]; <https://github.com/frederic-mahe/mumu>),
155 representative sequences were pairwise compared to the 18S rDNA database EukRibo [29],
156 using a global pairwise alignment approach (`usearch_global` `vsearch`'s command), and
157 taxonomically assigned to their best hit (<https://github.com/frederic-mahe/stampa/>). The
158 filtered occurrence table contains 460 147 swarms (here after referred to as amplicon
159 sequence variants (ASVs)), representing 1 403 019 176 reads, collected from 15 562 samples.

160 As for the usage of the EukBank occurrence table for the analysis, the raw number of
161 reads was rarefied to 10 000 reads per sample. A total of 1 715 samples from the ocean
162 surface (depth < 10 m) with spatiotemporal metadata were retained. These came from several
163 ocean sampling projects such as *Tara* Oceans [1], Malaspina [30], and Australian
164 Microbiome [31]. Occurrences in sequencing replicates from *Tara* Oceans were averaged.
165 Samples from *Tara* Oceans were size fractionated into several classes (e.g. 0.8–5, 5–20, 20–180,
166 and 180–2000 μm), but most samples from other projects were not size fractionated (0.2–3 μm

167 or $> 0.2 \mu\text{m}$). The samples from four size fractions mainly targeting piconano-plankton (0.2–
168 $3 \mu\text{m}$, $> 0.2 \mu\text{m}$, $0.8\text{--}5 \mu\text{m}$, and $> 0.8 \mu\text{m}$) were relatively similar in taxonomic composition
169 (Fig. S3) and were used in this study to maximize the number of samples available for
170 analysis. They were averaged inside each of the 653 bins that match one-by-one with the 5°
171 square monthly satellite data grid. Although more than one sample from different size
172 fractions, sampling locations and times were assigned to a single bin, samples in a same bin
173 were more similar compared to samples from different bins (Fig. S4). Hereafter, we call these
174 bins “samples”.

175

176 **Spatial resampling**

177 A total of 653 metabarcoding samples from previous processing were further filtered using
178 the following procedure. Samples with missing satellite data values owing to bad weather or
179 other reasons were removed. Samples from locations where the sea floor was shallower than
180 200 m were classified as coastal samples and were removed following previous
181 recommendations [32] using a global relief model [33]. Samples were thinned so that they
182 were separated by a minimum of 200 km from each other, using the R package spThin [34].
183 This procedure resulted in 177 samples available for analysis (Fig. S5).

184

185 **Network inference**

186 ASVs were selected by their occurrence to reduce the number of ASVs to those similar to
187 previous studies that analyzed network structures [35, 36]. Two hundreds and eight ASVs
188 with a minimum occurrence larger than 0.2% (20 reads) in at least 10% of samples (18
189 samples) were retained (Figs. S6). ASV read counts were centered log-ratio (clr) transformed
190 [37]. An ecological network was inferred based on co-occurrence patterns using the Julia
191 package FlashWeave [38] with the settings “heterogeneous=False”, “sensitive=True”, and

192 “alpha=0.05” as in previous studies [38, 39]. FlashWeave is a package for calculating partial
193 Pearson’s correlation coefficient between ASV pairs using a recursive approach. The nodes
194 in the obtained network were ASVs, and the edges were made based on correlations between
195 ASV pairs. Only positive correlations (edges) were considered here. For network community
196 detection in the network, Fast Greedy, Infomap, Label Propagation, Leading Eigenvector,
197 Leiden, Louvain, Spinglass, and Walktrap algorithms were applied using the R package
198 “igraph” (<https://igraph.org/>). To measure the structure of the detected community division,
199 we used the modularity index Q as defined by the following equation:

$$Q = \frac{1}{2S} \sum_{u,v} \left(\sigma(u, v) - \frac{k_u k_v}{2S} \right) \delta(C_u, C_v)$$

200 where u, v are nodes (ASVs), $\sigma(u, v)$ is an edge weight (partial correlation coefficient)
201 between u and v , S is the sum of all edge weights, k_u is a weighted degree of node u , C_u is a
202 community to which node u belongs, and $\delta(x, y)$ is 1 if $x=y$ and 0 otherwise [40].

203

204 **Edge satisfaction**

205 We defined an “edge satisfaction index” to determine which community dominated each
206 sample. If C is a community and i is a sample, then the edge satisfaction index of C and i is
207 defined by

$$ES_{C,i} = \frac{\sum_{u,v \in C} \sigma(u, v) \min(p_i(u), p_i(v))}{\sum_{u,v \in C} \sigma(u, v)}$$

208 where u, v are nodes, $\sigma(u, v)$ is an edge weight between u and v , $p_i(u)$ is a weight of node u ,
209 which is the sigmoid transformation of the clr-transformed read count of ASV u in sample i .

210 Briefly, this index measures the ratio of the number of edges between existing nodes in a
211 given sample and the number of all the edges within a given community. The nodes and
212 edges have a weight between 0 to 1 (because only positive correlations were considered). The
213 edge satisfaction index is thus also between 0 to 1.

214

215 **Machine learning and cross-validation**

216 Several machine learning algorithms were used to train predictive models of the
217 representative community based on satellite-derived data. In addition to the satellite-derived
218 parameters, spatial parameters (longitude and latitude) were also tested for their ability in the
219 prediction. The sine and cosine of the longitude were used as independent parameters
220 because it is circular (-180° and 180° are the same). K-nearest Neighbors, Naïve Bayes,
221 Multilayer Perceptron, Random Forest, and Support Vector Machine (SVM) were applied
222 using the Python package “scikit-learn” (<https://scikit-learn.org/>). In the training process for
223 all the methods except Random Forest, the satellite-derived and spatial parameters were
224 standardized by subtracting the mean and scaling to unit variance. The hyperparameters that
225 were tuned with the grid search are shown in [Table S1](#). Both leave-one-out cross-validation
226 and buffered cross-validation [41] were used to measure the model accuracy. In the buffered
227 cross-validation, a test sample is chosen like leave-one-out, but samples inside a buffer region
228 among the test sample were excluded from training samples. The buffer was set to 2,000 km
229 radius. In each fold of the training, hyperparameters were chosen with exhaustive search
230 using the implementation of grid search in scikit-learn. The class prediction output of each
231 method was used to measure accuracy, and output probabilities were used to calculate the
232 receiver operating characteristic (ROC) curve.

233

234 **Time series prediction**

235 The predictive model was trained again with all 177 samples. A 5-fold grid search was used
236 to choose hyperparameters. To reduce the computational cost, a grid cell at the center of each
237 12 by 12 grids square was chosen. In other words, a grid cell was chosen per every 1° square

238 monthly grid cells because the original grid is 5° square. The trained model was applied to
239 this 1° square grid data set for January 2003 to December 2021.

240

241 **Results**

242

243 **2-D map of points with 17 satellite-derived parameters**

244 We generated a 2-D map of points with 17 satellite-derived parameters using UMAP to
245 observe the parameter ranges (Fig. 1). More than eight thousand points were used to train
246 UMAP. These points were randomly selected from all possible locations and times to
247 document the shape of the “continents” in UMAP, which represent the possible range of
248 values of the satellite parameters (Fig. S2). Points associated with the EukBank
249 metabarcoding samples were scattered among most of the regions in the UMAP continents
250 (excluding a region indicated by an arrow). This result indicates that the metabarcoding data
251 covered a wide range of the parameter space and are suitable for being analyzed in terms of
252 their relationship with satellite data, although the number of samples was not large.

253

254 **Network inference and community detection**

255 The ecological network based on ASV co-occurrence patterns was inferred using the
256 FlashWeave algorithm. In the network, 560 positive edges (correlation coefficients > 0)
257 between 208 ASVs were detected (Fig. 2A). We applied several community detection
258 algorithms on the network. The communities detected by Leiden and Spinglass algorithms
259 had the highest modularity index (0.55) (Fig. S7). In the following analysis, the communities
260 detected by the Leiden algorithm [42] were used because it captured the macro structure
261 better than the others (i.e., there were no small communities) (Fig. S7). Among the six
262 detected communities, community 1 was well separated from the other five communities,

263 which formed one super community having highly aggregated community structure (Fig.
264 2B). In the super community, communities 2 and 5 were strongly connected with
265 communities 3 and 6, respectively (Fig. 2B).

266 The taxonomic breakdown of each community is shown in Table 1 and Fig. S8. The
267 well-separated community 1 mainly consisted of Dinoflagellata (mainly Dinophyceae), but
268 Dictyochophyceae (silicoflagellates) and Prymnesiophyceae (haptophytes) were also
269 included. Other five communities had different characteristics in terms of taxonomy.
270 Communities 5 and 6 were dominated by Dinoflagellata (mainly MALV-I and MALV-II),
271 but communities 2 and 3 contained some Arthropoda. Community 4 consisted of half
272 Dinoflagellata and half a variety of other taxa. See Data S1 for taxonomic annotation and
273 assigned community for each ASVs.

274

275 **Representative community of samples**

276 The newly proposed edge satisfaction index was used to measure the completeness of the
277 communities in each sample (see methods). Briefly, this index is one in case all edges within
278 a given community exist, and it is zero in case no edges exist. Fig. 3A shows the edge
279 satisfaction index of all the samples and communities. Notably, community 1 was an
280 exclusively assigned community for some of the samples. The community with the highest
281 edge satisfaction index was considered as the representative community of the sample. The
282 geographic distribution of the representative communities is shown in Fig. 3B. Community 1
283 was associated with high latitude regions, including the Arctic and the Southern Oceans.
284 Communities 3 and 6 were mainly seen in tropical regions of the Pacific and the Indian
285 Oceans, respectively. The other three communities were associated with mid-latitude regions.

286 In the 2-D map of satellite parameter space, samples formed clusters of representative
287 communities (Fig. S9). For example, communities 1 and 5 dominated the bottom of the small

288 and large continents of the map, respectively. This distribution implies a relationship between
289 the satellite parameters and the representative communities.

290

291 **Prediction performance**

292 We applied several machine learning algorithms to classify the representative communities
293 based on satellite parameters. Among five machine learning methods we used, SVM
294 achieved the highest prediction accuracy and micro-average area under the ROC curve
295 (ROC-AUC) (Table S2). Using leave-one-out cross-validation, the accuracy and the ROC-
296 AUC of SVM were 0.67 and 0.90, respectively (Figs. 4A and 4B). Using buffered cross-
297 validation, which excluded neighbors of a test sample from training samples, the measures
298 were reduced to 0.54 and 0.83, respectively (Figs. 4C and 4D).

299 We compared the prediction performance when different sets of satellite-derived and
300 spatial parameters were used (Table 2, Figs S10 and S11). For the prediction only using
301 spatial parameters (latitude and sine/cosine of longitude), the ROC-AUC dropped from 0.91
302 to 0.59 (close to 0.50, i.e., random prediction) when we changed the cross-validation method
303 from leave-one-out to buffered. In contrast, there was a small decrease from 0.90 to 0.83 for
304 the prediction with the satellite-derived parameters. This result indicates the advantage of
305 using satellite-derived parameters to classify the representative communities when spatial
306 biases are appropriately controlled. The prediction performance with only SST or [Chl] was
307 not as good compared to the one with all satellite-derived parameters, but it was improved
308 when SST and [Chl] were combined. Adding other five environmental parameters to SST and
309 [Chl] further improved the performance but it was still slightly worse than that with all
310 satellite-derived parameters (including SST, environmental parameters and R_{rs}).

311

312 **Time series prediction**

313 A full SVM model of community prediction was trained with all 177 samples. A 5-fold grid
314 search selected the linear kernel and the L2 penalty parameter $C=1.0$ for the full model. The
315 SST was the most important parameter in the full model, followed by photosynthetically
316 available radiation, and several channels of R_{rs} (Fig. S12). The chosen threshold of the
317 probabilistic output of SVM was 0.28, which gave the highest F1 score in cross-validation
318 (Fig. S13).

319 We applied the full SVM model to predict a 19-year time series, from January 2003 to
320 December 2021, of community distributions (Movie S1). The global community distributions
321 in each season of year 2021 are shown in Fig. 5. Communities 1 and 5 are located at high-
322 and mid-latitudes, respectively. Communities 3 and 6 are tropical. Community 2 fills the gap
323 between communities 5 and 3. Community 4 shows the pattern related to warm currents. It is
324 related to the region of the extensions of the Kuroshio and Gulf Stream in the late autumn and
325 early winter of the northern hemisphere (November–January) and that of the Brazil, Agulhas,
326 and East Australian Currents in the late autumn and early winter of the southern hemisphere
327 (May–July).

328

329 Discussion

330 Here, representative plankton community networks inferred from rDNA 18S V4 global
331 metabarcoding data from EukBank were successfully predicted from satellite data using a
332 machine learning approach. The outputs of this model are the plankton communities in a way
333 similar to the taxonomic group output of the PHYSAT model [15] rather than a quantitative
334 abundance output like the PhytoDOAS model [12, 13]. Our method has two advantages over
335 these latter models. First, the representative community output from our method included all
336 the taxonomic range of microbial eukaryotes from phytoplankton and zooplankton to
337 heterotrophic protists. There have been some attempts to capture zooplankton abundance

338 from satellite data [19, 43], but it is still difficult to capture a global view over the wide
339 taxonomic range. Our network approach can be extended to prokaryotes and viruses (as they
340 are strongly associated with eukaryotic communities [44, 45]), which are also difficult to
341 observe from satellites due to their small size and lack of optical properties.

342 Second, the output of our model was directly connected with the ASVs inferred from
343 the metabarcoding data. An ASV is a unit with a high taxonomic resolution that is
344 operationally treated as a “species”; thus, the representative communities integrate high
345 taxonomic resolution information. For example, dinoflagellates were treated as one group in
346 the PhytoDOAS model [12], whereas they were represented by 136 ASVs that were
347 classified into one of the six different communities in this study (Table 1, Fig. S8 and Data
348 S1). Although the high taxonomic resolution of metabarcoding data is fascinating, handling
349 their limited number of samples and high dimensionality lead several limitations. We used
350 only four size fractions mainly targeting piconano-plankton (0.2–3 μm , > 0.2 μm , 0.8–5 μm ,
351 and > 0.8 μm) to maximize the number of samples available for analysis. By this procedure,
352 taxonomies only observed in larger size fractions (e.g. diatoms) were not included in the
353 network (Fig. S3A). We choose the criteria of ASV selection and the resolution of
354 community detection (e.g. resolution parameter of Leiden algorithm) to reduce the high
355 dimensionality of the metabarcoding data. This choice of parameters naturally reduced the
356 precision and true diversity of detected communities. A benchmarking test of community
357 detection will overcome this limitation in the future.

358 Our results indicated that the performance of the prediction based on SST and [Chl]
359 was relatively high (Table 2, Figs. S10 and S11). This is not unexpected as SST and [Chl]
360 correlate with microbial community structure in the ocean [1]. We also showed that the
361 prediction performance with all 17 satellite-derived parameters (including SST,
362 environmental parameters and R_{rs}) was higher than with only on SST and/or [Chl] (Table 2,

363 **Figs. S10 and S11**). This result suggested the advantage of using additional environmental
364 parameters and R_{rs} to predict communities, but the improvement of the performance was not
365 large. Hyperspectral R_{rs} from future global satellite missions such as PACE [46] will likely
366 improve prediction performance.

367 We used 177 samples (after an appropriate filtration), which was relatively small for
368 applying a machine learning approach. This may explain why the linear SVM was the best
369 predicting algorithm for our problem. More complex and nonlinear algorithms such as
370 Multilayer Perceptron, Random Forest, and kernel SVM overfitted the training dataset during
371 full model training (**Fig. S14**). Although it is not easy to increase the number of samples
372 because of cost, labor, and weather limitations for sampling cruises, the current number,
373 1,715 (and only 177 after binning and thinning) out of 10,772 samples contained in the
374 metabarcoding dataset is quite limited. Thus, plankton samples from the ocean surface should
375 be strategically collected in the future. The 2-D map of satellite-derived parameters for the
376 ocean (**Fig. 1**) can provide guidance for such strategical guided sampling. If there is a one-to-
377 one connection between the satellite data and the microbial community assemblies, regions
378 without EukBank samples in the map will relate to communities that have yet to be observed
379 and will be an important target for future sampling. A region on the map indicated by an
380 arrow is an example of this kind of unexplored region (**Fig. 1**). We found that this region is
381 mainly consisted of points associated with high latitudinal regions of the southern hemisphere
382 and the North Pacific Ocean (**Fig. S15**).

383 The time series prediction of communities using the trained model revealed the
384 spatiotemporal distribution of each community (**Fig. 5 and Video S1**). Generally, community
385 distributions had a rough correspondence with Longhurst biomes [47] (**Fig. S16**). Community
386 1 corresponded with the “polar” biome, community 5 corresponded with the “westerlies”
387 biome, and communities 3 and 6 corresponded with the “trades” biome. The same

388 distribution can be seen in the representative community of each sample used to train the
389 model (Fig. 3B). Considering the latitudinal self-organization previously observed and
390 described in plankton community networks [22], this correspondence showed that the newly
391 proposed edge satisfaction index could appropriately capture the representative communities.
392 Community 4 had a seasonal spatiotemporal distribution possibly related to the extensions of
393 the western boundary currents (Fig. 5 and Video S1). A previous study showed that seasonal
394 changes in environmental variables (phosphate, nitrate, silicate, and dissolved inorganic
395 carbon) were the highest in the extension of the Kuroshio among other regions in the Pacific
396 basin [48]. Furthermore, clear seasonal variations in the abundance of cyanobacterial
397 diazotrophs were observed in the same region [49]. Community 4 connected the two well-
398 connected pairs (communities 2 and 3, 5 and 6) of the super community in the network (Fig.
399 2B) and had relatively high taxonomic diversity (Table 1 and Fig. S8). In a simulation of
400 emergent phytoplankton in the ocean, areas downstream of the western boundary currents
401 showed high species diversity [50]. The target of our model is the global spatiotemporal
402 distribution of communities, while our network-based approach will be applicable to satellite
403 observations of local ecosystems (e.g., Hawaii and Bermuda).

404 In this study, we inferred the ecological network of ASVs using a global
405 metabarcoding dataset and identified six distinct communities. We applied SVM to train the
406 predictive model of these communities based on satellite data and obtained an accuracy of
407 67%. The spatiotemporal distribution of these communities was shown by applying the
408 model to 19 years of global satellite data. Our model was able to predict communities that
409 included phytoplankton, zooplankton, and heterotrophic protists. The network-oriented
410 approach used in this study can be easily extended to identify the distribution of prokaryotes
411 and viruses. Given the ability of the model to predict the spatiotemporal dynamics of
412 plankton communities from space, our combined network-based and machine learning

413 approach provides a particularly useful tool to monitor and survey the impact of
414 environmental and climate change on plankton communities at both local and global scale.

415

416 **Data Availability Statement**

417 Newly sequenced *Tara* Oceans 18S V4 data have been deposited to EMBL/EBI-ENA:
418 PRJEB6610 (*Tara* Oceans), PRJEB9737 (TARA Oceans Polar Circle).

419

420 **References**

- 421 1. Ibarbalz FM, Henry N, Brandão MC, Martini S, Busseni G, Byrne H, et al. Global Trends
422 in Marine Plankton Diversity across Kingdoms of Life. *Cell* 2019; **179**: 1084-1097.e21.
- 423 2. Mouw CB, Hardman-Mountford NJ, Alvain S, Bracher A, Brewin RJW, Bricaud A, et al.
424 A Consumer's Guide to Satellite Remote Sensing of Multiple Phytoplankton Groups in
425 the Global Ocean. *Front Mar Sci* 2017; **4**: 41.
- 426 3. Bracher A, Bouman HA, Brewin RJW, Bricaud A, Brotas V, Ciotti AM, et al. Obtaining
427 Phytoplankton Diversity from Ocean Color: A Scientific Roadmap for Future
428 Development. *Front Mar Sci* 2017; **4**: 55.
- 429 4. Mirkovic T, Ostroumov EE, Anna JM, van Grondelle R, Govindjee, Scholes GD. Light
430 Absorption and Energy Transfer in the Antenna Complexes of Photosynthetic
431 Organisms. *Chem Rev* 2017; **117**: 249–293.
- 432 5. Hirata T, Hardman-Mountford NJ, Brewin RJW, Aiken J, Barlow R, Suzuki K, et al.
433 Synoptic relationships between surface Chlorophyll- *a* and diagnostic pigments specific
434 to phytoplankton functional types. *Biogeosciences* 2011; **8**: 311–327.
- 435 6. Uitz J, Claustre H, Morel A, Hooker SB. Vertical distribution of phytoplankton
436 communities in open ocean: An assessment based on surface chlorophyll. *J Geophys*
437 *Res* 2006; **111**: C08005.

- 438 7. Brewin RJW, Devred E, Sathyendranath S, Lavender SJ, Hardman-Mountford NJ. Model
439 of phytoplankton absorption based on three size classes. *Appl Opt* 2011; **50**: 4535.
- 440 8. Roy S, Sathyendranath S, Bouman H, Platt T. The global distribution of phytoplankton
441 size spectrum and size classes from their light-absorption spectra derived from satellite
442 data. *Remote Sens Environ* 2013; **139**: 185–197.
- 443 9. Devred E, Sathyendranath S, Stuart V, Platt T. A three component classification of
444 phytoplankton absorption spectra: Application to ocean-color data. *Remote Sens
445 Environ* 2011; **115**: 2255–2266.
- 446 10. Li Z, Li L, Song K, Cassar N. Estimation of phytoplankton size fractions based on
447 spectral features of remote sensing ocean color data: PSF FROM OCEAN COLOR
448 FEATURES. *J Geophys Res Oceans* 2013; **118**: 1445–1458.
- 449 11. Kostadinov TS, Siegel DA, Maritorena S. Global variability of phytoplankton functional
450 types from space: assessment via the particle size distribution. *Biogeosciences* 2010; **7**:
451 3239–3257.
- 452 12. Sadeghi A, Dinter T, Vountas M, Taylor BB, Altenburg-Soppa M, Peeken I, et al.
453 Improvement to the PhytoDOAS method for identification of coccolithophores using
454 hyper-spectral satellite data. *Ocean Sci* 2012; **8**: 1055–1070.
- 455 13. Bracher A, Vountas M, Dinter T, Burrows JP, Röttgers R, Peeken I. Quantitative
456 observation of cyanobacteria and diatoms from space using PhytoDOAS on
457 SCIAMACHY data. *Biogeosciences* 2009; **6**: 751–764.
- 458 14. Alvain S, Moulin C, Dandonneau Y, Bréon FM. Remote sensing of phytoplankton groups
459 in case 1 waters from global SeaWiFS imagery. *Deep Sea Res Part Oceanogr Res Pap*
460 2005; **52**: 1989–2004.
- 461 15. Alvain S, Moulin C, Dandonneau Y, Loisel H. Seasonal distribution and succession of
462 dominant phytoplankton groups in the global ocean: A satellite view:

- 463 PHYTOPLANKTON GROUPS - A SATELLITE VIEW. *Glob Biogeochem Cycles*
464 2008; **22**: GB3001.
- 465 16. Hirata T, Aiken J, Hardman-Mountford N, Smyth TJ, Barlow RG. An absorption model
466 to determine phytoplankton size classes from satellite ocean colour. *Remote Sens*
467 *Environ* 2008; **112**: 3153–3159.
- 468 17. Reygondeau G, Longhurst A, Martinez E, Beaugrand G, Antoine D, Maury O. Dynamic
469 biogeochemical provinces in the global ocean: DYNAMIC BIOGEOCHEMICAL
470 PROVINCES. *Glob Biogeochem Cycles* 2013; **27**: 1046–1058.
- 471 18. Johnsen S. Hidden in Plain Sight: The Ecology and Physiology of Organismal
472 Transparency. *Biol Bull* 2001; **201**: 301–318.
- 473 19. Basedow SL, McKee D, Lefering I, Gislason A, Daase M, Trudnowska E, et al. Remote
474 sensing of zooplankton swarms. *Sci Rep* 2019; **9**: 686.
- 475 20. Larsen PE, Field D, Gilbert JA. Predicting bacterial community assemblages using an
476 artificial neural network approach. *Nat Methods* 2012; **9**: 621–625.
- 477 21. Faust K, Raes J. Microbial interactions: from networks to models. *Nat Rev Microbiol*
478 2012; **10**: 538–550.
- 479 22. Chaffron S, Delage E, Budinich M, Vintache D, Henry N, Nef C, et al. Environmental
480 vulnerability of the global ocean epipelagic plankton community interactome. *Sci Adv*
481 2021; **7**: eabg1921.
- 482 23. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and
483 Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* 2020;
484 <https://doi.org/10.48550/arXiv.1802.03426>.
- 485 24. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source
486 tool for metagenomics. *PeerJ* 2016; **4**: e2584.

- 487 25. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
488 *EMPNet.journal* 2011; **17**: 10–12.
- 489 26. Mahé F, Czech L, Stamatakis A, Quince C, de Vargas C, Dunthorn M, et al. Swarm v3:
490 towards tera-scale amplicon clustering. *Bioinformatics* 2021; **38**: 267–269.
- 491 27. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity
492 and speed of chimera detection. *Bioinformatics* 2011; **27**: 2194–2200.
- 493 28. Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, et al. Algorithm
494 for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates.
495 *Nat Commun* 2017; **8**: 1188.
- 496 29. Berney C. EukRibo: a manually curated eukaryotic 18S rDNA reference database. 2022;
497 <https://doi.org/10.5281/zenodo.6327891>. Zenodo.
- 498 30. Salazar G, Cornejo-Castillo FM, Borrull E, Díez-Vives C, Lara E, Vaqué D, et al.
499 Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic
500 prokaryotes. *Mol Ecol* 2015; **24**: 5692–5706.
- 501 31. Brown MV, van de Kamp J, Ostrowski M, Seymour JR, Ingleton T, Messer LF, et al.
502 Systematic, continental scale temporal monitoring of marine pelagic microbiota by the
503 Australian Marine Microbial Biodiversity Initiative. *Sci Data* 2018; **5**: 180130.
- 504 32. Righetti D, Vogt M, Gruber N, Psomas A, Zimmermann NE. Global pattern of
505 phytoplankton diversity driven by temperature and environmental variability. *Sci Adv*
506 2019; **5**: eaau6253.
- 507 33. Amante C. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and
508 Analysis. 2009; <https://doi.org/10.7289/V5C8276M>. National Geophysical Data Center,
509 NOAA.

- 510 34. Aiello-Lammens ME, Boria RA, Radosavljevic A, Vilela B, Anderson RP. spThin: an R
511 package for spatial thinning of species occurrence records for use in ecological niche
512 models. *Ecography* 2015; **38**: 541–545.
- 513 35. Mikolajczak A, Maréchal D, Sanz T, Isenmann M, Thierion V, Luque S. Modelling
514 spatial distributions of alpine vegetation: A graph theory approach to delineate
515 ecologically-consistent species assemblages. *Ecol Inform* 2015; **30**: 196–202.
- 516 36. Baldassano SN, Bassett DS. Topological distortion and reorganized modular structure of
517 gut microbial co-occurrence networks in inflammatory bowel disease. *Sci Rep* 2016; **6**:
518 26087.
- 519 37. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are
520 Compositional: And This Is Not Optional. *Front Microbiol* 2017; **8**: 2224.
- 521 38. Tackmann J, Matias Rodrigues JF, von Mering C. Rapid Inference of Direct Interactions
522 in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data.
523 *Cell Syst* 2019; **9**: 286-296.e8.
- 524 39. Meng L, Endo H, Blanc-Mathieu R, Chaffron S, Hernández-Velázquez R, Kaneko H, et
525 al. Quantitative Assessment of Nucleocytoplasmic Large DNA Virus and Host
526 Interactions Predicted by Co-occurrence Analyses. *mSphere* 2021; **6**: e01298-20.
- 527 40. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks.
528 *Phys Rev E* 2004; **70**: 066111.
- 529 41. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillerá-Arroita G, et al. Cross-
530 validation strategies for data with temporal, spatial, hierarchical, or phylogenetic
531 structure. *Ecography* 2017; **40**: 913–929.
- 532 42. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-
533 connected communities. *Sci Rep* 2019; **9**: 5233.

- 534 43. Behrenfeld MJ, Gaube P, Della Penna A, O'Malley RT, Burt WJ, Hu Y, et al. Global
535 satellite-observed daily vertical migrations of ocean animals. *Nature* 2019; **576**: 257–
536 261.
- 537 44. Proding F, Endo H, Takano Y, Li Y, Tominaga K, Isozaki T, et al. Year-round
538 dynamics of amplicon sequence variant communities differ among eukaryotes,
539 *Imitervirales*, and prokaryotes in a coastal ecosystem. *FEMS Microbiol Ecol* 2021; **97**:
540 fiab167.
- 541 45. Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, et al. Biogeography of
542 marine giant viruses reveals their interplay with eukaryotes and ecological functions.
543 *Nat Ecol Evol* 2020; **4**: 1639–1649.
- 544 46. Werdell PJ, Behrenfeld MJ, Bontempi PS, Boss E, Cairns B, Davis GT, et al. The
545 Plankton, Aerosol, Cloud, Ocean Ecosystem Mission: Status, Science, Advances. *Bull*
546 *Am Meteorol Soc* 2019; **100**: 1775–1794.
- 547 47. Longhurst A, Sathyendranath S, Platt T, Caverhill C. An estimate of global primary
548 production in the ocean from satellite radiometer data. *J Plankton Res* 1995; **17**: 1245–
549 1271.
- 550 48. Yasunaka S, Nojiri Y, Nakaoka S, Ono T, Whitney FA, Telszewski M. Mapping of sea
551 surface nutrients in the North Pacific: Basin-wide distribution and seasonal to
552 interannual variability. *J Geophys Res Oceans* 2014; **119**: 7756–7771.
- 553 49. Cheung S, Nitani R, Tsurumoto C, Endo H, Nakaoka S, Cheah W, et al. Physical
554 Forcing Controls the Basin-Scale Occurrence of Nitrogen-Fixing Organisms in the
555 North Pacific Ocean. *Glob Biogeochem Cycles* 2020; **34**: e2019GB006452.
- 556 50. Follows MJ, Dutkiewicz S. Modeling Diverse Communities of Marine Microbes. *Annu*
557 *Rev Mar Sci* 2011; **3**: 427–451.
- 558

559

560

561 **Acknowledgments**

562 We thank the *Tara* Oceans consortium, the EukBank consortium, and the people and
563 sponsors who supported the *Tara* Oceans Expedition (<http://www.embl.de/tara-oceans/>) for
564 making the data accessible. This is contribution number XXX of the *Tara* Oceans Expedition
565 2009–2013. Computational time was provided by the Supercomputer System, Institute for
566 Chemical Research, Kyoto University. This work was supported by JSPS/KAKENHI (Nos.
567 18H02279 and 19H05667 to H.O.), and the Collaborative Research Program of the Institute
568 for Chemical Research, Kyoto University (2020-29 to K.T.). We thank Leonie Seabrook,
569 PhD, from Edanz (<https://jp.edanz.com/ac>) for editing a draft of this manuscript.

570

571 **Competing Interests**

572 The authors declare no competing interests.

573

574 **Author Contributions**

575 H.K. designed the study, performed most of the bioinformatics analyses and wrote the initial
576 manuscript. H.E., R.N., K.T., and H.O. contributed to the design of the work and supervised
577 H.K. N.H., C.B., F.M., and C.d.V. performed the amplicon sequence data processing and
578 annotation. J.P., K.L., O.B., and P.W. treated biological samples and performed sequencing.
579 R.E.H., S.C., L.K.-B., E.B., and C.B. provided expertise in marine biology. *Tara* Oceans
580 Coordinators (S.G.A., M.B., P.B., E.B., C.B., G.C., C.d.V., G.G., L.G., N.G., P.H., D.I., O.J.,
581 S.K., L.K.-B., E.K., F.N., H.O., N.P., S.P., C.S., S.S., L.S., M.B.S., S.S., and P.W.)
582 contributed to expeditionary infrastructure needed for global ocean sampling, sample
583 processing, and data production. All authors contributed to the interpretation of data and
584 finalization of the manuscript.

585

586 **The *Tara* Oceans Coordinators and Affiliations**

587 Silvia G. Acinas¹, Marcel Babin², Peer Bork^{3,4,5}, Emmanuel Boss⁶, Chris Bowler⁷, Guy
588 Cochrane⁸, Colomban de Vargas⁹, Gabriel Gorsky¹⁰, Lionel Guidi^{10,11}, Nigel Grimsley^{12,13},
589 Pascal Hingamp¹⁴, Daniele Iudicone¹⁵, Olivier Jaillon¹⁶, Stefanie Kandels¹⁷, Lee Karp-Boss⁶,
590 Eric Karsenti^{7,17}, Fabrice Not¹⁸, Hiroyuki Ogata¹⁹, Nicole Poulton²⁰, Stéphane Pesant²¹,
591 Christian Sardet^{10,22}, Sabrina Speich^{23,24}, Lars Stemmann¹⁰, Matthew B. Sullivan^{25,26},
592 Shinichi Sunagawa²⁷, and Patrick Wincker¹⁶

593

594 ¹Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC),
595 Barcelona, Catalonia, Spain.

596 ²Département de biologie, Québec Océan and Takuvik Joint International Laboratory
597 (UMI3376), Université Laval (Canada) - CNRS (France), Université Laval, Québec, QC,
598 G1V 0A6, Canada.

599 ³Structural and Computational Biology, European Molecular Biology Laboratory,
600 Meyerhofstrasse 1, 69117 Heidelberg, Germany.

601 ⁴Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany.

602 ⁵Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg,
603 Germany.

604 ⁶School of Marine Sciences, University of Maine, Orono, Maine 04469, USA.

605 ⁷Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole
606 Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005
607 Paris, France.

608 ⁸European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI),
609 Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

- 610 ⁹CNRS, UMR 7144, EPEP & Sorbonne Universités, UPMC Université Paris 06, Station
611 Biologique de Roscoff, 29680 Roscoff, France.
- 612 ¹⁰Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d’océanographie de
613 Villefranche (LOV), Observatoire Océanologique, 06230 Villefranchesur-Mer, France.
- 614 ¹¹Department of Oceanography, University of Hawaii, Honolulu, HI 96822, USA.
- 615 ¹²CNRS, UMR 7232, BIOM, Avenue de Pierre Fabre, 66650 Banyuls-sur-Mer, France.
- 616 ¹³Sorbonne Universités Paris 06, OOB UPMC, Avenue de Pierre Fabre, 66650 Banyuls-sur-
617 Mer, France.
- 618 ¹⁴Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, France.
- 619 ¹⁵Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.
- 620 ¹⁶Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry,
621 Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France
- 622 ¹⁷European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg Germany.
- 623 ¹⁸CNRS, UMR 7144, Sorbonne Universités, UPMC Université Paris 06, Station Biologique
624 de Roscoff, 29680 Roscoff, France.
- 625 ¹⁹Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan.
- 626 ²⁰Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA.
- 627 ²¹European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome
628 Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
- 629 ²²CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer,
630 France.
- 631 ²³Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané,
632 France.
- 633 ²⁴Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole
634 Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France.

635 ²⁵Department of Microbiology, The Ohio State University, Columbus, OH 43214, USA.

636 ²⁶Department of Civil, Environmental and Geodetic Engineering, The Ohio State University,
637 Columbus, OH 43214, USA.

638 ²⁷Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics,
639 ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

Table 1. Taxonomic breakdown of plankton communities.

| Taxogroup 1 ^a | Taxogroup 2 ^a | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|--------------------------|--------------------------|---------|----|----|----|----|----|-------|
| Dinoflagellata | Dinophyceae | 16 | 16 | 3 | 3 | 6 | | 44 |
| | MALV-I | 4 | 1 | 6 | 7 | 19 | 6 | 43 |
| | MALV-II | 5 | 1 | | 6 | 15 | 11 | 38 |
| | MALV-III | 2 | | 4 | | 1 | | 7 |
| | MALV-V | | | | | 1 | 3 | 4 |
| Metazoa | Arthropoda | 1 | 10 | 3 | | | | 14 |
| | Cnidaria | | 3 | | | | | 3 |
| | Urochordata | | | 1 | | | | 1 |
| Ochrophyta | Dictyochophyceae | 2 | | | 1 | | 1 | 4 |
| | Pelagophyceae | 1 | | | | | | 1 |
| | MOCH-2 | | | | 1 | | | 1 |
| | Diatomeae | | | 1 | | | | 1 |
| Chlorophyta | Mamiellophyceae | 1 | | | 3 | | | 4 |
| | Chloropicophyceae | | | 2 | | | | 2 |
| Prymnesiophyceae | Prymnesiophyceae | 3 | 1 | | 1 | | | 5 |
| Radiolaria | Acantharea | | 2 | | | | | 2 |
| | RAD-B-clade | | | 1 | 1 | | | 2 |
| | Spumellaria | | | | | | 1 | 1 |
| | MAST-01 | MAST-01 | 1 | | | 2 | | 2 |
| Picozoa | Picozoa | 2 | | | 1 | 2 | | 5 |
| Sagenista | MAST-04 | | | | 2 | | 2 | 4 |
| | MAST-07 | | | | | | 1 | 1 |
| Cryptomonadales | Cryptomonadales | 2 | | | 1 | | | 3 |
| Opalozoa | Nanomonadea | | | | 1 | | 1 | 2 |
| | Bicosoecida | | | | | 1 | | 1 |
| core-kathablepharids | core-kathablepharids | 1 | | 1 | | | | 2 |
| Ciliophora | Spirotrichea | 1 | | | | | | 1 |
| Centroplasthelida | Centroplasthelida | | | | | 1 | | 1 |
| Unclassified | Unclassified | 1 | | 1 | 2 | 2 | | 6 |
| | Total | 43 | 34 | 23 | 32 | 48 | 28 | 208 |

a: Taxonomic level in the EukRibo.

Table 2. Comparison of prediction performance using different sets of satellite-derived and spatial parameters.

Support vector machine were used in all the parameter sets.

| Parameter set | Leave-one-out cross-validation | | Buffered cross-validation | |
|----------------------------------|--------------------------------|----------------------|---------------------------|----------------------|
| | Accuracy | ROC-AUC ^a | Accuracy | ROC-AUC ^a |
| All satellite-derived parameters | 0.67 | 0.90 | 0.54 | 0.83 |
| Latitude, Longitude ^b | 0.68 | 0.91 | 0.29 | 0.59 |
| SST | 0.40 | 0.79 | 0.28 | 0.72 |
| [Chl] | 0.43 | 0.72 | 0.23 | 0.62 |
| SST, [Chl] | 0.52 | 0.86 | 0.47 | 0.82 |
| SST, Environmental parameters | 0.58 | 0.88 | 0.50 | 0.83 |

a: Micro-average area under the ROC curve.

b: Sine and cosine of longitude were used as parameters.

Table S1. Machine learning methods and their hyperparameters.

| Method | Scikit-learn implementation | Hyperparameters |
|--|---|--|
| K-nearest Neighbors | sklearn.neighbors.KNeighborsClassifier | Number of neighbors (1, 2, ..., 9) |
| Naïve Bayes (Gaussian) | sklearn.naive_bayes.GaussianNB | Smoothing parameter (10^{-4} , 10^{-3} , ..., 10^3) |
| Multilayer Perceptron (one hidden layer) | sklearn.neural_network.MLPClassifier | Activation function (identity, logistic, tanh, ReLU), Hidden layer size (10, 20, 30, 40), L2 penalty parameter (10^{-6} , 10^{-5} , ..., 10^3) |
| Random Forest | sklearn.ensemble.RandomForestClassifier | Number of trees (10, 100, 1000, 10000) |
| Support Vector Machine | sklearn.svm.SVC | Kernel (linear, RBF), Coefficient of RBF kernel (10^{-4} , 10^{-3} , ..., 10^3), L2 penalty parameter (10^{-4} , 10^{-3} , ..., 10^3) |

Table S2. Comparison of machine learning methods.
All the satellite-derived parameters were used in the validations.

| Method | Leave-one-out cross-validation | | Buffered cross-validation | |
|------------------------|--------------------------------|----------------------|---------------------------|----------------------|
| | Accuracy | ROC-AUC ^a | Accuracy | ROC-AUC ^a |
| K-nearest Neighbors | 0.56 | 0.86 | 0.47 | 0.76 |
| Naïve Bayes | 0.55 | 0.85 | 0.50 | 0.81 |
| Multilayer Perceptron | 0.60 | 0.88 | 0.54 | 0.82 |
| Random Forest | 0.63 | 0.89 | 0.50 | 0.81 |
| Support Vector Machine | 0.67 | 0.90 | 0.54 | 0.83 |

a: Micro-average area under the ROC curve.

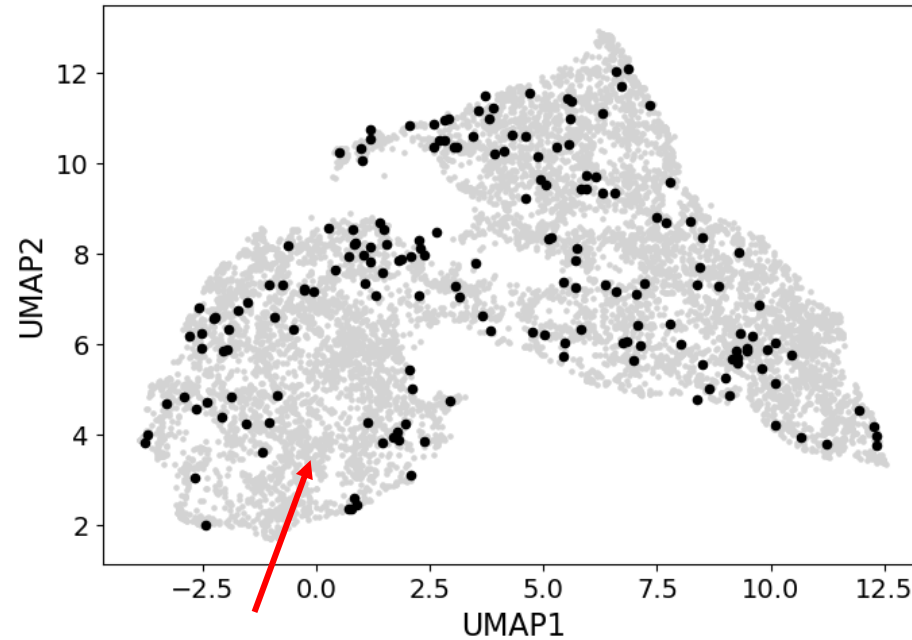


Figure 1. Metabarcoding samples used for prediction model training. Points associated with metabarcoding samples projected on the 2-D map of the satellite-derived parameter space (black points). Grey points are randomly selected grid cells used for learning the map. Red arrow indicates the region without metabarcoding samples.

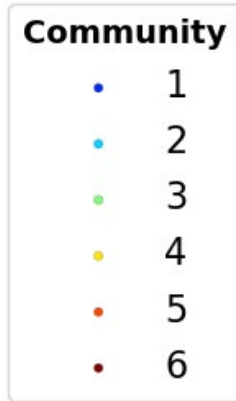
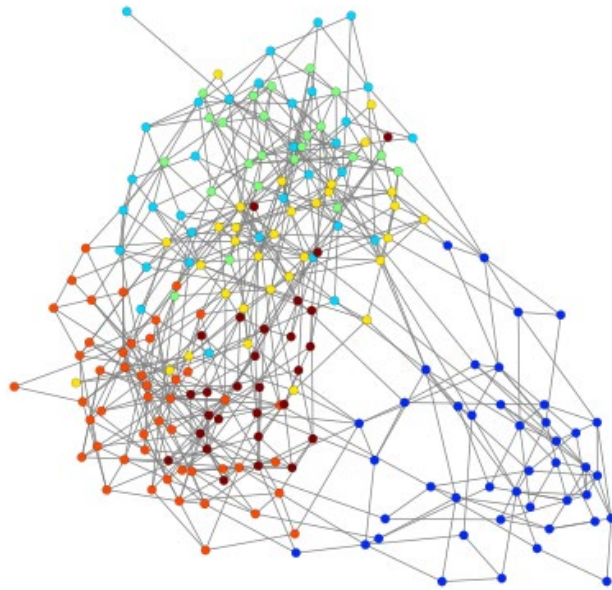
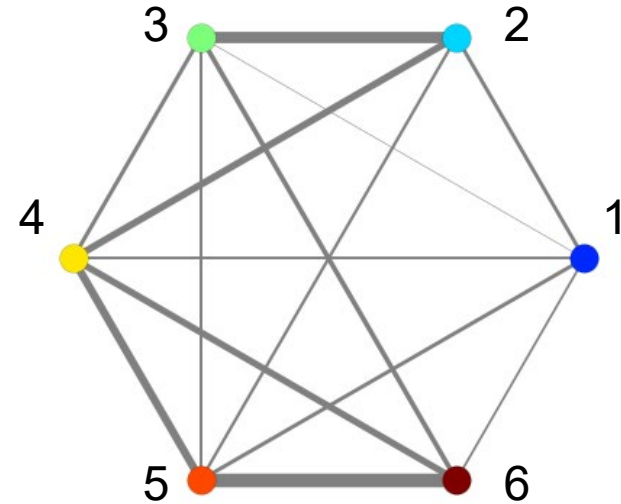
A**B**

Figure 2. Plankton network inferred using metabarcoding data.

A A force-directed representation of the network. Nodes (plankton ASVs) are colored by belonging to a community. **B** A graph representing the community connection of the same network. The width of the edges is proportional to the number of inter-community edges.

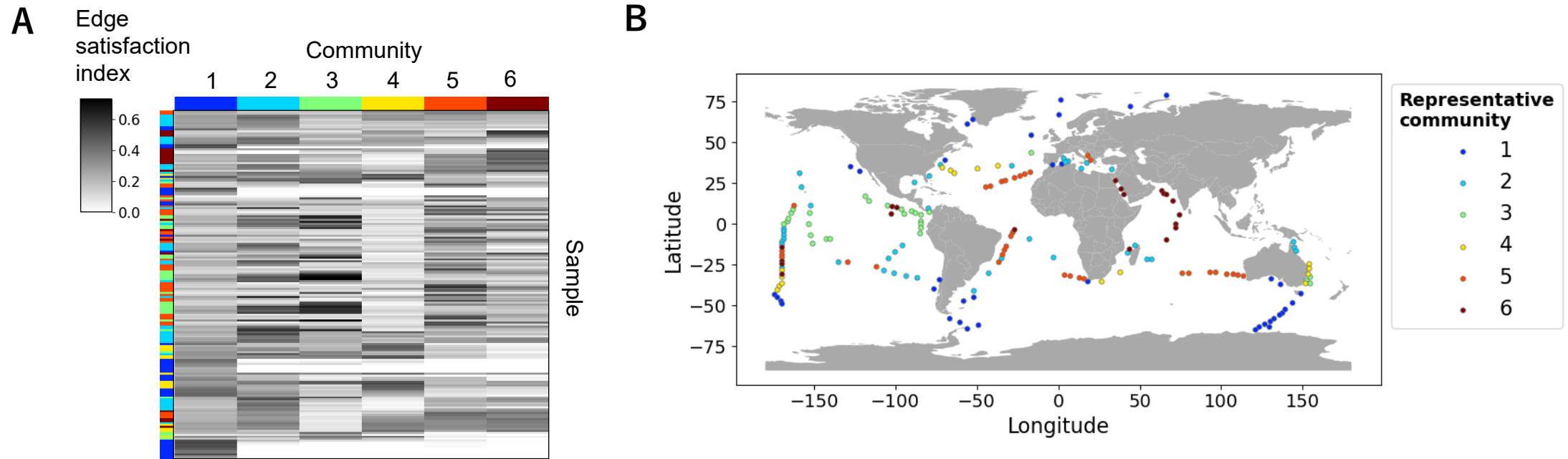


Figure 3. Representative community of samples.

A A heatmap of edge satisfaction index. The left most column shows representative community of each sample. **B** Geographic distribution of representative communities.

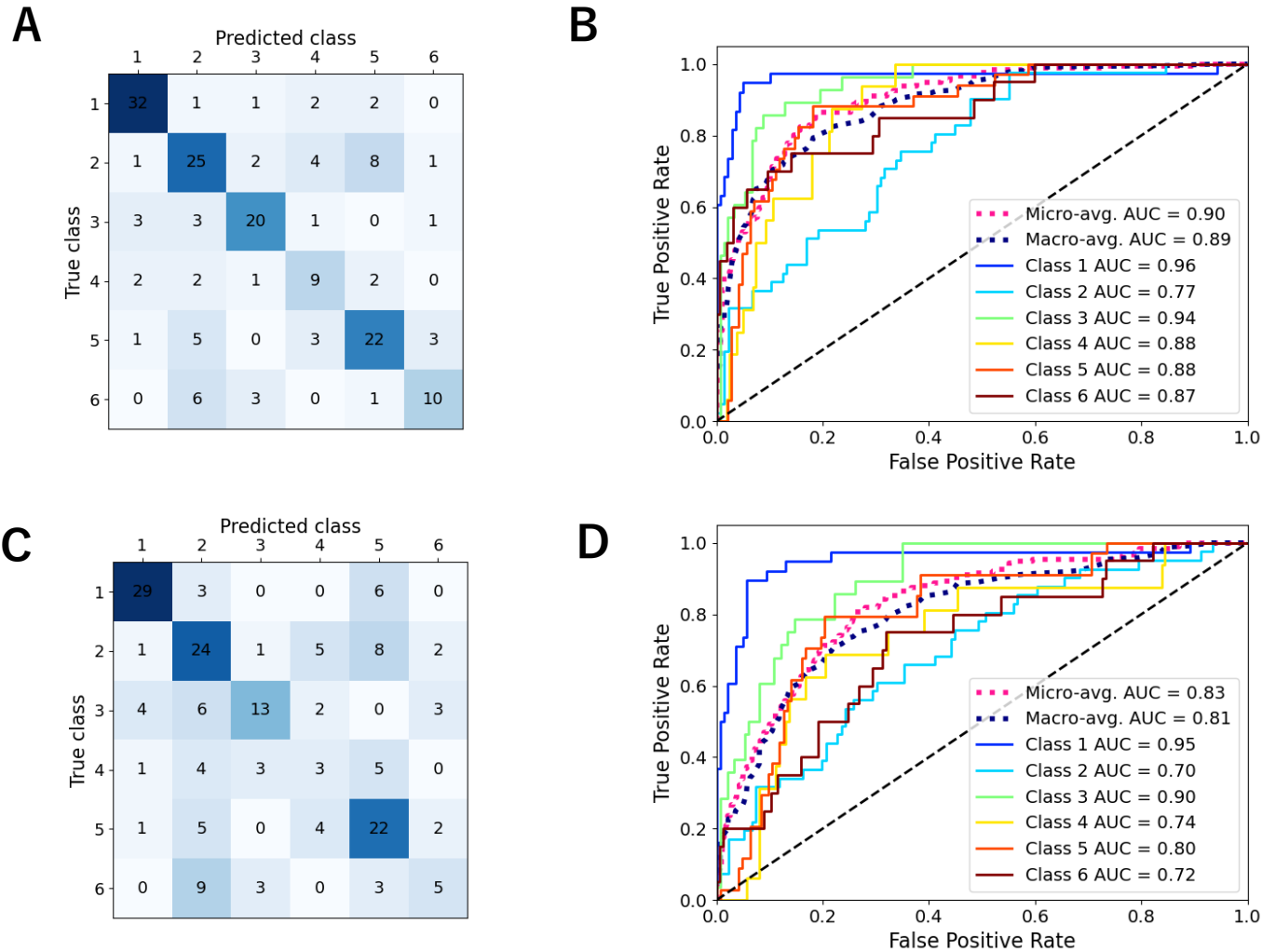


Figure 4. Performance of support vector machine on community prediction based on satellite-derived parameters. Performance of SVM using all the satellite-derived parameters. **A-B** The confusion matrix (**A**) and the ROC curve (**B**) in the condition of leave-one-out cross-validation. **C-D** The confusion matrix (**C**) and the ROC curve (**D**) in the condition of buffered cross-validation.

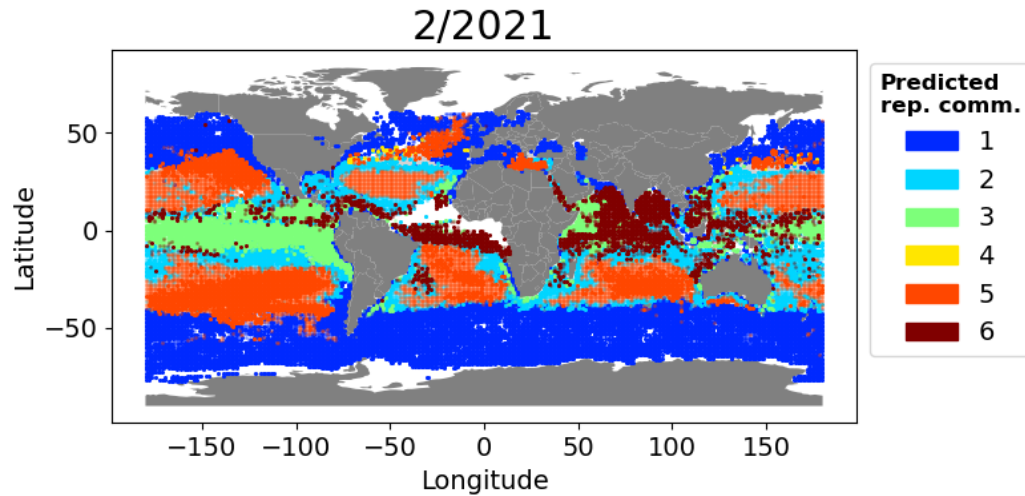
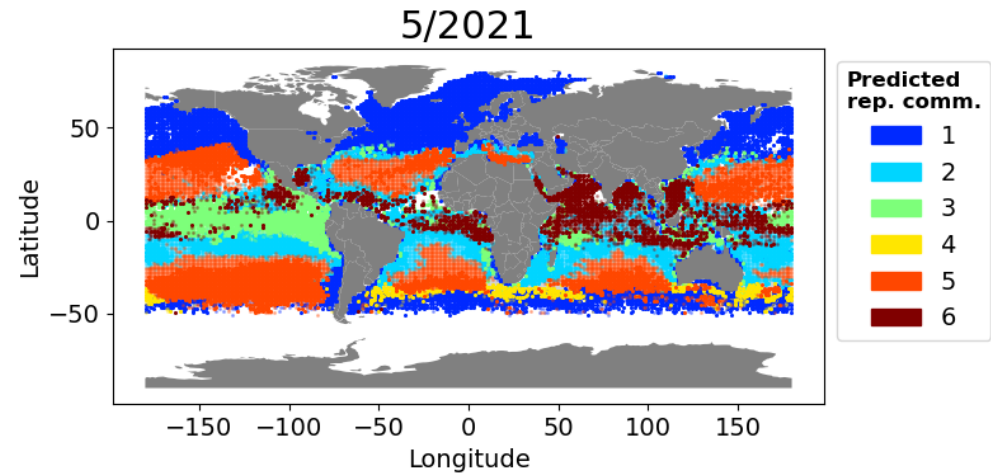
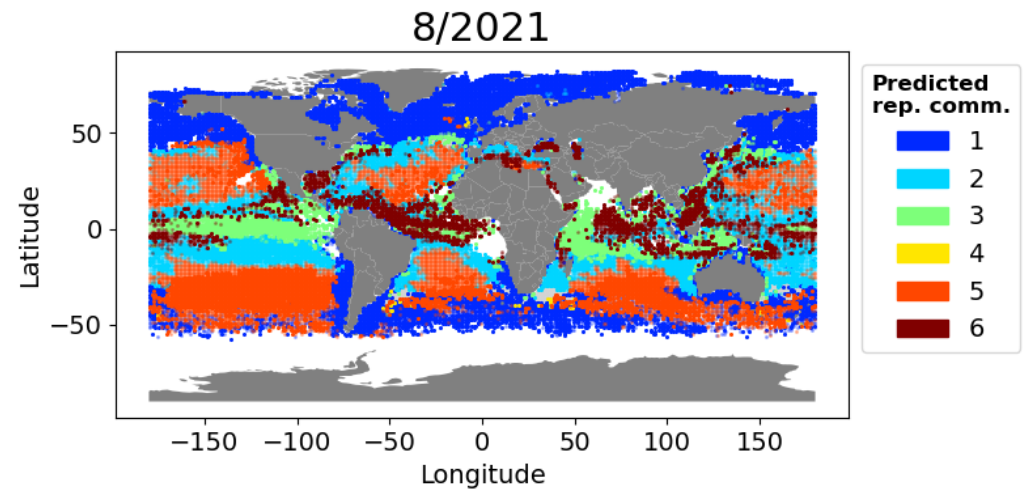
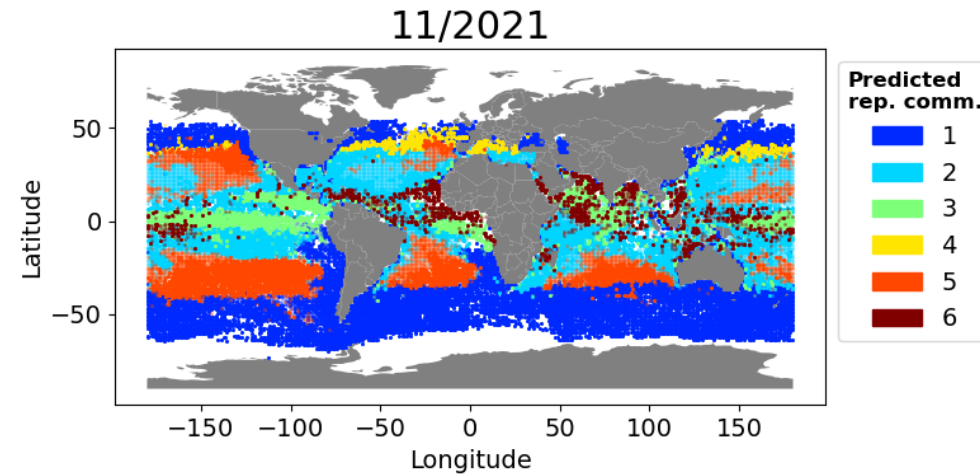
A**B****C****D**

Figure 5. Spatiotemporal distribution of communities predicted from satellite-derived parameters. Representative community distribution in February (**A**), May (**B**), August (**C**), and November (**D**) of year 2021 predicted from satellite-derived parameters. When multiple communities were predicted to be representative in the same point, community with the highest probability is show in pale color. Grey point means no community was predicted to be representative.

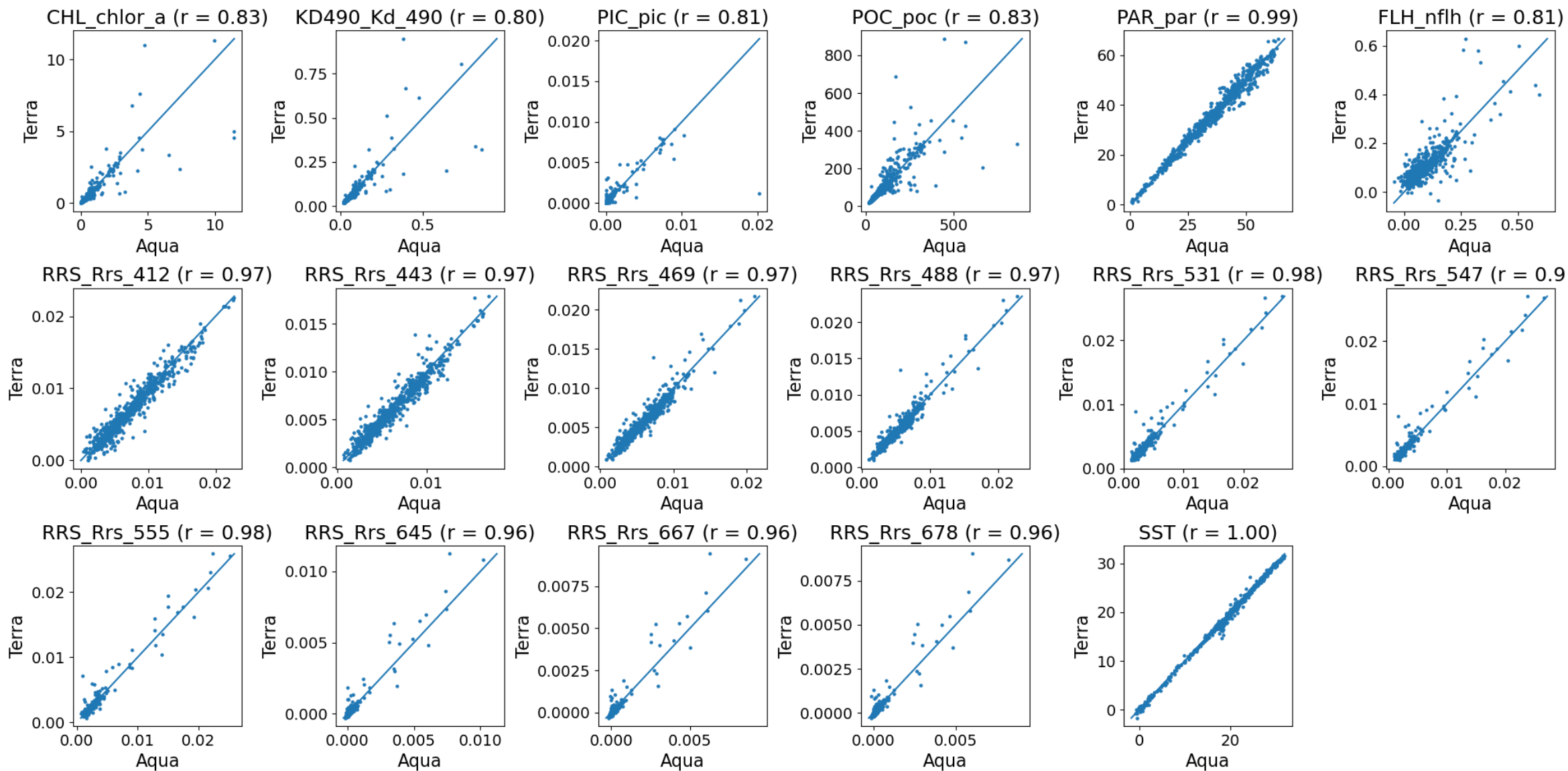


Figure S1. Comparison of satellite-derived parameters acquired by the Aqua and the Terra satellite.

Pearson's correlation coefficient was shown for each parameter pairs.

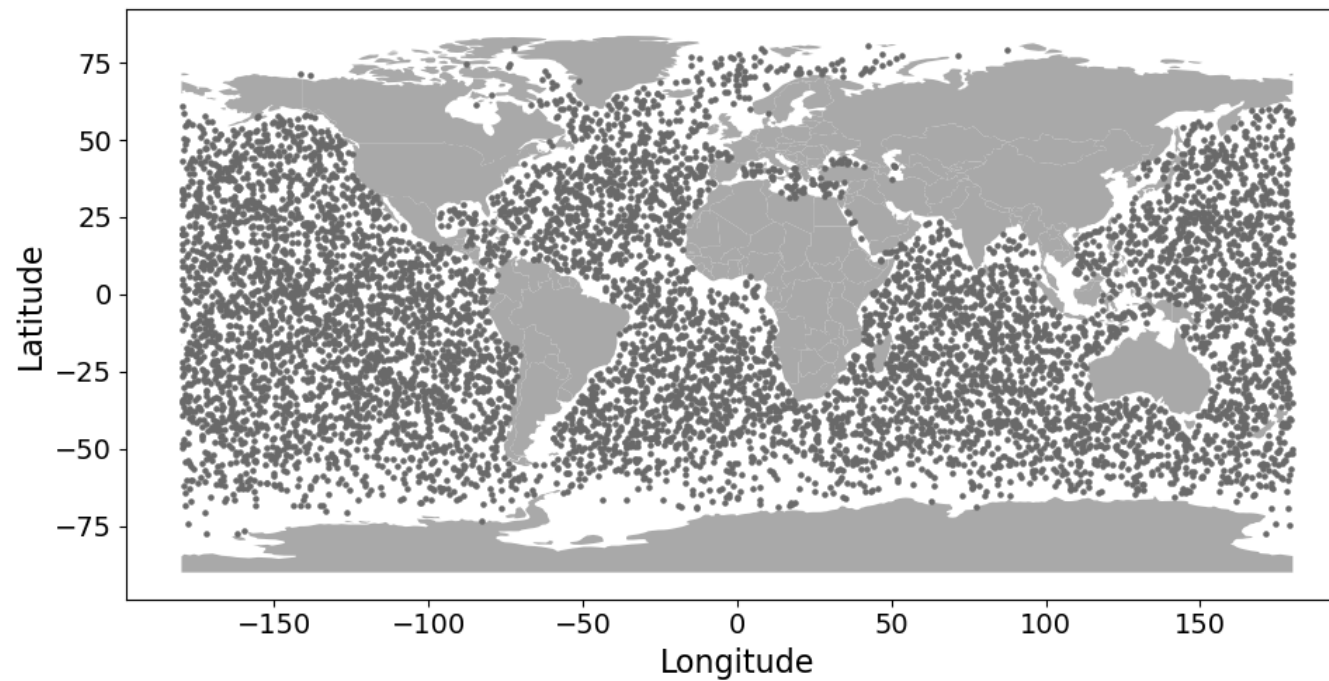


Figure S2. Location of satellite-derived parameter samples used for learning UMAP. Dark grey points are location of grid cells used to learn a map with UMAP. Sampling month was also randomly selected for each grid cell.

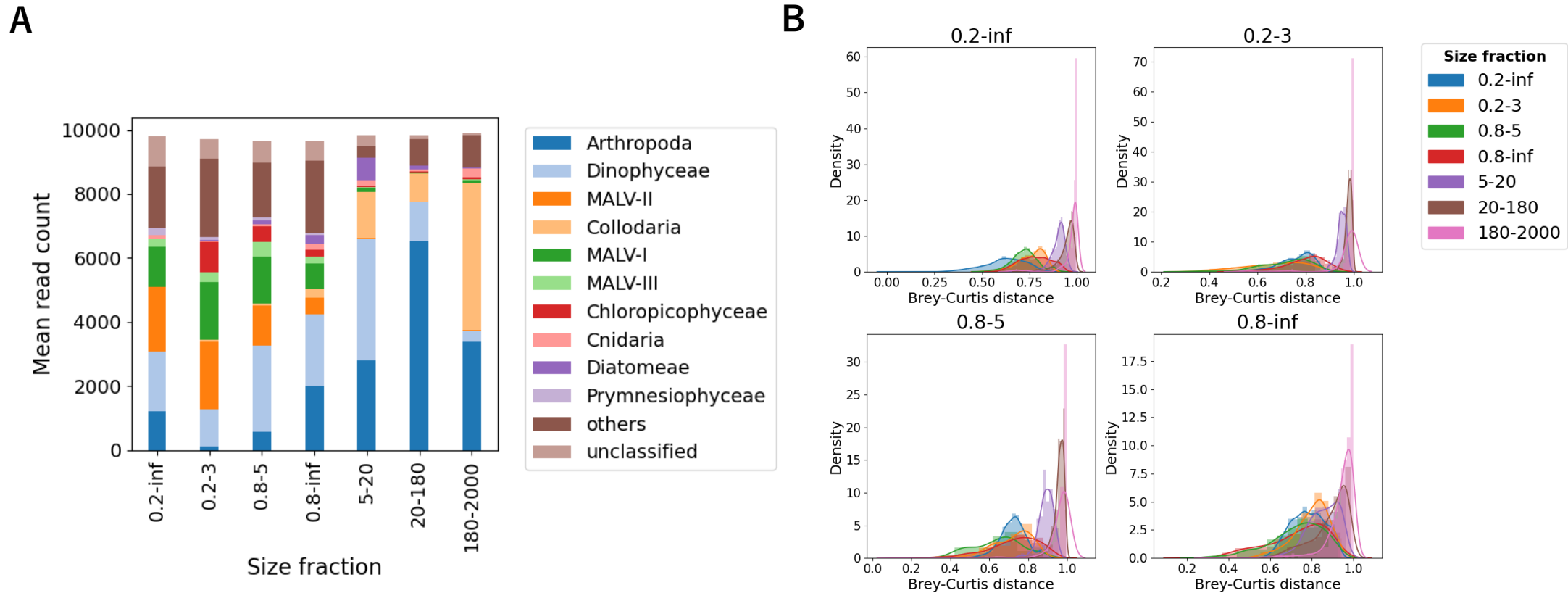


Figure S3. Comparison of size fractions in the South Pacific Subtropical Gyre. Taxonomic difference of size fractions was checked among samples from the South Pacific Subtropical Gyre, which contains samples from all major size fractions. **A** Mean taxonomic composition of each size fraction. Taxonomic level is “taxogroup 2” in the EukRibo. **B** Comparison of taxonomic distance of intra- and inter-size-fraction samples. Brey-Curtis distance was calculated based on ASV read count.

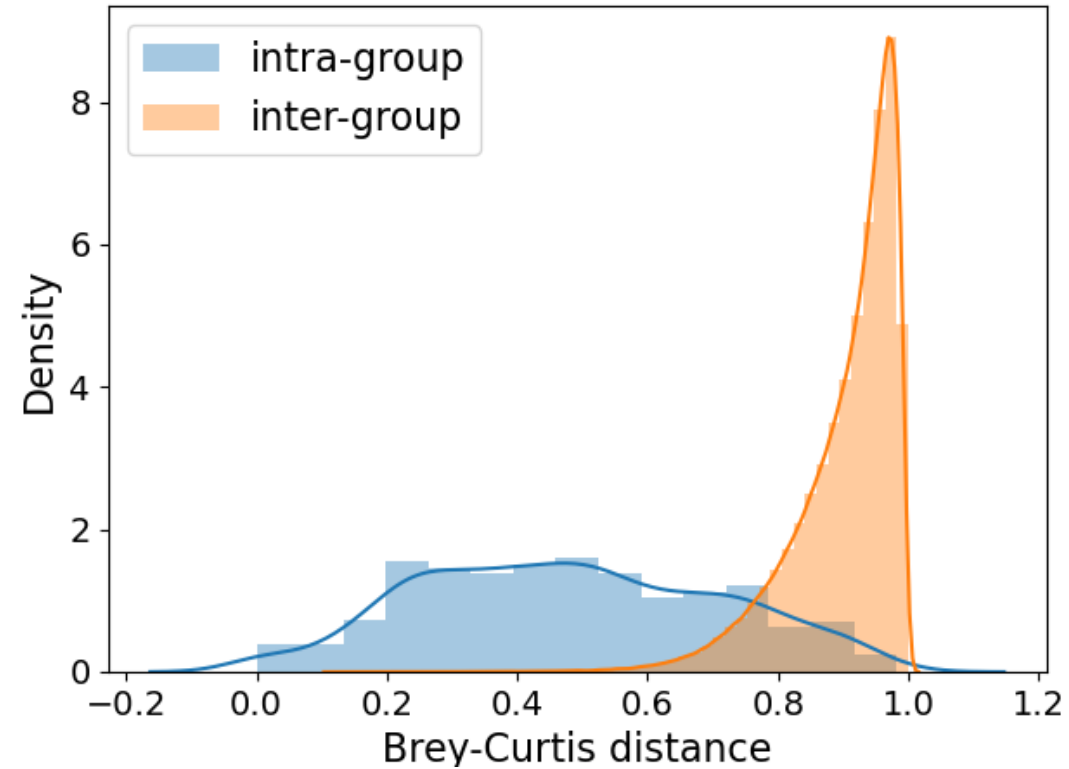


Figure S4. Comparison of intra- and inter-bin sample distance. Brey-Curtis distance was calculated based on ASV read count. Intra-bin sample distance is small enough compared to inter-bin sample distance in the most cases.

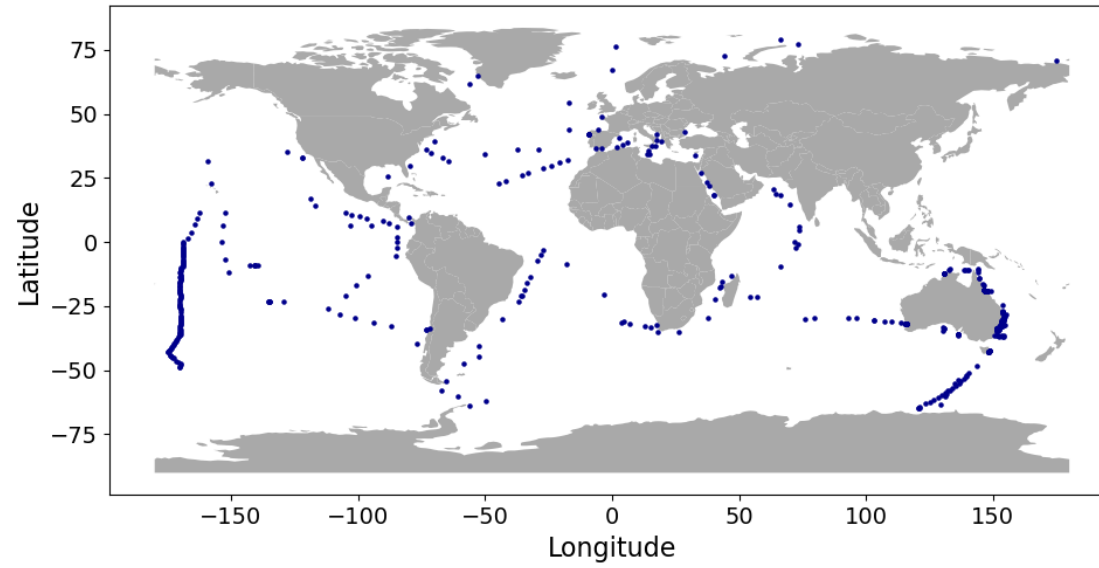
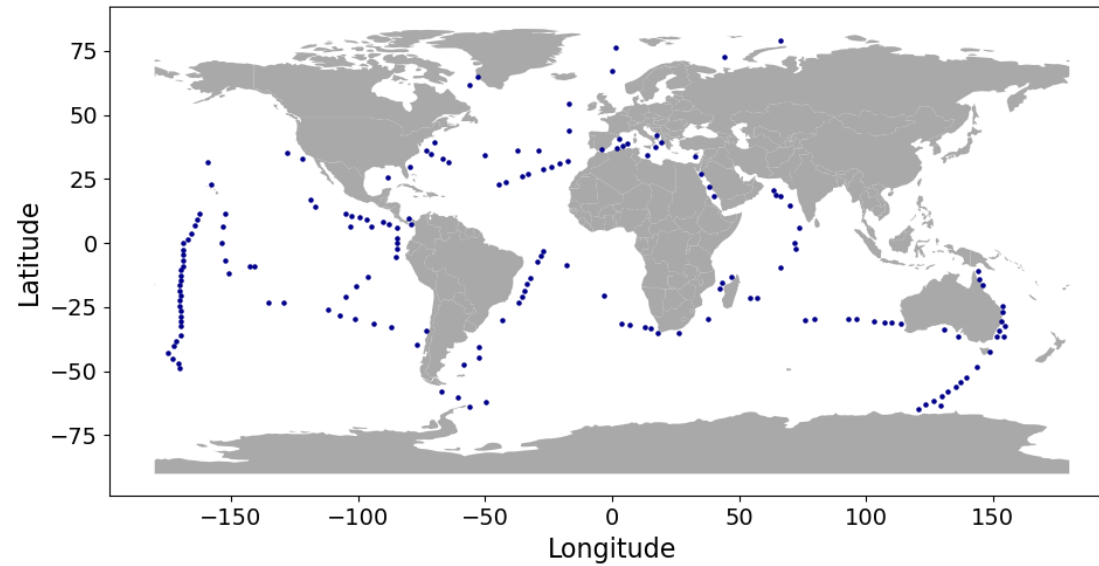
A**B**

Figure S5. Spatial resampling of metabarcoding data.

A Geographic location of 653 metabarcoding samples (bins) before spatial resampling. **B** 177 samples retained and used for the analysis.

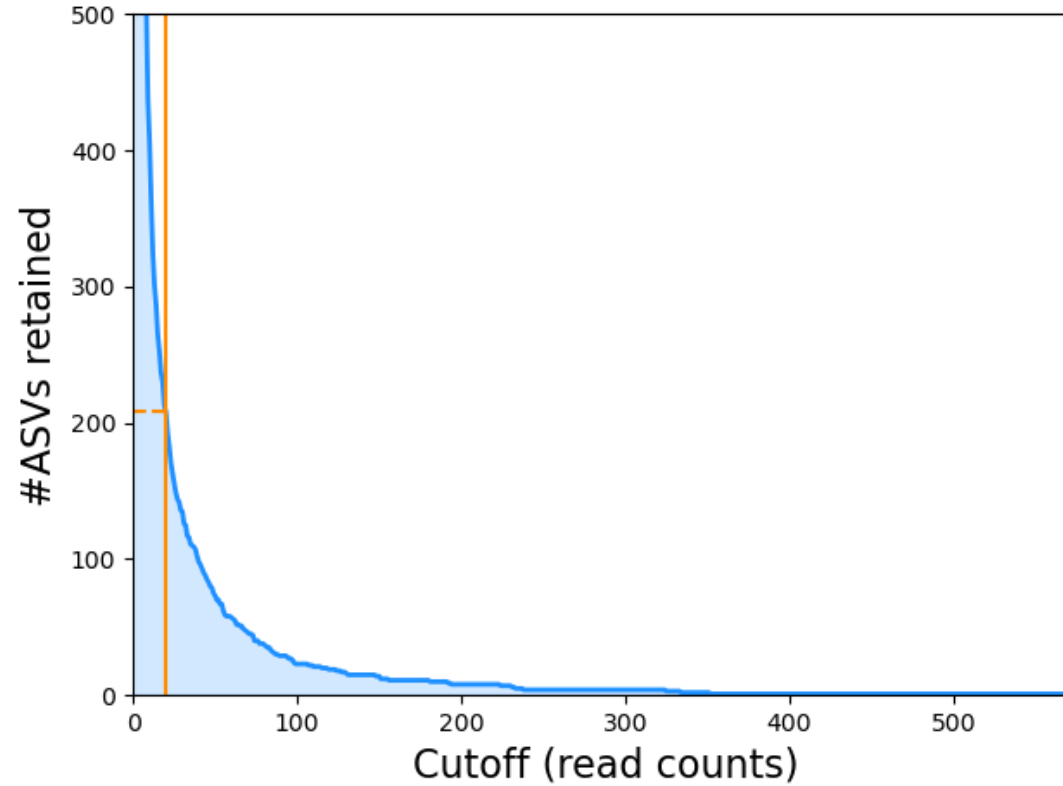
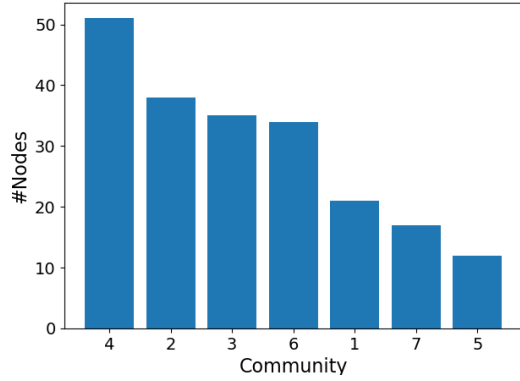
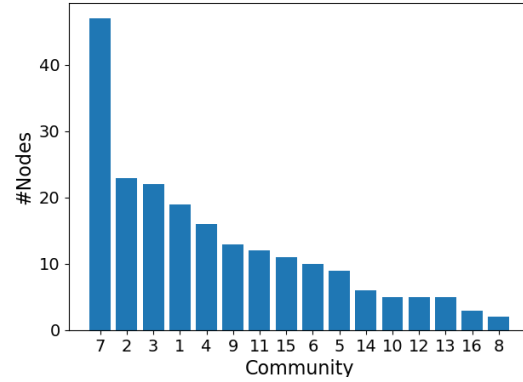


Figure S6. Number of ASVs retained with changing occurrence cutoff. Blue curve shows the number of ASVs retained with given cutoff used for selection. Orange line is the chosen cutoff (20 reads).

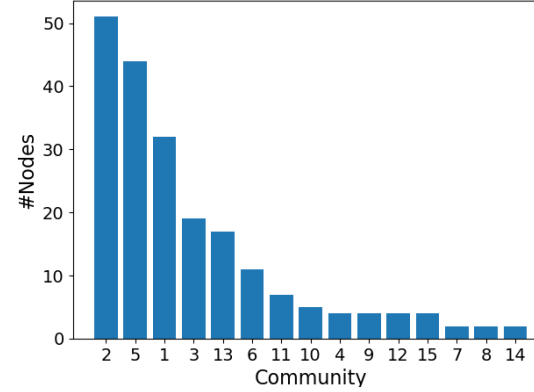
Fast Greedy
Modularity = 0.52



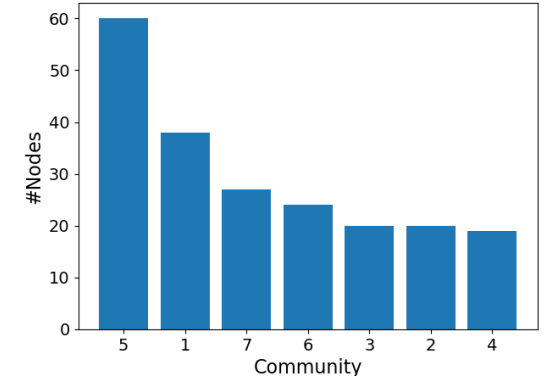
Infomap
Modularity = 0.51



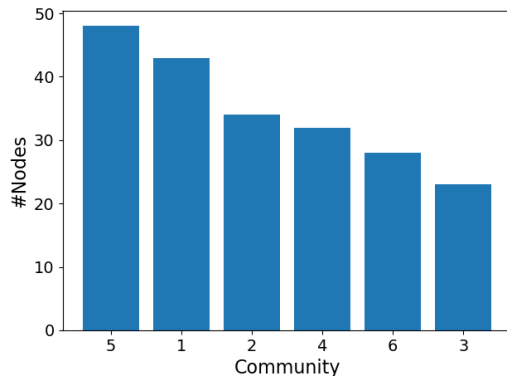
Label Propagation
Modularity = 0.49



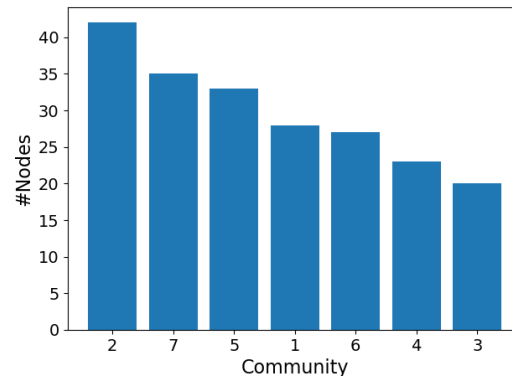
Leading Eigenvector
Modularity = 0.50



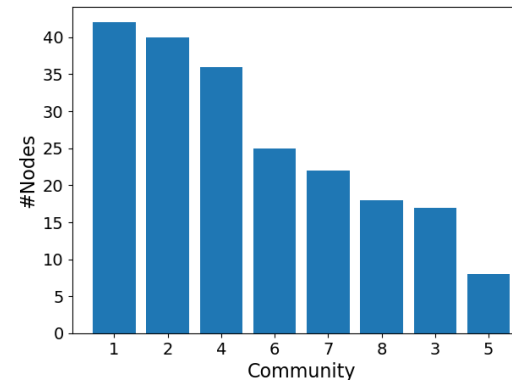
Leiden
Modularity = 0.55



Louvain
Modularity = 0.50



Spinglass
Modularity = 0.55



Walktrap
Modularity = 0.52

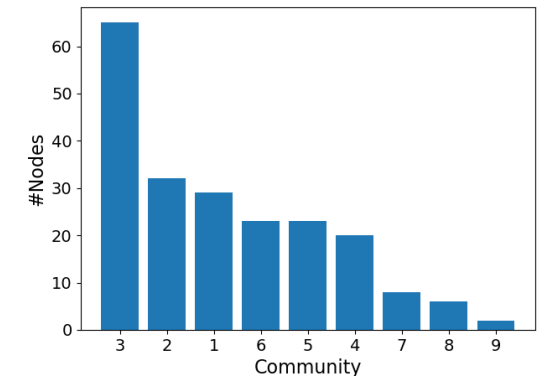


Figure S7. For each algorithm, modularity index and size of the communities. Modularity index of the community division by each algorithm and size of detected communities by it.

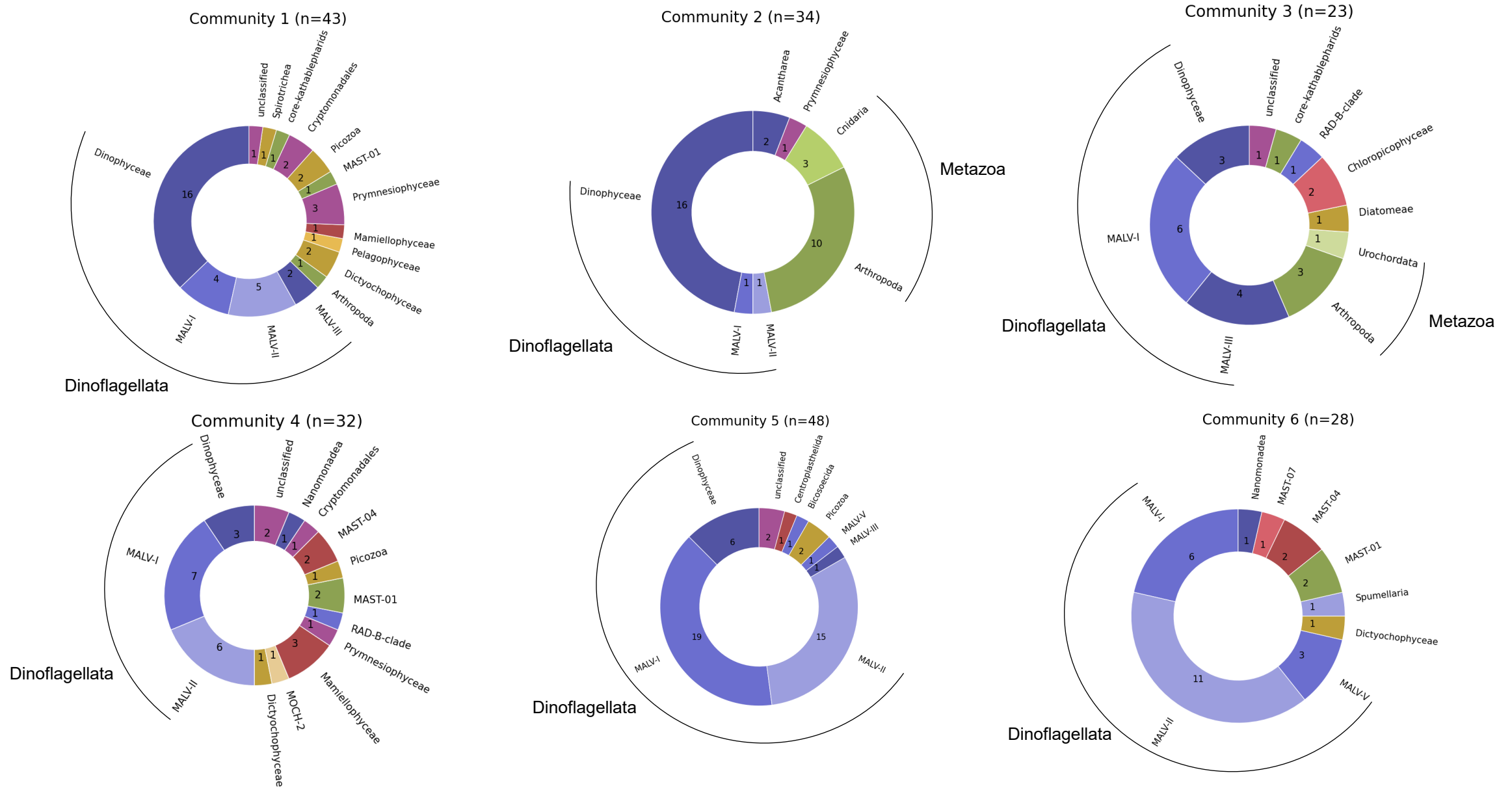


Figure S8. Taxonomic breakdown of plankton communities shown in circle charts. Taxonomic level is “taxogroup 2” in the EukRibo.

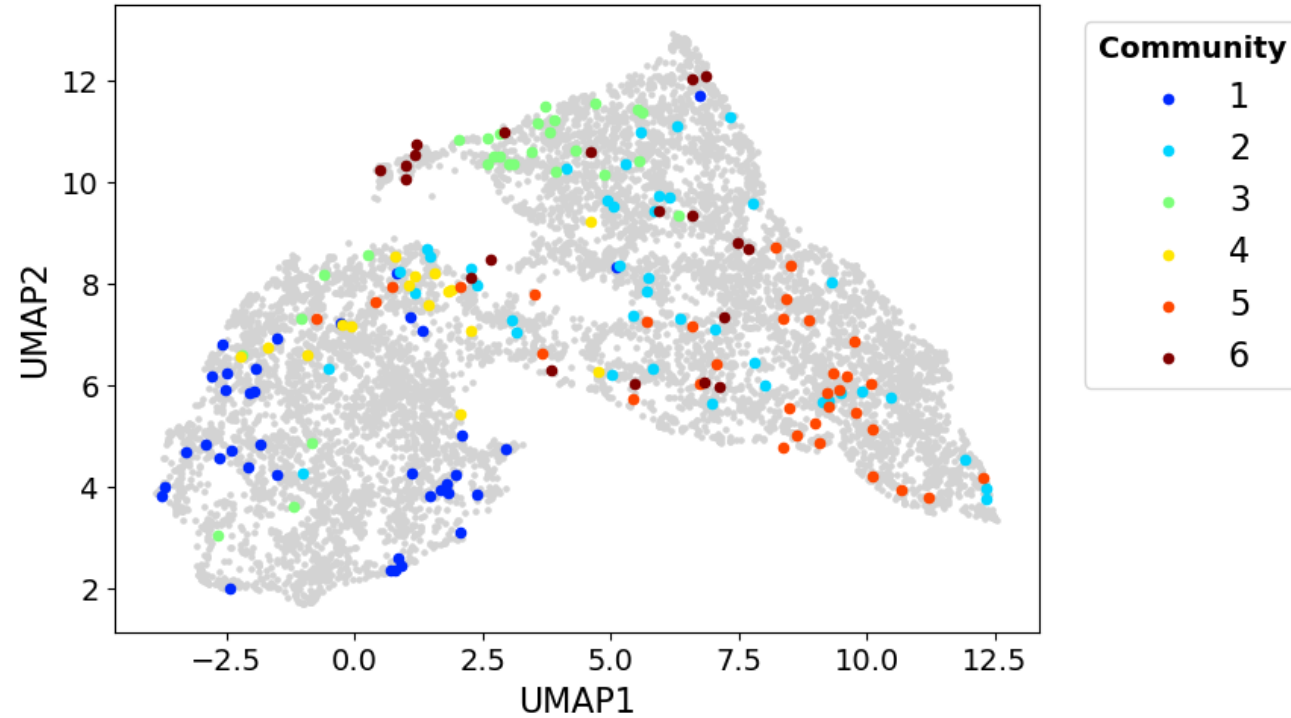
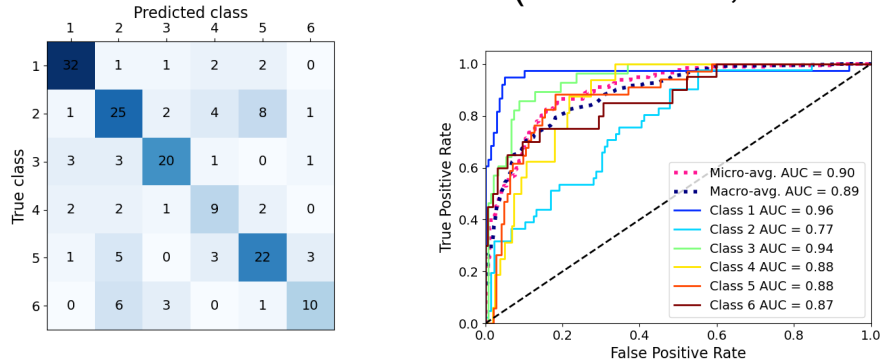


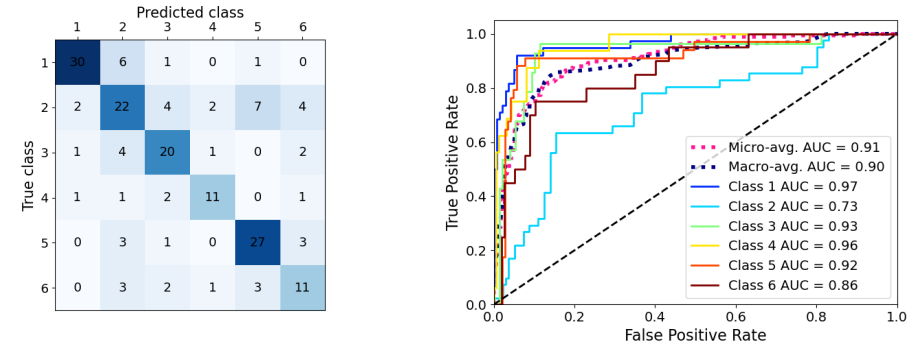
Figure S9. Representative community distribution in the satellite-derived parameter space.

Metabarcoding samples projected on the 2-D map of the satellite-derived parameter space colored by representative community. Grey points are randomly selected grid cells used for learning the map.

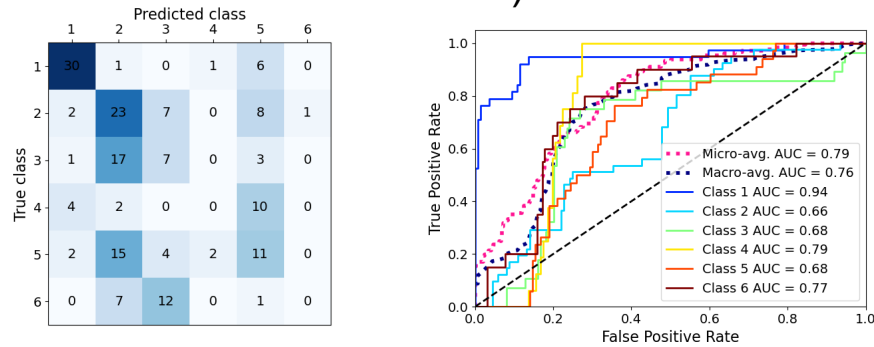
Satellite-derived parameters (Acc = 0.67, AUC = 0.90)



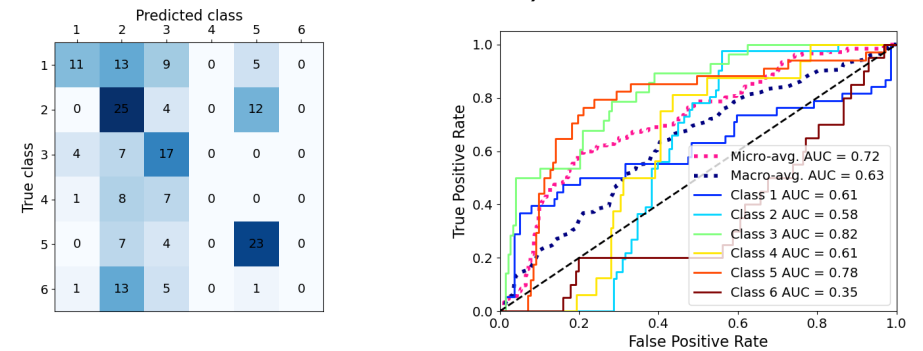
Latitude, Longitude (Acc = 0.68, AUC = 0.91)



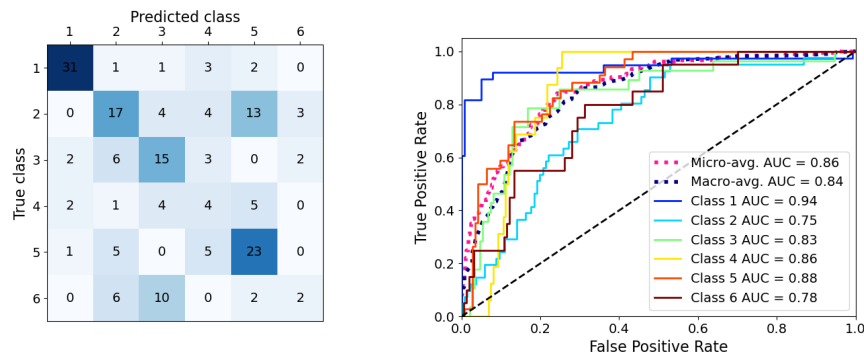
SST (Acc = 0.40, AUC = 0.79)



[Chl] (Acc = 0.43, AUC = 0.72)



SST, [Chl] (Acc = 0.52, AUC = 0.86)



SST, Environmental parameters (Acc = 0.58, AUC = 0.88)

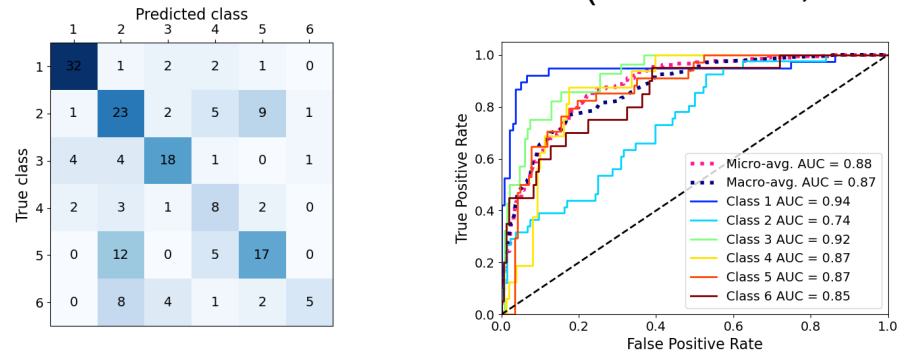
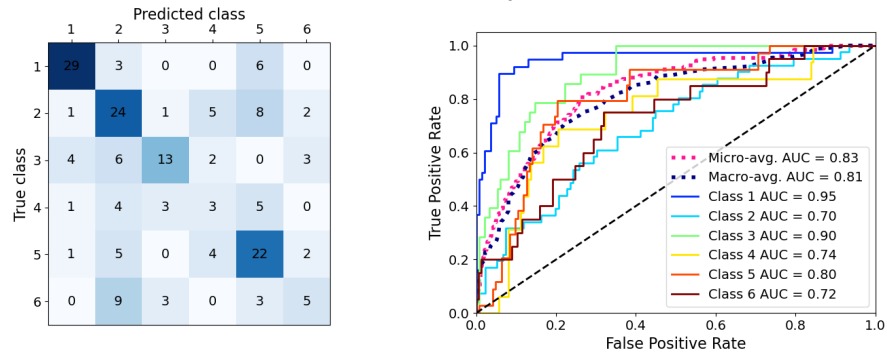
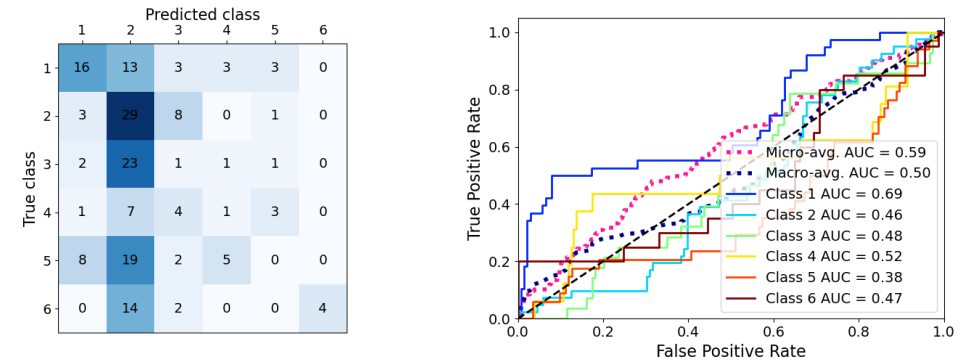


Figure S10. Comparison of prediction performance using different sets of satellite-derived and spatial parameters (leave-one-out cross-validation).

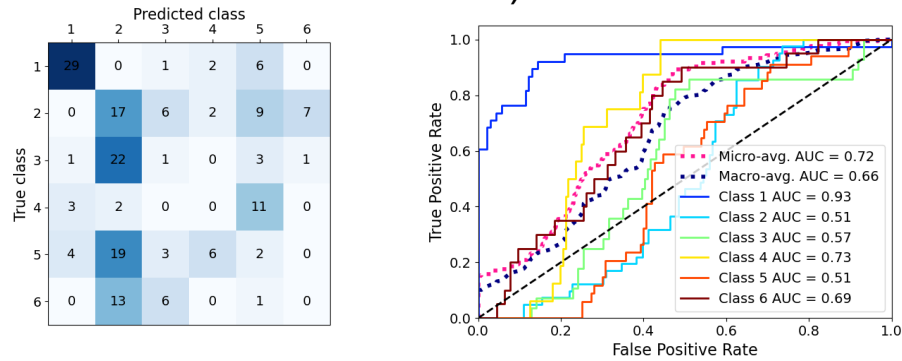
Satellite-derived parameters (Acc = 0.54, AUC = 0.83)



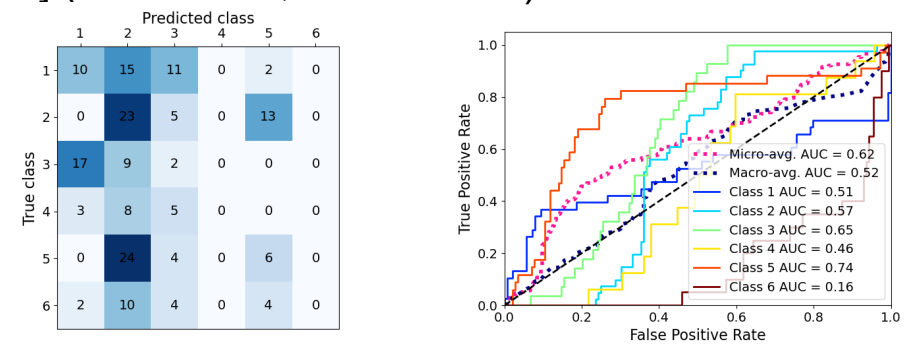
Latitude, Longitude (Acc = 0.29, AUC = 0.59)



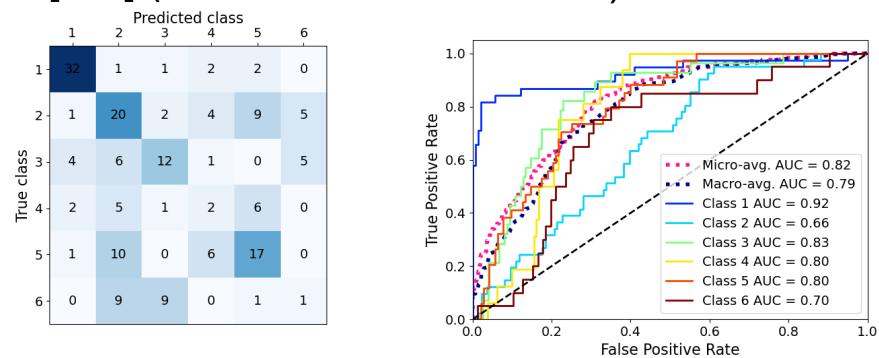
SST (Acc = 0.28, AUC = 0.72)



[Chl] (Acc = 0.23, AUC = 0.62)



SST, [Chl] (Acc = 0.47, AUC = 0.82)



SST, Environmental parameters (Acc = 0.50, AUC = 0.83)

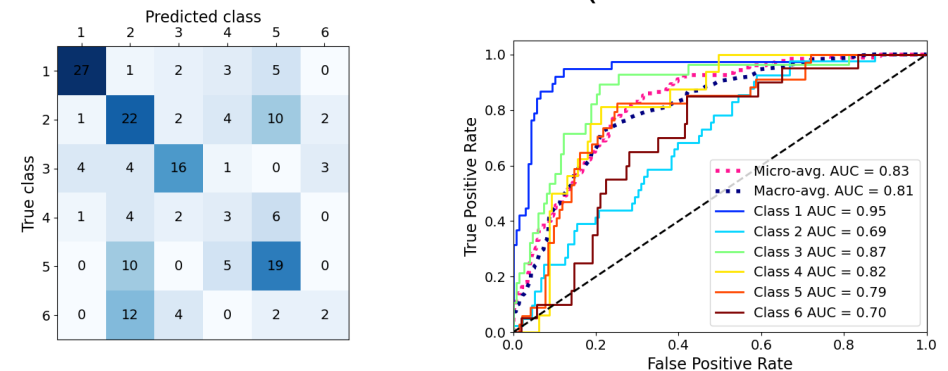


Figure S11. Comparison of prediction performance using different sets of satellite-derived and spatial parameters (buffered cross-validation).

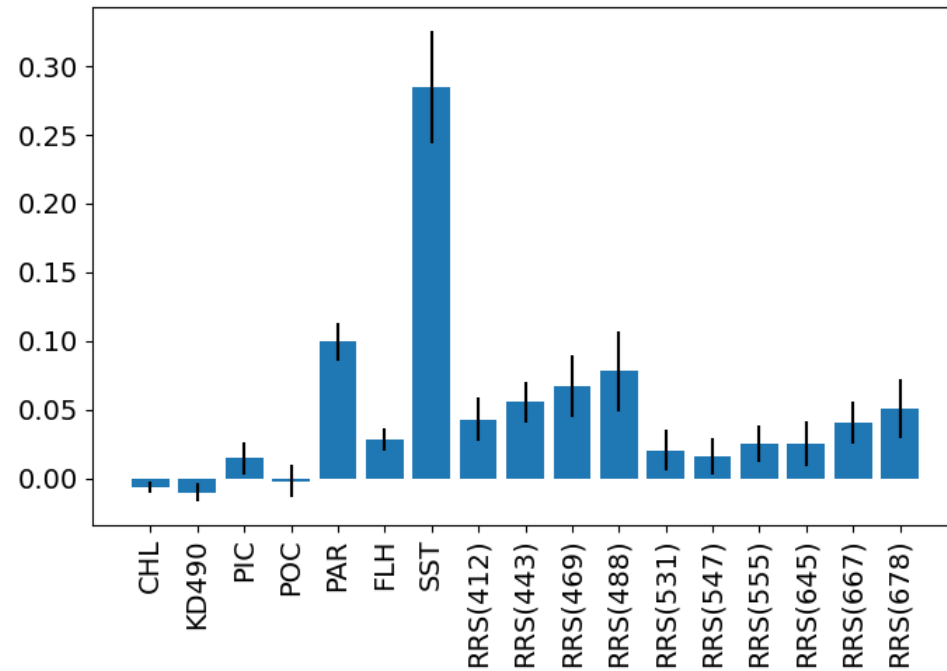


Figure S12. Permutation importance of each parameter in the full SVM model.

Blue bars show mean of parameter importance over 5 times repeats. Error bars show standard deviation over repeats. CHL: chlorophyll *a* concentration, KD490: diffuse attenuation coefficient for downwelling irradiance at 490 nm, PIC/POC: particulate organic/inorganic carbon concentration, PAR: photosynthetically available radiation, FLH: normalized fluorescence line height, SST: sea surface temperature, RRS: remote sensing reflectance

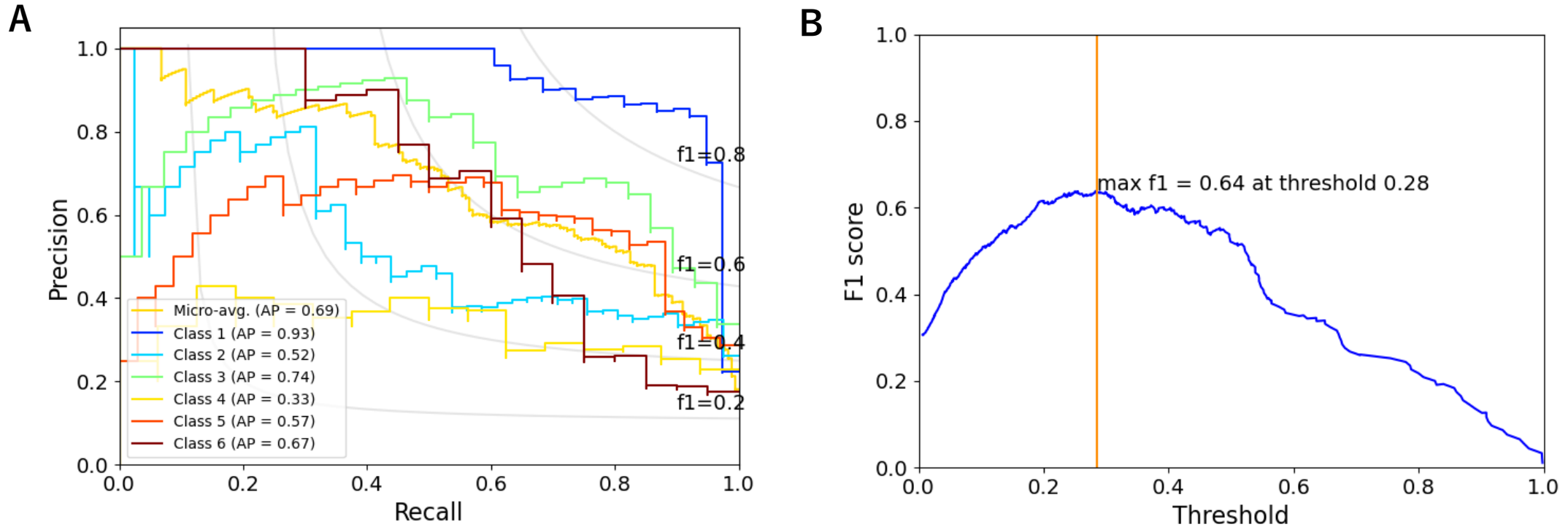


Figure S13. Precision, recall, and F1 score of SVM on community prediction based on satellite-derived parameters.

A The precision-recall curve in the condition of leave-one-out cross-validation same as Figure 4B. **B** F1 score versus threshold of probabilistic output of SVM. Orange line shows the threshold making highest F1 score.

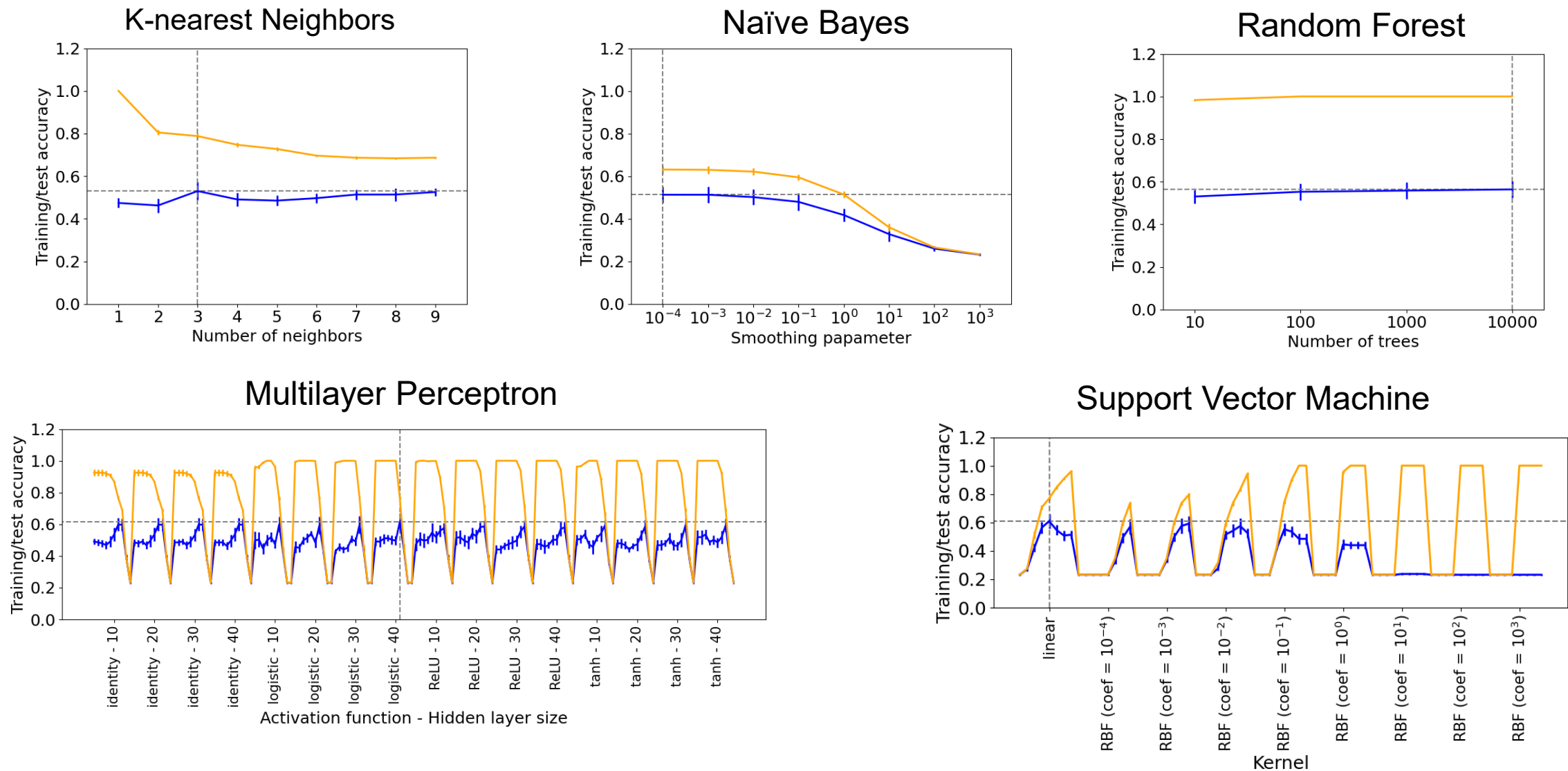


Figure S14. Grid search results in the training with all samples.

Orange and blue lines show training and test accuracy, respectively. Gray dashed lines show the parameter with the best test accuracy. Ten L2 penalty parameters (10^{-6} , 10^{-5} , ..., 10^3 ; from left to right) were tested for each setting of Multilayer Perceptron and eight (10^{-4} , 10^{-3} , ..., 10^3) were tested for Support Vector Machine.

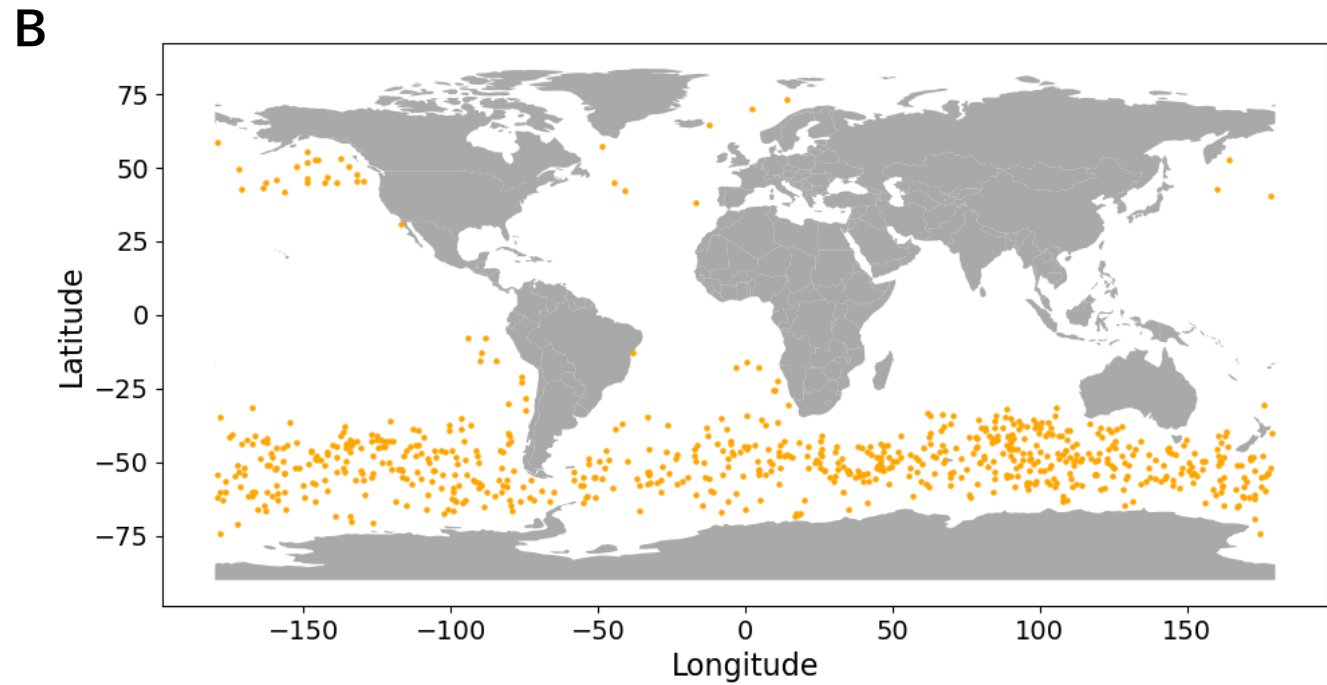
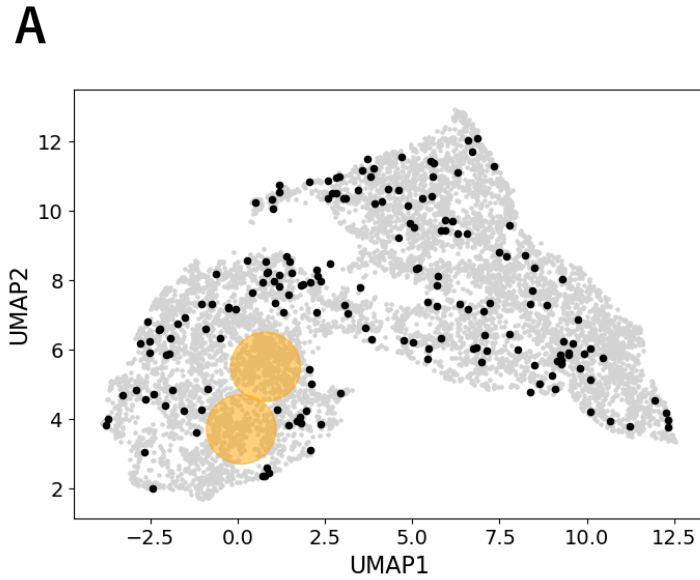


Figure S15. Location of samples in an unexplored region of the satellite-derived parameter space.

A 2-D map of the satellite-derived parameter space. An unexplored region is shown in orange. **B** Geographic location of samples in the unexplored region of the parameter space (orange points).

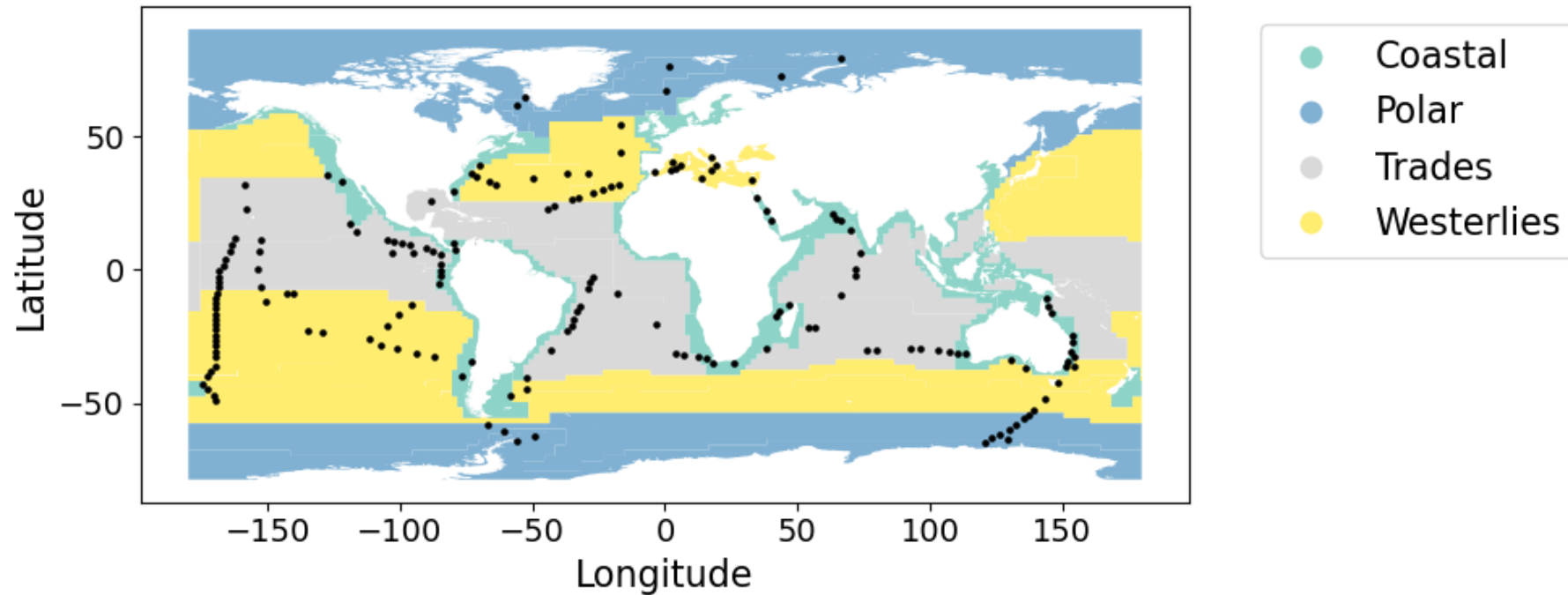


Figure S16. Map of Longhurst biome.

Black points are metabarcoding samples. The shape file of the Longhurst biomes was downloaded from Marine Regions (<https://www.marineregions.org>).