Geophysical Research Letters®

RESEARCH LETTER

10.1029/2021GL093455

Testing the Skill of a Species Distribution Model Using a 21st Century Virtual Ecosystem

L. R. Bardon[1,2], B. A. Ward[2], S. Dutkiewicz[3], and B. B. Cael[4]

[1]University of Southern California, Los Angeles, CA, USA, [2]University of Southampton, Southampton, UK, [3]Massachusetts Institute of Technology, Cambridge, MA, USA, [4]National Oceanography Centre, Southampton, UK

Key Points:

- We use a correlative species distribution model to predict the global plankton biogeography of a trait-based ecosystem model
- Predictive skill varies across test cases, with functional group, and spatiotemporally, with poor end-of-century performance
- Key sources of uncertainty are traced to sampling biases in observations, and the temporal variability in target-predictor relationships

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

L. R. Bardon and B. B. Cael,
lbardon@usc.edu;
cael@noc.ac.uk

Citation:

Bardon, L. R., Ward, B. A., Dutkiewicz, S., & Cael, B. B. (2021). Testing the skill of a species distribution model using a 21st century virtual ecosystem. *Geophysical Research Letters*, *48*, e2021GL093455. https://doi.org/10.1029/2021GL093455

Received 18 MAR 2021
Accepted 1 OCT 2021

Author Contributions:

Conceptualization: B. A. Ward, S. Dutkiewicz, B. B. Cael
Data curation: S. Dutkiewicz
Formal analysis: L. R. Bardon
Funding acquisition: B. A. Ward, S. Dutkiewicz
Investigation: L. R. Bardon
Methodology: L. R. Bardon, B. B. Cael
Resources: S. Dutkiewicz, B. B. Cael
Software: L. R. Bardon
Supervision: S. Dutkiewicz, B. B. Cael
Validation: L. R. Bardon
Visualization: L. R. Bardon

Abstract

Plankton communities play an important role in marine food webs, in biogeochemical cycling, and in Earth's climate; yet observations are sparse, and predictions of how they might respond to climate change vary. Correlative species distribution models (SDM's) have been applied to predicting biogeography based on relationships to observed environmental variables. To investigate sources of uncertainty, we use a correlative SDM to predict the plankton biogeography of a 21st century marine ecosystem model (Darwin). Darwin output is sampled to mimic historical ocean observations, and the SDM is trained using generalized additive models. We find that predictive skill varies across test cases, and between functional groups, with errors that are more attributable to spatiotemporal sampling bias than sample size. End-of-century predictions are poor, limited by changes in target-predictor relationships over time. Our findings illustrate the fundamental challenges faced by empirical models in using limited observational data to predict complex, dynamic systems.

Plain Language Summary

Marine plankton communities play a central role within 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. However, the oceans are vast, and samples are sparse, so global distributions are not well known. Statistical species distribution models (SDM's) have been developed that predict global distributions based on their relationships with observed environmental variables. They appear to perform well at summarizing present day distributions, and are increasingly being used to predict ecosystem changes throughout the 21st century. But it is not guaranteed that such models remain valid over time. Rather than wait 100 years to find out, we applied a statistical SDM to a complex virtual ocean, and trained it using virtual observations that match real-world ocean samples. This allows us to jump forward to the end-of-century to test the accuracy of our predictions. The SDM performed well at qualitatively predicting “present day” plankton distributions but yielded poor end-of-century predictions. Our case study emphasizes both the importance of environmental variable selection, and of changes in the underlying relationships between environmental variables and plankton distributions, in terms of model validity over time.

1. Introduction

Plankton underpin global ocean food webs and fisheries, mediate marine biogeochemical cycles, and affect climate (Falkowski et al., 2008; Fenchel, 1988; Guidi et al., 2016; Hutchinson, 1961; Marinov et al., 2008). Their global biogeography interacts with the ocean's inventory of nutrient elements, and its capacity to sequester CO_2 (Cermeño et al., 2008; Falkowski et al., 1998; Fuhrman, 2009; Guidi et al., 2009). 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 (Graff et al., 2016; Rutherford et al., 1999; Tittensor et al., 2010). 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.

Species distribution models (SDMs) (sometimes interchangeably referred to as ecological niche models) have been widely used to predict biogeographic distributions and fundamental niche parameters in terrestrial ecosystems, and have seen a recent surge of popularity in marine ecosystem context (Benedetti

Writing – original draft: L. R. Bardon, B. A. Ward, S. Dutkiewicz, B. B. Cael
Writing – review & editing: L. R. Bardon, B. A. Ward, S. Dutkiewicz, B. B. Cael

et al., 2021; Flombaum et al., 2020; Melo-Merino et al., 2020; Righetti et al., 2019). While mechanistic variants exist, the most popular implementations of SDM seek to identify the relationships between known geographic distributions of species' and sets of environmental variables. These relationships that are typically used by SDM developers to characterize biogeography in terms of where a species could, or could not, occur (Melo-Merino et al., 2020). Correlations are extracted using a variety of empirical methods, from classical statistics to bleeding-edge machine-learning (ML), or a hybridized ensemble thereof. For example, one might seek to characterize the relationships between measures of plankton concentrations (e.g., cell counts, gene markers or biomass) and simultaneously measured environmental factors (e.g., temperature, Chl-a, nutrient concentrations). The fitted model can then be used together with satellite or large synthesis database measurements to make diagnostic predictions of plankton. When the resulting SDM performs well relative to the measured data sets, predictions of species presence/absence or concentrations are then scaled globally (e.g., see Agusti et al., 2019; Barton et al., 2013; Irwin et al., 2012; Tang & Cassar, 2019).

However, a series of assumptions and uncertainties are incorporated into correlative SDMs, many of which go unchallenged or inadequately addressed by SDM developers. While an exhaustive overview of these assumptions and uncertainties is beyond the scope of the current work (see Wiens et al., 2009, for a thorough assessment), some are especially pertinent to marine microbial biogeography. For example, we cannot be certain that the environmental variables included in the model are a true and complete reflection of species' niche requirements', or whether some excluded or as-yet-unmeasured dimensions might better account for the observed distributions. Additionally, it is difficult to separate correlation from causation in such complex, dynamic and highly coupled systems. Our model might highlight sea surface temperature (SST) as the primary driver of abundance; yet it remains possible that separate factors coupled to SST—perhaps underwater solar radiation penetration or nutrient supply rates—are instead more directly linked to abundance. Thus, in this scenario, and adopting the terminology of Holder and Gnanadesikan (2021), the relationship between SST and abundance might be described as "apparent" while the relationship between underwater solar radiation and abundance as "intrinsic." This disconnect between cause and effect can be further complicated by trade-offs in the choice of empirical model used to build the SDM, see for example the inverse relationship between predictive skill and interpretability in machine-learning models (Carvalho et al., 2019).

There is a growing body of research that builds correlative SDMs with a variety of statistical and machine-learning models, and uses them to predict global plankton biogeography from sparse observational data, both in the present day, and many decades into the future (e.g., Benedetti et al., 2021; Flombaum et al., 2020; Ibarbalz et al., 2019; Righetti et al., 2019). Some of the results generated by such models have been novel and surprising, diverging significantly from other methodological approaches, such as trait-based mechanistic models (e.g., Cabré et al., 2015; Dutkiewicz et al., 2009, 2014; Ward et al., 2014). This is particularly true of predicting end-of-century distributions. For instance, the neural-network-derived correlative SDM developed in Flombaum et al. (2020) predicts an increase in picophytoplankton biomass in the future subtropical oceans, in direct contrast to mechanistic ecosystem models in Dutkiewicz et al. (2013) and Marinov et al. (2010). While it is not possible to comment on which particular modeling regime best approximates the real global oceans of 2100, identifying and describing potential sources of error would be nonetheless be beneficial for improving accuracy and guiding interpretation.

Here we set up an idealized testbed to assess the predictive skill of an SDM built on Generalized Additive Models (GAMs) (Hastie & Tibshirani, 1986) using the output from a mechanistic global scale ecosystem model, the "Darwin" model (Dutkiewicz et al., 2021), as a "ground truth." To assess the effect of known spatiotemporal biases in real-world observational data sets, we sample Darwin model outputs both randomly, and to mimic historical ocean measurements. The resulting SDM is evaluated in its ability to capture the virtual ocean's emergent biogeography in the present day "*spatial predictions*" and by the end-of-century "*temporal predictions*." Any predictions that diverge significantly from the ground-truthed virtual ocean are explored from the perspective of the assumptions and uncertainties inherent to SDM's, and of the more fundamental challenges inherent to all empirical models applied in similar contexts.

At the outset, we stress that our intention here is not to raise a false dichotomy whereby one particular methodological approach is pitted against another to decide a "winner." Nor are we making any claim as to the accuracy of the Darwin model in its ability to faithfully predict plankton abundance and diversity in the real ocean. Rather, the following case study is designed to assess how a correlative SDM might fare in predicting a complex but well-understood microbial ecosystem (see e.g., Dutkiewicz et al., 2020) embedded in a dynamic, self-consistent model of the Earth's ocean through time.

2. Materials and Methods

We performed a suite of tests using a widely applied implementation of GAMs (Servén & Brummitt, 2018) as our SDM and the Darwin model, a dynamic marine microbial ecosystem model coupled to an Earth system model (Dutkiewicz et al., 2021; Sokolov, 2005). Our decision to use GAMs as the empirical framework underlying our correlative SDM was informed by the work of Righetti et al. (2019), who demonstrated that GAMs perform comparably to Random Forest and Generalized Linear Models in a range of relevant predictive tasks, while offering a high degree of both interpretability and flexibility.

To train the GAMs, we sample the Darwin model at the same places and times as in a large ocean measurement data set used for similar purposes (Martiny & Flombaum, 2020). The resulting GAMs SDM is then used to predict Darwin model plankton biogeography. To quantify how spatiotemporal bias in the training data set affects predictive skill, we train an additional set of GAMs using a data set of the same size, but sampled uniformly randomly across the virtual ocean's surface, and uniformly randomly over the same period of time. To quantify the effect of training set sample size on predictive skill, we generate 54 additional random-sample training sets, in 18 different sample sizes. We evaluate the ability of the SDM to predict the global biogeography of the different plankton functional groups in the simulation, both during the 22-year period over which measurements were taken (i.e., spatial extrapolation), and during the last 22 years of the 21st century (i.e., both spatial and temporal extrapolation).

2.1. Numerical Model Simulation

The Darwin model ecosystem used here includes 51 plankton populations across 7 functional groups (2 prokaryotes (pro), 2 picoeukaryotes (pico), 5 coccolithophores (cocco), 5 diazotrophs (diazo), 11 diatoms (diatom), 10 mixotrophic dinoflagellates (dino), and 16 zooplankton (zoo)). It is described further in the Supporting Information S1, and in greater details in Dutkiewicz et al. (2020).

2.2. Ecosystem and Environmental Variables

Surface-level plankton abundance data and environmental parameters were extracted from Darwin simulation outputs, where surface in this context refers to the 10 m thick surface grid box. The ecosystem data contain 51 separate plankton biomasses, arranged into seven functional groups (as described above). A number of environmental variables have frequently been integrated into correlative SDMs to predict abundance and diversity, and have thus been included here. They are: sea surface temperature (SST), photosynthetically active radiation (PAR), phosphate (PO_4), nitrate (NO_3), silicate (Si), and iron (Fe). We sampled both the plankton abundance data and the environmental predictor variables from the 3,586 spatiotemporal cells that encompass the representative ocean measurement coordinates, and from the 3,586 randomly selected spatiotemporal cells. We sample the model output from the beginning of 1991 to the end of 2012 and consider this as a substitute to 1987–2008 (see Supporting Information S1). To validate predictions, we also consider whole-ocean surface data over the same period, and for the final 22 years of the simulation, from 2079 to 2100.

2.3. Building the Correlative SDM

We used the standard "LinearGAM" model of the freely available PyGAM package (Servén & Brummitt, 2018). LinearGAM incorporates a Gaussian distribution function with an identity link function, and fits predictor functions using penalized B-splines. These components impose smoothness to prevent overfitting, and enable the automatic fitting of nonlinear relationships. For an initial set of results, we set the number of permitted splines to 20 for each predictor variable. We note that our results are not sensitive to the choice of this parameter (see "Model Comparison & Sensitivity Tests" in Supporting Information S1). At the outset, we attempted to resolve and make predictions for individual plankton tracers, but the resulting models proved to be highly unstable, so we instead choose to proceed by summing the abundance data for each functional group, and training GAMs accordingly. The resulting partial dependency plots were examined for unexpected behaviors, or any clear indications of over or underfitting. The resulting GAMs SDM was then used to make predictions for the global surface ocean plankton biomasses during 1987–2008 and 2079–2100. Please see the Supporting Information S1 for details on model comparison and sensitivity tests.

2.4. Correlation Analyses

To accompany to SDM, we also performed a range of simpler correlation analyses. These act as a visual aid to better understand how these relationships might change in time and space. We first calculate the Pearson's Correlation Coefficient (ρ) for each functional group-predictor pair, and the Spearman's Rank Correlation Coefficient (ρ_s). Respectively, these popular methods detect the strength of linear associations between variables, and the strength of correlation in monotonic relationships. A commonly used method for addressing skew or capturing scaling relationships is the log-transform, which we apply to all data sets before recalculating ρ. Finally, we use the more recent distance correlations method of Székely et al. (2007). This technique captures the strength of both linear and nonlinear associations and avoids the need to make assumptions about variable distributions or linearity. We plot the correlation matrices for the main 3,586 cell test cases, both measurements-derived and randomly sampled, in 1987–2008, and at the same locations in 2079–2100. We explore the effect of sample size on the derived correlations by increasing the number of randomly sampled cells to 12,894, and finally to 25,683 cells.

3. Results

3.1. Spatial Predictions

We first describe the results of predicting plankton biogeography during the historical measurement period (1987—2008) (Figure 1). We find that predictive ability varies considerably across functional groups. There are fewer instances of our SDM incorrectly predicting presence (false positive) or absence (false negative) biomass for prokaryotes, picophytoplankton, and coccolithophores (16—19% of all location-month pairs) than for diatoms, diazotrophs, and dinoflagellates (26—31%), with zooplankton in between (21%). Where biomass is present and is predicted as such, the SDM's predictive ability for biomass concentration also varies substantially between functional groups (Figure 2); the SDM accounts for as much as 71% of the variance in biomass (diazotrophs) and as little as 41% (zooplankton). These patterns are reflected also in the mean relative differences and the balanced accuracy.

Patterns of overprediction of biomass occur across most of the oceans. For prokaryotes, picoeukaryotes, dinoflagellates, and zooplankton, this is especially evident in the Arctic (see Figures S1c, S2c, S5c, and S6c in Supporting Information S1). For these groups, we also see consistent underprediction in most of the Indian Ocean and in the Eastern Equatorial Pacific.

In general, the SDM shows a tendency to overestimate biomass ranging between 9% and 21% on average (picoeukaryotes and zooplankton, respectively), with a median overprediction of ≥16%. Despite this, there are some notable instances in the current context where the model performs well. Spatial predictions for coccolithophores, prokaryotes and diazotrophs all yield R^2 values that range between 0.62 and 0.71 (Figures 1e, S1e and S5e in Supporting Information S1). Diazotrophs fare particularly well in this regime, with a mean overprediction of 10%, an R^2 of 0.71, and the best visual, qualitative match of biogeography overall (although we note that the median overprediction in this case is a substantial 194%) (Figures S3c and S3e in Supporting Information S1). Overall, the SDM trained on data from historical measurement locations appear to be able to reproduce qualitative biogeographic patterns from spatial predictions well, but quantitative performance is variable, with a broad tendency toward overprediction. Notably, the greatest predictive errors more often occur in the undersampled regions of the ocean, such as the Arctic and Indian Oceans.

3.2. Temporal Predictions

The SDM's predictive ability is substantially reduced when extrapolating to the future ocean (see Figures 1 and 2). Rates of false positives and negatives in presence/absence do not uniformly change across functional groups: the cosmopolitan groups whose ranges expand poleward experience the least overall change,

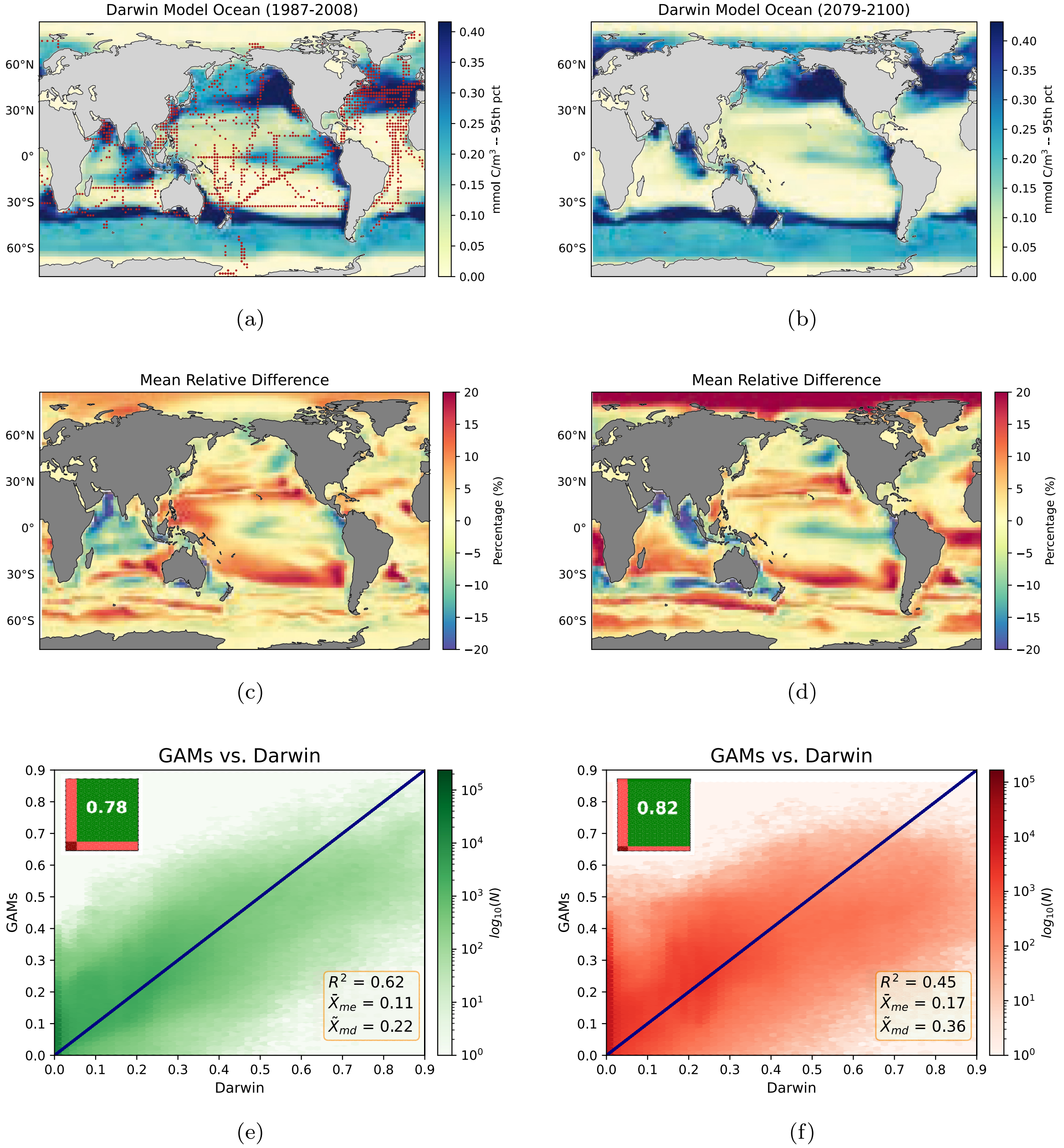

Figure 1. (a) Mean coccolithophore surface biomass (1987–2008) from the Darwin model. Red points indicate spatial location of training set datapoints, derived from ocean measurement data. (b) As per (a) for the years 2079–2100. (c) Relative (percent) difference between mean coccolithophore surface biomass from the Darwin model and the Generalized Additive Models Species Distribution Model (GAMs SDM) (1987–2008). (d) As per (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 SDM predictions versus 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/m^3) (green box), GAMs SDM 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 ($\bar{X}_{me}$ given as $(mean_{predicted} - mean_{actual})/mean_{actual}$), and relative difference of the medians ($\tilde{X}_{md}$ given as $(median_{predicted} - median_{actual})/median_{actual}$). (f) As per (e) but for 2079–2100. See Supporting Information S1 for other functional groups.

increasing by between 3% and 11% in prokaryotes, dinoflagellates, and coccolithophores, with a decrease of 5% for picophytoplankton. The SDM's ability to correctly predict presence/absence is further reduced for the groups with a more confined biogeography, increasing by between 14% and 23% for diazotrophs, zooplankton, and diatoms. We see a substantial increase in false negative occurrences for diatoms (to 29%), the group whose biogeographic range contracts most. Where biomass is present and is predicted as such, the SDM's

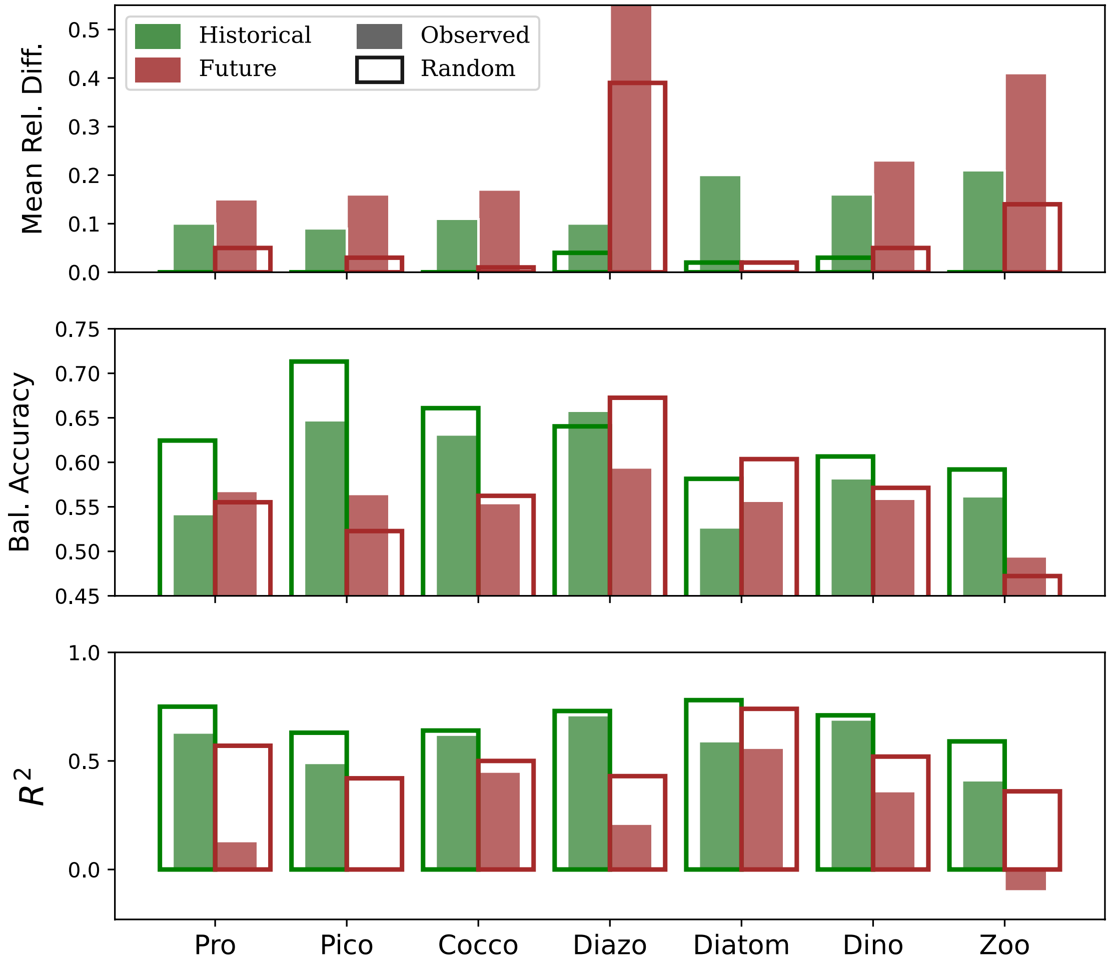

Figure 2. Comparing Darwin model "true" biomasses with Generalized Additive Models Species Distribution Models (GAMs SDM) predictions for each functional group in 1987–2008 (historical) and 2079–2100 (future), and from measurements-derived and randomly sampled training sets. *Top to Bottom:* (a) Relative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

predictive ability was reduced for all functional groups, with the fraction of variance accounted for by the SDM reducing by between 17% and 50%. The prediction for zooplankton is worse than just assuming a globally uniform constant biomass (i.e., $R^2 < 0$). We see a marked increased in mean relative differences compared to the "spatial" predictions, accompanied by a reduction in balanced accuracy for all groups besides diatoms (Figure 2). Diatoms are the only group for which the fraction of variance accounted for does not decrease substantially, only from $R^2 = 0.59$ to $R^2 = 0.56$ (Figure S4 in Supporting Information S1). Thus, the predictive ability for diatom biomass where it is present is not greatly reduced, despite the SDM's substantial overprediction of the contraction of diatoms' biogeography. This is not sensitive to varying the absence/presence cut-off value by an order or magnitude in either direction (Table S1 in Supporting Information S1).

Spatial patterns of prediction errors of coccolithophores, prokaryotes, picoeukaryotes, dinoflagellates, and zooplankton are largely similar to those for the historical period, except the North Atlantic is now underpredicted for all groups besides diazotrophs (Figures 1, S1, S2 and S4–S6 in Supporting Information S1). Diatom biomass is notably underpredicted in the Southern Ocean and Northern Atlantic (Figure S4 in Supporting Information S1). Meanwhile, diazotroph biomass is notably overpredicted throughout the Atlantic Ocean, the Arctic, bands of the subtropical Pacific and Indian Ocean (Figure S3 in Supporting Information S1). Excluding diatoms, the overall tendency toward overprediction is exacerbated for all groups.

3.3. Model Trained on Randomized Locations

Here we compared the above results with those produced when the GAMs SDM was trained on randomly sampled data sets (Figure 2). Interestingly, the broad spatial patterns of where overprediction and underprediction occurs do not change much when training the SDM on randomly distributed data (Figures S8 and S9 in Supporting Information S1). Nonetheless, predictive abilities increase, biases are reduced, and balanced accuracy increases in both the spatial and temporal cases (Figure 2). The fraction of variance accounted for by the SDM increases by 2–19% when using random data to predict historical biogeography, but increase from 5% to 46% when using random data to predict future biogeography. The magnitude of the biases also decreases—average biases are within 3–4% in the historical case using random data. The median bias for all groups is still that of overprediction, with most groups in the range of ≥17% compared to ≥30% for measurements-derived predictions. Diatoms and diazotrophs have a markedly higher bias in both measurements-derived and random cases, of ≥194% and ≥162%, and ≥65% and ≥35%. In the future case, using random data reduces biases for all groups, though does not eliminate them. We also found that the predictive ability of the SDM was only weakly dependent on sample size (where sample size here refers to the number of grid cell-month pairs that are sampled), with predictive ability appearing to plateau with increasing sample size (Figure S14 in Supporting Information S1).

The results using random training data sets suggest that historical measurement biases reduce the predictive ability of the SDM more than the sample size of the training data set. Predictive ability can be improved by subsampling or weighting one's training data set to reduce spatiotemporal biases, although the coarse resolution of the Darwin model—and thus reduced variability as a result of correlated observations—relative to the real ocean may contribute to this plateauing effect.

4. Discussion

Broadly, our SDM captures large-scale spatial patterns of plankton biogeography, but struggles to make robust quantitative predictions, particularly when the model is trained on historical ocean measurement data. The emergent relationships between predictor variables and plankton abundances change spatially, seasonally and over the longer term, as demonstrated both by the variable nature of the partial dependence plots (Figures 3a, 3b, S10 and S11 in Supporting Information S1), and by the change in correlation strengths (Figures 3c–3f and S12 in Supporting Information S1). The latter offer a particularly powerful illustration of the changes in apparent relationships between biomass and environmental predictors in the measurements-derived sample space, assessed over the same period of time one hundred years into the future (Figures 3c and 3d). It is important to note that we should expect these differences to be exaggerated in the real world, where the system is significantly more complex.

Our results also demonstrate how spatiotemporal sampling bias can significantly alter the patterns of apparent relationships between environmental predictors and plankton biomass. The association strengths identified in the measurements-derived sample vary considerably from those found in the random sample of equivalent size (see Figure 3c versus Figure 3e). This finding is robust across a range of sample sizes, where almost identical patterns of correlations are seen in the 3,586 cell case as in the 25,683 cell case, as well as across several methods of deriving correlations (see Figure S12 in Supporting Information S1). Nonetheless, the spatial patterns of over and underprediction are not merely the result of spatiotemporal measurement biases. We see general agreement in these broad qualitative patterns between the predictions generated from measurements-derived and random samples (Figures 1c, 1d and S1–S6, S8, S9 in Supporting Information S1). Ocean measurement biases may explain some element of the tendency toward overestimation of historical biogeography/abundances; perhaps because measurements have more often been made in places with higher than average abundances. In all cases, training the statistical model on a nonbiased data set reduces the severity of over and underprediction, especially for spatial predictions (Figure S8e and S9e in Supporting Information S1). But the same broad biogeographic patterns remain, indicating that the SDM is failing to effectively capture changes over time, despite its relatively robust performance according to the broad-brush strokes of summary statistics (Figures S4e and S4f in Supporting Information S1).

The fraction of variance that the SDM can account for saturates with sample size well below 100% (see Figure S14 in Supporting Information S1), perhaps implying a potential ceiling on predictive ability. Nonetheless,

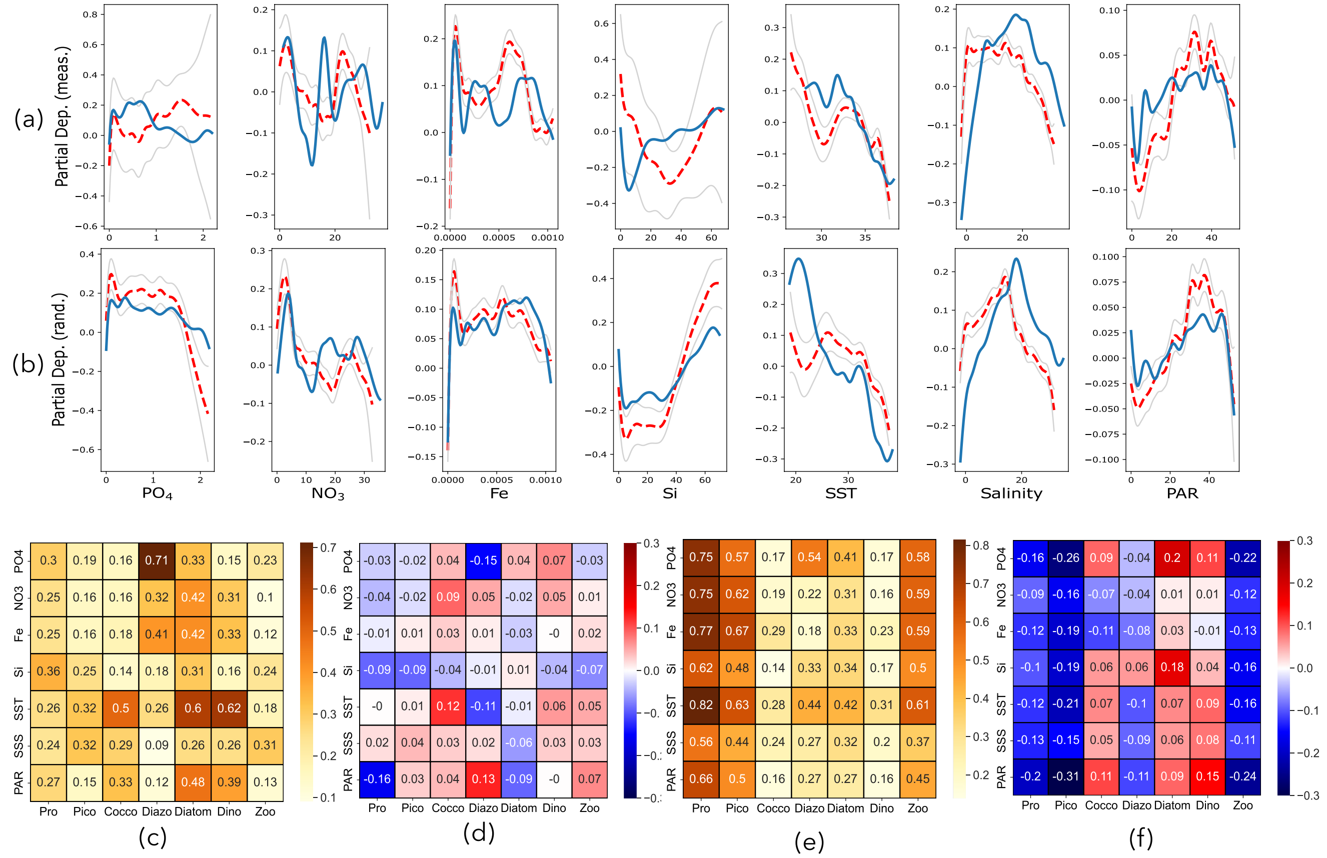

Figure 3. *Changing Relationships*: (a) Partial dependence plots of coccolithophore biomass (mmol C/m^3) as a function of each predictor, centered around the median (PO_4, NO_3, Fe, Si in mmol X/m^3, sea surface temperature (SST) in °*C*, SSS in PSU, photosynthetically active radiation (PAR) in E/m^2/day). Plotted using data from 3,586 Darwin surface ocean cells at measurements-derived locations spanning 1987–2008 (dashed red line) and at the same locations from 2079 to 2100 (blue line). Gray lines indicate 95% confidence interval for the 1987–2008 case. (b) As per (a), but using data from 3,586 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 (b) and those found at the same locations from 2079 to 2100. (e, f) As per (c, d), but for the equivalently sized, randomly sampled training set.

a number of optimizations could be implemented to improve predictive skill. First, we note that an unrepresentative training set presence/absence ratio compared to the population can lead to an unreliable representation of presence/absence in the resulting predictions. To avoid this possibility, researchers working with real observational data will sometimes employ resampling techniques (e.g., Wei & Dunbrack, 2013). By contrast, our experimental design permits us to test our outcomes alongside a range of representative, randomly sampled data sets spanning the surface ocean. These unbiased samples *are* representative of the presence/absence ratios of the population, and thus act as a control for our observations-derived test case. Given the broadly similar patterns of over and underprediction found across test cases, we do not employ resampling techniques here.

Related also to the more flexible nature of our study in comparison to correlative SDMs built from real-world observations, is the manner in which we approach training, validation and testing data sets. Here, we use whole-ocean Darwin model output as our test set for evaluating overall performance. Given model response to sensitivity tests, and GAM's natural robustness to overfitting as a result of predictor function regularization, we do not explicitly employ a validation set. Model skill could be improved with parameter fine-tuning, especially in the spatial predictions test case. But it is less clear whether fine-tuning for performance in the Darwin model ocean of 1987–2008 would improve end-of-century predictions. Additionally, we speculate that our decision to train the GAMs SDM using the entire measurements-derived sample might itself yield improvements relative to splitting the samples into training, testing and validation subsamples.

The median overestimations of the GAMs SDM compared to the Darwin "ground truth," even when using randomly sampled training data, also implies that these predicted abundance distributions are less skewed than the Darwin model distributions, which are, in turn, less skewed than distributions in the real ocean. That is not to say, however, that all correlative SDMs will yield equivalent outcomes, regardless of the empirical models at their cores. Recent work by Rudy et al. (2017) demonstrates that empirical methods can reliably extract the underlying mechanistic equations that govern a dynamical system. Similarly, Holder and Gnanadesikan (2021) evaluate random forest (RF) and neural-network ensembles (NNE) in their ability to resolve the underlying intrinsic relationships between plankton biomass and environmental predictors, from the apparent relationships in the data. They demonstrate variability in predictive skill across different empirical test cases, and find that NNE's yield overall superior performance; particularly in the case where plankton growth rates respond rapidly to environmental change, as might be expected in many real-world ocean environments. These hybrid methods represent a potential step toward building more skillful and descriptive models.

Although improvements to overall predictive skill might be made, the assumptions and uncertainties inherent to correlative SDMs may apply more fundamental constraints. For instance, questions still remain as to whether the environmental data included in the model reflect the true niche requirements of the target species'. In addition, using environmental correlates of distribution to predict abundance elsewhere in space and time implies that the distributions in the training data are at equilibrium, which may not be the case.

Empirical methods that extract the intrinsic drivers of plankton abundance and distribution (as derived in laboratory settings) might also yield improvements to the predictive capabilities of correlative SDMs; particularly if factors such as spatiotemporal sampling bias and spatial autocorrelation in ocean measurements can also be accounted for. However, this would not guarantee improvements to multidecadel predictions of how plankton communities might respond to climate change; we cannot assume that a specie's niche envelope is fixed and immutable over time, as there are very many degrees of freedom and coupling in real-world interactions between plankton individuals, communities, and the wider ecosystem and environment. In addition to the controlling influence of e.g., nutrient supply rate, physical transport processes and level of top-down pressure, plankton are also able to adapt genetically, epigenetically and plastically to change. With their short generation times and high biodiversity, we might expect that even intrinsic relationships could change over the course of a century. This is especially likely in such a dynamic, randomly perturbed, and far-from-equilibrium environment, where conditions are ideal for unpredictable emergent phenomena to arise. By contrast, all such elements within the Darwin Model are simplified by design, and intrinsic relationships are held steady over time, such that the spatiotemporal variability in apparent relationships seen here are the product of many fewer sources of complexity, right down to how climate change proceeds (a known quantity in the Darwin Model, and yet another significant source of uncertainty in the real world).

We focus here on deriving our SDM using a statistical learning model that, for reasons outlined in Section 2, we believe makes for an excellent case study. Our investigation has allowed us to better clarify the strengths and limitations of such an approach, as applied in the current context. Owing to the complexity and ever-changing nature of the system, some of these limitations could be fundamental and unavoidable, particularly when extrapolating far beyond the training regime.

Methodologically, the broader approach we have presented of applying an empirical model to output from a numerical model may be useful for addressing a number of additional questions. These might include evaluating how best to empirically model whole-ecosystem properties, such as diversity, from observations, or assessing where and when to make new observations to maximize information content about global plankton biogeography. But, as our results here have demonstrated and reinforced, it is important to be aware of the strengths and limitations of this approach, especially when dealing with a high degree of complexity over time.

5. Conclusion

In summary, our results suggest that correlative SDMs like the one developed 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, their predictions are especially sensitive to the spatiotemporal bias in historical measurements, and can tend toward overprediction if not properly accounted for. In addition, such models demonstrably struggle to predict future plankton biomass because the spatial and temporal complexity of the physical, chemical and biological interactions that characterize the system give rise to a variability that cannot be accurately predicted decades ahead of time from correlations in contemporary data. The changes in relationship between environmental variables and the plankton abundances demonstrated in the current work could be greatly exaggerated in correlative SDMs that tackle the significantly more complex task of predicting real-world plankton biogeography using sparse observational data.

Data Availability Statement

The physical model used in the Darwin simulation is available at http://mitgcm.org. The generic ecosystem code is available at https://gitlab.com/jahn/gud, while the equations and documentation can be found at https://darwin3.readthedocs.io/en/latest/phys_pkgs/darwin.html. The specific modifications for the setup of the Darwin model used here, and all parameter values are available at https://doi.org/10.7910/DVN/U. The code used to process and analyze the Darwin output data, and to generate the current results, is available at https://github.com/leebardon/stats-biogeo-2021. The Darwin model output used in the current study is available at https://dataverse.harvard.edu/dataverse/gud-igsm; in particular, the biomass of the functional groups of plankton at https://doi.org/10.7910/DVN/RPL6PT; and the environmental variables at https://doi.org/10.7910/DVN/LQH9PX (Dutkiewicz, 2021a, 2021b). A collection of preprocessed Darwin output data, for use with the codebase at https://github.com/leebardon/stats-biogeo-2021, can be found at https://doi.org/10.7910/DVN/DT7POF (Bardon, 2021).

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 No. 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.

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. *Frontiers in Marine Science*, *5*, 506. https://doi.org/10.3389/fmars.2018.00506

Bardon, L. (2021). *Processed gud igsm surface ecosystem and environmental data (1991–2012; 2079–2100)*. Harvard Dataverse. https://doi.org/10.7910/DVN/DT7POF

Barton, A. D., Pershing, A. J., Litchman, E., Record, N. R., Edwards, K. F., Finkel, Z. V., et al. (2013). The biogeography of marine plankton traits. *Ecology Letters*, *16*(4), 522–534. https://doi.org/10.1111/ele.12063

Benedetti, F., Vogt, M., Elizondo, U., Righetti, D., Zimmermann, N. E., & Gruber, N. (2021). Major restructuring of marine plankton assemblages under global warming. *Nature Communications*, *12*, 5226. https://doi.org/10.1038/s41467-021-25385-x

Cabré, A., Marinov, I., & Leung, S. (2015). Consistent global responses of marine ecosystems to future climate change across the IPCC AR5 earth system models. *Climate Dynamics*, *45*(5), 1253–1280. https://doi.org/10.1007/s00382-014-2374-3

Carvalho, D. V., Pereira, E. M., & Cardosa, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*, 832. https://doi.org/10.3390/electronics8080832

Cermeño, P., Dutkiewicz, S., Harris, R. P., Follows, M., Schofield, O., & Falkowski, P. G. (2008). The role of nutricline depth in regulating the ocean carbon cycle. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(51), 20344–20349. https://doi.org/10.1073/pnas.0811302106

Dutkiewicz, S. (2021a). *Gud igsm monthly mean plankton group surface biomass*. Harvard Dataverse. https://doi.org/10.7910/DVN/RPL6PT

Dutkiewicz, S. (2021b). *Gud igsm monthly surface environmental factors*. Harvard Dataverse. https://doi.org/10.7910/DVN/LQH9PX

Dutkiewicz, S., Boyd, P. W., & Riebesell, U. (2021). Exploring biogeochemical and ecological redundancy in phytoplankton communities in the global ocean. *Global Change Biology*, *27*(6), 1196–1213. https://doi.org/10.1111/gcb.15493

Dutkiewicz, S., Cermeno, P., Jahn, O., Follows, M. J., Hickman, A. E., Taniguchi, D. A. A., & Ward, B. A. (2020). Dimensions of marine phytoplankton diversity. *Biogeosciences*, *17*(3), 609–634. https://doi.org/10.5194/bg-17-609-2020

Dutkiewicz, S., Follows, M. J., & Bragg, J. G. (2009). Modeling the coupling of ocean ecology and biogeochemistry. *Global Biogeochemical Cycles*, *23*, GB4017. https://doi.org/10.1029/2008GB003405

Dutkiewicz, S., Scott, J. R., & Follows, M. J. (2013). Winners and losers: Ecological and biogeochemical changes in a warming ocean. *Global Biogeochemical Cycles*, *27*, 463–477. https://doi.org/10.1002/gbc.20042

Dutkiewicz, S., Ward, B. A., Scott, J. R., & Follows, M. J. (2014). Understanding predicted shifts in diazotroph biogeography using resource competition theory. *Biogeosciences*, *11*(19), 5445–5461. https://doi.org/10.5194/bg-11-5445-2014

Falkowski, P. G., Barber, R. T., & Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science*, *281*(5374), 200–206. https://doi.org/10.1126/science.281.5374.200

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science*, *320*(5879), 1034–1039. https://doi.org/10.1126/science.1153213

Fenchel, T. (1988). Marine plankton food chains. *Annual Review of Ecology and Systematics*, *19*(1), 19–38. https://doi.org/10.1146/annurev.es.19.110188.000315

Flombaum, P., Wang, W.-L., Primeau, F. W., & Martiny, A. C. (2020). Global picophytoplankton niche partitioning predicts overall positive response to ocean warming. *Nature Geoscience*, *13*(2), 116–120. https://doi.org/10.1038/s41561-019-0524-2

Fuhrman, J. A. (2009). Microbial community structure and its functional implications. *Nature*, *459*(7244), 193–199. https://doi.org/10.1038/nature08058

Graff, J., Westberry, T., Milligan, A., Brown, M., Dall'Olmo, G., Reifel, K., & Behrenfeld, M. (2016). Photoacclimation of natural phytoplankton communities. *Marine Ecology Progress Series*, *542*, 51–62. https://doi.org/10.3354/meps11539

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., et al. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*(7600), 465–470. https://doi.org/10.1038/nature16942

Guidi, L., Stemmann, L., Jackson, G. A., Ibanez, F., Claustre, H., Legendre, L., et al. (2009). Effects of phytoplankton community on production, size, and export of large aggregates: A world-ocean analysis. *Limnology & Oceanography*, *54*(6), 1951–1963. https://doi.org/10.4319/lo.2009.54.6.1951

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297–310. https://doi.org/10.1214/ss/1177013604

Holder, C., & Gnanadesikan, A. (2021). Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations?—A proof-of-concept study. *Biogeosciences*, *18*, 1941–1970. https://doi.org/10.5194/bg-18-1941-2021

Hutchinson, G. E. (1961). *The paradox of the plankton*. The American Naturalist.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., et al. (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell*, *179*(5), 1084–1097. https://doi.org/10.1016/j.cell.2019.10.008

Irwin, A. J., Nelles, A. M., & Finkel, Z. V. (2012). Phytoplankton niches estimated from field data. *Limnology & Oceanography*, *57*(3), 787–797. https://doi.org/10.4319/lo.2012.57.3.0787

Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., & Appeltans, W. (2019). Globally consistent quantitative observations of planktonic ecosystems. *Frontiers in Marine Science*, *6*, 196. https://doi.org/10.3389/fmars.2019.00196

Marinov, I., Doney, S. C., & Lima, I. D. (2010). Response of ocean phytoplankton community structure to climate change over the 21st century: Partitioning the effects of nutrients, temperature and light. *Biogeosciences*, *7*(12), 3941–3959. https://doi.org/10.5194/bg-7-3941-2010

Marinov, I., Gnanadesikan, A., Sarmiento, J. L., Toggweiler, J. R., Follows, M., & Mignone, B. K. (2008). Impact of oceanic circulation on biological carbon storage in the ocean and atmospheric pCO2. *Global Biogeochemical Cycles*, *22*, GB3007. https://doi.org/10.1029/2007GB002958

Martiny, A., & Flombaum, P. (2020). *Global observations prochlorococcus, synechococcus, and picoeukaryotic phytoplankton with ancillary environmental data from 1987 to 2008*. https://doi.org/10.1575/1912/bco-dmo.793451.1

Melo-Merino, S. M., Reyes-Bonilla, H., & Lira-Noriega, A. (2020). Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. *Ecological Modelling*, *415*, 108837. https://doi.org/10.1016/j.ecolmodel.2019.108837

Righetti, D., Vogt, M., Gruber, N., Psomas, A., & Zimmermann, N. E. (2019). Global pattern of phytoplankton diversity driven by temperature and environmental variability. *Science Advances*, *5*(5), eaau6253. https://doi.org/10.1126/sciadv.aau6253

Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2017). Data-driven discovery of partial differential equations. *Science Advances*, *3*, e1602614. https://doi.org/10.1126/sciadv.1602614

Rutherford, S., D'Hondt, S., & Prell, W. (1999). Environmental controls on the geographic distribution of zooplankton diversity. *Nature*, *400*(6746), 749–753. https://doi.org/10.1038/23449

Servén, D., & Brummitt, C. (2018). pygam: Generalized additive models in python. *Zendo*.

Sokolov, A., Schlosser, C. A., Dutkiewicz, S., Paltsev, S., Kicklighter, D. W., Jacoby, H. D., et al. (2005). *The MIT integrated global system model (IGSM) version 2: Model description and baseline evaluation. Joint Program Report Series* (Rep. 124, p. 46). Retrieved from http://globalchange.mit.edu/publication/14579

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, *35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505

Tang, W., & Cassar, N. (2019). Data-driven modeling of the distribution of diazotrophs in the global ocean. *Geophysical Research Letters*, *46*, 12258–12269. https://doi.org/10.1029/2019GL084376

Tittensor, D. P., Mora, C., Jetz, W., Lotze, H. K., Ricard, D., Berghe, E. V., & Worm, B. (2010). Global patterns and predictors of marine biodiversity across taxa. *Nature*, *466*(7310), 1098–1101. https://doi.org/10.1038/nature09329

Ward, B. A., Dutkiewicz, S., & Follows, M. J. (2014). Modelling spatial and temporal patterns in size-structured marine plankton communities: Top–down and bottom–up controls. *Journal of Plankton Research*, *36*(1), 31–47. https://doi.org/10.1093/plankt/fbt097

Wei, Q., & Dunbrack, R. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One*, *8*, e67863. https://doi.org/10.1371/journal.pone.0067863

Wiens, J. A., Stralberg, D., Jongsomjit, D., Howell, C. A., & Snyder, M. (2009). Niches, models, and climate change: Assessing the assumptions and uncertainties. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 19729–19736. https://doi.org/10.1073/pnas.0901639106