Phytoplankton niches estimated from field data

Andrew J. Irwin,[a,1,*] Andrew M. Nelles,[a] Zoe V. Finkel[b]

[a]Mathematics and Computer Science, Mount Allison University, Sackville, New Brunswick, Canada

[b]Environmental Science Program, Mount Allison University, Sackville, New Brunswick, Canada

*Corresponding author: airwin@mta.ca

[1]Present address: Earth, Atmosphere and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts

Running head: Estimating phytoplankton niches

Acknowledgements

This work was only possible because of the long-term work of many scientists working on the Continuous Plankton Recorder (CPR) project at Sir Alister Hardy Foundation for Ocean Science. We thank Andrew Barton for discussions and assistance with CPR data. Two anonymous referees suggested changes which greatly improved the manuscript. Funding was provided to AJI and ZVF by National Science and Engineering Research Council Canada.

Abstract

We combine phytoplankton occurrence data for 119 species from the continuous plankton recorder with climatological environmental variables in the North Atlantic to obtain ecological response functions of each species using the MaxEnt statistical method. These response functions describe how the probability of occurrence of each species changes as a function of environmental conditions and can be reduced to a simple description of phytoplankton realized niches using the mean and standard deviation of each environmental variable, weighted by its response function. Although there was substantial variation in the realized niche among species within groups, the envelope of the realized niches of North Atlantic diatoms and dinoflagellates are mostly separate in niche space.

Introduction

The coming century will bring global and regional changes in climate with numerous effects on marine ecosystems. Changes in climate will affect the distribution and productivity of phytoplankton in the ocean (Behrenfeld et al. 2006; Finkel et al. 2010). One tool for anticipating these changes is the simulation of phytoplankton communities in global ocean models (Follows et al. 2007; Barton et al. 2010). The diversity of phytoplankton is vast, so it is difficult to know if the species usually incorporated into such models adequately represent phytoplankton communities as a whole. The characterization of phytoplankton in these models relies heavily on physiological parameters obtained from laboratory experiments and may not be representative of the niches occupied by phytoplankton functional groups (Anderson 2005).

We propose that a broad perspective on phytoplankton strategies is needed to investigate how phytoplankton will respond to climate change, how the vast diversity of phytoplankton should be simplified for modeling purposes, and how the problem of adaptation to future scenarios should be approached. Large field programs that sample substantial diversity in the phytoplankton are a source of data that can be used to address these questions. By examining the range of responses to present environmental conditions, we hope to gain an understanding of how phytoplankton communities will respond to future climate scenarios. The future ocean will be different from the present, but many of these changes will be a matter of degree, e.g., how much stratified ocean, the timing of the shoaling of the mixed layer, and the location of phytoplankton blooms. It may be possible to predict primary productivity and phytoplankton community composition at the functional group level by treating the future ocean as a rearrangement of conditions from the contemporary ocean.

Our goal is to mine the continuous plankton recorder (CPR) and climatological environmental data in the North Atlantic to determine functions describing the probability of species occurrence and use these functions to characterize the realized niche for each species. A species' fundamental niche is the hypervolume in a space of environmental variables where it can persist as a consequence of its physiology and its realized niche is a subset of the fundamental niche where it is found and includes the effects of ecological interactions (Hutchinson 1957; Kylafis and Loreau 2011). Niches have already been identified for two species of copepod using CPR data (Beaugrand and Helaouet 2008). Obtaining a description of the realized niche for many phytoplankton species will allow us to observe the variability within functional groups and the extent of differences between functional groups on average. Variability across taxa in the response functions can be used

to guide the development of phytoplankton parameterizations in ocean simulation models. The probability models could be used to develop statistical models predicting the distribution of phytoplankton taxa in future ocean climate scenarios.

Methods

The CPR survey conducted by the Sir Alister Hardy Foundation for Ocean Science is the largest multi-decade plankton monitoring program in the world, with over 200,000 samples and 2.5 million plankton abundance counts from 1946-2004, and an additional 5000 samples each year (Barnard et al. 2004; Beaugrand 2004; Richardson et al. 2006). The CPR sampling device is towed by ships of opportunity at their conventional operating speeds at a standard depth of 7 m and has changed little since the initiation of the survey. Water enters through an opening 1.27 cm wide and flows down a tunnel that increases in cross-section, decreasing water pressure to minimize damage to the organisms and plankton are filtered onto a constantly moving band of silk (270 μm mesh size). The filtering silk is covered by a second band of silk and is wound onto a spool into a storage tank containing formalin. This silk mesh captures many larger phytoplankton species, predominantly diatoms and dinoflagellates, as well as many smaller phytoplankton with diameters as small as 10 μm. Clogging due to mucilage or increased loading associated with high densities of plankton may contribute to the number of small plankton captured. Despite these challenges the CPR has been shown to capture a consistent fraction of the in situ abundance of the species assayed (Richardson et al. 2006). Each CPR sample represents a transect of 10 nautical miles, approximately 3 m^3 of water filtered, and all phytoplankton on the silk are identified by trained experts under light microscopes (54X–450X). The CPR survey recommends that the species counts be interpreted and analyzed as a semi-quantitative estimate of abundance. Furthermore, the taxonomy data is collected along transects through large survey regions, with sampling effort varying across the North Atlantic and over time (Richardson et al. 2006). In this study, we use observations of phytoplankton taxa only as evidence of the presence of the species in a particular month and the corresponding standard CPR study area.

We analyzed a subset of the CPR data with 187 phytoplankton species, consisting primarily of diatoms (97 species) and dinoflagellates (86 spp.), with monthly observations over 60 years (1947-2006) and 41 standard CPR survey regions. The sampling effort of the CPR survey varied by year, month, and region, with most of the variation across years, some geographic variation, and relatively little month-to-month variation. We discarded 64 species with fewer than 10 observations in the entire dataset, plus the one cyanophyte, one silicoflagellate, and two prasinophytes as these groups were represented by very few species. The presence data for any particular species are very sparse since most species not observed in most locations and times. The total amount of data is very large: there are 153,450 species observations over a total of 69 diatom and 50 dinoflagellate species used and 45% of the 29,520 (=12·60·41) combinations of months, years, and regions were sampled. Although the number of observations of each species varies greatly from 10 to 9038, nearly all species were identified if present in each sample throughout the time-series.

We characterize phytoplankton realized niches using statistical models combining the CPR taxonomic data with environmental variables. The environmental data we used in the model are sea surface temperature (SST, °C), salinity, nitrate, phosphate, and silicate concentration (μmol L^{-1}) in the upper 10 m from the World Ocean Atlas 2009 (Antonov et al. 2010; Garcia et al. 2010; Locarnini

et al. 2010); sea-surface photosynthetically available irradiance (PAR, μmol m^{-2} s^{-1}) and light attenuation at 490 nm (k_{490}, m^{-1}) from the SeaWiFS project (oceancolor.gsfc.nasa.gov); and mixed layer depth (MLD, m) (De Boyer Montégut et al. 2004). We computed the mean irradiance over the mixed layer as

$$\frac{1}{\mathrm{MLD}}\int_0^{\mathrm{MLD}} \mathrm{PAR}\, e^{-k_{490}z}dz = \frac{\mathrm{PAR}}{k_{490}\,\mathrm{MLD}}(1 - e^{-k_{490}\mathrm{MLD}}) \qquad (1)$$

where z is the integration variable representing depth (m) in the mixed layer. All data are averaged over the 41 standard survey areas from the CPR project (www.sahfos.ac.uk/data-archive/standard-areas.aspx) and averaged to produce monthly climatologies. Although some data are available for many years (e.g., SST, PAR), most data are only available as monthly climatologies, and we did not want to include yearly variation in some variables and not others as this might bias the results and complicate their interpretation. We include each year of CPR data as a separate set of observations and pair these taxonomic data with monthly climatological environmental data.

We analyze the combined taxonomic and environmental data to establish a functional relationship between environmental conditions and the probability of observing a particular phytoplankton species. The MaxEnt method (Phillips et al. 2006; Phillips and Dudík 2008) is a statistical machine-learning technique that uses species presence data together with coincident environmental data and the full distribution of environmental data to estimate these functional relationships. No data on species abundance or absence are required. The name 'MaxEnt' refers to the entropy maximization performed when estimating functional relationships. Briefly, probability distribution functions are estimated for the suite of observed environmental conditions (predictor variables) known as background data, $f(x)$ and, using the species presence data, for environmental conditions in which the individual species are known to occur, $f_1(x)$. The presence data are separated into training and testing subsets. The conditional probability of finding species i, $\mathrm{P}(y_i = 1 \mid x)$, in a particular environment, x, is evaluated using Bayes' theorem:

$$\mathrm{P}(y_i = 1 \mid x) = \mathrm{P}(y_i = 1)\, f_1(x) / f(x), \qquad (2)$$

where $\mathrm{P}(y_i = 1)$ is the probability that species i would be found in a random sample, without regard to the environment, and represents the overall prevalence of species i. With abundance-only data this prevalence is not known. The predicted logistic probabilities of the MaxEnt method set $\mathrm{P}(y_i = 1)$ so that the probability of presence of a species at a site where that species is found is 1/2 on average (Elith et al. 2011). Two species found in the same environments will have the same predicted probabilities, even if the relative abundance of the two species is very different; this is a consequence of using presence only data. Here we are characterizing phytoplankton response curves, and looking for differences among species responses to contrasting environments, without regard to abundance.

The functions $f(x)$ and $f_1(x)$ are sums of piecewise linear and quadratic functions and products of these functions. The version of MaxEnt we used (3.3.3e; www.cs.princeton.edu/~schapire/maxent/) permitted discontinuous step (threshold) functions, but we did not use them as they often produced physically unrealistic response functions, likely as a result of

over-fitting to data. We used the default regularization parameters to avoid the problem of overfitting sparse data (Elith et al. 2011), but some tendency to overfit the data may remain (Warren and Seifert 2011). We produced two sets of models: one containing all of the environmental predictors, for the purpose of predicting species presence probabilities and assessing model utility, and a second set of models using just one environmental predictor at a time for the purpose of characterizing the species response to each environmental condition individually as recommended by the MaxEnt software and tutorial (Phillips et al. 2004; Buermann et al. 2008; Martínez-Freiría et al. 2008).

Since there are correlations between the predictors, the response functions from the full multivariate model can be difficult to interpret. Response functions generated from a multivariate model must be averaged over the other variables, producing a marginal response function, to permit visualization but this averaging hides the correlated effects of the other variables. For example, nitrate and phosphate concentrations are strongly positively correlated and phosphate concentration does not vary much at a constant nitrate concentration. So a marginal response function for nitrate concentration will include the effects of phosphate concentration. The univariate response function for nitrate concentration, on the other hand, illustrates the effect of changing nitrate concentration without accounting for the changing phosphate concentration. For example, for *Coscinodiscus wailesii*, the marginal response curves for temperature, mean irradiance, nitrate concentration, and phosphate concentration are all flat (results not shown), apparently due to correlations with silicate concentration. The signal of all those environmental variables is included in the response curve for silicate concentration. The univariate response functions show more easily interpreted results; increased probability of occurrence with increasing macronutrient concentrations and decreasing mean irradiance, and a broad but not flat temperature response function. We examined the means of the marginal response functions from the multivariate model and found that with pairs of highly correlated environmental variables (e.g., nitrate and phosphate), the MaxEnt model would frequently attribute the vast majority of the variation to only a few variables leaving the response for the other variables nearly flat. Phillips et al. (2004) caution that this may happen: "from a flat profile we may not conclude that the species distribution does not depend on the corresponding variable since variables may be correlated and MaxEnt will sometimes pick only one of the correlated variables."

The method does not require uniform sampling effort in time or space, although inferences about environments with less data may be less reliable, and extrapolations outside observed environmental conditions are unwise. When sampling, the observation of a species confirms presence, but non-observation does not confirm absence as the species may be rare or difficult to sample. Using presence data to predict the presence of the species as a function of environmental data is intended to make efficient use of the presence data without using potentially unreliable absence data. Biases in sampling, whether a result of changes in effort (mostly inter-annual and geographic for the CPR) or sampling effectiveness (some species may be less readily sampled on the silk mesh) can affect the modeled niche, but these biases should be greatly reduced by not using absence data.

The models produced by the MaxEnt method are complex and should be examined from several perspectives. We report a measure of the quality of the model fit, the relative importance of each predictor, sample response functions, and summary statistics characterizing the niche of each species.

The receiver operating characteristic (ROC) curve (Fig. 1) summarizes the predictions of a MaxEnt model and is further summarized by a single number, the area under the ROC curve (AUC). Each point on the ROC curve corresponds to a threshold probability dividing a prediction of species presence from absence. The predicted presences are used to estimate a true positive and false positive rate for each prediction. For example, if the threshold is low (e.g., species presence predicted if the probability > 0.1) then the true positive rate will be close to 100% but the false positive rate will also be high. An ideal ROC curve will be horizontal across the top of the plot with an area of 1 under the curve. With presence data used by MaxEnt there are no known species absence data, so the false positive rate is approximate and inferred from predictions of species presence in the background data, where observations are not known. We use the MaxEnt estimation of the maximum possible AUC as a baseline for the AUC; the maximum AUC is less than 1 due to corrections for errors in the false positive rate, and sometimes less than the observed AUC. A random model will have AUC = ½ as it is essentially a coin toss used to rank environmental suitability (Fig. 1, dotted line). The AUC is computed for both the training data and the data reserved for testing.

For each species, the importance of each environmental predictor can be estimated by randomly permuting the values of that variable across known presence and background data. The model is reevaluated on the permuted data, and the resulting drop in AUC on the test data, normalized to percentages, is used as a measure of the importance of that variable to the model. Since some environmental predictors are correlated, the contributions should be interpreted with caution. Based on the results from the entire suite of phytoplankton models, and from the diatoms and dinoflagellates separately, we estimate the most important predictors by averaging the importance and the rank importance for each model and computing how frequently a predictor is ranked as the most important of the 7 considered.

The MaxEnt model provides estimates of the effect of each environmental variable on the probability of presence for each species, called response functions. We use logistic scaling of the MaxEnt model output so that predictions are confined to the range 0 to 1 and can be interpreted roughly as the probability of species presence. Some response functions are approximately Gaussian across the environmental conditions and well summarized by a mean and variance, others are truncated on the right or left, skewed left or right, bimodal, or approximately flat over a wide interval (Fig. 2). Examining response functions for dozens of species would be challenging, so we use two parameters to summarize the response functions: the mean (μ), and standard deviation (σ), of each environmental variable (x), weighted by the univariate response function and defined by

$$\mu = \frac{\int x f(x)\, dx}{\int f(x)\, dx} \qquad (3)$$

and

$$\sigma^2 = \frac{\int (x-\mu)^2 f(x)\, dx}{\int f(x)\, dx} \qquad (4)$$

We call these parameters the mean univariate realized niche (or simply mean niche, μ), and the

breadth of the niche (σ). The vector of μ and σ for all environmental variables represents the environmental conditions in which a species is found and we interpret them as a simple description of a species' realized niche. Distilling the response functions down to these two parameters facilitates the comparison of a large number of species, but the degree of approximation involved varies by species. Since many of the environmental variables are correlated (macronutrients, temperature, irradiance), the mean univariate realized niches must be interpreted with caution as they will include information both from the direct response to the focal variable and the response associated with changes in correlated variables. At the functional group level, we test the null hypothesis that the mean of the environmental response functions for each environmental variable is the same for diatoms and dinoflagellates using a *t*-test with Welch correction for unequal variances (R Development Core Team 2011).

We use the ratio of the breadth of the niche to the range of environmental conditions in the North Atlantic to characterize a species as either a specialist or generalist, depending on whether the breadth of the niche at half its maximum value extends over less or more than 40% of the full range of the predictor in the North Atlantic; a specialist species will have $2.35\ \sigma\,(\text{predictor range})^{-1} < 0.40$. While this threshold is somewhat arbitrary, it provides a consistent guide to comparing the degree of specialization or generalization to different environmental conditions. We estimate a 95% confidence interval on the mean, μ, and breadth, σ, of the niche for each species and environmental variable using 500 models for each species. The replicate models are constructed with random resampling of the data (bootstrapping) as built into the MaxEnt software. Finally, since the CPR data extend over 60 years and we are using climatological environmental conditions, we were concerned that the mean environments for each species may have appeared to change because the phytoplankton data have year-to-year variation, but the environmental data do not. We computed MaxEnt models for the first 30 years (1947-1976) and the last 30 years (1977-2006) and compared the mean niches for all species found in both periods.

Results

A simple measure of the skill of a MaxEnt model is the area under the receiver operating characteristic curve (AUC). For our 119 models, the AUC ranged from 60-99% with a median of 84%. Since we do not have species absence data, we can not assess the false positive rate (Fig. 1) exactly and as a consequence the best possible AUC will be less than 100%. Species observed more frequently in the CPR dataset tend to have smaller AUC because of the difficulty in estimating the false positive rate. The AUC for testing data were in all cases within a few percent of the AUC on training data and the estimated maximum possible AUC. All our MaxEnt models do a good job of predicting the observed presence data according to the AUC metric.

We can assess how important an environmental variable is to a model by permuting the data for that variable at random and observing the percentage decrease in the AUC. We report 3 different measures of the average importance of each environmental variable: The percentage decrease in AUC when the environmental variable is omitted, which we call the importance, averaged over species; the mean rank (1-7) of each environmental variable's importance; and the percentage of species for which that environmental variable is most important (Table 1). We report each statistic

for all species, for diatoms, and for dinoflagellates. Generally, $(MLD)^{-1}$ was the most important predictor, followed by salinity, and followed closely by SST and mean irradiance, but there are notable differences between diatoms and dinoflagellates. Salinity was more important for dinoflagellates than diatoms while mean irradiance was more important for diatoms. Macronutrient concentrations (nitrate, phosphate, and silicate) were the least important variables, accounting for less than 25% of the aggregate importance and less than 10% for each nutrient.

The phytoplankton realized niches are described by the mean and breadth of the niche for each environmental variable (Fig. 3). The 95% confidence intervals on some points are quite large, but close examination indicates that many species' niches are well estimated and do not overlap the majority of other species' niches. Diatoms and dinoflagellates have distinct niches for each environmental variable although there is some overlap between the cloud of points for each functional group. Most dinoflagellate species are found in warmer and saltier waters with lower nutrient concentration and higher mean irradiance in the mixed layer, compared with diatoms (Figs. 2, 3). The mean niche differs between the diatoms and dinoflagellates for each environmental variable (t-test, $p < 10^{-5}$, (R Development Core Team 2011). For some variables (SST, phosphate) the number of species with a particular mean niche seems to be approximately proportional to the abundance of habitat (histograms in Fig. 3A, D) while for the other variables (particularly salinity, nitrate and silicate concentration, and mean irradiance) the most abundant environments often have among the fewest species with a corresponding mean niche.

Many species are generalists (high σ) so that their univariate niches extend over much of the observed range of each environmental variable (above the dotted line in Fig. 3). There are no low temperature or low salinity specialists, but there are some specialists at high temperatures or high salinities, most of which are dinoflagellates. All nutrient specialists have mean niche towards the low end of nutrient concentrations observed and are almost all dinoflagellates. Mean irradiance specialists are all diatoms and as for nutrient concentration, all have mean niches at the lower resource levels. Dinoflagellates have broader salinity niches than diatoms at the same mean niche, but the reverse is true for mean irradiance, where diatoms have broader niches.

Species with mean niches at intermediate environmental conditions tend to have broader niches than species with mean niches at the high or low extremes of the environments sampled (Fig. 3). Species with broad niches are relatively insensitive to variation in that environmental variable and have mean niches near the middle of the range of observed environments. Species with mean niche near the edge of the observed range of environments are likely to have narrow niches. Response functions for SST are frequently not Gaussian, and can be skewed for cold (e.g., *Ceratium arcticum*) or warm (e.g., *Ceratium belone*) water species, supporting the observation of narrower niches in extreme environments. The observed SST range from –1.2°C to 21.8°C ($\sigma \sim 7$), so the species with the smallest mean SST niche (5°C) is not forced to be a specialist.

A species' niche for one environmental variable may be related to its niche for another variable but such relationships are obscured when examining one variable at a time (Fig. 3). Furthermore, relationships between niches for pairs of environmental variables may be common across all species in the study or vary across functional groups. Comparing the species' mean niches two variables at a time makes it possible to observe relationships between niches for those environmental variables (Fig. 4). We overlay these points on a grey background of the observed

environmental data to facilitate comparison between the mean niches for each species and the range of variation of the corresponding environmental variables. Diatoms and many dinoflagellates are clustered near the center of the salinity-SST graph, while the high salinity specialists stand out clearly from the cloud of species (Fig. 4A). There is an anti-correlation between mean nitrate and temperature niches and between mean nitrate and mean irradiance niches (Fig. 4B, C) which mirrors the anti-correlations between these pairs of variables in the environment. Unlike temperature and salinity, the cloud of points representing the mean niches for temperature, irradiance, and nitrate concentration are against the edge or even outside the range of observed environments. Few species have mean nitrate and mean phosphate niches which deviate from the N:P Redfield ratio of 16:1 (Fig. 4D) although diatoms seem to thrive in environments with N:P > 16. The ratio of mean nitrate to mean silicate niches is about 1.6:1 (reduced major axis regression), with some dinoflagellates found at relatively high N:Si (Fig. 4E). Species mean niches are in the middle of the background cloud for N:Si as opposed to being at the high N edge of the N:P cloud. Diatoms and dinoflagellates are separated by both mean irradiance in the mixed layer and by the reciprocal of the mixed layer depth (Fig. 4F). (We use $(\mathrm{MLD})^{-1}$ instead of MLD since the distribution of MLD is more strongly skewed than that of $(\mathrm{MLD})^{-1}$.) While these variables are autocorrelated, there is also an anti-correlation between sea-surface PAR and MLD which spreads the niches out over a wider range for mean irradiance, and using both variables simultaneously improves the separation of the functional groups. As with other pairs of variables, except nitrate and silicate concentration, there is a clear bias in the mean niches away from the most abundant, but less desirable, environmental conditions.

The mean niches for each species were similar in the first half of the study compared to the mean niches in the second half, with the largest deviations (more than 15%) appearing in species with fewer than 1000 observations. The correlations between mean niches from 1947-1976 and 1977-2006 were strong for most environmental variables for the 69 species with at least 40 observations in each time period; the R^2 were 0.85, 0.82, 0.75, 0.71, 0.73, and 0.60 for SST, phosphate, silicate, nitrate, mean irradiance, and salinity, respectively. There have been documented changes in the relative abundance of diatoms and dinoflagellates (Leterme et al. 2005) and in the relative timing of the maximum density of phytoplankton and zooplankton (Edwards and Richardson 2004). Our analysis shows that the species occurrence niches are robust on the half-century time scale when sufficient data are available.

Discussion

Components of the fundamental niche for some factors such as nitrate concentration, irradiance, and temperature have been compiled for a few species from laboratory studies (Sarthou et al. 2005; Litchman et al. 2007; Litchman and Klausmeier 2008). Even with all the work that has been done, we have only begun to sample the tremendous taxonomic diversity known to be present in the ocean and the myriad biotic and abiotic factors that may influence phytoplankton niches. Furthermore, it is not clear how successful lab-based physiological traits will be in predicting phytoplankton distributions in the field. Here we use a statistical approach to extract the realized niches of 119 phytoplankton species from the field using the CPR dataset (153,450 observations of species presence) and monthly climatologies of environmental conditions. We estimate the probability distribution of each species' presence as a function of each environmental variable and

further summarize each probability distribution by its mean and breadth (Fig. 3), which we refer to as the species' realized univariate environmental niche. Below we discuss the most important environmental predictors for individual species' presence, variation in the mean and breadth of the realized niche with variation in environmental conditions, and taxonomic differences in the niches of diatoms and dinoflagellates.

Salinity, mixed layer depth, mean irradiance in the mixed layer, and temperature are the most important predictors of individual phytoplankton species presence, with macronutrient concentrations the least important (Table 1). Laboratory experiments have established that there are significant differences in growth rate across species as a function of salinity (Braarud 1951; Balzano et al. 2011), temperature (Eppley 1972; Brand et al. 1981; Moisan et al. 2002), irradiance (Geider et al. 1986; Rodriguez et al. 2005; Schwaderer et al. 2011) and nutrient concentrations (Eppley et al. 1969; Hein et al. 1995). One might expect nutrient concentrations to be most informative in determining the presence or absence of a particular phytoplankton species (Litchman et al. 2007) as much of the North Atlantic is nitrate limited. There are several potential reasons for this inversion of expectations: salinity, irradiance in the mixed layer, and temperature may directly affect phytoplankton community structure due to species-specific physiological responses to these environmental conditions; the physical factors (salinity, MLD, and SST) may be correlated with important predictive factors not included in the model such as nutrient supply rate, short-term environmental variability, and grazing; and SST is often highly correlated with nutrient supply rates to phytoplankton in the surface (Kamykowski et al. 2002). The importance of temperature, salinity and irradiance has been noted in other studies: temperature and salinity tended to be better predictors than macronutrient concentrations for phytoplankton community structure (1984-2001) in the Baltic Sea (Gasiunaite et al. 2005) and the distributions of *Procholorococcus* ecotypes follow changes in temperature and irradiance consistent with physiological responses (Johnson et al. 2006).

The phytoplankton niches estimated here are non-random subsets of the environment conditions (histograms in Fig. 3 and grey background in Fig. 4) prevalent in the North Atlantic, confirming the importance of physiological differences and ecological interactions in shaping the taxonomic structure of phytoplankton communities. If species were randomly distributed with respect to environmental conditions the mean of the species' niches should aggregate towards the center of the environmental data. In contrast, species' mean niches are distributed across much, but not all, of the environmental data. In addition niche breadth for the individual environmental variables tends to vary systematically across several of the environmental conditions examined. For example niche breadths tend to be narrower under lower temperatures (often correlated with low nutrient supply rates), higher salinities, lower macronutrient concentrations, and very low irradiances in the mixed layer (Fig. 3), perhaps reflecting increased niche specialization under resource scarcity (although some of this may be due to edge effects, *see* Results). A high proportion of species have their mean niche for nutrient concentrations under high concentrations, especially relative to the frequency distribution of these environmental conditions (Fig. 4). The higher diversity of species found under higher nutrient concentrations, and lower temperatures, might be a function of the higher biomasses that can be supported when nutrient concentrations and supply rates are high.

Correlation in the mean and breadth of the realized niches for each individual environmental variable can sometimes be attributed to underlying co-variation in the distribution of environmental

conditions in the North Atlantic. For example, dinoflagellates have nearly the same silicate : nitrate niche ratios as diatoms despite having no requirement for silicate. This pattern is caused by the co-variation of Si concentration with the concentration and supply rates of nitrate and phosphate. Macronutrient concentrations are highly correlated with one another and inversely correlated with sea surface temperature and irradiance in the mixed layer. As a consequence species specializing in low concentrations of one macronutrient are likely to specialize in low concentrations of the other macronutrients as well, and will likely be exposed to higher temperatures and higher irradiances in the mixed layer. Among the nutrient specialists, 10 species do not specialize on a temperature, but the majority (18 species) do specialize on temperature as well. The specializations are not randomly associated: 13 species (11% of all species) are specialists for 4 to 6 of the variables examined, but a simulation of randomly assorting specializations predicts this would happen only in 0.5% of the species. There are physiological challenges for phytoplankton species under low nutrient concentrations, low irradiances, and extreme salinities and temperatures that limit the distribution of niches. For example, in these data we see that there are no phytoplankton species' exploiting the abundant low light habitat below ~5 μmol m^{-2} s^{-1} (Fig. 4); this is below the compensation point for most phytoplankton.

The niches for each species clearly separate out the diatoms and dinoflagellates, reflecting the fundamental physiological and ecological differences between diatoms and dinoflagellates that structure communities. Diatoms are more likely to be found in waters with colder temperatures, lower salinity, higher macronutrient concentrations, and lower mean irradiance, compared with dinoflagellates. Although these niches are fundamentally ecological, our results mirror physiological differences quantified from laboratory experiments (Litchman et al. 2007) such as growth inhibition in dinoflagellates in response to shear and turbulent mixing (Peters and Marrase 2000) and tolerance of low irradiance in diatoms (Geider et al. 1986). Dinoflagellates are much more likely to specialize in one or more variable (29 specialist species or 58%) than the diatoms (13 specialist species or 19%), consistent with previous studies (Smayda and Reynolds 2001). The differences between and variation within the groups indicates that there is evolutionary selection acting to separate strategies at the functional group and species level, but that fundamental physiological characteristics of the functional groups constrain the niche variation. Although there is a strong signal separating these functional groups on these predictors, there is still a great deal of variability within the groups and substantial overlap between the niches at the interface between the two groups. These results provide additional evidence that the use of diatoms and dinoflagellates as separate functional groups within ecological and biogeochemical models has a rational basis but that the variability within the groups should be explored (Follows et al. 2007) as guided by these results.

Our analysis appears to document trade-offs in the niches of several pairs of environmental data across species: the niches for nitrate concentration and mean irradiance and nitrate concentration and temperature are anti-correlated (Fig. 4). Since this is a statistical and observational study, attributing the cause of these correlations to physiological characteristics or environmental forcing is difficult. There are physiological trade-offs between maximum uptake rate and half-saturation constants for nitrate and between R^* (a measure of competitive ability at equilibrium) and maximum growth rate in phytoplankton (Litchman et al. 2007). Since warmer temperatures are associated with increased stability and lower nitrate concentrations in the ocean, we might expect species with low

half-saturation constants and low R^* to dominate warmer waters; conversely lower temperatures should favor species with greater maximum uptake rates to exploit increased nutrient flux from mixing and reduce the importance of half-saturation constants and R^*. Thus there may be a physiological basis, grounded in nutrient uptake kinetics, for niche differentiation we observe in SST. We observe that species with low nitrate niches tend to have high irradiance niches and vice versa, and this distinction separates the diatoms (low light) and dinoflagellates (low nitrate concentrations). Mixotrophic and heterotrophic dinoflagellates have access to nutrient resources which are unavailable to diatoms, but may require extra energy for handling, while under low light, autotrophic phytoplankton require more nitrogen for their pigment protein complexes, indicating a physiological basis for an irradiance-nitrate trade-off (Harrison et al. 1990). While this trade-off appears in the environment as well, there is a great deal of habitat simultaneously low in nitrate concentration and mean irradiance which is not used by dinoflagellates or diatoms. The absence of diatoms and dinoflagellates with niches low in both macronutrients and irradiance is a consequence of the fact that many of these species are relatively resource intensive compared to many of the low-resource species not observed by the CPR (Raven et al. 2006). Diatoms and dinoflagellates are forced into a trade-off between light and nitrate concentration by the available habitat and physiological adaptations contributing to the fundamental differences between these functional groups.

We have characterized the realized niches for a diverse collection of diatoms and dinoflagellates in the North Atlantic. Diatoms tend to be found in cooler waters with deeper mixed layers and lower mean irradiance, while dinoflagellates tend to favor warmer waters with shallower mixed layers and higher mean irradiance. Although diatoms and dinoflagellates are clearly broadly distinguishable by their niches, there is a great deal of variation among species within these large taxonomic groups. These niches could be used to predict biogeographic distributions of phytoplankton species and anticipate shifts in community structure that may result from changing environmental conditions and climate. Important caveats are that the emergence of novel habitats, the introduction of invading phytoplankton or grazers or mismatch between predators and prey (Edwards and Richardson 2004) could alter ecological niches and complicate future predictions (Edwards et al. 2001; Reid et al. 2007; Williams and Jackson 2007). Differences in niches between grazers and primary producers have led to phenological shifts and cascading consequences throughout the food web as climate changes (Edwards and Richardson 2004). Predictive models of phytoplankton responses to changing climate should include some of this diversity or they will run the risk of making unrealistic predictions limited by a lack of diversity in species responses to the environment.

References

Anderson, T. R. 2005. Plankton functional type modelling: Running before we can walk? J. Plankton Res. **27:** 1073-1081.

Antonov, J. I., D. Seidov, T. P. Boyer, R. A. Locarnini, A. V. Mishonov, H. E. Garcia, O. K. Baranova, M. M. Zweng, and D. R. Johnson. 2010. Word ocean atlas 2009, volume 2, p. 184. *In* S. Levitus [ed.], Salinity. NOAA Atlas NESDIS.

Balzano, S., D. Sarno, and W. H. C. F. Kooistra. 2011. Effects of salinity on the growth rate and morphology of ten *Skeletonema* strains. J. Plankton Res. **33:** 937-945.

Barnard, R., S. Batten, G. Beaugrand, C. Buckland, D. V. P. Conway, M. Edwards, J. Finlayson, L. W. Gregory, N. C. Halliday, A. W. G. John, D. G. Johns, A. D. Johnson, T. D. Jonas, J. A. Lindley, J. Nyman, P. Pritchard, P. C. Reid, A. J. Richardson, R. E. Saxby, J. Sidey, M. A. Smith, D. P. Stevens, C. M. Taylor, P. R. G. Tranter, A. W. Walne, M. Wootton, C. O. M. Wotton, and J. C. Wright. 2004. Continuous plankton records: Plankton atlas of

the north Atlantic Ocean (1958-1999). II. Biogeographical charts. Mar. Ecol. Prog. Ser. **Supp:** 11-75, doi: 10.3554/mepscpr011
Barton, A. D., S. Dutkiewicz, G. Flierl, J. Bragg, and M. J. Follows. 2010. Patterns of diversity in marine phytoplankton. Science **327:** 1509-1511.
Beaugrand, G. 2004. The North Sea regime shift: Evidence, causes, mechanisms and consequences. Elsevier - Progress in Oceanography **60:** 245-262.
Beaugrand, G., and P. Helaouet. 2008. Simple procedures to assess and compare the ecological niche of species. Mar. Ecol. Prog. Ser. **363:** 29-37.
Behrenfeld, M. J., R. O'Malley, D. A. Siegel, C. McClain, J. Sarmiento, G. Feldman, A. Milligan, P. G. Falkowski, R. Letelier, and E. Boss. 2006. Climate-driven trends in contemporary ocean productivity. Nature **444:** 752-755
Braarud, T. 1951. Salinity as an ecological factor in marine phytoplankton. Physiologia Planetarum **4:** 28-34.
Brand, L. E., L. S. Murphy, R. R. L. Guillard, and H.-T. Lee. 1981. Genetic variability and differentiation in the temperature niche component of the diatom *Thalassiosira pseudonana*. Mar. Biol. **62:** 103-110.
de Boyer Montégut, C., G. Madec, A. S. Fischer, A. Lazar, and D. Ludicone. 2004. Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. J. Geophys. Res. **109**: C12003, doi: 10.1029/2004JC002378
Buermann, W., S. Saatchi, T. B. Smith, B. R. Zutta, J. A. Chaves, B. Mila, and C. H. Graham. 2008. Predicting species distributions across the Amazonian and Andean regions using remote sensing data. J. Biogeography **35**: 1160-1176, doi: 10.1111/j.1365-2699.2007.01858.x
Edwards, M., A. W. G. John, D. G. Johns, and P. C. Reid. 2001. Case history and persistence of the non-indigenous diatom *Coscinodiscus wailesii* in the north-east Atlantic. Journal of the Marine Biology Association of the United Kingdom **81**: 207-211, doi: 10.1017/S0025315401003654
Edwards, M., and A. J. Richardson. 2004. Impact of climate change on marine pelagic phenology and trophic mismatch. Nature **430:** 881-884.
Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. **17:** 43-57.
Eppley, R. W. 1972. Temperature and phytoplankton growth in the sea. Fish. Bull. **70:** 1063-1085.
Eppley, R. W., J. N. Rogers, and J. J. Mccarthy. 1969. Half-saturation constant for uptake of nitrate and ammonium by marine phytoplankton. Limnol. Oceanogr. **14:** 912-920.
Finkel, Z. V., J. Beardall, K. Flynn, A. S. Quigg, T. A. V. Rees, and J. A. Raven. 2010. Phytoplankton in a changing world: cell size and elemental stoichiometry. J. Plankton Res. **32:** 119-137.
Follows, M. J., S. Dutkiewicz, S. Grant, and S. W. Chisholm. 2007. Emergent biogeography of microbial communities in a model ocean. Science **315:** 1843-1846.
Garcia, H. E., R. A. Locarnini, T. P. Boyer, J. I. Antonov, M. M. Zweng, O. K. Baranova, and D. R. Johnson. 2010. World ocean atlas, volume 4, p. 398. *In* S. Levitus [ed.], Nutrients (phosphate, nitrate, silicate). NOAA Atlas NESDIS.
Gasiunaite, Z. R., A. C. Cardosa, A.-S. Heiskanen, P. Henriksen, P. Kauppila, I. Olenina, R. Pilkaityte, I. Purina, A. Razinkovas, A. Sagert, H. Schubert, and N. Wasmund. 2005. Seasonality of coastal phytoplankton in the Baltic Sea: Influence of salinity and eutrophication. Elsevier - Estuarine Coastal and Shelf Science **65:** 239-252.
Geider, R. J., B. A. Osborne, and J. A. Raven. 1986. Growth, photosynthesis and maintenance metabolic cost in the diatom *Phaeodactylum tricornutum* at very low light levels. J. Phycol. **22:** 39-48.
Harrison, P. J., P. A. Thompson, and G. S. Calderwood. 1990. Effects of nutrient and light limitation on the biochemical composition of phytoplankton. J. Appl. Phycol. **2:** 45-56.
Hein, M., M. Folager Pedersen, and K. Sand-Jensen. 1995. Size-dependent nitrogen uptake in micro- and macroalgae. Mar. Ecol. Prog. Ser. **118:** 247-253.
Hutchinson, G. E. 1957. Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology **22**: 415–427.
Johnson, Z. I., E. R. Zinser, A. Coe, N. P. Mcnulty, E. M. S. Woodward, and S. W. Chisholm. 2006. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science **311:** 1737-1740.
Kamykowski, D., S.-J. Zentara, J. M. Morrison, and A. C. Switzer. 2002. Dynamic global patterns of nitrate, phosphate, silicate, and iron availability and phytoplankton community composition from remote sensing data. Global Biogeochem. Cycles **16:** 1077, doi: 10.1029/2001GB001640
Kylafis, G., and M. Loreau. 2011. Niche construction in the light of niche theory. Ecology Letters **14**: 82–90, doi: 10.1111/j.1461-0248.2010.01551.x
Leterme, S. C., M. Edwards, L. Seuront, M. J. Attrill, P. C. Reid, and A. W. G. John. 2005. Decadal basin-scale changes in diatoms, dinoflagellates, and phytoplankton color across the north Atlantic. Limnology and Oceanography **50:** 1244-1253.
Litchman, E., and C. A. Klausmeier. 2008. Trait-based community ecology of phytoplankton. Annual Reviews of Ecology, Evolution, and Systematics **39:** 615-639.

Litchman, E., C. A. Klausmeier, O. M. Schofield, and P. G. Falkowski. 2007. The role of functional traits and trade-offs in structuring phytoplankton communities: Scaling from cellular to ecosystem level. Ecology Letters **10:** 1-12.

Locarnini, R. A., A. V. Mishonov, J. I. Antonov, T. P. Boyer, H. E. Garcia, O. K. Baranova, M. M. Zweng, and D. R. Johnson. 2010. World ocean atlas 2009, volume 1, p. 184. *In* S. Levitus [ed.], Temperature. NOAA Atlas NESDIS.

Martínez-Freiría, F., N. Sillero, M. Lizana, and J. C. Brito. 2008. GIS-based niche models identify environmental correlates sustaining a contact zone between three species of European vipers. Divers. Distrib. **14**: 452-461, doi: 10.1111/j.1472-4642.2007.00446.x

Moisan, J. R., T. A. Moisan, and M. R. Abbot. 2002. Modelling the effect of temperature on the maximum growth rates of phytoplankton populations. Ecol. Model. **153:** 197-215.

Peters, F., and C. Marrase. 2000. Effects of turbulence on plankton: An overview of experimental evidence and some theoretical considerations. Mar. Ecol.-Prog. Ser. **205:** 291-306.

Phillips, S. J., M. Dudík, and R. E. Schapire. 2004. A maximum entropy approach to species distribution modeling. ICML '04 proceedings of the twenty-first international conference on machine learning. ACM, 83, doi: 10.1145/1015330.1015412

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. **190:** 231-259.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. Ecography **31:** 161-175.

R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Raven, J. A., Z. V. Finkel, and A. J. Irwin. 2006. Picophytoplankton: bottom-up and top-down controls on ecology and evolution. Vie et Milieu **55:** 209-215.

Reid, P. C., D. G. Johns, M. Edwards, M. Starr, M. Poulin, and P. Snoeijs. 2007. A biological consequence of reducing Arctic ice cover: Arrival of the Pacific diatom *Neodenticula seminae* in the North Atlantic for the first time in 800 000 years. Global change Biology **13:** 1910-1921.

Richardson, A. J., A. W. Walne, A. W. G. John, T. D. Jonas, J. A. Lindley, D. W. Sims, D. Stevens, and M. Witt. 2006. Using continuous plankton recorder data. Prog. Oceanogr. **68:** 27-74.

Rodriguez, F., E. Derelle, L. Guillou, F. Le Gall, D. Vaulot, and H. Moreau. 2005. Ecotype diversity in the marin picoeukaryote Ostrococcus (Chlorophyta, Prasinophyceae). Environ. Microbiol. **7:** 853-859.

Sarthou, G., K. R. Timmermans, S. Blain, and P. Treguer. 2005. Growth physiology and fate of diatoms in the ocean: A review. Journal of Sea Research **53:** 25-42.

Schwaderer, A. S., K. Yoshiyama, P. D. T. Pinto, N. G. Swenson, C. A. Klausmeier, and E. Litchman. 2011. Eco-evolutionary differences in light utilization traits and distributions of freshwater phytoplankton Limnol. Oceanogr. **56:** 589-598.

Smayda, T. J., and C. S. Reynolds. 2001. Community assembly in marine phytoplankton: Application of recent models to harmful dinoflagellate blooms. J. Plankton Res. **23:** 447-461.

Warren, D. L., and S. N. Seifert. 2011. Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. Ecological Applications **21**: 335-342.

Williams, J. W., and S. T. Jackson. 2007. Novel climates, no-analog communiities, and ecological surprises. Front. Ecol. Environ. **5:** 475-482.

Table 1

Summary of importance measures for 7 different environmental variables, averaged over all species, and averaged within diatoms and dinoflagellates.

	Mean importance (%)			Mean rank importance (1-7; 1 most important)			Proportion of species ranking this variable as most important (%)		
	all	diatoms	dinoflagellates	all	diatoms	dinoflagellates	all	diatoms	dinoflagellates
$(MLD)^{-1}$	25.8	28.3	22.4	2.9	2.6	3.2	31.9	31.9	32.0
Salinity	19.5	16.2	24.1	3.2	3.4	2.8	23.5	17.4	32.0
Mean irradiance	17.9	22.2	11.9	3.4	2.9	4.0	17.6	23.2	10.0
SST	14.9	13.4	17.0	3.7	3.9	3.4	15.1	14.5	16.0
Nitrate	7.9	7.8	8.0	4.5	4.5	4.6	3.4	2.7	4.0
Phosphate	7.4	5.5	10.0	5.2	5.4	4.9	2.5	2.9	2.0
Silicate	6.6	6.6	6.6	5.1	5.2	5.1	5.9	7.2	4.0

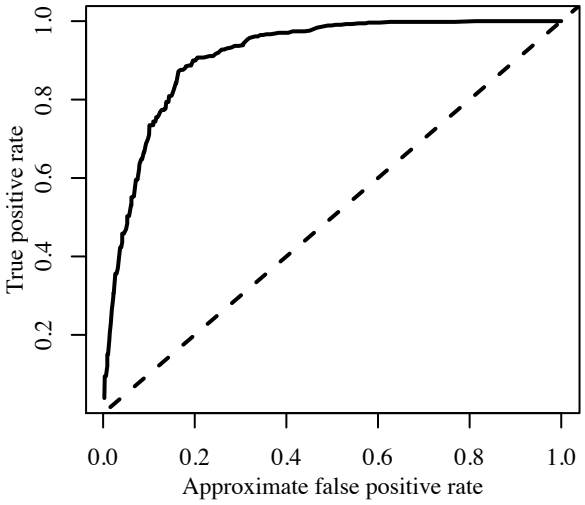
Figure captions

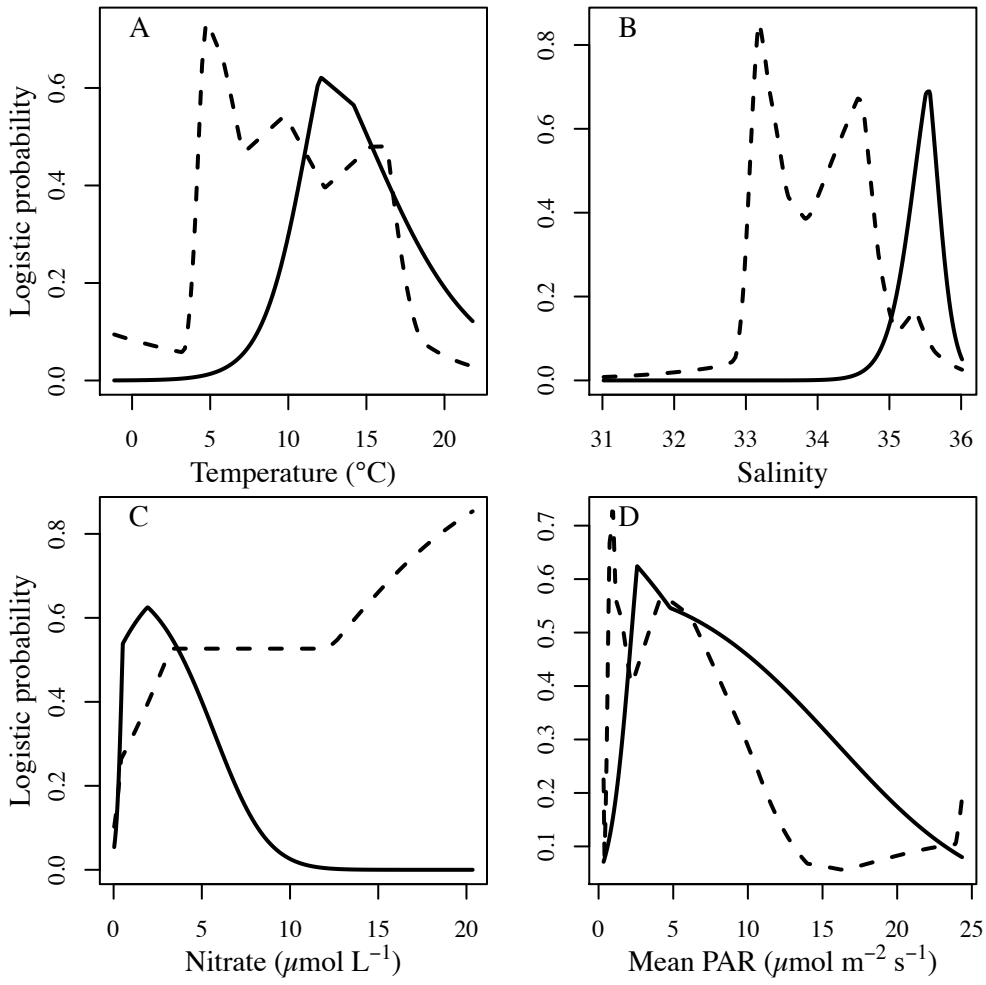
Figure 1. Receiver operating characteristic curve for MaxEnt model of *Ceratium hexacanthum*. The solid line compares the true positive classification rate as an implicit function of the approximate false positive rate using a range of probability thresholds to convert predicted logistic probabilities to presence or absence predictions. The dotted line is a random coin-flipping model. The area under the curve is 0.91.

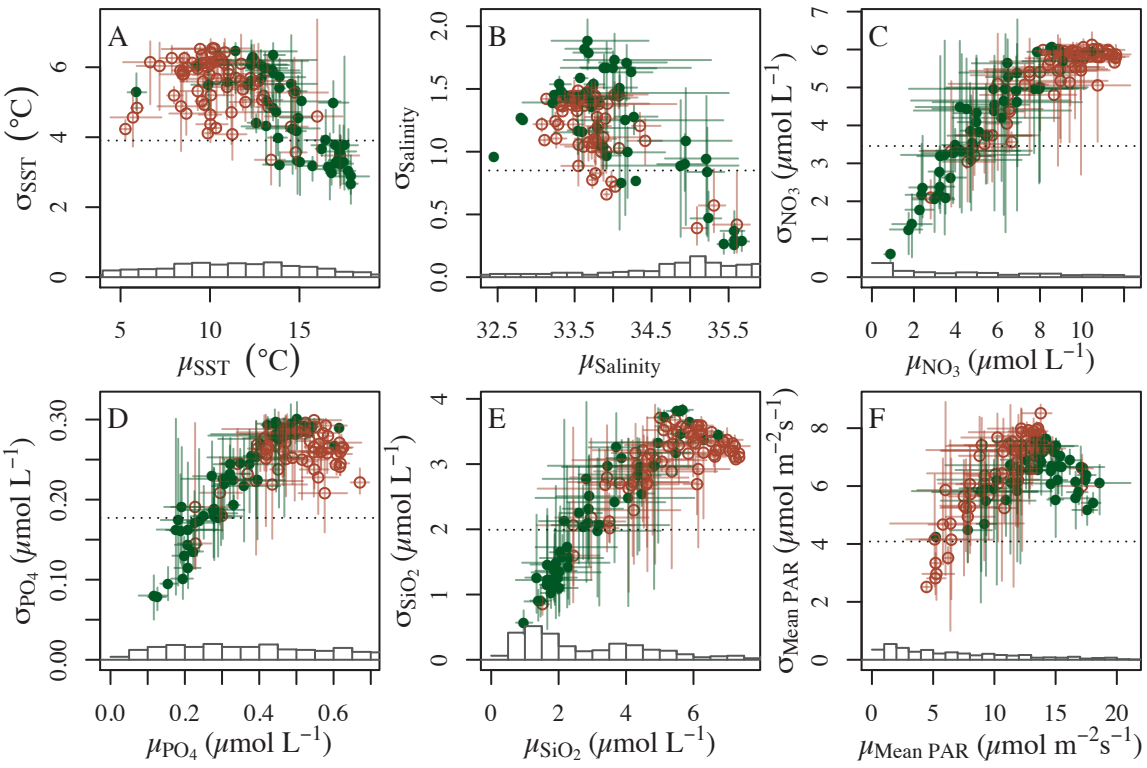
Figure 2. Univariate ecological response functions for the dinoflagellate *Ceratium compressum* (solid line) and the diatom *Odontella sinensis* (dashed line), reported as the logistic probability of occurrence as a function of (A) sea surface temperature (°C), (B) salinity, (C) nitrate concentration in the upper 10 m (μmol L^{-1}), and (D) mean irradiance in the mixed layer (μmol m^{-2} s^{-1}).

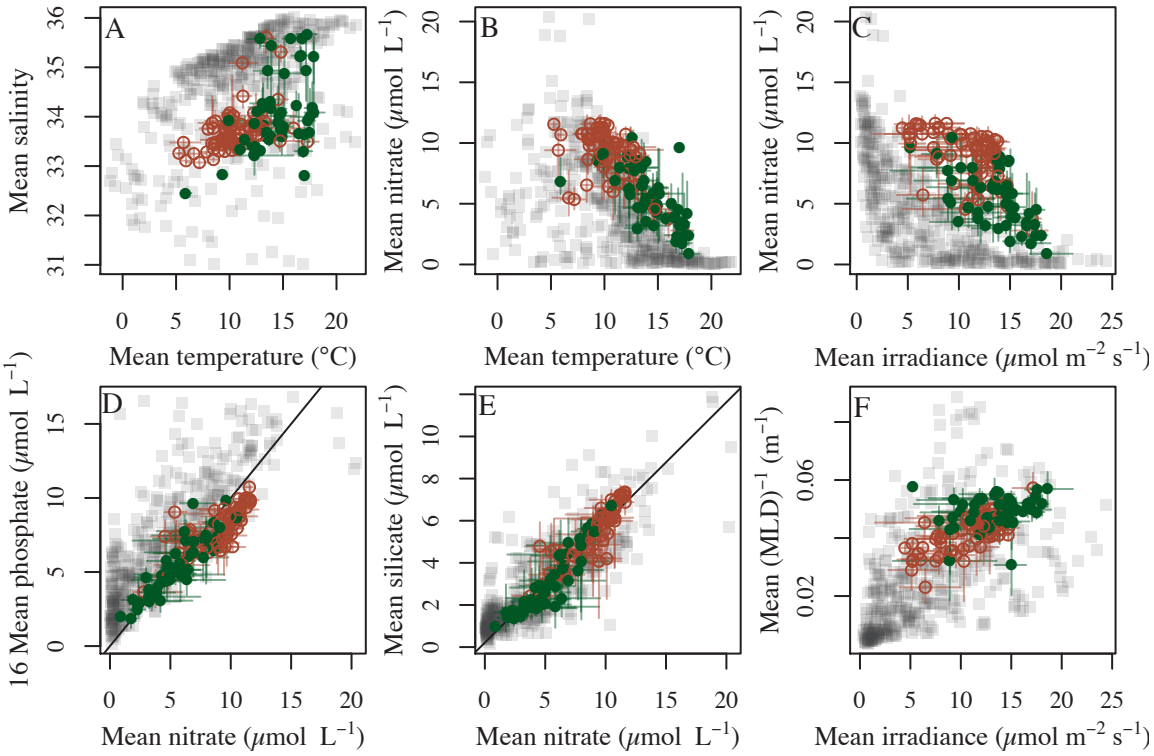
Figure 3. Mean and breadth of the univariate realized niche (Eqs. 3-4) for (A) SST (°C), (B) surface salinity, (C) surface nitrate (μmol L^{-1}), (D) surface phosphate (μmol L^{-1}), (E) surface silicate (μmol L^{-1}), and (F) mean irradiance over the mixed layer (μmol m^{-2} s^{-1}) for 69 diatoms (open red circles) and 50 dinoflagellates (filled green circles). Colored lines indicate the 95% confidence interval on each parameter from bootstrap resampling. The histograms at the bottom of each figure shows the relative abundance of background data. The dashed line separates generalist (above the line) and specialist species (*see* Methods).

Figure 4. Mean niches of pairs of environmental variables for each species (diatoms in open red circles, dinoflagellates in filled green circles) against the background distribution of climatological environments (overlapping grey squares). Pairs of variables are (A) SST (°C) and salinity, (B) SST and nitrate concentration (μmol L^{-1}), (C) mean irradiance (μmol m^{-2} s^{-1}) and nitrate, (D) nitrate and 16·phosphate (μmol L^{-1}) with a 1:1 Redfield line, (E) nitrate and silicate concentration (μmol L^{-1}) with a regression line through the niches, (F) mean irradiance and $(MLD)^{-1}$ (m^{-1}). Colored lines indicate the 95% confidence interval on each parameter.


True positive rate
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Approximate false positive rate


A
B
C
D
Logistic probability
Temperature (°C)
Salinity
Nitrate (μmol L−1)
Mean PAR (μmol m−2 s−1)


A
B
C
D
E
F
σSST (°C)
μSST (°C)
σSalinity
μSalinity
σNO3 (μmol L−1)
μNO3 (μmol L−1)
σPO4 (μmol L−1)
μPO4 (μmol L−1)
σSiO2 (μmol L−1)
μSiO2 (μmol L−1)
σMean PAR (μmol m−2s−1)
μMean PAR (μmol m−2s−1)


A
B
C
D
E
F
Mean salinity
Mean nitrate (μmol L−1)
Mean temperature (°C)
Mean irradiance (μmol m−2 s−1)
16 Mean phosphate (μmol L−1)
Mean silicate (μmol L−1)
Mean nitrate (μmol L−1)
Mean (MLD)−1 (m−1)