How predictable is plankton biogeography using statistical learning methods?

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

[1]University of Southampton, UK
[2]Massachusetts Institute of Technology, Cambridge, MA, USA
[3]National Oceanography Centre, Southampton, UK

Key Points:

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

Corresponding author: Lee Bardon, `l.r.bardon@soton.ac.uk`

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 CO_2 (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

follow (Holder & Gnanadesikan, 2021), in referring to underlying mechanistic relationships as 'intrinsic', and the emergent correlations between variables in the data as 'apparent'. In the context of predicting plankton biogeography, statistical 'niche models' might seek to extract the apparent 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). These relationships can then be used together with satellite or large synthesis database measurements to make diagnostic predictions of plankton abundance. The sparse data are typically separated into a training dataset for model development and a testing dataset to evaluate performance. When the statistical models perform well relative to the measured datasets, predictions of species presence/absence or concentrations can then be scaled globally (e.g. (Tang & Cassar, 2019; Barton et al., 2013; Irwin et al., 2012; Agusti et al., 2019)).

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

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

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

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

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

2 Materials & Methods

We performed a suite of tests using a widely applied implementation of GAMs (Servén & Brummitt, 2018) and the 'Darwin Model', a dynamic marine microbial ecosystem model coupled to an Earth system model ((Dutkiewicz et al., 2021), (Sokolov, 2005)). To train the GAMs, we sample the Darwin model at the same places and times as in a large ocean measurement dataset used for similar purposes (Martiny & Flombaum, 2020). The resulting GAMs are then used to predict Darwin Model plankton biogeography. To quantify how spatiotemporal bias in the training dataset affects the GAMs' predictive skill, we train an additional set of GAMs using a dataset of the same size, but sampled uniformly randomly across the ocean's surface, and uniformly randomly over the same period of time. To quantify the effect of training set sample size on the GAMs' predictive skill, we generate 54 additional random-sample training sets, in 18 different sample sizes. We evaluate the ability of the GAMs 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 pico-eukaryotes (pico), 5 coccolithophores (cocco), 5 diazotrophs (diazo), 11 diatoms (diatom), 10 mixotrophic dinoflagellates (dino) and 16 zooplankton (zoo)). Individual populations correspond to different size classes within functional groups, with all size classes covering a range of $0.6-2425$ μm equivalent spherical diameter. Functional groups have distinct allometric relationships for growth, grazing, and sinking parameters (see (Dutkiewicz et al., 2020)). The model ecosystem is embedded within the Massachusetts Institute of Technology Integrated Global System Model (IGSM) (Prinn, 2013; Sokolov, 2005) which includes modules for the physics, chemistry, and biogeochemistry of the atmosphere, land and ocean. The ocean component has a $2^\circ \times 2.5^\circ$ resolution grid and 22 vertical layers (10m thickness at surface to 500m at bottom). The simulation is forced with observed greenhouse gas emissions from $1860-1990$ and then with a high emissions scenario that is analogous to the IPCC's Representative Concentration Pathway 8.5, from $1990-2110$. This perturbation results in $\sim 3^\circ$C sea surface temperature warming by 2100, sea ice retreat, increased stratification, and an altered overturning circulation. The IGSM has been used to examine changes in marine biogeochemstry and ecology in previous studies (Dutkiewicz et al. 2013; 2019) but with a simpler version of the ecosystem model. The current more complex ecosystem has also been used in previous studies, but only for the current day's ocean (Dutkiewicz et al., 2021; Sonnewald et al., 2020; Kuhn et al., 2019). This model and previous model validation for the current day demonstrates the output compares well with observations along both axes of size and functional type (e.g. (Dutkiewicz et al., 2021, 2020)).

2.2 Ecosystem and Environmental Variables

Surface-level plankton abundance data and environmental parameters were extracted from the Darwin model simulation output, where surface in this context refers to the $10m$ thick surface grid box. The ecosystem data contains 51 separate plankton biomasses, arranged into seven functional groups (as described above). A number of environmental variables have been used by statistical models 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 3586 spatiotemporal cells that encompass the representative ocean measurement coordinates, and from the 3586 randomly selected spatiotemporal cells. Note that the model simulation used for the current analysis nominally starts in 1991 and extends to 2100. As such, we sample the model output from the beginning of 1991 to the end of 2012 and consider this as a substitute to $1987-2008$ in this context. This is justified because the dynamic model's internal variability does not match real-world interannual variability in terms of timing, though does capture the magnitudes (e.g. there are El Niño events, but these do not occur in the same years as the real ocean). To validate predictions, we also consider whole-ocean surface data over the same period, and for the final 22 years of the simulation, from $2079 - 2100$.

2.3 Training the Statistical Learning Model

A variety of statistical learning algorithms have already been applied to ocean measurement data, and used to make predictions about the future state of the ocean microbiome (see e.g. (Righetti et al., 2019; Flombaum et al., 2020)). Indeed, the methods and results of (Righetti et al., 2019) act both as a guide to the current work, and as a contributing factor in our decision to use GAMs as our 'representative' statistical learning method. This is due to the (Righetti et al., 2019) finding that GAMs perform comparably to Random Forest and Generalised Linear Models in a range of predictive tasks, while offering a high degree of both interpretability and flexibility. Additionally, GAMs are of intermediate complexity between classical statistical regression models, and more sophisticated machine learning methods, making them both accessible and potentially attractive to a wide range of researchers.

Here, we used the standard 'LinearGAM' model of the freely available PyGAM package (Servén & Brummitt, 2018), incorporating a Gaussian distribution function with an identity link function. Feature functions are built using penalised B-splines that impose smoothness to avoid over-fitting, and enable the automatic fitting of nonlinear relationships, while maintaining additivity. 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'). Rather than attempt to resolve and make predictions for individual plankton tracers, we instead sum the abundance data for each functional group, and train GAMs accordingly. The resulting partial dependency plots are examined for unexpected behaviours, or any clear indications of over or under-fitting. We thus use the relationships identified by the GAMs to make predictions for the global surface ocean plankton biomasses during 1987-2008 and 2079-2100.

2.4 Model Comparison & Sensitivity Tests

We define presence/absence as modelled biomass being above/below a cutoff threshold (10^{-5} mmol C/m^3), but find that patterns in the resulting predictions are not sensitive to the choice of this threshold (Table S4).

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

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

We also perform a range of simpler correlation analyses, to build a broader picture of the emergent relationships between functional group biomass and predictors. These act as a visual aid to better understand how these relationships might change in time and space, and as a basic cross-reference for GAMs-derived partial dependence plots of the training sets. 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 datasets before recalculating ρ. However, this method of broadly applying a single transformation is not optimal. A more robust approach would be to examine the distribution of each target-predictor relationship individually, before an appropriate transformation is selected. Nonetheless, even this more optimal method runs the risk propagating transformation uncertainty into the resulting confidence interval.

With these limitations in mind, we also determine correlations using 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 3586 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 GAMs incor-

rectly 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, GAMs' predictive ability for biomass concentration also varies substantially between functional groups (Figure 2); the GAMs account 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 occurs across most of the oceans. For prokaryotes, picoeukaryotes, dinoflagellates and zooplankton, this is especially evident in the Arctic (see Figures (c) of S1, S2, S5, S6). For these groups, we also see consistent underprediction in most of the Indian Ocean and in the Eastern Equatorial Pacific. Meanwhile, diatoms are substantially overpredicted in most of the mid- and high-latitudes in the Northern Hemisphere but perform relatively well in the subtropics (Figure S4(c)). Diazotrophs yield the best overall performance, with only a small amount of overprediction in the subtropical Atlantic, and overprediction in the transition zone latitudes poleward of the subtropics (Figure S3(c)).

In general the GAMs show a tendency to overestimate biomass in the spatial predictions regime. Overestimation ranges between 9−21% 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 GAMs perform well. Spatial predictions for coccolithophores, prokaryotes and diazotrophs all yield R^2 values that range between 0.62 and 0.71 (Figures 1(e), S1(e), S5(e)). 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 S3(c) and S3(e)). On the whole, GAMs 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 towards overprediction. Notably, the greatest predictive errors more often occur in the undersampled regions of the ocean, such as the Arctic and Indian Oceans, but are by no means confined to these regions. For instance in the highly sampled North Atlantic predictions for diatoms and diazotrophs was also poor.

3.2 Temporal Predictions

GAMs' 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, increasing by between 3% and 11% in prokaryotes, dinoflagellates and coccolithophores, with a decrease of 5% for picophytoplankton. GAMs' 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, GAMs' predictive ability was reduced for all functional groups. In most cases, this reduction is substantial, with the fraction of variance accounted for by the GAMs reducing by between 17 and 50%, such that 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). Thus, the predictive ability for diatom biomass where it is present is not greatly reduced, despite the GAMs 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).

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, S4, S5, S6). Diatom biomass is notably underpredicted in the Southern Ocean and Northern Atlantic (Figure S4). Meanwhile, diazotroph biomass is notably overpredicted throughout the Atlantic Ocean, the Arctic, bands of the subtropical Pacific and Indian Ocean (Figure S3). Excluding diatoms, the overall tendency towards overprediction is exacerbated for all groups, increasing by 57% for prokaryotes, picoeukaryotes, coccolithophores, and dinoflagellates, by 20% for zooplankton, and by 49% for diazotrophs. Median overpredictions also increase for all groups besides diatoms.

3.3 Model Trained on Randomised Locations

Here we compared the above results with those produced when the GAMs were trained on randomly sampled datasets (Figure 2). Interestingly, the broad spatial patterns of where overprediction and underprediction occurs do not change much when training GAMs on randomly distributed data, as opposed to the ocean observation locations (Figures S8 and S9). 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 GAMs increases by 2−19% when using random data to predict historical biogeography, but increase from 5−46% when using random data to predict future biogeography. The most notable differences are for prokaryotic, picoeukaryotic, and zooplankton biomass in the future case. 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 $\geq$17% compared to $\geq$30% for measurements-derived predictions. Diatoms and diazotrophs have a markedly higher bias in both measurements-derived and random cases, of $\geq$194% and $\geq$162%, and $\geq$65% and $\geq$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 GAMs was only weakly dependent on sample size (where sample size here refers to the number of grid cell-month pairs that are sampled)(Figure 4), with predictive ability appearing to plateau with increasing sample size.

The results using random training datasets suggest that historical measurement biases reduce the predictive ability of GAMs more than the sample size of the training dataset. Predictive ability can be improved by subsampling or weighting one's training dataset to reduce biases in space and time, 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.

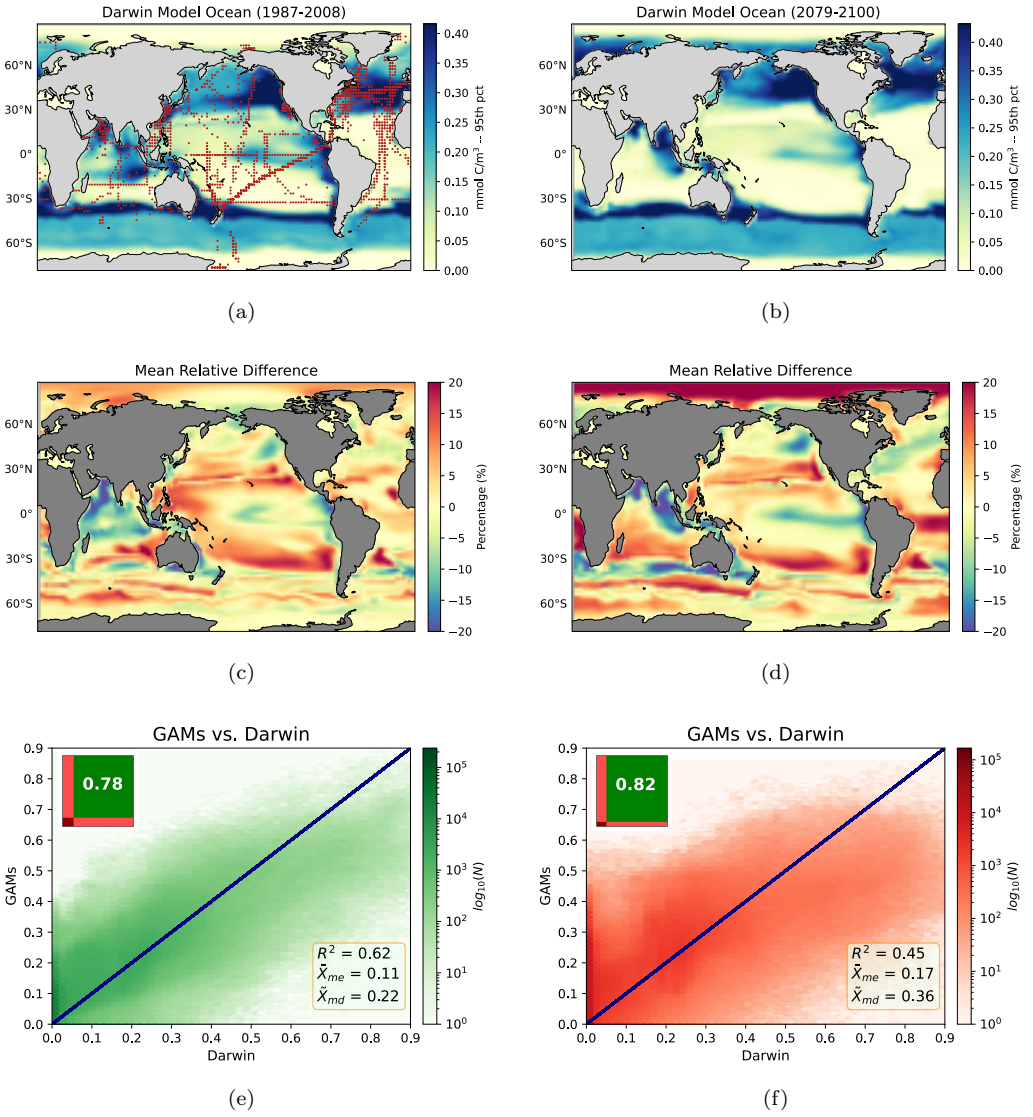

Figure 1: **(a)** Mean coccolithophore surface biomass (1987 - 2008) from the Darwin model. Red points indicate spatial location of training set datapoints, derived from ocean measurement data. **(b)** As per 1(a) for the years 2079 - 2100. **(c)** Relative (percent) difference between mean diatom surface 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/m^3)(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 ($\tilde{X}_{md}$). **(f)** As per 1(e) but for 2079-2100. See Supplemental Materials for other functional groups.

354

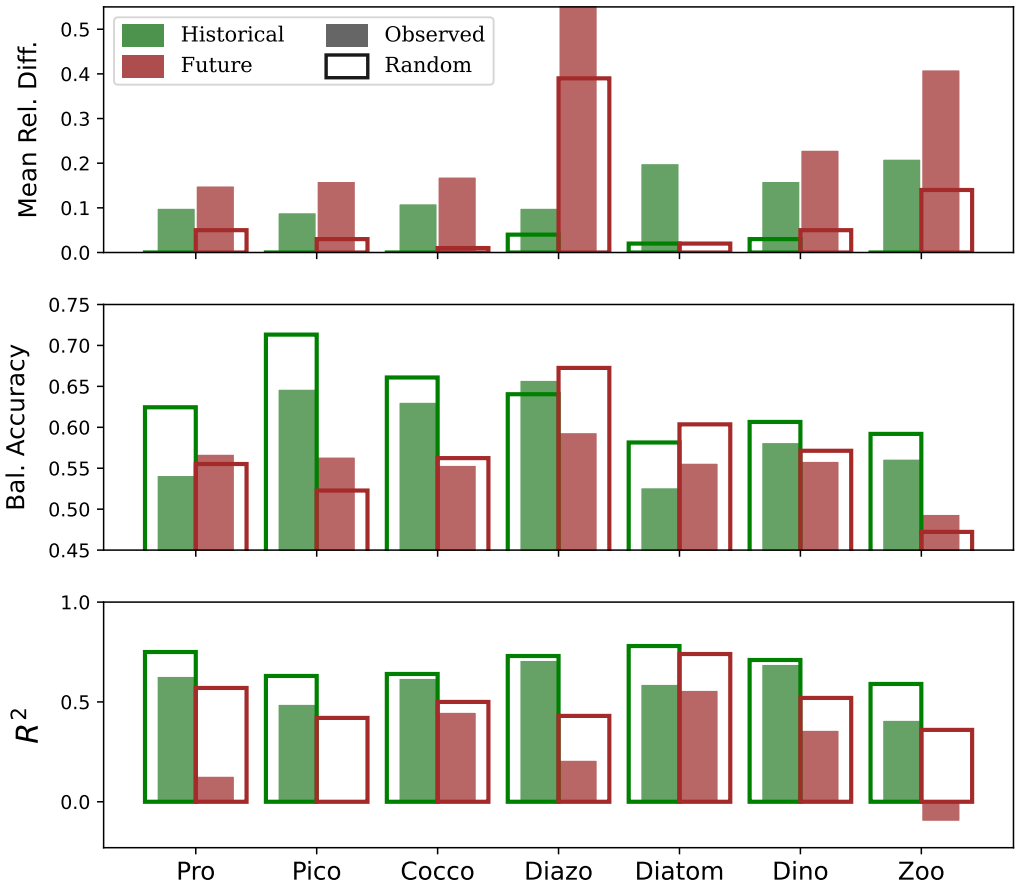

Figure 2: Comparing Darwin model 'true' biomasses with GAMs predictions for each functional group in 1987-2008 (historical) and 2079-2100 (future), and from measurements-derived and randomly-sampled training sets. *Top to Bottom:* **(a)** Relative differences of the means, given by $(GAMs_{mean} - Darwin_{mean})/Darwin_{mean}$.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

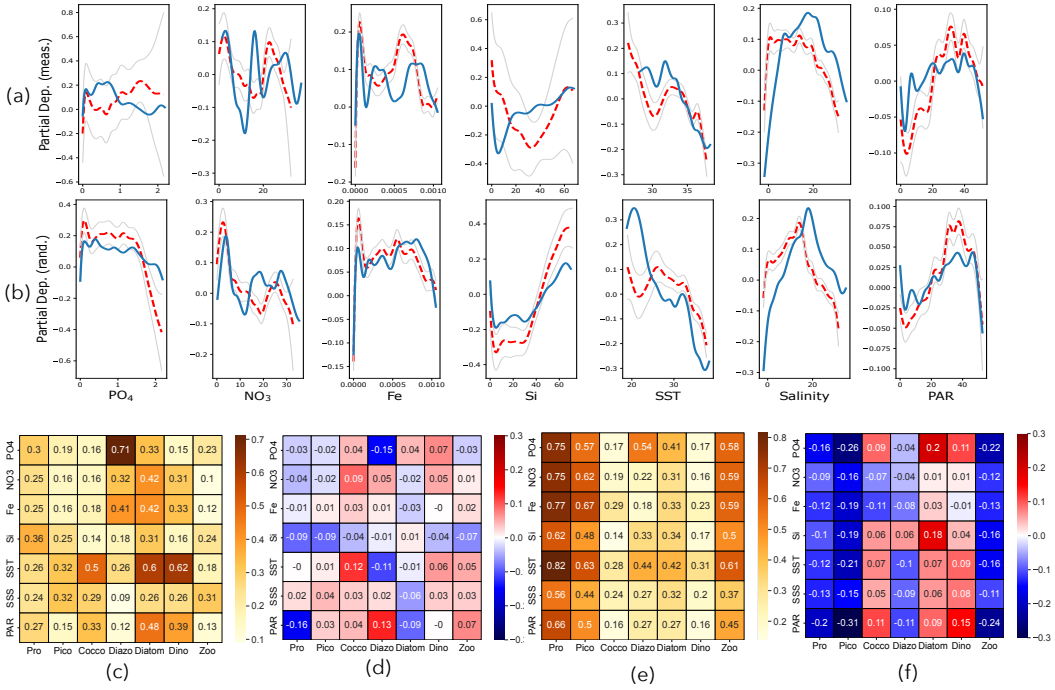

Figure 3: *Changing Relationships*: **(a)** Partial dependence plots of coccolithophore biomass (mmol C/m^3) as a function of each predictor, centred around the median (PO_4, NO_3, Fe, Si in mmol X/m^3, SST in $°C$, SSS in PSU, PAR in E/m^2/day). Plotted using data from 3586 Darwin surface ocean cells 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.

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

For example, there are many more degrees of freedom in real-world interactions between plankton individuals, communities, 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 and epigenetically to change. With their short generation times and high biodiversity, we might reasonably expect intrinsic relationships to 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 by necessity, 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.

Additionally, our results also demonstrate how spatial sampling bias can significantly alter the patterns of apparent relationships between environmental predictors and

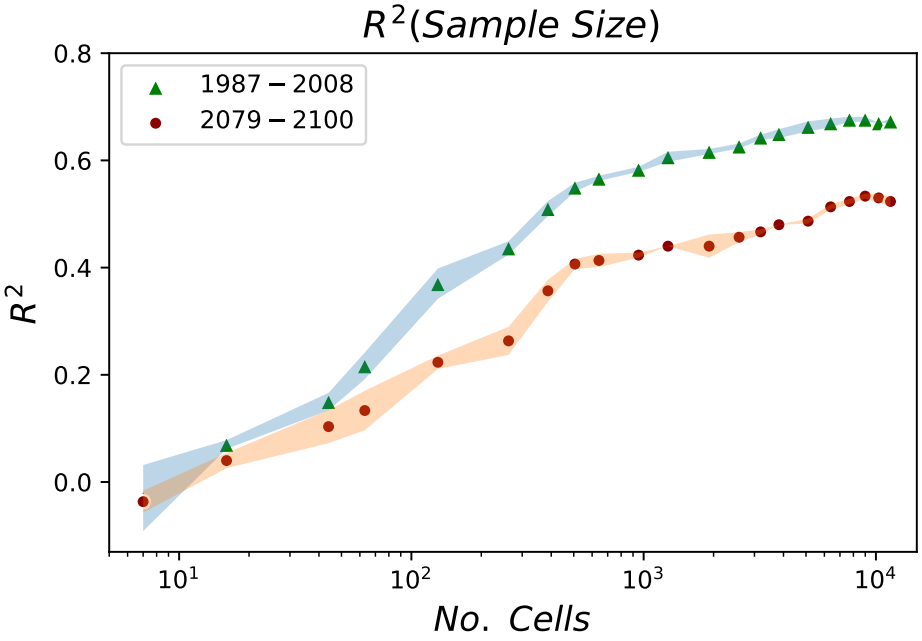

Figure 4: R^2 of GAMs model prediction as a function of sample size. Points are the mean R^2 value for coccolithophore predictions from three independent randomly-generated training sets for each of the 18 sample sizes, ranging from *N=63* to *N=11,557*. Shading is the standard deviation.

plankton biomass. The association strengths identified in the measurements-derived sample vary considerably from those found in the randomly sample of equivalent size (see Figure 3(c) vs. 3(e)). Importantly, this finding is robust across a range of random sample sizes, where almost identical patterns of correlations are seen in the 3586 cell case as in the 25,683 cell case, and robust across several methods of deriving correlations (see Figure $S12$). Nonetheless, the spatial patterns of over and under-prediction derived from the GAMs are not merely the result of spatiotemporal measurement biases. We see remarkable agreement in these broad qualitative patterns between the predictions generated from measurements-derived and random samples ((c) and (d) of Figures 1, and S1−6, and Figures $S8$ and $S9$). Ocean measurement biases may explain some element of the tendency towards 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 non-biased dataset reduces the severity of over and underprediction, especially for spatial predictions (Figure S8(e) and S9(e)). But the same broad biogeographic patterns of over and underprediction remain, indicating that the GAMs are still failing to effectively capture changes over time, despite their relatively robust performance according to the broad brush strokes of summary statistics (Figure S4(e) and S4(f)).

With that in mind, a number of optimisations could be made to improve predictive skill in real-world applications. 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)) to account for this effect. By contrast, our experimental de-

sign permitted us the unusual opportunity of testing our outcomes alongside a range of representative, randomly-sampled datasets 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, but we encourage their application in real-world settings.

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

We focus here on a particular type of statistical learning method that, for reasons outlined in Materials & Methods, 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 complex and ever-changing nature of the system, some of these limitations could be fundamental and unavoidable, particularly when extrapolating far beyond the training regime. Indeed, the median overestimation by the GAMs, even when using randomly sampled training data, implies that the predicted abundance distributions are less skewed than the Darwin model distributions, which are, in turn, less skewed than distributions in the the real ocean. But we stress that these observations do not extend to data-driven methods writ large. The recent work of (Holder & Gnanadesikan, 2021) evaluates random forests (RF) and neural network ensembles (NNE) in their ability to resolve the intrinsic relationships between plankton biomass and predictors, as extracted in a laboratory setting, from the apparent relationships in the data. They demonstrate variability in model predictive skill across different test scenarios, and find that NNE's yield overall superior performance, particularly in the case where plankton growth rates respond rapidly to environmental change. However, while these more sophisticated machine-learning models might yield an improvement in predictive skill, this can come at the cost of interpretability. Nonetheless, recent work by (Rudy et al., 2017) has shown that it is possible to use data-driven methods to directly extract the mechanistic equations that describe a dynamical system. This is an extremely promising advance toward hybrid methods that can provide both high levels of predictive skill, and an underlying description of the drivers of change.

Methodologically, the approach we have presented of applying a statistical model to output from a numerical model may be useful for addressing a number of additional questions. These might include evaluating how best to statistically model whole-ecosystem properties, such as diversity, from observations, or assessing where and when to make new observations to maximise information content about global plankton biogeography.But, as our results here have demonstrated and reinforced, it is important to be aware 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 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,

Z. V., ... Ward, B. A. (2013). The biogeography of marine plankton traits. *Ecology Letters*, *16*(4), 522–534.

Cabré, A., Marinov, I., & Leung, S. (2015). Consistent global responses of marine ecosystems to future climate change across the IPCC AR5 earth system models. *Clim. Dyn.*, *45*(5), 1253–1280.

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

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. *PNAS*, *105*(51), 20344–20349.

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.

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.

Dutkiewicz, S., Follows, M. J., & Bragg, J. G. (2009). Modeling the coupling of ocean ecology and biogeochemistry. *Global Biogeochemical Cycles*, *23*(4).

Dutkiewicz, S., Scott, J. R., & Follows, M. J. (2013). Winners and losers: Ecological and biogeochemical changes in a warming ocean. *Global Biogeochemical Cycles*, *27*(2), 463–477.

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.

Falkowski, P. G., Barber, R. T., & Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science*, *281*(5374), 200–206.

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science*, *320*(5879), 1034–1039.

Fenchel, T. (1988). Marine plankton food chains. *Ann. Rev. Eco. Sys.*, *19*(1), 19–38.

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.

Fuhrman, J. A. (2009). Microbial community structure and its functional implications. *Nature*, *459*(7244), 193–199.

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*.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., ... Gorsky, G. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*(7600), 465–470.

Guidi, L., Stemmann, L., Jackson, G. A., Ibanez, F., Claustre, H., Legendre, L., ... Gorskya, G. (2009). Effects of phytoplankton community on production, size, and export of large aggregates: A world-ocean analysis. *Limnology and Oceanography*, *54*(6), 1951–1963.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297–310.

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.

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

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., ... Zinger, L. (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell*, *179*(5), 1084–1097.e21.

Irwin, A. J., Nelles, A. M., & Finkel, Z. V. (2012). Phytoplankton niches estimated from field data. *Limnology and Oceanography*, *57*(3), 787–797.

Kuhn, A. M., Dutkiewicz, S., Jahn, O., Clayton, S., Rynearson, T. A., Mazloff, M. R., & Barton, A. D. (2019). Temporal and spatial scales of correlation in marine phytoplankton communities. *Journal of Geophysical Research: Oceans*, *124*(12), 9417–9438.

Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., ... Appeltans, W. (2019). Globally consistent quantitative observations of planktonic ecosystems. *Front. Mar. Sci.*, *6*.

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.

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*(3).

Martiny, A., & Flombaum, P. (2020). Global observations prochlorococcus, synechococcus, and picoeukaryotic phytoplankton with ancillary environmental data from 1987 to 2008.

Prinn, R. G. (2013). Development and application of earth system models. *PNAS*, *110*, 3673–3680.

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.

Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2017). Data-driven discovery of partial differential equations. *Science Advances*, *3*, e1602614.

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

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

Sokolov, A. (2005). The MIT integrated global system model (IGSM) version 2: Model description and baseline evaluation. , 46.

Sonnewald, M., Dutkiewicz, S., Hill, C., & Forget, G. (2020). Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Science Advances*, *6*(22), 2375-2548.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794.

Tang, W., & Cassar, N. (2019). Data-driven modeling of the distribution of diazotrophs in the global ocean. *Geophysical Research Letters*, *46*(21), 12258–12269.

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.

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.

Wei, Q., & Dunbrack, R. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PloS One*, *8*, e67863.