

1 **How predictable is plankton biogeography using**
2 **statistical learning methods?**

3 **L.R. Bardon^{1,2}, B.A. Ward², S. Dutkiewicz³, and B.B. Cael⁴**

4 ¹University of Southern California, Los Angeles, CA, USA

5 ²University of Southampton, UK

6 ³Massachusetts Institute of Technology, Cambridge, MA, USA

7 ⁴National Oceanography Centre, Southampton, UK

8 **Key Points:**

- 9 • We use a statistical-learning model to predict the plankton biogeography of a 21st
10 Century marine ecosystem model
- 11 • The model consistently reproduces broad qualitative patterns, but quantitative
12 predictions are less robust
- 13 • Predictive skill varies with functional group and spatiotemporally, with poor end-
14 of-century performance

Corresponding author: Lee Bardon, lbardon@usc.edu

Abstract

[Plankton play an important role in marine food webs, in biogeochemical cycling, and in moderating Earth’s climate. Their possible responses to climate change are of broad scientific and social interest; yet observations are sparse, and mechanistic and statistical methods yield diverging predictions. Here, we evaluate a statistical learning method using output from a 21st Century marine ecosystem model as a ‘ground truth’. The model is sampled to mimic historical ocean observations, and Generalised Additive Models (GAMs) are used to predict the simulated plankton biogeography in space and time. Predictive skill varies across test cases, and between functional groups, and errors are more attributable to spatiotemporal sampling bias than to sample size. Overall, the GAMs yield poor end-of-century predictions. Given that statistical methods are unable to capture changes in relationships between variables over time, we advise caution in their application and interpretation, particularly when modelling complex, dynamic systems.]

Plain Language Summary

[Marine plankton communities play a central role within the Earth’s climate system, with important processes often divided among different ‘functional groups’. Changes in the relative abundance of these groups can therefore impact on ecosystem function. Sophisticated statistical models have been developed to map the global distribution of major functional groups, based on their relationships with observed environmental variables. They appear to do a good job of summarising present-day distributions, and are increasingly being used to predict ecosystem changes throughout the 21st century. However, it is not guaranteed that such models remain valid when extrapolating over time. Rather than wait 100 years to find out, we applied such a statistical model to a complex virtual ocean. This allows to immediately jump forward to the end-of-century to test the accuracy of our predictions. We trained the model using virtual observations that match real-world ocean samples in time and place. The statistical model performed well at qualitatively predicting ‘present day’ plankton distributions but yielded poor predictions for the end of the century. The model is unable to account for changes in the underlying relationships between environmental variables and plankton distributions that occur over time. These results suggest that statistical techniques must be applied with caution when attempting to predict the future state of complex systems.]

1 Introduction

Plankton underpin global ocean food webs and fisheries, mediate marine biogeochemical cycles, and affect climate (Fenchel, 1988; Falkowski et al., 2008; Marinov et al., 2008; Guidi et al., 2016; Hutchinson, 1961). Their global biogeography interacts with the ocean’s inventory of nutrient elements, and its capacity to sequester CO₂ (Cermeño et al., 2008; Guidi et al., 2009; Fuhrman, 2009; Falkowski et al., 1998). Understanding present and possible future biogeographic patterns of plankton communities is therefore a key component of marine microbial research. These biogeographic patterns are affected by numerous environmental factors, including supplies of nutrients and light, ambient temperature, grazing pressure, physical circulation and water column structure, and the seasonality and variability of these drivers (Tittensor et al., 2010; Rutherford et al., 1999; Graff et al., 2016). Despite substantial efforts by observational oceanographers e.g. (Lombard et al., 2019), the vastness of the global ocean and the challenges of measuring complex microscopic plankton communities makes data-limitation inevitable.

Empirical methods have often been applied to making predictions from sparse observational data, from classical statistical models, to more sophisticated machine-learning (ML) methods. Their focus is not typically on extracting the underlying mechanisms that govern the behaviour of a system, but to prognostically identify correlations in data that may then be leveraged to make accurate predictions. To clarify this distinction, we here

65 follow (Holder & Gnanadesikan, 2021), in referring to underlying mechanistic relation-
66 ships as ‘intrinsic’, and the emergent correlations between variables in the data as ‘ap-
67 parent’. In the context of predicting plankton biogeography, statistical ‘niche models’
68 might seek to extract the apparent relationships between measures of plankton concen-
69 trations (e.g. cell counts, gene markers or biomass) and simultaneously measured envi-
70 ronmental factors (e.g. temperature, Chl-a, nutrient concentrations). These relationships
71 can then be used together with satellite or large synthesis database measurements to make
72 diagnostic predictions of plankton abundance. The sparse data are typically separated
73 into a training dataset for model development and a testing dataset to evaluate perfor-
74 mance. When the statistical models perform well relative to the measured datasets, pre-
75 dictions of species presence/absence or concentrations can then be scaled globally (e.g.
76 (Tang & Cassar, 2019; Barton et al., 2013; Irwin et al., 2012; Agusti et al., 2019)).

77 Data-driven methods have also been used in the specific case of predicting future
78 patterns of diversity and climate-change-driven trends in biogeography (Righetti et al.,
79 2019; Flombaum et al., 2020; Ibarbalz et al., 2019). However, their predictions have con-
80 flicted with those produced by the dynamic Earth system models used in coupled cli-
81 mate change predictions, and dynamic trait-based ecosystem models (e.g. (Ward et al.,
82 2014; Dutkiewicz et al., 2009, 2014; Cabré et al., 2015)). For instance, the neural-network-
83 derived quantitative niche model developed in (Flombaum et al., 2020) predicts
84 an increase in picophytoplankton biomass in the future subtropical oceans, in direct con-
85 trast to Earth system models, e.g. (Dutkiewicz et al., 2013; Marinov et al., 2010). Given
86 the complexity of the problem and the paucity of observational data, it is difficult to as-
87 sess which of these diverging outcomes is most likely.

88 For instance, one could argue that the output of statistical models is more trust-
89 worthy, as they do not depend on the current state of theoretical knowledge, which may
90 be incomplete. Nor do they risk the loss of important information through over-simplifying
91 system structures, components, and their interactions. However, the predictive skill of
92 statistical methods is dependant on the quality, quantity and type of available data, and
93 the suitability of a given model to the task at hand. Interpreting their outputs can also
94 prove challenging, particularly with respect to the nuanced task of separating correla-
95 tion from causation. For example, the statistical model might identify a correlation be-
96 tween sea surface temperature (SST) and plankton biomass; yet it is uncertain whether
97 SST is the primary driver of abundance, or whether separate factors coupled to SST –
98 perhaps underwater solar radiation penetration or nutrient supply rates – are more sig-
99 nificant. This can be further exacerbated by an inverse relationship between predictive
100 skill and interpretability (Carvalho et al., 2019).

101 Regardless of methodological approach, predicting unknown states of a complex
102 and dynamic system is a notoriously challenging problem. With that in mind, we em-
103 phasise that our intention here is not to broadly compare and contrast mechanistic and
104 statistical methodologies. Rather, the goal of the current work is to help minimise some
105 of this uncertainty by evaluating the performance of a popular statistical model when
106 the true global state of the system over time is known. Specifically, we set up an idealised
107 testbed to assess the predictive capabilities of Generalised Additive Models (GAMs, (Hastie
108 & Tibshirani, 1986)) using the output from an Earth system model (the ‘Darwin Model’)
109 ((Dutkiewicz et al., 2021)) as a ‘ground truth’. Darwin model output is sampled in pat-
110 terns that mimic historical ocean measurements, and at random, and the samples are
111 used to train the GAMs. In this manner, we evaluate the GAMs’ ability to capture the
112 dynamic model’s emergent biogeography in the present day ‘*spatial predictions*’ and by
113 the end-of-century ‘*temporal predictions*’. At the outset, we stress that we are not mak-
114 ing any claim as to the accuracy of the Darwin Model, nor its ability to faithfully pre-
115 dict future plankton abundance in the real world. But, as a self-consistent global ocean
116 model, with a complex, well-understood ecosystem (see e.g. (Dutkiewicz et al., 2020))

117 that is subject to perturbation by climate change, we feel that it represents a unique and
118 valuable testing analogue for the current purposes.

119 A fundamental question that we aim to address is whether the relationships be-
120 tween plankton biogeography and the considered environmental factors will be the same
121 in the present day versus the end of the century. If the statistical model accurately re-
122 produces the end-of-century biogeography of the dynamic model ocean, then we can be
123 more confident that the apparent relationships extracted from the training data are closely
124 tied to the intrinsic drivers of global plankton abundance, and that these relationships
125 remain stable over space and time.

126 2 Materials & Methods

127 We performed a suite of tests using a widely applied implementation of GAMs (Servén
128 & Brummitt, 2018) and the ‘Darwin Model’, a dynamic marine microbial ecosystem model
129 coupled to an Earth system model ((Dutkiewicz et al., 2021), (Sokolov, 2005)). To train
130 the GAMs, we sample the Darwin model at the same places and times as in a large ocean
131 measurement dataset used for similar purposes (Martiny & Flombaum, 2020). The re-
132 sulting GAMs are then used to predict Darwin Model plankton biogeography. To quan-
133 tify how spatiotemporal bias in the training dataset affects the GAMs’ predictive skill,
134 we train an additional set of GAMs using a dataset of the same size, but sampled uni-
135 formly randomly across the ocean’s surface, and uniformly randomly over the same pe-
136 riod of time. To quantify the effect of training set sample size on the GAMs’ predictive
137 skill, we generate 54 additional random-sample training sets, in 18 different sample sizes.
138 We evaluate the ability of the GAMs to predict the global biogeography of the differ-
139 ent plankton functional groups in the simulation, both during the 22-year period over
140 which measurements were taken (i.e. spatial extrapolation), and during the last 22 years
141 of the 21st century (i.e. both spatial and temporal extrapolation).

142 2.1 Numerical Model Simulation

143 The Darwin model ecosystem used here includes 51 plankton populations across
144 7 functional groups (2 prokaryotes (pro), 2 pico-eukaryotes (pico), 5 coccolithophores (cocco),
145 5 diazotrophs (diazo), 11 diatoms (diatom), 10 mixotrophic dinoflagellates (dino) and
146 16 zooplankton (zoo)). Individual populations correspond to different size classes within
147 functional groups, with all size classes covering a range of 0.6–2425 μm equivalent spher-
148 ical diameter. Functional groups have distinct allometric relationships for growth, graz-
149 ing, and sinking parameters (see (Dutkiewicz et al., 2020)). The model ecosystem is em-
150 bedded within the Massachusetts Institute of Technology Integrated Global System Model
151 (IGSM) (Prinn, 2013; Sokolov, 2005) which includes modules for the physics, chemistry,
152 and biogeochemistry of the atmosphere, land and ocean. The ocean component has a
153 $2^\circ \times 2.5^\circ$ resolution grid and 22 vertical layers (10m thickness at surface to 500m at bot-
154 tom). The simulation is forced with observed greenhouse gas emissions from 1860–1990
155 and then with a high emissions scenario that is analogous to the IPCC’s Representative
156 Concentration Pathway 8.5, from 1990 – 2110. This perturbation results in $\sim 3^\circ\text{C}$ sea
157 surface temperature warming by 2100, sea ice retreat, increased stratification, and an
158 altered overturning circulation. The IGSM has been used to examine changes in marine
159 biogeochemistry and ecology in previous studies (Dutkiewicz et al. 2013; 2019) but with
160 a simpler version of the ecosystem model. The current more complex ecosystem has also
161 been used in previous studies, but only for the current day’s ocean (Dutkiewicz et al.,
162 2021; Sonnewald et al., 2020; Kuhn et al., 2019). This model and previous model val-
163 idation for the current day demonstrates the output compares well with observations along
164 both axes of size and functional type (e.g. (Dutkiewicz et al., 2021, 2020)).

165 **2.2 Ecosystem and Environmental Variables**

166 Surface-level plankton abundance data and environmental parameters were extracted
 167 from the Darwin model simulation output, where surface in this context refers to the 10m
 168 thick surface grid box. The ecosystem data contains 51 separate plankton biomasses, ar-
 169 ranged into seven functional groups (as described above). A number of environmental
 170 variables have been used by statistical models to predict abundance and diversity, and
 171 have thus been included here. They are: sea surface temperature (SST), photosynthet-
 172 ically active radiation (PAR), phosphate (PO_4), nitrate (NO_3), silicate (Si) and iron (Fe).
 173 We sampled both the plankton abundance data and the environmental predictor vari-
 174 ables from the 3586 spatiotemporal cells that encompass the representative ocean mea-
 175 surement coordinates, and from the 3586 randomly selected spatiotemporal cells. Note
 176 that the model simulation used for the current analysis nominally starts in 1991 and ex-
 177 tends to 2100. As such, we sample the model output from the beginning of 1991 to the
 178 end of 2012 and consider this as a substitute to 1987–2008 in this context. This is jus-
 179 tified because the dynamic model’s internal variability does not match real-world inter-
 180 annual variability in terms of timing, though does capture the magnitudes (e.g. there
 181 are El Niño events, but these do not occur in the same years as the real ocean). To val-
 182 idate predictions, we also consider whole-ocean surface data over the same period, and
 183 for the final 22 years of the simulation, from 2079 – 2100.

184 **2.3 Training the Statistical Learning Model**

185 A variety of statistical learning algorithms have already been applied to ocean mea-
 186 surement data, and used to make predictions about the future state of the ocean micro-
 187 biome (see e.g. (Righetti et al., 2019; Flombaum et al., 2020)). Indeed, the methods and
 188 results of (Righetti et al., 2019) act both as a guide to the current work, and as a con-
 189 tributing factor in our decision to use GAMs as our ‘representative’ statistical learning
 190 method. This is due to the (Righetti et al., 2019) finding that GAMs perform compa-
 191 rably to Random Forest and Generalised Linear Models in a range of predictive tasks,
 192 while offering a high degree of both interpretability and flexibility. Additionally, GAMs
 193 are of intermediate complexity between classical statistical regression models, and more
 194 sophisticated machine learning methods, making them both accessible and potentially
 195 attractive to a wide range of researchers.

196 Here, we used the standard ‘LinearGAM’ model of the freely available PyGAM pack-
 197 age (Servén & Brummitt, 2018), incorporating a Gaussian distribution function with an
 198 identity link function. Feature functions are built using penalised B-splines that impose
 199 smoothness to avoid over-fitting, and enable the automatic fitting of nonlinear relation-
 200 ships, while maintaining additivity. For an initial set of results, we set the number of per-
 201 mitted splines to 20 for each predictor variable. We note that our results are not sen-
 202 sitive to the choice of this parameter (see ‘Model Comparison & Sensitivity Tests’). Rather
 203 than attempt to resolve and make predictions for individual plankton tracers, we instead
 204 sum the abundance data for each functional group, and train GAMs accordingly. The
 205 resulting partial dependency plots are examined for unexpected behaviours, or any clear
 206 indications of over or under-fitting. We thus use the relationships identified by the GAMs
 207 to make predictions for the global surface ocean plankton biomasses during 1987-2008
 208 and 2079-2100.

209 **2.4 Model Comparison & Sensitivity Tests**

210 We define presence/absence as modelled biomass being above/below a cutoff thresh-
 211 old (10^{-5} mmol C/m³), but find that patterns in the resulting predictions are not sen-
 212 sitive to the choice of this threshold (Table S4).

213 The R^2 value of the GAMs predictions against the ‘ground-truth’ simulation val-
 214 ues is given as $R^2 = 1 - SS_{res}/SS_{tot}$, where SS_{res} is the residual sum of squares and
 215 SS_{tot} is the total sum of squares. While R^2 is a widely-used statistic in regression anal-
 216 yses, it does not by itself provide a complete picture of goodness of fit. We therefore also
 217 examine the mean and median relative differences, defined here as $\bar{X}_{me} = (mean_{predicted} -$
 218 $mean_{actual})/mean_{actual}$ and $\tilde{X}_{md} = (median_{predicted} - median_{actual})/median_{actual}$,
 219 as an indicator of bias. We also consider the false positive and false negative fractions,
 220 i.e. the fraction of grid cells where the GAMs incorrectly predict, respectively, present
 221 and absent biomass. Finally, we performed the above analyses with the logarithm of biomass
 222 concentrations and found that our results were not sensitive to this choice. Overall, we
 223 found that coccolithophores yielded the median performance in terms of goodness of fit
 224 with respect to spatial extrapolations. As such, this group is featured in the main body
 225 of this work, while results for the other six functional groups are reported in the sup-
 226 plements.

227 GAM sensitivity was investigated by varying the number of splines used in perform-
 228 ing the fits; first by halving to 10, and then doubling to 40. While the resulting partial
 229 dependency plots revealed a clear change to the smoothness of the fit, as expected, we
 230 found that the resulting statistics were not appreciably impacted. To investigate the ef-
 231 fect of sample size on the overall predictive power of the GAMs, we vary the number of
 232 randomly-sampled cells from a minimum of 100 (reducing to 63 ocean cells), to a max-
 233 imum of 20,000 (reducing to 11,557 ocean cells), using 18 different test cases. Each sam-
 234 ple size test case consists of three independent random samples, with the mean value be-
 235 ing reported along with the standard deviation (Figure 4).

236 We also perform a range of simpler correlation analyses, to build a broader picture
 237 of the emergent relationships between functional group biomass and predictors. These
 238 act as a visual aid to better understand how these relationships might change in time
 239 and space, and as a basic cross-reference for GAMs-derived partial dependence plots of
 240 the training sets. We first calculate the Pearson’s Correlation Coefficient (ρ) for each func-
 241 tional group-predictor pair, and the Spearman’s Rank Correlation Coefficient (ρ_s). Re-
 242 spectively, these popular methods detect the strength of linear associations between vari-
 243 ables, and the strength of correlation in monotonic relationships. A commonly used method
 244 for addressing skew or capturing scaling relationships is the log-transform, which we ap-
 245 ply to all datasets before recalculating ρ . However, this method of broadly applying a
 246 single transformation is not optimal. A more robust approach would be to examine the
 247 distribution of each target-predictor relationship individually, before an appropriate trans-
 248 formation is selected. Nonetheless, even this more optimal method runs the risk prop-
 249 agating transformation uncertainty into the resulting confidence interval.

250 With these limitations in mind, we also determine correlations using the more re-
 251 cent distance correlations method of (Székely et al., 2007). This technique captures the
 252 strength of both linear and nonlinear associations and avoids the need to make assump-
 253 tions about variable distributions or linearity. We plot the correlation matrices for the
 254 main 3586 cell test cases, both measurements-derived and randomly-sampled, in 1987-
 255 2008, and at the same locations in 2079-2100. We explore the effect of sample size on
 256 the derived correlations by increasing the number of randomly-sampled cells to 12,894,
 257 and finally to 25,683 cells.

258 3 Results

259 3.1 Spatial Predictions

260 We first describe the results of predicting plankton biogeography during the his-
 261 torical measurement period (1987 – 2008) (Figure 1). We find that predictive ability
 262 varies considerably across functional groups. There are fewer instances of GAMs incor-

263 rectly predicting presence (false positive) or absence (false negative) biomass for prokary-
 264 otes, picophytoplankton and coccolithophores (16–19% of all location-month pairs) than
 265 for diatoms, diazotrophs, and dinoflagellates (26–31%), with zooplankton in between
 266 (21%). Where biomass is present and is predicted as such, GAMs’ predictive ability for
 267 biomass concentration also varies substantially between functional groups (Figure 2); the
 268 GAMs account for as much as 71% of the variance in biomass (diazotrophs) and as lit-
 269 tle as 41% (zooplankton). These patterns are reflected also in the mean relative differ-
 270 ences and the balanced accuracy.

271 Patterns of overprediction of biomass occurs across most of the oceans. For prokary-
 272 otes, picoeukaryotes, dinoflagellates and zooplankton, this is especially evident in the Ar-
 273 ctic (see Figures (c) of S1, S2, S5, S6). For these groups, we also see consistent underpre-
 274 diction in most of the Indian Ocean and in the Eastern Equatorial Pacific. Meanwhile,
 275 diatoms are substantially overpredicted in most of the mid- and high-latitudes in the North-
 276 ern Hemisphere but perform relatively well in the subtropics (Figure S4(c)). Diazotrophs
 277 yield the best overall performance, with only a small amount of overprediction in the sub-
 278 tropical Atlantic, and overprediction in the transition zone latitudes poleward of the sub-
 279 tropics (Figure S3(c)).

280 In general the GAMs show a tendency to overestimate biomass in the spatial pre-
 281 diction regime. Overestimation ranges between 9–21% on average (picoeukaryotes and
 282 zooplankton, respectively), with a median overprediction of $\geq 16\%$. Despite this, there
 283 are some notable instances in the current context where the GAMs perform well. Spa-
 284 tial predictions for coccolithophores, prokaryotes and diazotrophs all yield R^2 values that
 285 range between 0.62 and 0.71 (Figures 1(e), S1(e), S5(e)). Diazotrophs fare particularly
 286 well in this regime, with a mean overprediction of 10%, an R^2 of 0.71, and the best vi-
 287 sual, qualitative match of biogeography overall (although we note that the median over-
 288 prediction in this case is a substantial 194%) (Figures S3(c) and S3(e)). On the whole,
 289 GAMs trained on data from historical measurement locations appear to be able to re-
 290 produce qualitative biogeographic patterns from spatial predictions well, but quantita-
 291 tive performance is variable, with a broad tendency towards overprediction. Notably, the
 292 greatest predictive errors more often occur in the undersampled regions of the ocean, such
 293 as the Arctic and Indian Oceans, but are by no means confined to these regions. For in-
 294 stance in the highly sampled North Atlantic predictions for diatoms and diazotrophs was
 295 also poor.

296 3.2 Temporal Predictions

297 GAMs’ predictive ability is substantially reduced when extrapolating to the future
 298 ocean (see Figures 1 and 2). Rates of false positives and negatives in presence/absence
 299 do not uniformly change across functional groups: the cosmopolitan groups whose ranges
 300 expand poleward experience the least overall change, increasing by between 3% and 11%
 301 in prokaryotes, dinoflagellates and coccolithophores, with a decrease of 5% for picophy-
 302 toplankton. GAMs’ ability to correctly predict presence/absence is further reduced for
 303 the groups with a more confined biogeography, increasing by between 14% and 23% for
 304 diazotrophs, zooplankton and diatoms. We see a substantial increase in false negative
 305 occurrences for diatoms (to 29%), the group whose biogeographic range contracts most.
 306 Where biomass is present and is predicted as such, GAMs’ predictive ability was reduced
 307 for all functional groups. In most cases, this reduction is substantial, with the fraction
 308 of variance accounted for by the GAMs reducing by between 17 and 50%, such that the
 309 prediction for zooplankton is worse than just assuming a globally uniform constant biomass
 310 (i.e. $R^2 < 0$). We see a marked increased in mean relative differences compared to the
 311 ‘spatial’ predictions, accompanied by a reduction in balanced accuracy for all groups be-
 312 sides diatoms (Figure 2).

313 Diatoms are the only group for which the fraction of variance accounted for does
 314 not decrease substantially, only from $R^2 = 0.59$ to $R^2 = 0.56$ (Figure S4). Thus, the
 315 predictive ability for diatom biomass where it is present is not greatly reduced, despite
 316 the GAMs substantial overprediction of the contraction of diatoms' biogeography. This
 317 is not sensitive to varying the absence/presence cut-off value by an order or magnitude
 318 in either direction (Table S1).

319 Spatial patterns of prediction errors of coccolithophores, prokaryotes, picoeukary-
 320 otes, dinoflagellates and zooplankton are largely similar to those for the historical pe-
 321 riod, except the North Atlantic is now underpredicted for all groups besides diazotrophs
 322 (Figures 1, S1, S2, S4, S5, S6). Diatom biomass is notably underpredicted in the South-
 323 ern Ocean and Northern Atlantic (Figure S4). Meanwhile, diazotroph biomass is notably
 324 overpredicted throughout the Atlantic Ocean, the Arctic, bands of the subtropical Pa-
 325 cific and Indian Ocean (Figure S3). Excluding diatoms, the overall tendency towards over-
 326 prediction is exacerbated for all groups, increasing by 57% for prokaryotes, picoeukary-
 327 otes, coccolithophores, and dinoflagellates, by 20% for zooplankton, and by 49% for di-
 328 azotrophs. Median overpredictions also increase for all groups besides diatoms.

329 3.3 Model Trained on Randomised Locations

330 Here we compared the above results with those produced when the GAMs were trained
 331 on randomly sampled datasets (Figure 2). Interestingly, the broad spatial patterns of
 332 where overprediction and underprediction occurs do not change much when training GAMs
 333 on randomly distributed data, as opposed to the ocean observation locations (Figures
 334 S8 and S9). Nonetheless, predictive abilities increase, biases are reduced, and balanced
 335 accuracy increases in both the spatial and temporal cases (Figure 2). The fraction of vari-
 336 ance accounted for by the GAMs increases by 2–19% when using random data to pre-
 337 dict historical biogeography, but increase from 5–46% when using random data to pre-
 338 dict future biogeography. The most notable differences are for prokaryotic, picoeukary-
 339 otic, and zooplankton biomass in the future case. The magnitude of the biases also de-
 340 creases – average biases are within 3–4% in the historical case using random data. The
 341 median bias for all groups is still that of overprediction, with most groups in the range
 342 of $\geq 17\%$ compared to $\geq 30\%$ for measurements-derived predictions. Diatoms and dia-
 343 zotrophs have a markedly higher bias in both measurements-derived and random cases,
 344 of $\geq 194\%$ and $\geq 162\%$, and $\geq 65\%$ and $\geq 35\%$. In the future case, using random data re-
 345 duces biases for all groups, though does not eliminate them. We also found that the pre-
 346 dictive ability of the GAMs was only weakly dependent on sample size (where sample
 347 size here refers to the number of grid cell-month pairs that are sampled)(Figure 4), with
 348 predictive ability appearing to plateau with increasing sample size.

349 The results using random training datasets suggest that historical measurement
 350 biases reduce the predictive ability of GAMs more than the sample size of the training
 351 dataset. Predictive ability can be improved by subsampling or weighting one's training
 352 dataset to reduce biases in space and time, although the coarse resolution of the Dar-
 353 win model – and thus reduced variability as a result of correlated observations – rela-
 354 tive to the real ocean may contribute to this plateauing effect.

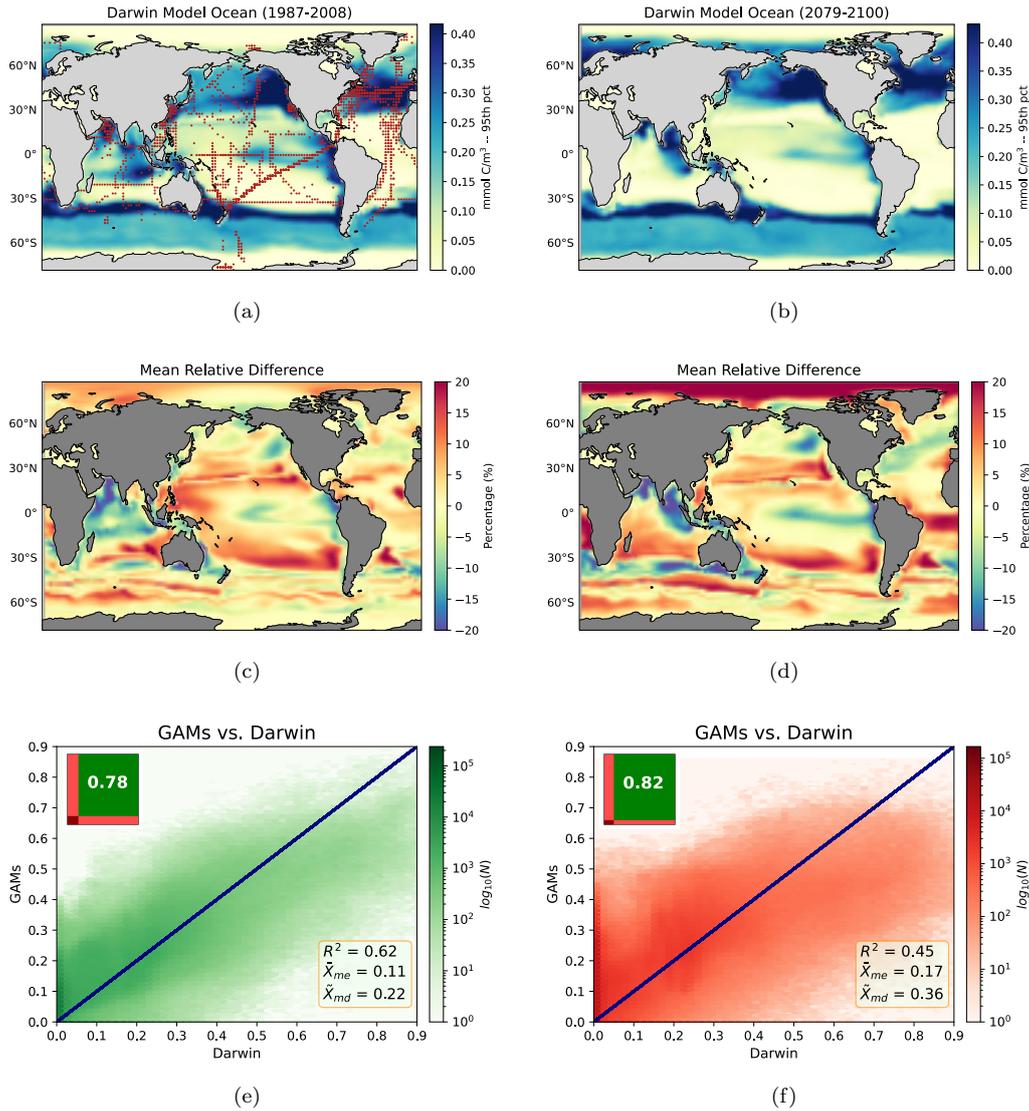


Figure 1: **(a)** Mean coccolithophore surface biomass (1987 - 2008) from the Darwin model. Red points indicate spatial location of training set datapoints, derived from ocean measurement data. **(b)** As per 1(a) for the years 2079 - 2100. **(c)** Relative (percent) difference between mean diatom surface biomass from the Darwin model and the GAMS (1987 - 2008) **(d)** As per 1(c) for the years 2079 - 2100. For direct visual comparison, we first calculate the 5th and 95th percentile of the relative difference values for both the spatial and temporal predictions, then scale symmetrically to whichever of these values is the greatest, in either direction. **(e)** Hexagonally binned scatterplot of 1987-2008 GAMS predictions vs 1987-2008 Darwin model. Colorbar shows log-scaled density of observations. *Top inset:* Fraction of data above the presence/absence threshold (10^{-5} mmol C/m³)(green box), GAMS below threshold (left, light red), Darwin below threshold (bottom, light red), both below threshold (dark red). *Bottom inset:* The R^2 , relative difference of the means (\bar{X}_{me}), and relative difference of the medians (\bar{X}_{md}). **(f)** As per 1(e) but for 2079-2100. See Supplemental Materials for other functional groups.

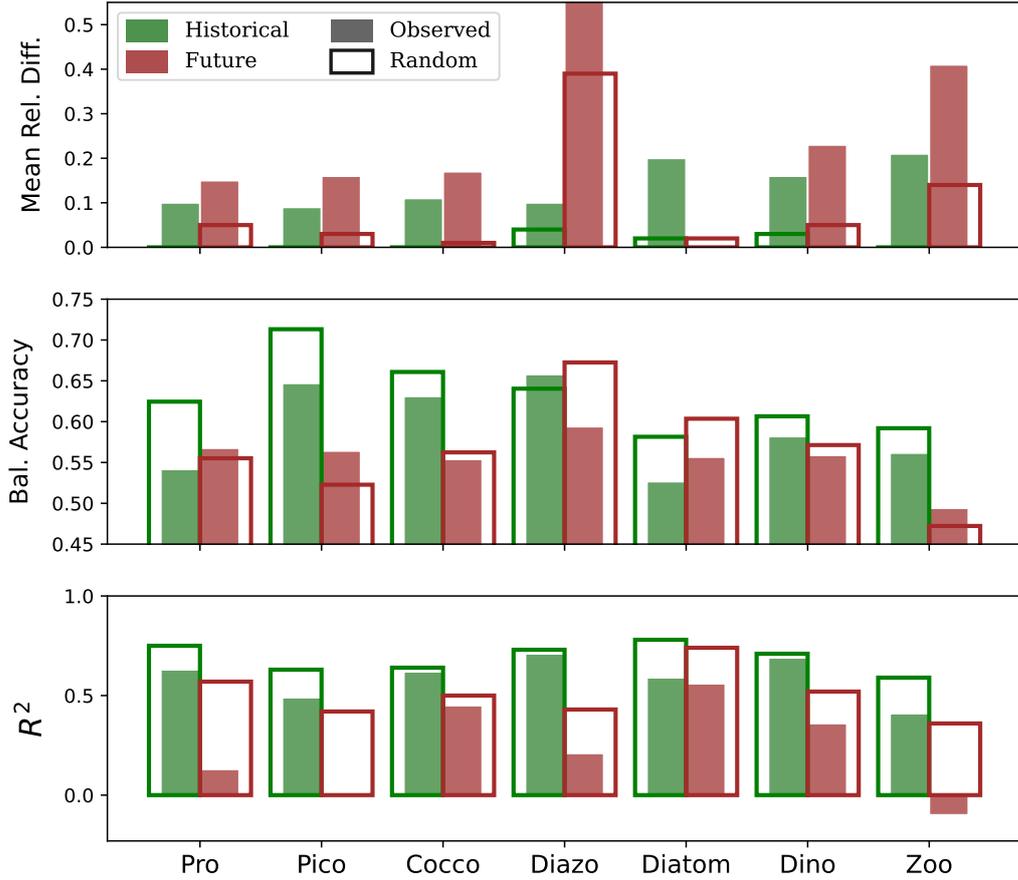


Figure 2: Comparing Darwin model ‘true’ biomasses with GAMs predictions for each functional group in 1987-2008 (historical) and 2079-2100 (future), and from measurements-derived and randomly-sampled training sets. *Top to Bottom: (a)* Relative differences of the means, given by $(GAMs_{mean} - Darwin_{mean})/Darwin_{mean}$. *(b)* Balanced accuracy, given by $(sensitivity + specificity)/2$. *(c)* R^2

4 Discussion

Broadly, our results suggest that statistical models – as applied in the current context – can qualitatively capture large-scale spatial patterns of plankton biogeography, but struggle to make robust quantitative predictions. This is particularly evident when the model is trained on historical ocean measurement data, and used to predict future plankton biogeography as a response to climate change. The fraction of variance that GAMs can account for saturates with sample size well below 100%, implying a ceiling on GAMs’ predictive ability. The emergent relationships between predictor variables and plankton abundances change spatially, seasonally and over the longer term. This is demonstrated by the variable nature of the partial dependence plots (Figure 3(a)–(b) and Figures S10 and S11), and by the change in correlation strengths identified by each of the independent methods used in generating the correlation matrices (Figure 3(c)–(f) and Figure S12). The correlation matrices offer an especially powerful visual demonstration of these points; we clearly see the change in apparent relationships between biomass and environmental predictors in the measurement-derived sample space, assessed over the same period of time, one hundred years into the future (Figure 3(c) and 3(d)). It’s im-

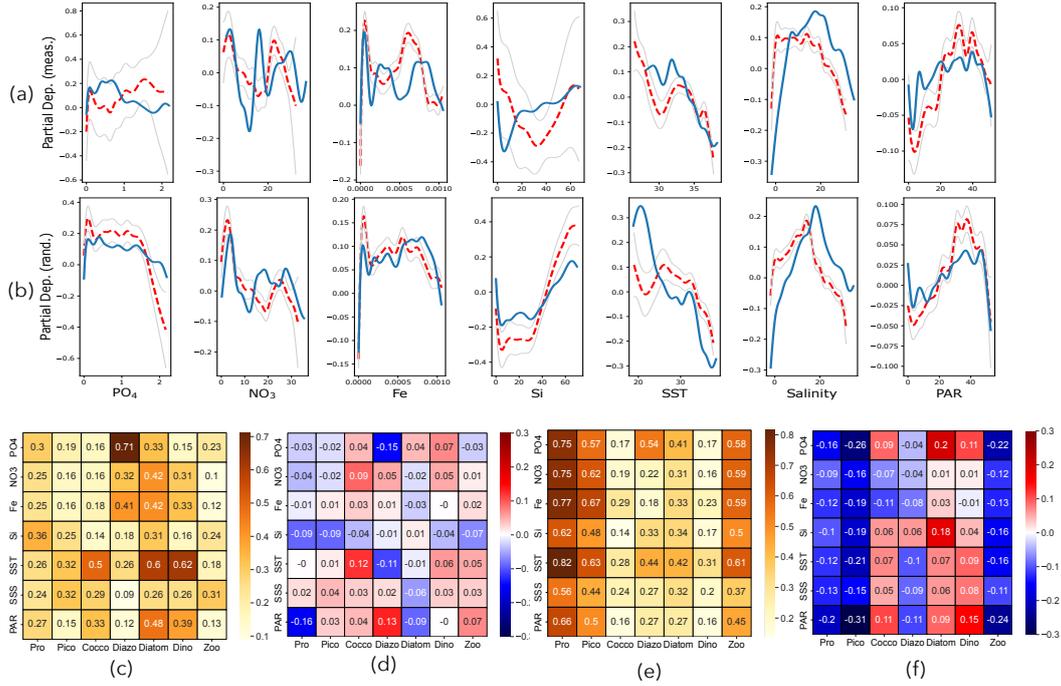


Figure 3: *Changing Relationships*: (a) Partial dependence plots of coccolithophore biomass (mmol C/m^3) as a function of each predictor, centred around the median (PO_4 , NO_3 , Fe, Si in mmol X/m^3 , SST in $^\circ\text{C}$, SSS in PSU, PAR in $\text{E/m}^2/\text{day}$). Plotted using data from 3586 Darwin surface ocean cells at measurements-derived locations spanning 1987-2008 (dashed red line) and at the same locations from 2079-2100 (blue line). Grey lines indicate 95% confidence interval for the 1987-2008 case. (b) As per 3(a), but using data from 3586 randomly sampled cells. (c) Correlation heatmap for the measurements-derived training set, 1987-2008, generated using the distance correlations method of (Székely et al., 2007). (d) Difference between correlation strengths derived in 3(b) and those found at the same locations from 2079-2100. (e) and (f) As per 3(c) and 3(d), but for the equivalently-sized, randomly-sampled training set.

372 important to note that we should expect these differences to be exaggerated in the real world,
 373 where the system is significantly more complex.

374 For example, there are many more degrees of freedom in real-world interactions be-
 375 tween plankton individuals, communities, the wider ecosystem and environment. In ad-
 376 dition to the controlling influence of e.g. nutrient supply rate, physical transport pro-
 377 cesses and level of top down pressure, plankton are also able to adapt genetically and
 378 epigenetically to change. With their short generation times and high biodiversity, we might
 379 reasonably expect intrinsic relationships to change over the course of a century. This is
 380 especially likely in such a dynamic, randomly-perturbed, and far-from-equilibrium en-
 381 vironment, where conditions are ideal for unpredictable emergent phenomena to arise.
 382 By contrast, all such elements within the Darwin Model are simplified by design and by
 383 necessity, and intrinsic relationships are held steady over time, such that the spatiem-
 384 poral variability in apparent relationships seen here are the product of many fewer sources
 385 of complexity.

386 Additionally, our results also demonstrate how spatial sampling bias can signifi-
 387 cantly alter the patterns of apparent relationships between environmental predictors and

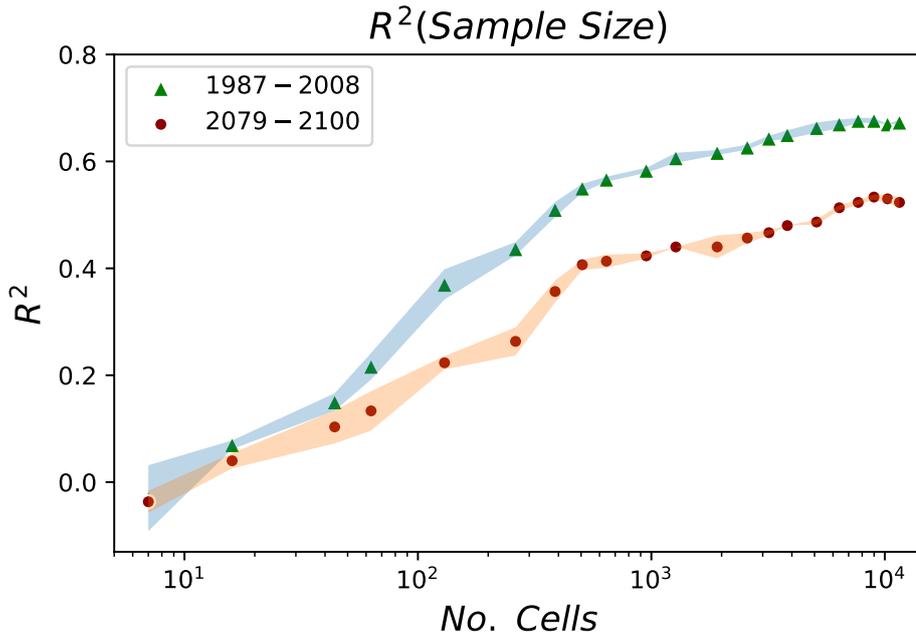


Figure 4: R^2 of GAMs model prediction as a function of sample size. Points are the mean R^2 value for coccolithophore predictions from three independent randomly-generated training sets for each of the 18 sample sizes, ranging from $N=63$ to $N=11,557$. Shading is the standard deviation.

388 plankton biomass. The association strengths identified in the measurements-derived sam-
 389 ple vary considerably from those found in the randomly sample of equivalent size (see
 390 Figure 3(c) vs. 3(e)). Importantly, this finding is robust across a range of random sam-
 391 ple sizes, where almost identical patterns of correlations are seen in the 3586 cell case
 392 as in the 25,683 cell case, and robust across several methods of deriving correlations (see
 393 Figure S12). Nonetheless, the spatial patterns of over and under-prediction derived from
 394 the GAMs are not merely the result of spatiotemporal measurement biases. We see re-
 395 markable agreement in these broad qualitative patterns between the predictions gener-
 396 ated from measurements-derived and random samples ((c) and (d) of Figures 1, and S1-
 397 6, and Figures S8 and S9). Ocean measurement biases may explain some element of the
 398 tendency towards overestimation of historical biogeography/abundances; perhaps because
 399 measurements have more often been made in places with higher than average abundances.
 400 In all cases, training the statistical model on a non-biased dataset reduces the severity
 401 of over and underprediction, especially for spatial predictions (Figure S8(e) and S9(e)).
 402 But the same broad biogeographic patterns of over and underprediction remain, indi-
 403 cating that the GAMs are still failing to effectively capture changes over time, despite
 404 their relatively robust performance according to the broad brush strokes of summary statis-
 405 tics (Figure S4(e) and S4(f)).

406 With that in mind, a number of optimisations could be made to improve predic-
 407 tive skill in real-world applications. First, we note that an unrepresentative training set
 408 presence/absence ratio compared to the population can lead to an unreliable represen-
 409 tation of presence/absence in the resulting predictions. To avoid this possibility, researchers
 410 working with real observational data will sometimes employ resampling techniques (e.g.
 411 (Wei & Dunbrack, 2013)) to account for this effect. By contrast, our experimental de-

412 sign permitted us the unusual opportunity of testing our outcomes alongside a range of
413 representative, randomly-sampled datasets spanning the surface ocean. These unbiased
414 samples are representative of the presence/absence ratios of the population, and thus act
415 as a control for our observations-derived test case. Given the broadly similar patterns
416 of over and underprediction found across test cases, we do not employ resampling tech-
417 niques here, but we encourage their application in real-world settings.

418 Related also to the more flexible nature of our study in comparison to statistical-
419 learning models applied to real-world observations, is the manner in which we approach
420 training, validation and testing datasets. In some cases, ML practitioners working with
421 real-world observations might reserve a proportion of the training set for model valida-
422 tion, as well as an independent, but similarly-distributed, dataset for performance test-
423 ing. A validation set allows for optimisation via the fine-tuning of model parameters, and
424 for the avoidance of over-fitting, while the test set permits evaluation of model skill. Here,
425 we use whole-ocean Darwin Model output as our test set for evaluating overall perfor-
426 mance. Given model response to sensitivity tests, and GAMs natural robustness to over-
427 fitting, we do not explicitly employ a validation set. Model skill could be improved with
428 parameter fine-tuning, especially in the spatial predictions test case. It is less clear whether
429 fine-tuning for GAMs performance using a training set sampled from the Darwin Model
430 ocean of 1987-2008 would have a positive effect on end-of-century predictions, as this would
431 depend on the direction of drift between the statistical model and the ground truth over
432 time. Additionally, we speculate that our decision to train the GAMs using the entire
433 measurements-derived sample might itself yield improvements relative to splitting the
434 samples into training, testing and validation subsamples.

435 We focus here on a particular type of statistical learning method that, for reasons
436 outlined in Materials & Methods, we believe makes for an excellent case study. Our in-
437 vestigation has allowed us to better clarify the strengths and limitations of such an ap-
438 proach, as applied in the current context. Owing to the complex and ever-changing na-
439 ture of the system, some of these limitations could be fundamental and unavoidable, par-
440 ticularly when extrapolating far beyond the training regime. Indeed, the median over-
441 estimation by the GAMs, even when using randomly sampled training data, implies that
442 the predicted abundance distributions are less skewed than the Darwin model distribu-
443 tions, which are, in turn, less skewed than distributions in the the real ocean. But we
444 stress that these observations do not extend to data-driven methods writ large. The re-
445 cent work of (Holder & Gnanadesikan, 2021) evaluates random forests (RF) and neu-
446 ral network ensembles (NNE) in their ability to resolve the intrinsic relationships between
447 plankton biomass and predictors, as extracted in a laboratory setting, from the appar-
448 ent relationships in the data. They demonstrate variability in model predictive skill across
449 different test scenarios, and find that NNE's yield overall superior performance, partic-
450 ularly in the case where plankton growth rates respond rapidly to environmental change.
451 However, while these more sophisticated machine-learning models might yield an improve-
452 ment in predictive skill, this can come at the cost of interpretability. Nonetheless, recent
453 work by (Rudy et al., 2017) has shown that it is possible to use data-driven methods to
454 directly extract the mechanistic equations that describe a dynamical system. This is an
455 extremely promising advance toward hybrid methods that can provide both high levels
456 of predictive skill, and an underlying description of the drivers of change.

457 Methodologically, the approach we have presented of applying a statistical model
458 to output from a numerical model may be useful for addressing a number of additional
459 questions. These might include evaluating how best to statistically model whole-ecosystem
460 properties, such as diversity, from observations, or assessing where and when to make
461 new observations to maximise information content about global plankton biogeography. But,
462 as our results here have demonstrated and reinforced, it is important to be aware strengths
463 and limitations of this approach, especially when dealing with a high degree of complex-
464 ity over time.

5 Conclusion

In summary, our results suggest that statistical models like the one explored here can be powerful tools for extrapolating from sparse measurement sets to capture the qualitative spatial patterns of plankton biomass in the present-day ocean. However, these biomass predictions are especially sensitive to the spatiotemporal bias in historical measurements, and can tend towards overprediction if not properly accounted for. In addition, such models demonstrably struggle to predict future plankton biomass because the inherently complex and dynamic nature of the system generates variability in the relationships between predictors and biomass over time; variability that cannot be captured by statistical methods. This model drift effect could be exaggerated when attempting to address the substantially more complex problem of predicting real-world plankton biogeography using sparse observational data. Of course, this is a challenge that applies equally to all methods that may be applied to its possible resolution. Our results nonetheless help to constrain the strengths and limitations of statistical learning models in this context, and when applied to a wide range of broadly similar problems.

Acknowledgments

Ward acknowledges support from a Royal Society University Research Fellowship. Dutkiewicz acknowledges support from the Simons Collaboration on Computational Biogeochemical Modelling of Marine Ecosystems (CBIOMES)(Grant Id: 549931) and from the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Cael acknowledges support from the National Environmental Research Council (NE/R015953/1) and the Horizon 2020 Framework Programme (820989). The work reflects only the authors' view; the European Commission and their executive agency are not responsible for any use that may be made of the information the work contains. Finally, the authors' would like to thank the two anonymous reviewers for their insightful comments, which have yielded substantial improvements to the final version of this manuscript.

Code Availability. The physical model used here is available through <http://www.mitgcm.org>, and the generic ecosystem code is available through <http://www.gitlab.com/jahn/gud>. The specific modifications for the setup used here are available via Harvard Dataverse at <http://www.dataverse.harvard.edu/dataverse/>. Note that a more up-to-date version of the ecosystem model used here is available at <http://www.github.com/darwinproject/darwin/>. The code used to process and analyse the data, and to produce the results for this manuscript, is available at <https://github.com/teatauri/stats-biogeo-2021>.

Data Availability. The Darwin Model output used in the current study is available at <http://www.dataverse.harvard.edu/dataverse/>. The dataset will have a doi, and will be hosted through the Harvard Dataverse Darwin project site. The extracted and processed Darwin surface data will also be made similarly available.

References

- Agusti, S., Lubián, L. M., Moreno-Ostos, E., Estrada, M., & Duarte, C. M. (2019). Projected changes in photosynthetic picoplankton in a warmer subtropical ocean. *Front. Mar. Sci.*, 5.
- Barton, A. D., Pershing, A. J., Litchman, E., Record, N. R., Edwards, K. F., Finkel,

- 514 Z. V., ... Ward, B. A. (2013). The biogeography of marine plankton traits.
515 *Ecology Letters*, 16(4), 522–534.
- 516 Cabré, A., Marinov, I., & Leung, S. (2015). Consistent global responses of ma-
517 rine ecosystems to future climate change across the IPCC AR5 earth system
518 models. *Clim. Dyn.*, 45(5), 1253–1280.
- 519 Carvalho, D. V., Pereira, E. M., & Cardosa, J. S. (2019). Machine learning inter-
520 pretability: A survey on methods and metrics. *Electronics*, 8, 832.
- 521 Cermeño, P., Dutkiewicz, S., Harris, R. P., Follows, M., Schofield, O., & Falkowski,
522 P. G. (2008). The role of nutricline depth in regulating the ocean carbon cycle.
523 *PNAS*, 105(51), 20344–20349.
- 524 Dutkiewicz, S., Boyd, P. W., & Riebesell, U. (2021). Exploring biogeochemical
525 and ecological redundancy in phytoplankton communities in the global ocean.
526 *Global Change Biology*, 27(6), 1196–1213.
- 527 Dutkiewicz, S., Cermeño, P., Jahn, O., Follows, M. J., Hickman, A. E., Taniguchi,
528 D. A. A., & Ward, B. A. (2020). Dimensions of marine phytoplankton diver-
529 sity. *Biogeosciences*, 17(3), 609–634.
- 530 Dutkiewicz, S., Follows, M. J., & Bragg, J. G. (2009). Modeling the coupling of
531 ocean ecology and biogeochemistry. *Global Biogeochemical Cycles*, 23(4).
- 532 Dutkiewicz, S., Scott, J. R., & Follows, M. J. (2013). Winners and losers: Ecologi-
533 cal and biogeochemical changes in a warming ocean. *Global Biogeochemical Cy-
534 cles*, 27(2), 463–477.
- 535 Dutkiewicz, S., Ward, B. A., Scott, J. R., & Follows, M. J. (2014). Understanding
536 predicted shifts in diazotroph biogeography using resource competition theory.
537 *Biogeosciences*, 11(19), 5445–5461.
- 538 Falkowski, P. G., Barber, R. T., & Smetacek, V. (1998). Biogeochemical controls
539 and feedbacks on ocean primary production. *Science*, 281(5374), 200–206.
- 540 Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that
541 drive earth’s biogeochemical cycles. *Science*, 320(5879), 1034–1039.
- 542 Fenchel, T. (1988). Marine plankton food chains. *Ann. Rev. Eco. Sys.*, 19(1), 19–
543 38.
- 544 Flombaum, P., Wang, W.-L., Primeau, F. W., & Martiny, A. C. (2020). Global pi-
545 cophytoplankton niche partitioning predicts overall positive response to ocean
546 warming. *Nature Geoscience*, 13(2), 116–120.
- 547 Fuhrman, J. A. (2009). Microbial community structure and its functional implica-
548 tions. *Nature*, 459(7244), 193–199.
- 549 Graff, J., Westberry, T., Milligan, A., Brown, M., Dall’Olmo, G., Reifel, K., &
550 Behrenfeld, M. (2016). Photoacclimation of natural phytoplankton com-
551 munities. *Marine Ecology Progress Series*, 542.
- 552 Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., ... Gorsky,
553 G. (2016). Plankton networks driving carbon export in the oligotrophic ocean.
554 *Nature*, 532(7600), 465–470.
- 555 Guidi, L., Stemmann, L., Jackson, G. A., Ibanez, F., Claustre, H., Legendre, L.,
556 ... Gorsky, G. (2009). Effects of phytoplankton community on production,
557 size, and export of large aggregates: A world-ocean analysis. *Limnology and
558 Oceanography*, 54(6), 1951–1963.
- 559 Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*,
560 1(3), 297–310.
- 561 Holder, C., & Gnanadesikan, A. (2021). Can machine learning extract the mech-
562 anisms controlling phytoplankton growth from large-scale observations? – a
563 proof-of-concept study. *Biogeosciences*, 18, 1941–1970.
- 564 Hutchinson, G. E. (1961). The paradox of the plankton. *Amer. Naturalist*.
- 565 Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., ...
566 Zinger, L. (2019). Global trends in marine plankton diversity across kingdoms
567 of life. *Cell*, 179(5), 1084–1097.e21.

- 568 Irwin, A. J., Nelles, A. M., & Finkel, Z. V. (2012). Phytoplankton niches estimated
569 from field data. *Limnology and Oceanography*, *57*(3), 787–797.
- 570 Kuhn, A. M., Dutkiewicz, S., Jahn, O., Clayton, S., Rynearson, T. A., Mazloff,
571 M. R., & Barton, A. D. (2019). Temporal and spatial scales of correlation in
572 marine phytoplankton communities. *Journal of Geophysical Research: Oceans*,
573 *124*(12), 9417–9438.
- 574 Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., . . . Ap-
575 peltans, W. (2019). Globally consistent quantitative observations of planktonic
576 ecosystems. *Front. Mar. Sci.*, *6*.
- 577 Marinov, I., Doney, S. C., & Lima, I. D. (2010). Response of ocean phytoplankton
578 community structure to climate change over the 21st century: partitioning the
579 effects of nutrients, temperature and light. *Biogeosciences*, *7*(12), 3941–3959.
- 580 Marinov, I., Gnanadesikan, A., Sarmiento, J. L., Toggweiler, J. R., Follows, M., &
581 Mignone, B. K. (2008). Impact of oceanic circulation on biological carbon
582 storage in the ocean and atmospheric pCO₂. *Global Biogeochemical Cycles*,
583 *22*(3).
- 584 Martiny, A., & Flombaum, P. (2020). Global observations prochlorococcus, syne-
585 chococcus, and picoeukaryotic phytoplankton with ancillary environmental
586 data from 1987 to 2008.
- 587 Prinn, R. G. (2013). Development and application of earth system models. *PNAS*,
588 *110*, 3673–3680.
- 589 Righetti, D., Vogt, M., Gruber, N., Psomas, A., & Zimmermann, N. E. (2019).
590 Global pattern of phytoplankton diversity driven by temperature and environ-
591 mental variability. *Science Advances*, *5*(5), eaau6253.
- 592 Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2017). Data-driven dis-
593 covery of partial differential equations. *Science Advances*, *3*, e1602614.
- 594 Rutherford, S., D’Hondt, S., & Prell, W. (1999). Environmental controls on the geo-
595 graphic distribution of zooplankton diversity. *Nature*, *400*(6746), 749–753.
- 596 Servén, D., & Brummitt, C. (2018). pygam: Generalized additive models in python.
597 *Zenodo*.
- 598 Sokolov, A. (2005). The MIT integrated global system model (IGSM) version 2:
599 Model description and baseline evaluation. , 46.
- 600 Sonnewald, M., Dutkiewicz, S., Hill, C., & Forget, G. (2020). Elucidating ecological
601 complexity: Unsupervised learning determines global marine eco-provinces.
602 *Science Advances*, *6*(22), 2375-2548.
- 603 Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing depen-
604 dence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794.
- 605 Tang, W., & Cassar, N. (2019). Data-driven modeling of the distribution of di-
606 azotrophs in the global ocean. *Geophysical Research Letters*, *46*(21), 12258–
607 12269.
- 608 Tittensor, D. P., Mora, C., Jetz, W., Lotze, H. K., Ricard, D., Berghe, E. V., &
609 Worm, B. (2010). Global patterns and predictors of marine biodiversity across
610 taxa. *Nature*, *466*(7310), 1098–1101.
- 611 Ward, B. A., Dutkiewicz, S., & Follows, M. J. (2014). Modelling spatial and tem-
612 poral patterns in size-structured marine plankton communities: top-down and
613 bottom-up controls. *Journal of Plankton Research*, *36*(1), 31–47.
- 614 Wei, Q., & Dunbrack, R. (2013). The role of balanced training and testing data sets
615 for binary classifiers in bioinformatics. *PloS One*, *8*, e67863.