Ecological Modelling 269 (2013) 1–8

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Which environmental factors control phytoplankton populations? A Bayesian variable selection approach

Crispin M. Mutshinda[a,*], Zoe V. Finkel[b], Andrew J. Irwin[a]

[a] Mathematics & Computer Science, Mount Allison University, 67 York Street, Sackville, NB E4L 1E6, Canada
[b] Environmental Science Program, Mount Allison University, Sackville, NB, Canada

ARTICLE INFO

Article history:
Received 16 December 2012
Received in revised form 26 June 2013
Accepted 29 July 2013

Keywords:
Bayesian variable selection
Environmental forcing
Multicollinearity
Overfitting
Phytoplankton

ABSTRACT

The structure of phytoplankton communities is thought to influence total productivity, trophic structure and the export of carbon below the mixed layer. Community structure is determined by a complex interaction between the physiological characteristics of each species, environmental conditions, resource availability, competition among species, and numerous loss terms. This complexity makes it very difficult to predict how changes in environmental conditions will alter the structure of phytoplankton communities. Here we develop a hierarchical Bayesian model with variable selection to identify how temperature, salinity, irradiance, and macronutrient concentrations determine the abundance of the 67 dominant identified species at Station CARIACO in the Caribbean Sea. This approach allows us to overcome the statistical challenge presented by the highly correlated environmental variables. Approximately three-quarters of the variables for each species have little effect on phytoplankton abundance. About half of the species decline in abundance with increasing temperature. Diatom species' abundances are much more likely to respond to changes in irradiance and nitrate concentration than dinoflagellates and dinoflagellate species' abundances are more likely to respond to changes in salinity.

1. Introduction

Phytoplankton are the base of the marine food web and are an integral part of the global carbon cycle, sequestering carbon in the deep ocean (Field et al., 1998; Falkowski et al., 2000). Changes in climate, notably ocean temperature and availability of nutrients, are expected during the next century (Boyd et al., 2010; Finkel et al., 2010). Each phytoplankton species has its own characteristic ecophysiological traits which determine how it responds to the environment (Hutchinson, 1957; Litchman and Klausmeier, 2008; Schwaderer et al., 2011). As environmental conditions change, they become more or less favorable to the individual species in a community. At Station CARIACO in the southern Caribbean Sea, gradual changes in many factors including temperature and macronutrient concentrations have been documented over the past 15 years (Taylor et al., 2012).

Niche differentiation among the phytoplankton means that at a site such as Station CARIACO, some species will find near optimal conditions for growth, while others will be close to either their upper or lower tolerances for temperature, salinity, irradiance, or macronutrient concentrations. As a consequence, each individual species can be expected to exhibit an increase, decrease, or no change in response to increases or decreases in each environmental variable. A practical way to study these anticipated changes for many species is through the analysis of time-series data with detailed taxonomic information on the phytoplankton communities (Irwin et al., 2012).

A straightforward regression analysis of the abundance of many phytoplankton species as a function of environmental conditions is unlikely to succeed. The physical variables, temperature and salinity, and the availability of resources are highly correlated and so estimated effects will also be correlated. Including a variable in a statistical analysis which has no effect on a particular species, but is correlated with an important variable will increase the uncertainties in estimated regression parameters and diminish the utility of the model. It is quite possible to have a regression model fit the data well with a high R^2 or F-statistic, while none of the regression coefficients are significantly different from zero.

A solution to this problem is to omit variables which are not important on a species by species basis using Bayesian variable selection (BVS). A new indicator variable is introduced for each potential explanatory variable, to enable each predictor to be included or excluded for each species. The indicators are assigned independent Bernoulli priors with success probability 0.5, so that before the data are incorporated into the model, we are

* Corresponding author. Tel.: +1 506 364 2633.
E-mail addresses: cmutshinda@mta.ca, mutshinda@gmail.com, crissmwanza@yahoo.fr (C.M. Mutshinda).

0304-3800/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2013.07.025

Table 1
Summary of the environmental variables from the Cariaco Ocean Time-series.

Variable	Units	Mean	S.D.	Range
Irradiance	mol m^{-2} d^{-1}	9.99	9.93	0.00–29.30
Temperature	°C	24.14	2.34	19.38–30.06
Salinity	psu	36.78	0.17	35.84–37.07
NO_3	μmol L^{-1}	2.27	3.59	0.00–29.80
PO_4	μmol L^{-1}	0.19	0.21	0.00–3.13
$SiOH_4$	μmol L^{-1}	2.23	3.43	0.00–63.94
pH	–	8.02	0.06	7.86–8.12

indifferent about whether each variable is or is not important. These prior probabilities are updated using the data to measure the strength of evidence that a variable should be included in the model.

Here we develop a statistical model to identify the variables which influence phytoplankton abundances at Station Cariaco. A hierarchical Bayesian model with variable selection circumvents the most troublesome aspects of a regression analysis with correlated predictors and makes this analysis possible. For each species we determine which environmental variables are important and estimate the linear effects on species log-abundance expected with a small change in each environmental condition. The analysis enables us to find similarities and differences among species and functional groups of phytoplankton in their responses to potential environmental changes. We summarize our results by sketching out the likely consequences for this phytoplankton community of small changes in environmental conditions.

2. Materials and methods

2.1. Description of the data

Phytoplankton abundances (cells L^{-1}) and environmental data were collected as part of the CARIACO Ocean time series (www.imars.usf.edu/CAR/) (Muller-Karger et al., 2001, 2004). The environmental variables used here are water temperature (°C), salinity, irradiance (mol m^{-2} d^{-1}) and the concentration (in μmol L^{-1}) of dissolved nutrients namely, nitrate, phosphate, and silicic acid. The data were recorded in 169 monthly cruises spanning November 1995 through January 2011 at seven different depths (1 m, 7 m, 15 m, 25 m, 35 m, 55 m, and 75 m). Irradiance was estimated from Sea-viewing Wide-Field-of-View (SeaWiFS) satellite-derived monthly sea-surface PAR (mol m^{-2} d^{-1}) and attenuated over depth using the diffuse attenuation coefficient k_{490} (m^{-1}) obtained from Giovanni (http://disc.sci.gsfc.nasa.gov/giovanni) from a $0.4^\circ \times 0.4^\circ$ box around the CARIACO Ocean Time Series station from October 1997 to December 2010. The monthly SeaWiFS average was used for months outside this period. Irradiance at depth was attenuated according to the Beer–Lambert law, and averaged over the mixed layer (0–25 m). The range and variability of different environmental variables at Station CARIACO over the study period are summarized in Table 1. We use phytoplankton abundance data from the 67 most abundant species from six higher taxonomic groups namely, diatoms, dinoflagellates, coccolithophorids, cyanobacteria, ciliates and silicoflagellates (Table 2). Phytoplankton abundances are frequently below the detection limit. Instead of counting all of the missing abundances as zero, we imputed them with half of the minimum observable threshold (5×10^{-3} cells L^{-1}), following (Mutshinda et al., 2013). The missing data for environmental variables were imputed with the mean value of the corresponding months based on the full dataset. All environmental variables were standardized to have zero mean and unit variance before the analysis.

Table 2
List of the most frequently observed phytoplankton species at Station CARIACO from November 1995 through January 2011, with taxonomic authorities and functional group classification. Names without authorities are not in algaebase.org or marinespecies.org.

Diatom
Bacteriastrum delicatulum Cleve
Bacteriastrum sp. Shadbolt
Cerataulina pelagica (Cleve) Hendey
Chaetoceros affinis Lauder
Chaetoceros anastomosans Grunow
Chaetoceros compressus Lauder
Chaetoceros decipiens Cleve
Chaetoceros didymus Ehrenberg
Chaetoceros lorenzianus Grunow
Chaetoceros sp. Ehrenberg
Chaetoceros sp.2
Cylindrotheca closterium (Ehrenberg) Reiman & Lewin
Dactyliosolen fragilissimus (Bergon) Hasle apud G.R. Hasle & Syvertsen
Eucampia zodiacus Ehrenberg
Guinardia delicatula (Cleve) Hasle
Guinardia flaccida (Castracane) Peragallo
Guinardia striata (Stolterfoth) Hasle
Haslea wawrikae (Husedt) Simonsen
Helicotheca tamesis Ricard
Hemiaulus hauckii Grunow in Van Heurck
Hemiaulus sinensis Greville
Lauderia annulata Cleve
Leptocylindrus danicus Cleve
Leptocylindrus mediterraneus (H. Peragallo) Hasle
Leptocylindrus minimus Gran
Navicula sp. Bory de Saint-Vincent
Navicula yarrensis Grunow
Nitzschia fluminensis Grunow
Nitzschia longissima (Brébisson in Kützing) Ralfs in Pritchard
Proboscia alata (Brightwell) Sundström
Pseudo-nitzschia pseudodelicatissima (Hasle) Hasle
Pseudo-nitzschia pungens (Grunow ex P.T. Cleve) Hasle
Pseudo-nitzschia seriata (P.T. Cleve) H. Peragallo in H. & M. Peragallo
Pseudo-nitzschia sp. H. Peragallo in H. & M. Peragallo
Pseudo-nitzschia subfraudulenta (Hasle) Hasle
Rhizosolenia hebetate J.W. Bailey
Rhizosolenia imbricate Brightwell
Rhizosolenia setigera Brightwell
Rhizosolenia styliformis Brightwell
Skeletonema costatum Greville (Cleve)
Thalassionema delicatula
Thalassionema frauenfeldii (Grunow)
Thalassionema nitzschioides (Grunow) Mereschkowsky
Thalassiosira gravida P.T. Cleve
Thalassiosira rotula Meunier
Thalassiosira sp. Cleve
Thalassiosira subtilis (Ostenfeld) Gran

Dinoflagellate
Gonyaulax polygramma Stein
Gymnodinium mitratum Schiller
Gymnodinium sp. Stein
Gyrodinium fusus (Meunier) Akselman
Gyrodinium sp. Kofoid & Swezy
Heterocapsa triquetra (Ehrenberg) Stein
Neoceratium lineatum (Ehrenberg) F. Gomez, D. Moreira & P. Lopez-Garcia
Prorocentrum gracile Schütt
Prorocentrum micans Ehrenberg
Scrippsiella sp. Balech ex A.R. Loeblich III
Scrippsiella trochoidea (Stein) Balech ex A.R. Loeblich III

Coccolithophore
Calcidiscus leptoporus (Murray & Blackman) Loeblich & Tappan
Calcidiscus sp. Kamptner
Calciopappus caudatus Gaarder & Ramsfjell
Calciosolenia murrayi Gran
Emiliania sp. Hay & Mohler, in Hay, Mohler, Roth, Schmidt & Boudreaux + *Gephyrocapsa* spp. Kamptner

Cyanobacteria
Synechococcus sp. Nägeli
Trichodesmium thiebautii Gomont

Silicoflagellate
Dictyocha fibula Ehrenberg

Ciliate
Mesodinium rubrum (Lohmann, 1908)

2.2. Preliminary analysis

We fit a linear regression of the natural logarithm of species abundances on the standardized environmental variables to test if a simple model could describe the variation in phytoplankton abundance. The linear regression model failed to reveal any pattern in the data, with none of the regression coefficients being statistically different from zero. This indicates a more elaborate model is required to analyze this complex dataset.

2.3. Model specification

Let $n_{s,k,d}$ and $x_{j,k,d}$ denote, respectively, the natural logarithm of the abundance (cells L^{-1}) of species s and the value of the jth environmental variable in sample (cruise) k at depth d. We assume that

$$n_{s,k,d} \sim N(\mu_{s,k,d}, \sigma_s^2), \tag{1}$$

where

$$\mu_{s,k,d} = \alpha_s + \sum_{j=1}^{J} \gamma_{s,j}\beta_{s,j}x_{j,k,d}. \tag{2}$$

In this equation, α_s is a species-specific intercept, $\beta_{s,j}$ is the effect of the jth environmental variable, x_j, on the log-abundance of species s, σ_s^2 is the error variance specific to species s, and $\gamma_{s,j}$ is an indicator variable that takes the value 1 if the jth environmental variable is a relevant predictor of the abundance of species s, and the value 0 otherwise.

The model is developed and fitted with a Bayesian approach (Gelman et al., 2004; McCarthy, 2007), which requires explicit statements of prior distributions on all unknown quantities. Letting $\beta_s = (\beta_{s,1}, \ldots, \beta_{s,j})$ denote the vector of environmental effects on the log-abundance of species s, we assume that $\beta_s \sim \text{MVN}(0, \Omega)$ *a priori*, where MVN(a, B) denotes the multivariate normal distribution with mean vector a and covariance matrix B. We further assume that $\Omega \sim InvWish(J, I_{J\times J})$, where $I_{J\times J}$ and $InvWish(k, R)$ denote, respectively, the $J \times J$ identity matrix and the inverse Wishart distribution with scale matrix R and k degrees of freedom. For model details on the Wishart distribution, see (Gelman et al., 2004, pp. 87–88). The priors on the remaining parameters are defined as follows: $\gamma_{s,j} \sim Bernoulli(0.5)$ independently for $1 \le s \le S$ and $1 \le j \le J$, $\alpha_s \sim N(0, \sigma_\alpha^2)$, $\sigma_\alpha^2 \sim Gamma(1, 1)$, $\sigma_s^2 \sim Gamma(a, b)$, $a \sim Gamma(1, 1)$, and $b \sim Gamma(1, 1)$.

The indicators $\gamma_{s,j}$ are the tools for variable selection, and the $Bernoulli(0.5)$ prior independently imposed on each of them assumes prior odds of 1:1 for inclusion *vs* exclusion of each environmental variable as a predictor of the abundance of any species. These prior odds are updated by the data into posterior odds to decide whether or not a variable, x_j, is an important predictor of the abundance of a given species s.

Let $H_{s,j}^1$ and $H_{s,j}^0$ denote, respectively, the hypotheses "x_j is an important predictor of the abundance of species s" and "x_j is not an important predictor of the abundance of species s". The amount of the support provided by the data in favor of the inclusion of x_j as a predictor of the abundance of species s can be evaluated by the Bayes factor (Kass and Raftery, 1995)

$$B_{s,j} = \frac{Pr(\gamma_{s,j} = 1|\text{data})/Pr(\gamma_{s,j} = 0|\text{data})}{Pr(\gamma_{s,j} = 1)/Pr(\gamma_{s,j} = 0)} \tag{3}$$

with $B_{s,j} > 1$ implying more support than assumed *a priori* and vice versa. The strength of evidence in favor of $H_{s,j}^1$ against $H_{s,j}^0$ is evaluated on the following scale (Jeffreys, 1961). $B_{s,j} < 1$: evidence against $H_{s,j}^1$ (i.e., support for $H_{s,j}^0$); $1 < B_{s,j} \le 3$: weak support for $H_{s,j}^1$ (against $H_{s,j}^0$); $3 < B_{s,j} \le 10$: substantial support for $H_{s,j}^1$; $10 \le B_{s,j} < 100$: strong support for $H_{s,j}^1$, and $B_{s,j} > 100$: decisive support for $H_{s,j}^1$.

If $Bernouli(0.5)$ priors are independently imposed on the indicators $\gamma_{s,j}$, the Bayes factor $B_{s,j}$, boils down to posterior odds $\Pr(\gamma_{s,j} = 1|\text{data})/\Pr(\gamma_{s,j} = 0|\text{data})$, and the Jeffreys' cut-off points 1, 3, and 10 correspond, respectively, to posterior inclusion probabilities 0.5, 0.75, and 0.91. We consider $\Pr(\gamma_{s,j} = 1|\text{data}) \ge 0.75$ corresponding to $B_{s,j} \ge 3$ under our prior specification as providing substantial evidence that the jth environmental variable is an important predictor of the abundance of species s. Symmetrically, we consider $\Pr(\gamma_{s,j} = 1|\text{data}) \le 0.25$ corresponding to $B_{s,j} \le 1/3$ as substantial evidence that the jth environmental variable is not an important predictor of the abundance of species s.

2.4. Model fitting and statistical analyses

We use Markov chain Monte Carlo simulation (MCMC) (Gilks et al., 1996), implemented in OpenBUGS (Thomas et al., 2006), to sample from the joint posterior of the model parameters. We ran 25,000 iterations of two parallel Markov chains and discarded from each Markov chain the first 5000 samples as burn-in, thinning the remainder to each 10th sample.

The variables presumed important by Bayesian variable selection are not always the ones with statistically non-zero coefficients, i.e., with 95% credible intervals excluding 0. A large posterior uncertainty, reflected in wide 95% credible intervals, may mean some posterior estimates are statistically zero even though they are selected as important variables. Bayesian variable selection will generally single out the influential predictors, even in cases where a high estimation uncertainty may prevent the corresponding coefficients from being statistically non-zero. We contrast our Bayesian variable selection results with an identical model except without the variable selection mechanism (Mutshinda et al., 2013). We refer to the model of Mutshinda et al. (2013) and to the model presented here as Model 1 and Model 2, respectively. We compare the set of predictors presumed important by BVS under Model 2 to the set of variables with statistically non-zero coefficients *a posteriori* from both Model 1 and Model 2. Let u and v denote two Boolean vectors of length n. We introduce the following matching index

$$\varphi_{u,v} = \sum_i \frac{I(u_i = v_i)}{n}, \tag{4}$$

where $I(\cdot)$ denotes the indicator function taking the value 1 when its argument is true, and 0 otherwise. For each environmental variable, we construct three Boolean vectors VS (variable selection), E1 (Model 1 effects) and E2 (Model 2 effects). VS is 1 if the variable is selected as an important predictor of the abundance of a particular species and is 0 otherwise. E1 is 1 if the 95% credible intervals of environmental effect of a particular species estimated from Model 1 excludes 0. E2 is similar to E1 but is based on the estimates from Model 2.

3. Results

The abundances of the 67 individual phytoplankton species studied from the Cariaco Ocean Time-series are determined by different combinations of environmental variables (Fig. 1). Irradiance, temperature, and salinity are each important for 21–27 species while any one of the macronutrients are only important for 7–12 species. For eight species, no variables are identified as important indicating that our environmental variables are not predictive of variation in abundance for these species. There is quite a bit of variation in the combinations of variables determined to be important for predicting the abundance of any particular species, but there is taxonomic structure beyond the species level, largely consistent with higher phylogenetic groupings.

A dendrogram constructed from a 0/1 matrix of variable importance for diatom and dinoflagellate species identifies four major

Table 3
Summary of predictor importance and sign of the effect on abundance: number and proportion of 42 diatom, 8 dinoflagellate, and 9 other species for which each environmental variable is important, and the number of species with positive or negative effects (species abundance increases or decreases) with increasing values for each variable.

Environmental variable	Variables important for # (%) species of			% of species	
	Diatom	Dinoflagellate	Other species	With + effect	With − effect
Irradiance	24 (57%)	0 (0%)	3 (33%)	0	100
Temperature	19 (45%)	4 (50%)	4 (44%)	11	89
Salinity	16 (38%)	5 (62%)	2 (22%)	78	22
Nitrate	12 (29%)	0 (0%)	1 (11%)	85	15
Phosphate	8 (19%)	0 (0%)	3 (33%)	73	27
Silicic acid	5 (12%)	3 (38%)	2 (22%)	70	30

clades (Fig. 2). Species in Clade 1 are all diatoms and none have irradiance as an important variable. Clade 2 is also all diatoms, but all have irradiance as an important variable and none have phosphate or silicic acid concentration as important variables. Clade 3 contains most of the dinoflagellate species and none of the species have irradiance or nitrate as important predictors. None of the species in Clade 4 have temperature or silicic acid as an important variable. None of the dinoflagellates have irradiance, nitrate or phosphate concentration identified as important, while more than half of the diatom species have irradiance as an important variable (Table 3). In general, dinoflagellates are likely to respond to changes in temperature, salinity, or silicic acid concentration and diatoms are likely to respond to changes in irradiance and nitrate concentration. There are not enough species of coccolithophores, cyanobacteria, ciliates, and silicoflagellates to report taxonomic trends with confidence.

Our model also indicates whether species increase or decrease in abundance following an increase in each important variable (Fig. 3). Increases in temperature and irradiance are overwhelmingly likely to lead to a decrease in abundance while increases in the other variables indicate increases in abundance for most species (Table 3). While increases in macronutrient concentration indicate increases in abundance for most species, two or three species are expected to have lower abundance under increased nutrient concentrations. The magnitudes of the environmental effects on log abundance vary across the diatoms and dinoflagellates. The effects of increasing temperature and irradiance are strongly negative for diatoms but these effects are generally weak among the dinoflagellates. Similarly, the broadly positive effects of increasing salinity and nitrate concentration on diatom abundance are much weaker in dinoflagellates. Interestingly three dinoflagellate species are expected to increase in abundance with an increase in silicic acid concentration, while most diatoms do not respond to silicic acid and among the five species that do, one is expected to decrease in abundance following an increase in silicic acid. We interpret counterintuitive results such as this as an ecological signal incorporating competition for resources and unassessed factors such as grazing rates.

The effects ($\beta_{s,j}$) of each environmental variable can be compared directly with an identical analysis except for the absence of the variable selection mechanism (Mutshinda et al., 2013). Bayesian variable selection omits unimportant variables from the model so that they cannot interfere with the estimation of the effects of the important variables. Just under three-quarters of the variables are identified as unimportant for each species (an average of 1.7 variables per species or 111 combinations of variables and species out of a total of 402 possible combinations are deemed important; Fig. 1). When unimportant correlated predictors are included in a regression, effects tend also to be correlated, so omitting unimportant predictors should reduce this spurious correlation. The effects of environmental predictors are similarly correlated between the models (Table 4) indicating, for example, that the effects of nitrate and phosphate concentration on log abundance are uncorrelated regardless of whether variable selection is used or not. Four notable exceptions are dramatic reductions in the magnitude of the correlation between the effects of temperature and nitrate concentration, temperature and silicic acid concentration, salinity and phosphate concentration, and salinity and irradiance. This indicates that removing unimportant variables has reduced spurious correlations that were present in the previous analysis and improved the model.

Table 4
Correlations, ρ, between the posterior distributions of model effects estimated using the Bayesian model with variable selection (top row), and without variable selection (bottom row).

	Temp	$SiOH_4$	PO_4	NO_3	Salinity
$SiOH_4$	−0.16 −0.32				
PO_4	−0.10 −0.17	−0.16 −0.16			
NO_3	−0.36 −0.65	0.20 0.27	−0.04 0.00		
Salinity	−0.50 −0.59	0.16 0.27	0.05 0.51	0.37 0.51	
Irradiance	0.41 0.50	−0.08 −0.10	−0.04 −0.01	−0.33 −0.30	−0.03 −0.24

The greatest disagreements between importance as determined by variable selection and significance as determined by 95% credible intervals on posterior mean effects are found for temperature (76% agreement) using the model without variable selection and for salinity (85% agreement) using the model in this paper (Table 5). Overall 34 species-variable combinations were identified as important but with effects not statistically different from zero under the model presented in this paper. It may also happen that Bayesian variable selection excludes a variable whose effect is statistically non-zero but whose impact on the response variable is negligible. This occurred only with 2 species-variable combinations, out of a total of 402 combinations examined.

Table 5
Matching probabilities of the Boolean vectors indicating the pertinence (1) or not (0) of each environmental variable as a predictor of species abundance. The pertinence of each variable is computed three ways: using the 95% credible intervals of environmental effects from the model without (E1) and with (E2) variable selection, and using a posterior inclusion threshold of 0.75 from the variable selection model (VS).

	Irradiance	Temperature	Salinity	Nitrate	Phosphate	Silicic acid
E1/E2	0.82	0.75	0.85	0.70	0.90	0.79
E1/VS	0.90	0.76	0.97	0.78	0.93	0.90
E2/VS	0.93	0.96	0.85	0.93	0.91	0.90

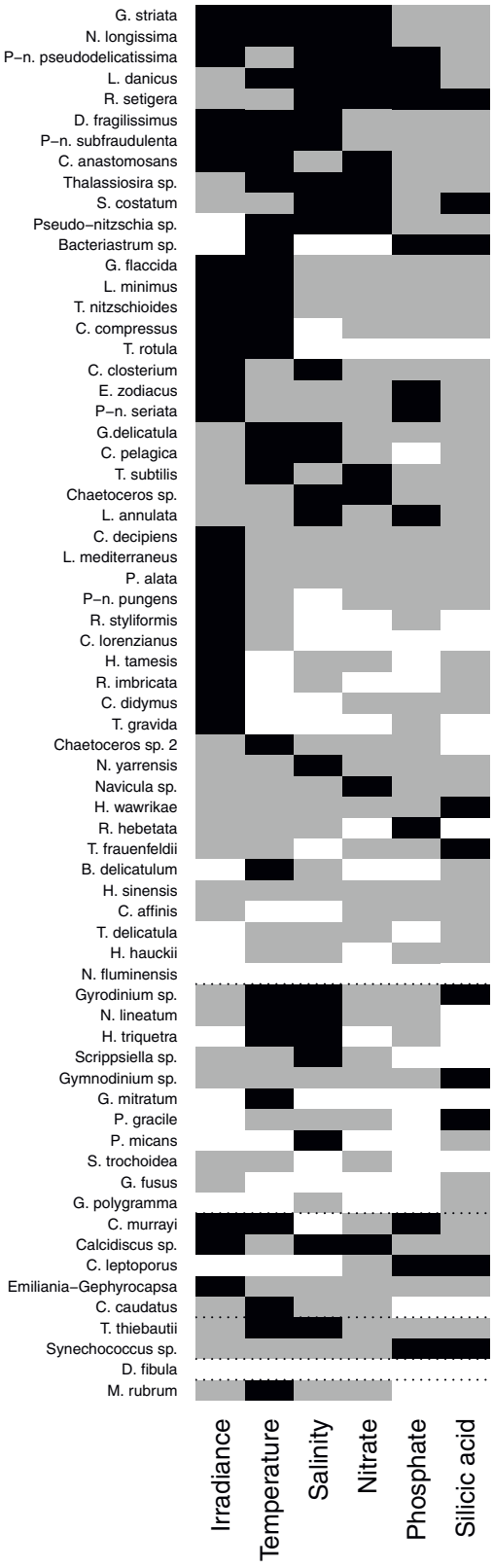

Fig. 1. Bayesian variable selection results.
Heat map of the variable selection results for the dominant phytoplankton species on a three point scale: important (black, posterior inclusion probability ≥ 0.75), not important (white, posterior inclusion probability ≤ 0.25), and uncertain (gray, all other cases). Species are listed according to the number of important variables within higher taxonomic groupings which are separated by horizontal dashed lines, with diatoms at the top, as in Table 2.

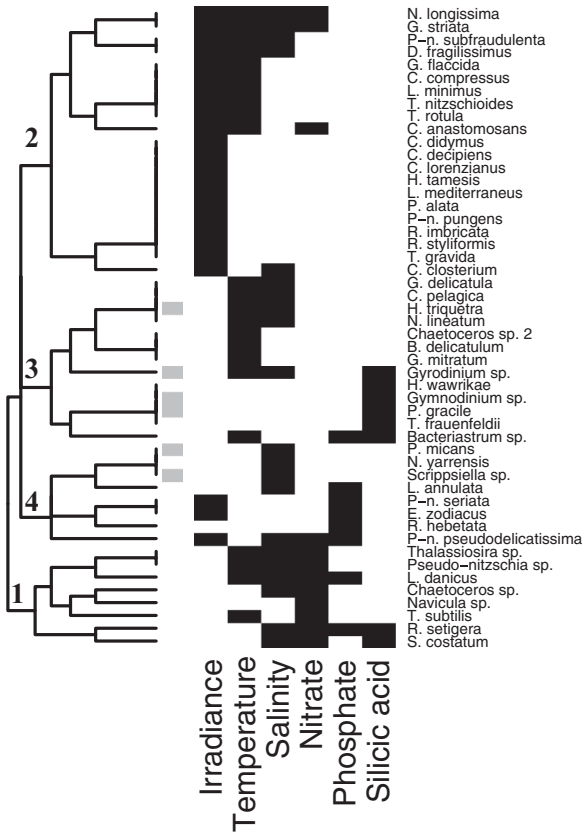

Fig. 2. Hierarchal clustering of diatoms and dinoflagellates based on variable selection results.
Dendrogram of diatoms and dinoflagellates based on the variable selection results shown in the binary heat map (important, black; excluded, white). Dinoflagellates are identified by a gray block in the left column closest to the dendrogram. The four major clades are identified with a number on the tree.

4. Discussion

Over the past fifteen years there has been a gradual warming in sea-surface temperature of about 1 °C at Station CARIACO, gradual decreases in macronutrient concentrations near the surface, decreases in chlorophyll concentration and rates of primary production and changes in the structure of the phytoplankton community (Taylor et al., 2012). The most dramatic change occurred in 2005, and resulted in a 30–100-fold decrease in the abundance of many phytoplankton species. These changes are likely due to a combination of changes in resource availability and trophic control, but the effects of each variable are not known. Many ecological studies use regression analyses to investigate the effects of particular explanatory variables on a response variable of interest; however, the complex nature of ecological data, in particular multicollinearity, presents challenges for standard statistical regression tools such as ordinary least squares fitting and statistical significance tests (Seip and Reynolds, 1995). A linear regression of the log species abundances on the environmental variables through ordinary least squares produced no significant results for any of the environmental variables, and yielded no insight into the environmental regulation of species abundances. Here we used a hierarchical Bayesian model with variable selection to identify which environmental variables are important determinants of the abundance of each individual species and how changes in the variables affect the abundance of individual phytoplankton species. A hierarchical Bayesian model without variable selection (Mutshinda et al., 2013) estimated effects on abundance similar to those reported

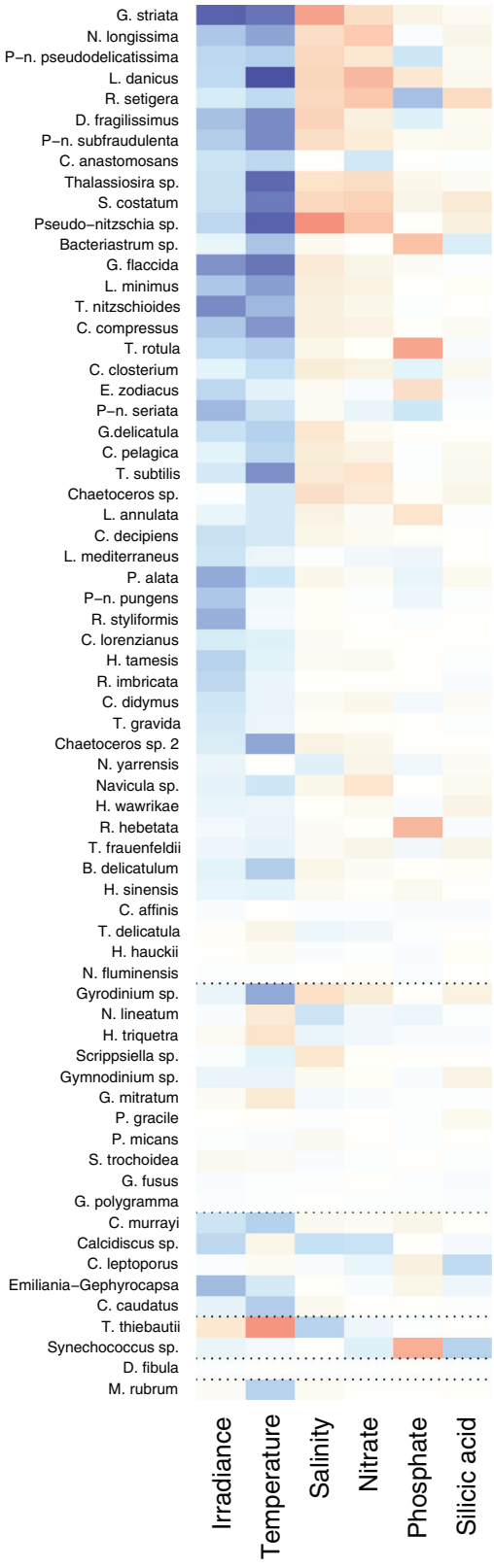

Fig. 3. Heat map of the environmental effects.
Heat map of the posterior mean environmental effects on the log abundances of the 67 dominant phytoplankton species at Station CARIACO estimated from the Bayesian model involving variable selection. The heat scale ranges from blue (negative) through white (mean effect 0) to red (positive) with the lighter colors representing mean effects closer to 0. Species are listed according to the number of important variables within higher taxonomic groupings which are separated by horizontal dashed lines, with diatoms at the top, as in Table 2.

here (Fig. 3), but did not provide a clear way to decide which variables are important for each species.

Bayesian variable selection shows that an average of 1.7 variables (of 6 tested) are important for each species so that overall only about 25% of the variables are important determinants of phytoplankton abundance. Irradiance, temperature, and salinity were the variables most likely to be important influences on phytoplankton abundance. There is taxonomic structure in both the variables selected as important for each species and the sign of the effects of important variables on phytoplankton abundance. We identified four major clades according to which variables were important determinants of abundance for diatoms and dinoflagellates (Fig. 2). Clades 1 and 2 are composed entirely of diatoms and are roughly characterized by irradiance being unimportant and nitrate concentration important (Clade 1) or by the importance of irradiance and the unimportance of nutrients (Clade 2). In Clade 3, irradiance and nitrate concentration are both unimportant, while in Clade 4, temperature and silicic acid are unimportant. The dinoflagellates were restricted to Clades 3 and 4, with most of them in Clade 3. The differences and rank order of the important environmental predictors for diatoms and dinoflagellates from the Cariaco ocean time-series are broadly consistent with patterns observed for diatoms and dinoflagellates from the North Atlantic CPR data, using a very different analysis (Irwin et al., 2012). Increases in temperature and irradiance almost always led to a decrease in abundance of the species at Station CARIACO (Fig. 3 and Table 3). As a result we expect changes in phytoplankton community structure with future changes in climate over the next century, with diatoms likely to be more strongly affected than dinoflagellates.

At Station CARIACO, irradiance is an important variable for more than half of the diatoms, with abundance decreasing with increasing irradiance, while irradiance is not important for any of the dinoflagellate species (Figs. 1 and 3). Irradiance is generally not limiting at this tropical site, so it seems likely that diatom abundance is being reduced by photoinhibition (Alderkamp et al., 2010; Raven, 2011). Alternatively, a negative association may result from increasing light attenuation resulting from increased phytoplankton abundance rather than the other way around (Mutshinda et al., 2013). Nitrate concentration is the most likely to be an important predictor among the macronutrients and its effect on diatom abundance is generally to increase abundance with increasing nitrate concentration (Fig. 3). This suggests that nitrate is more likely to limit diatom abundance at Station CARIACO than either phosphate or silicic acid, consistent with previous analyses (Lane-Serff and Pearce, 2009). Species with positive nitrate coefficients tend to have a phosphate coefficient that is either nearly zero or negative and vice versa (Fig. 3), suggesting that many diatoms are exhibiting limitation by only one of these two nutrients. Silicic acid is an important predictor for the abundance of only a few diatoms. This may seem surprising since diatoms require silicic acid to build their frustules, as corroborated by the fact that all silicic acid coefficients are non-negative, except one (Fig. 3); however, a vital resource may be non-informative about a species' abundance if it is not in limited supply, which may be the case for silicic acid at Station CARIACO (Scranton et al., 2006).

Irradiance, phosphate, and nitrate are not important predictors for any of the dinoflagellate species. A plausible reason why light, which is not limiting phytoplankton abundance at Station CARIACO, may be less informative for the dinoflagellates is that most dinoflagellates are motile and as such, can maintain a favorable position in the water column to avoid photoinhibition (Samuelsson and Richardson, 1982; Jacobson, 1999). Similarly, phosphate and nitrate concentration are not important for dinoflagellate abundances possibly because dinoflagellates have the ability to migrate into deep nutrient-rich waters to replenish

their internal nutrient provisions when surface waters become nutrient-depleted (Raven and Richardson, 1984; Jeong et al., 2010). Moreover, many dinoflagellates are mixotrophic as they can combine photosynthesis and the ingestion of prey (Stoecker, 1999; Leterme et al., 2005; Seong et al., 2006). It is surprising that silicic acid is important for 3 dinoflagellate species. Some dinoflagellates, particularly mixotrophic species can co-occur with competitively superior photoautotrophic diatoms in silicate-rich environments by relying on phagotrophy (Hansen et al., 1994; Tittel et al., 2003). This may result in a positive association between silicic acid concentration and the abundances of some dinoflagellates, such as *Gymnodinium* sp. (Fig. 3), which is a small dinoflagellate known to be heterotrophic (Jakobsen and Hansen, 1997).

Our analysis is based on observational data which confines our inferences to statements of association between species abundances and the environmental variables. Controlled experiments cannot replicate the complexity of natural ecological systems. On the other hand, a statistical analysis of observational data can reveal complex relationships between species abundances and environmental variables. An association between a resource and the abundance of a particular species does not necessarily imply the species' direct limitation or inhibition by the resource. For example, some non-silicate-dependent phytoplankton may thrive at low silicate concentrations when subjected to less competition from diatoms (Moncheva et al., 2001). This may be the explanation for the negative association between silicate concentration and the abundance of the cyanobacterium *Synechococcus* sp. On the other hand, a species' abundance may be positively associated with a resource that is limiting to its competitor. This may explain the positive relationship between silicic acid concentration and the abundances of the dinoflagellates *Gymnodinium* sp., *Gyrodinium* sp., and *Prorocentrum gracile*.

The predictors presumed influential by Bayesian variable selection do not necessarily coincide with the variables associated with statistically non-zero coefficients (Murray and Conner, 2009). About 8% of the variables are important but have statistically zero effect. Bayesian variable selection generally picks out the influential predictors regardless of the level of posterior uncertainty, which in many cases is a result of data scarcity. Conversely, Bayesian variable selection can exclude some variables associated with statistically non-zero coefficients if their impacts on the response variable are negligible, but this occurred rarely in our field data. The difficulty of designing suitable experiments for analyzing ecological data and the inherent complexity of the underlying processes imply that the data available are often inadequate to address the questions of interest through standard statistical methods. Bayesian variable selection helps identify the important predictors with less data than may be required to make effects statistically different from zero. It quantifies the evidence in favor of the inclusion of different predictors in the model, so that predictors can be ranked in terms of their importance. It also helps avoid the common fallacy of treating a statistically significant predictor as important, even though its impact on the response variable is negligible.

Bayesian variable selection has allowed us to identify the environmental variables associated with changes in the abundance of the dominant phytoplankton observed at Station CARIACO. The analysis has revealed taxonomic structure in the predictors identified as important (Fig. 2) and the effects of changes in those predictors (Fig. 3). Dinoflagellates are relatively insensitive to temperature and irradiance, the variables most likely to be important, while diatoms are expected to decrease in abundance with increases in those variables. The model used here is a valuable tool for analyzing complex observational data in which many species are simultaneously responding to a wide range of changes in their environment. Possible elaborations of this approach would employ non-linear mechanistic functions for species responses to environmental variables, allow for interactions among variables, or introduce additional predictors such as estimates of grazing rates.

Acknowledgements

CMM was supported by an Ace-net post-doctoral fellowship and the Marjorie-Young Bell fund. AJI and ZVF were supported by NSERC Canada. We are grateful to the scientific, technical, and administrative staff of the CARIACO time series program. The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2013.07.025.

References

Alderkamp, A.-C., de Baar, H.J., Visser, R.J., Arrigo, K.R., 2010. Can photoinhibition control phytoplankton abundance in deeply mixed water columns of the Southern Ocean? Limnol. Ocean. 55, 1248.

Boyd, P.W., Strzepek, R., Fu, F., Hutchins, D.A., 2010. Environmental control of open-ocean phytoplankton groups: now and in the future. Limnol. Ocean. 55, 1353.

Falkowski, P.G., Boyle, E., Canadell, J., Canfield, D., Elser, J.J., Gruber, N., Hibbard, K., Hogberg, P., Linder, S., Mackenzie, F.T., Moore, B., Pedersen, T., Rosenthal, Y., Seitzinger, S., Smetacek, V., Steffan, W., 2000. The global carbon cycle: a test of our knowledge of the Earth as a system. Science 290, 291–294.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P.G., 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281, 237–240.

Finkel, Z.V., Beardall, J., Flynn, K.J., Quigg, A., Rees, T.A.V., Raven, J.A., 2010. Phytoplankton in a changing world: cell size and elemental stoichiometry. J. Plankton Res. 32, 119–137.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis, 2nd ed. Chapman & Hall/CRC, London, UK.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, London, UK.

Hansen, B., Bjornsen, P.K., Hansen, P.J., 1994. The size ratio between planktonic predators and their prey. Limnol. Ocean. 39, 395–403.

Hutchinson, G.E., 1957. The multivariate niche. Cold Spr. Harb. Symp. Quant. Biol. 22, 415–421.

Irwin, A.J., Nelles, A.M., Finkel, Z.V., 2012. Phytoplankton niches estimated from field data. Limnol. Ocean. 57, 787.

Jacobson, D.M., 1999. A brief history of dinoflagellate feeding research. J. Eukaryot. Microbiol. 46, 376–381.

Jakobsen, H.H., Hansen, P.J., 1997. Prey size selection, grazing and growth response of the small heterotrophic dinoflagellate *Gymnodinium* sp. and the ciliate *Balanion comatum*—a comparative study. Mar. Ecol. Prog. Ser. 158, 75–86, http://dx.doi.org/10.3354/meps158075.

Jeffreys, H., 1961. Theory of Probability. Oxford University Press, Oxford, UK.

Jeong, H.J., Du Yoo, Y., Kim, J.S., Seong, K.A., Kang, N.S., Kim, T.H., 2010. Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. Ocean Sci. J. 45, 65–91.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.

Lane-Serff, G.F., Pearce, R.B., 2009. Modeling hydrography and marine sedimentation in the Cariaco Basin since the Last Glacial Maximum. J. Geophys. Res. Oceans 1978–2012 (114), C04003, http://dx.doi.org/10.1029/2008JC005076.

Leterme, S.C., Edwards, M., Seuront, L., Attrill, M.J., Reid, P.C., John, A.W.G., 2005. Decadal basin-scale changes in diatoms, dinoflagellates, and phytoplankton color across the North Atlantic. Limnol. Ocean. 50, 1244–1253.

Litchman, E., Klausmeier, C.A., 2008. Trait-based community ecology of phytoplankton. Annu. Rev. Ecol. Evol. Syst. 39, 615–639.

McCarthy, M.A., 2007. Bayesian Methods for Ecology. Cambridge University Press, New York, USA.

Moncheva, S., Gotsis-Skretas, O., Pagou, K., Krastev, A., 2001. Phytoplankton blooms in Black Sea and Mediterranean coastal ecosystems subjected to anthropogenic eutrophication: similarities and differences. Estuar. Coast. Shelf Sci. 53, 281–295.

Muller-Karger, F., Varela, R., Thunell, R., Astor, Y., Zhang, H., Luerssen, R., Hu, C., 2004. Processes of coastal upwelling and carbon flux in the Cariaco Basin. Deep Sea Res. Part II Top. Stud. Ocean. 51, 927–943.

Muller-Karger, F., Varela, R., Thunell, R., Scranton, M., Bohrer, R., Taylor, G., Capelo, J., Astor, Y., Tappa, E., Ho, T.-Y., 2001. Annual cycle of primary production in the Cariaco Basin: response to upwelling and implications for vertical export. J. Geophys. Res. 106, 4527–4542.

Murray, K., Conner, M.M., 2009. Methods to quantify variable importance: implications for the analysis of noisy ecological data. Ecology 90, 348–355, http://dx.doi.org/10.1890/07-1929.1.

Mutshinda, C.M., Troccoli-Ghinaglia, L., Finkel, Z.V., Müller-Karger, F.E., Irwin, A.J., 2013. Environmental control of the dominant phytoplankton in the Cariaco basin: a hierarchical Bayesian approach. Mar. Biol. Res. 9, 247–261.

Raven, J.A., 2011. The cost of photoinhibition. Physiol. Plant 142, 87–104, http://dx.doi.org/10.1111/j.1399-3054.2011.01465.x.

Raven, J.A., Richardson, K., 1984. Dinophyte flagella: a cost–benefit analysis. New Phytol. 98, 259–276.

Samuelsson, G., Richardson, K., 1982. Photoinhibition at low quantum flux densities in a marine dinoflagellate (*Amphidinium carterae*). Mar. Biol. 70, 21–26.

Schwaderer, A.S., Yoshiyama, K., de Tezanos Pinto, P., Swenson, N.G., Klausmeier, C.A., Litchman, E., 2011. Eco-evolutionary differences in light utilization traits and distributions of freshwater phytoplankton. Limnol. Ocean. 56, 589.

Scranton, M.I., McIntyre, M., Astor, Y., Taylor, G.T., Müller-Karger, F., Fanning, K., 2006. Temporal variability in the nutrient chemistry of the Cariaco Basin. In: Past and Present Water Column Anoxia. Springer, Dordrecht, The Netherland, pp. 139–160.

Seip, K.L., Reynolds, C.S., 1995. Phytoplankton functional attributes along trophic gradient and season. Limnol. Ocean. 40, 589–597.

Seong, K.A., Jeong, H.J., Kim, S., Kim, G.H., Kang, J.H., 2006. Bacterivory by co-occurring red-tide algae, heterotrophic nanoflagellates, and ciliates. Mar. Ecol. Prog. Ser. 322, 85.

Stoecker, D.K., 1999. Mixotrophy among Dinoflagellates. J. Eukaryot. Microbiol. 46, 397–401.

Taylor, G.T., Muller-Karger, F.E., Thunell, R.C., Scranton, M.I., Astor, Y., Varela, R., Ghinaglia, L.T., Lorenzoni, L., Fanning, K.A., Hameed, S., Doherty, O., 2012. Ecosystem responses in the southern Caribbean Sea to global climate change. Proc. Natl. Acad. Sci. U. S. A. 109, 19315–19320, http://dx.doi.org/10.1073/pnas.1207514109.

Thomas, A., O'Hara, B., Ligges, U., Sturtz, S., 2006. Making BUGS open. R News 6, 12–17.

Tittel, J., Bissinger, V., Zippel, B., Gaedke, U., Bell, E., Lorke, A., Kamjunke, N., 2003. Mixotrophs combine resource use to outcompete specialists: implications for aquatic food webs. Proc. Natl. Acad. Sci. U. S. A. 100, 12776–12781, http://dx.doi.org/10.1073/pnas.2130696100.