Hydrobiologia
https://doi.org/10.1007/s10750-021-04789-2

PRIMARY RESEARCH PAPER

Bayesian two-part modeling of phytoplankton biomass and occurrence

Crispin M. Mutshinda · Aditya Mishra · Zoe V. Finkel · Claire E. Widdicombe · Andrew J. Irwin

Received: 3 May 2021 / Revised: 6 December 2021 / Accepted: 17 December 2021

Abstract Phytoplankton biomass data often involve zero outcomes preventing a description by continuous distributions with positive support such as the lognormal distribution commonly used to describe ecological data. Two usual solutions: ignoring the zeroes and adding a small positive number to all outcomes, induce bias and reduce predictive power. To address these shortcomings, we design a Bayesian two-part model with a binary component for presence or absence and a continuous component involving a lognormal model for non-zero biomass. We specify two equations relating species-specific occurrence probabilities and expected log-biomasses when present to potential covariates, with spike-and-slab priors imposed on linear effects to selectively discard the irrelevant predictors. We analyze the biomass data of 74 phytoplankton (57 diatoms and 17 dinoflagellates) recorded weekly at Station L4 (Western English Channel, UK) between April 2003 and December 2009, along with measurements of abiotic covariates. Our results disclose different combinations of environmental predictors for the occurrence and the biomass of individual species. Overall, the occurrence of dinoflagellates is associated with higher temperature and irradiance levels compared to diatoms, with virtually no dependence on nutrient concentrations. Irradiance emerges as the key predictor of biomass when species are present. Optimum temperatures for biomass accumulation and temperature sensitivities vary widely among and within functional types. Compared to one-stage models based on usual zero-handling approaches, our two-part model stands out with higher prediction accuracy. The two-part modeling approach provides a valuable framework for decoupling the predictors of species occurrence and abundance from observational data.

Keywords Bayesian inference · Delta distribution · Hurdle model · Over dispersion · Semi-continuous data · Stochastic search variable selection

Handling Editor: Alex Elliott.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10750-021-04789-2.

C. M. Mutshinda (✉) · A. J. Irwin
Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada
e-mail: crispin.mutshinda@dal.ca

A. Mishra
Flatiron Institute, Simons Foundation, New York, NY 10010, USA

Z. V. Finkel
Department of Oceanography, Dalhousie University, Halifax, NS, Canada

C. E. Widdicombe
Plymouth Marine Laboratory, Prospect Place, Plymouth PL1 3DH, UK

Published online: 28 January 2022

Introduction

Phytoplankton are the foundation of the aquatic food web and key players in global biogeochemical cycles of carbon and other elements (Field et al., 1998). Long-term monitoring data are increasingly used to analyze the environmental controls of phytoplankton biomass distribution and community structure, and anticipate the community response to climate change (Mutshinda et al., 2013a,b, 2016). Phytoplankton biomass data are often recorded on a continuous scale and typically involve a substantial proportion of zeroes and few extremely large values, resulting in right-skewed and heteroskedastic distributions for the non-zero values (e.g., Clarke & Green, 1988; Fletcher et al., 2005; Martin et al., 2005). Continuously distributed nonnegative values with a proportion of zeroes are said to be semi-continuous (e.g., Min & Agresti, 2002; Olsen & Shaffer, 2001; Wang et al., 2020). The presence of a point mass at zero in semi-continuous data prevents their description by usual positive continuous distributions such as the lognormal distribution commonly used to describe ecological data on theoretical and empirical grounds (e.g., Preston, 1948; May, 1975; McGill, 2003; Sugihara 1980). There are two common solutions: (1) adding a small positive number to all outcomes to get rid of the zeroes and proceed with a positive continuous distribution such as the lognormal distribution and (2) ignoring the zeroes, both of which induce bias and undermine the model's predictive power. Appropriate modeling approaches are required to address these flaws without compromising data integrity.

A valuable strategy for handling the zeroes in semi-continuous data is to consider the data as arising from the interplay of two distinct processes: the first process determines whether an outcome is strictly positive or zero, and, conditional on the outcome being positive, the second process determines its actual value (Su et al., 2009). Two-part models accommodate excess zeroes and asymmetries in the distribution of strictly positive values by combining a Bernoulli distribution for the binary indicator of whether an observation is strictly positive or zero and a positive continuous distribution, typically the lognormal distribution or the Gamma distribution for the strictly positive outcomes. The two-part modeling approach allows for explanatory variables to influence the outcome through their role in the binary and/or the continuous model components. For count data, excess zeroes can be handled using zero-inflated models (Lambert, 1992) describing the response variable as a mixture of a Bernoulli distribution and a count distribution supported on nonnegative integers, typically the Poisson or the negative binomial distribution, or hurdle models (Cragg, 1971) where the conditional distribution of non-zero values does not support zero.

The aim of this paper is to illustrate the value of two-part models for decoupling the abiotic predictors of species occurrence and biomass from semi-continuous monitoring data. We will use this model to test if the abiotic variables that influence presence (biomass greater than 0) are distinct from the variables that influence the biomass of cells present in a sample. We will also show how the two-part model improves predictions compared to a model considering only strictly positive biomass data. Since two-part models fall in the category of zero-modified models, we start by providing a synopsis of zero-modified models before delving into the details of our model. A dataset is said to be zero-inflated (zero-deflated) with regard to a hypothetical probability distribution if it involves more (fewer) zeroes than expected under the presumed distribution. While zero-deflation is uncommon in practice, zero-inflated data abound across disciplines. Following Neelon et al. (2016), we use the term zero-modified data to collectively refer to zero-inflated and zero-deflated data.

The challenges posed by zero-modified data are well-documented (e.g., Amemiya, 1974; Feng et al., 2014; Xu et al., 2015). Zero-modified models have been proposed to handle zero-modified data primarily in the context of count data. Zero-modified models for count data can be separated into zero-inflated models (e.g., Lambert, 1992; Mwalili et al., 2008) and hurdle models (e.g., Cragg, 1971; Mullahy, 1986), which all involve a Bernoulli distribution for the binary indicator of whether an outcome is strictly positive or zero, and a conditional count model. The fundamental difference between zero-inflated and hurdle models lies in the fact that conditional distributions under zero-inflated models, as opposed to hurdle models, are regular (untruncated) count distributions such as the Poisson distribution or the negative binomial distribution supporting not only strictly positive values but also zeroes. The zeroes arising from the conditional distribution are called sampling zeroes, in contrast to so-called structural zeroes resulting from the Bernoulli component. On the other hand, conditional distributions under hurdle models are zero-truncated count distributions such as the truncated Poisson and truncated negative binomial distributions (Cameron &Trivedi, 1998). The likelihood function of a hurdle model is separable with regard to the parameters of the binary and the assumed zero-truncated count model, meaning that the log-likelihood can be written as the sum of the log-likelihoods of the two model components, which can be maximized separately. This explains why hurdle models are called two-part or two-step models.

Although most of the attention regarding zero-augmented models has been devoted to count data, semi-continuous data (i.e., zero-inflated nonnegative continuous data) arise frequently across disciplines, but zero-modified models for semi-continuous data have not received commensurate attention. A variable is said to be semi-continuous if it takes either the value 0 or any value between a finite lower bound ($b \geq 0$) and an upper bound ($B > b$) which need not be finite. There is no semi-continuous counterpart to a zero-inflated count model, but delta-distributions (Aitchison, 1955; Aitchison & Brown, 1957, p. 95; Rubec et al., 2016) have been proposed as semi-continuous counterparts of hurdle models. A delta-distribution is a two-part model for semi-continuous data involving a Bernoulli distribution for the indicator of whether an outcome is positive or zero and a conditional distribution that is continuous on the positive real line. Typical choices of conditional distributions include the Gamma distribution and the lognormal distribution leading, respectively to the delta-gamma (Stefánsson, 1996) and delta-lognormal (Maunder & Punt, 2004; Pennington, 1983) models.

In this paper we design, under the Bayesian framework (McCarthy, 2007; Gelman et al., 2013), a two-part model to disentangle the abiotic predictors of the occurrence and biomass patterns of phytoplankton species from semi-continuous monitoring data. The model combines a binary model for the indicator of whether a species has positive or zero biomass (the occurrence model) and a quantitative model using a lognormal distribution for the strictly positive biomasses (the biomass model). The occurrence and biomass models involve equations relating, respectively, the occurrence probability of each species at any time and its expected log-biomass when observed to abiotic covariates, with embedded stochastic search variable selection (SSVS) mechanism (George & McCulloch, 1993; Mutshinda et al., 2009, 2011) to selectively discard the irrelevant predictors. We use our model to analyze the abiotic predictors of the occurrence and biomass patterns of 74 phytoplankton species at Station L4 (Western English Channel, UK) from semi-continuous biomass data and coincident measurements of five abiotic covariates expected to affect the occurrence probability as well as the growth rate and biomass density (temperature, irradiance, nitrogen, silicate and phosphate concentrations) recorded weekly between April 2003 and December 2009. We carry out the model fitting to data by Markov chain Monte-Carlo (MCMC) simulation (Gilks et al., 1996) implemented in OpenBUGS (Thomas et al., 2006).

Materials and methods

Description of data

The data comprise weekly biomass data (mg C m^{-3}) for 74 phytoplankton species including 57 diatoms and 17 dinoflagellates (Table S1, Online Supplementary Materials) recorded at Station L4 (50° 15.00’ N, 4° 13.02’ W) between April 2003 and December 2009, and measurements of five abiotic covariates namely temperature (°C), photosynthetically active radiation (PAR;mol m^{-2}d^{-1}), and concentration (μmolL^{-1}) of dissolved inorganic nitrogen (nitrate + nitrite), silicate, and phosphate. Station L4 is located in the Western English Channel about 10 nautical miles south-west of Plymouth, UK, with a water column depth of approximately 50 m (Harris, 2010). It is a typical temperate coastal site with well mixed waters during the autumn and winter months under low sea surface temperatures and relatively high nutrient concentrations, whereas spring and summer months are characterized by a weak stratification of the water column accompanied with declining nutrients and increasing sea surface temperature which typically peaks at 18 °C in August (Widdicombe et al., 2010). Phytoplankton samples were identified to genus or species and counted by microscopy and abundance data converted to biomass using conversion factors based on empirically established carbon to volume relationships (Menden-Deuer & Lessard, 2000).

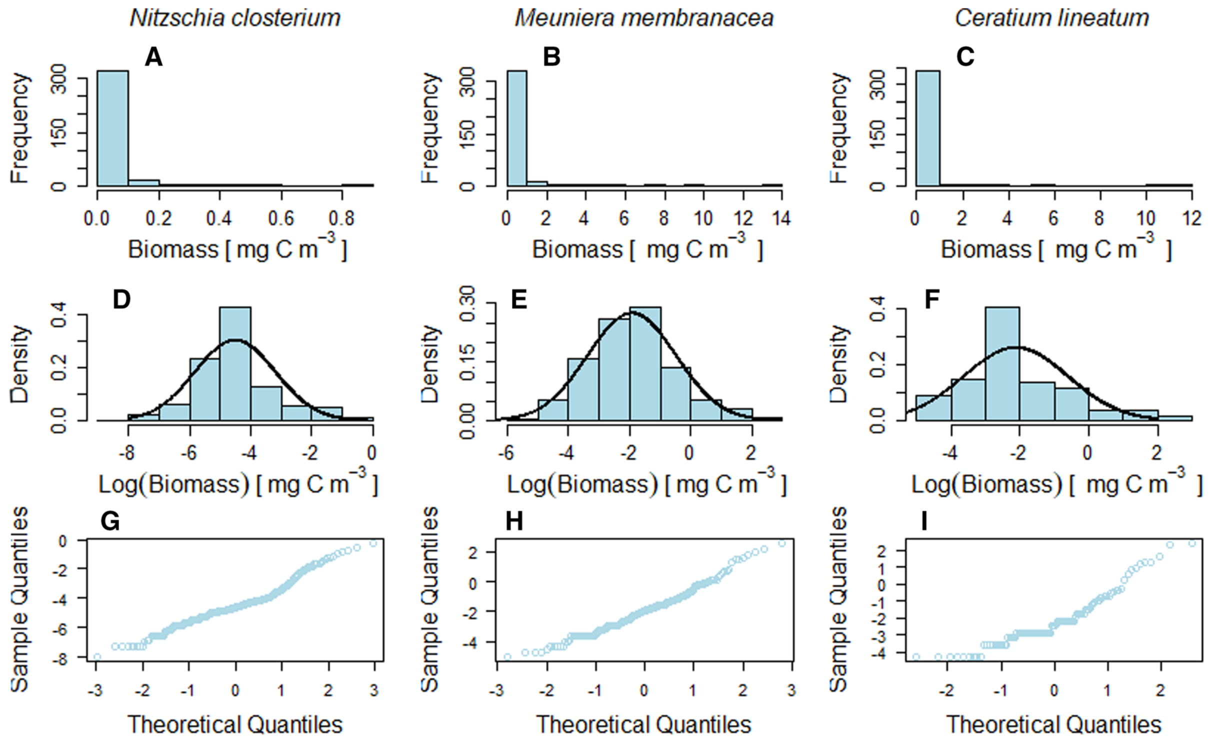


Fig. 1 Histograms of observed semi-continuous biomasses (top), histograms of the strictly positive biomass values on the natural logarithmic scale with overlaid fitted normal densities (middle), and normal Q-Q plots of log-biomasses (bottom) for the diatoms *Nitzschia closterium* (left) and *Meuniera membranacea* (center) and the dinoflagellate *Ceratium lineatum* (right)

Figure 1 illustrates the general structure of species-level biomass with data on three species: two diatoms (*Nitzschia closterium* and *Meuniera membranacea*) and a dinoflagellate (*Ceratium lineatum*). The histograms shown in panels A-C are L-shaped with a peak at zero indicating the zero-inflated nature of species-level biomass data. When the zeroes are omitted, the histograms of the strictly positive values on the natural logarithmic scale (panels D-F) and corresponding normal QQ plots (panels G-I) suggest that the lognormal distribution provides a reasonable description of the strictly positive values, lending support to the two-part delta-lognormal model for species-level biomass data.

Model specification

Let $p_{s,t}$ and $Y_{s,t}$ denote respectively the occurrence indicator and the observed biomass concentration (mg C m^{-3}) of species s at time t, so that $p_{s,t} = 1$ if $Y_{s,t} > 0$ and $p_{s,t} = 0$ if $Y_{s,t} = 0$. We model the probability density of $Y_{s,t}$ as

$$f\left(Y_{s,t}\right) = \left(1 - \pi_{s,t}\right)I\left(Y_{s,t} = 0\right) + \pi_{s,t}\, LN\left(\mu_{s,t}, \sigma_s^2\right) I\left(Y_{s,t} > 0\right) \qquad (1)$$

where $\pi_{s,t} = \Pr\left(p_{s,t} = 1\right)$, $I(.)$ is the indicator function taking the value 1 when its argument evaluates to true and the value 0 otherwise, and $\mathrm{LN}\left(\mu_{s,t}, \sigma_s^2\right)$ is the density function of the lognormal distribution with expected value $\exp\left(\mu_{s,t} + \sigma_s^2/2\right)$ and variance $\exp\left(2\mu_{s,t} + 2\sigma_s^2\right) - exp\left(2\mu_{s,t} + \sigma_s^2\right)$, implying that $y_{s,t} = \log\left(Y_{s,t} | Y_{s,t} > 0\right)$ is normally distributed with mean value $\mu_{s,t}$ and variance σ_s^2, where $\log(.)$ denotes the natural logarithm function.

Equation (1) describes the semi-continuous biomass data as arising from two entangled processes: a Bernoulli distribution (occurrence model component) governs the binary outcome of whether the outcome is strictly positive $\left(p_{s,t} = 1\right)$ or zero $\left(p_{s,t} = 0\right)$ and conditional on the binary outcome being 1, a lognormal distribution (biomass model component) describes the values of the strictly positive biomass. In order to explore the abiotic predictors of biomass distribution and community composition, we specify two separate models: the occurrence model and the biomass model relating respectively species-specific occurrence probabilities $\pi_{s,t} = \Pr(p_{s,t} = 1)$ and expected log-biomasses when present $\mu_{s,t} = \mathrm{E}[y_{s,t}|Y_{s,t} > 0]$ to abiotic covariates. It is worth emphasizing that the sets of covariates involved in the two model components are not required to be identical. We next describe the occurrence model and the biomass model which are both developed with a Bayesian approach.

The occurrence model

The occurrence model describes the occurrence indicator of species s at time t, $p_{s,t}$, by a Bernoulli random variable with success probability $\pi_{s,t}$, where the logit of $\pi_{s,t}$ depends linearly on the values assumed by the abiotic covariates at time t. Letting $Z_{j,t}$ denote the z-score of the j th environmental variable (temperature, irradiance, nitrogen, silicate, phosphate) at time t, we assume that

$$p_{s,t} \sim \text{ Bernoulli}\left(\pi_{s,t}\right) \tag{2}$$

$$logit\left(\pi_{s,t}\right) = \eta_s + \sum\nolimits_{j=1}^{J} \theta_{s,j} Z_{j,t} \tag{3}$$

where $\text{logit}\left(\pi_{s,t}\right) = \log\left(\frac{\pi_{s,t}}{1-\pi_{s,t}}\right)$, η_s and $\theta_{s,j}$ are respectively the intercept parameter specific to species s and the linear effect of the j th environmental variable on $\text{logit}\left(\pi_{s,t}\right)$.

The model specification is completed with explicit statements of prior distributions on all unknown quantities. We placed standard normal priors independently on the species-specific intercept parameters η_s, and assigned to each regression coefficient $\theta_{s,j}$ a hierarchical mixture prior with a "spike" and a "slab" Gaussian components (George & McCulloch, 1993; Mwanza 2010) to perform variable selection and identify promising subsets of environmental covariates. More specifically, we assume that

$$\theta_{s,j} \sim N(0, v_{s,j}) \tag{4}$$

$$v_{s,j} = \left(1 - \gamma_{s,j}\right) \times a + \gamma_{s,j} \times A \tag{5}$$

where the constants $a > 0$ and $A > 0$ representing the variances of the spike and the slab mixture components are respectively set to a very small and a large value to virtually exclude the irrelevant predictors from the model by constraining their coefficients to a narrow range around zero, while allowing the coefficients of relevant predictors to be freely estimated from data. The variable inclusion indicator $\gamma_{s,j}$ takes the value 1 when the jth covariate is required in the occurrence model of species s and the value 0 otherwise. This Bayesian variable selection method known as stochastic search variable selection (SSVS) was introduced by George & McCulloch (1993) in the context of linear regression, but has since been extended to other settings including generalized linear models. Finally, we assigned Bernoulli (0.5) priors on the variable inclusion indicators γ_{sj} independently for all species and all potential covariates, implying prior odds of 1:1 for including *versus* excluding each covariate in the occurrence models of individual species. The prior distributions of all unknown quantities are updated with the data information into posterior distributions which serve as basis for Bayesian inferences. As a rule of thumb for variable selection, we consider all predictors with posterior inclusion probability 0.75 or larger to be relevant and *vice-versa*. Compared to the prior inclusion probability of 0.5 assumed here, a posterior inclusion probability of 0.75 corresponds to a Bayes factor of 3 in favor of variable inclusion. On the Jeffrey's scale of evidence for interpreting Bayes factors (Jeffreys, 1961) as amended by Kass & Raftery (1995), a Bayes factor of 3 for hypothesis H_1 against H_2 implies positive evidence for H_1.

The biomass model

Conditionally on species s being observed at time t with biomass $Y_{s,t}$, the biomass model assumes that $Y_{s,t}$ follows a lognormal distribution with location parameter $\mu_{s,t}$ and scale parameter σ_s^2, where $\mu_{s,t}$ depends on abiotic variables. Letting $Z_{k,t}$ denote the z-score of the k th environmental variable (PAR, nitrogen, silicate, and phosphate) and T_t denote the observed temperature (°C) at time t, it follows that

$$[Y_{s,t}|Y_{s,t} > 0] \sim \mathrm{LN}(\mu_{s,t}, \sigma_s^2) \tag{6}$$

$$\mu_{s,t} = \alpha_s + \sum\nolimits_{k=1}^{K} \beta_{s,k} Z_{k,t} - \delta_s |T_t - \rho_s| \tag{7}$$

where $\beta_{s,k}$ is the linear effect of the k th environmental variable on the expected log-biomass of species s, representing the change in $\mu_{s,t}$ corresponding to the effect of a 1 standard deviation change in the corresponding variable, everything else being held constant. The parameter ρ_s represents the optimum temperature for the biomass of species s and $\delta_s > 0$ is the temperature sensitivity parameter quantifying the increase in the expected log-biomass of species s for a 1 °C change in temperature towards ρ_s, and alternatively, the decrease thereof for a 1 °C change in temperature away from ρ_s at average values of other abiotic variables. The species-specific intercept α_s represents the expected log-biomass of species s at its optimum temperature when all resources (irradiance and nutrients) are at their average values over the time series. Ecological data are often fraught with observation errors, which may result from the sampling techniques. In our model, observation error is lumped together with the residual environmental variance σ_s^2. However, we expect the observation error to be small since the sample size is large and the sampling technique remained the same throughout the study period.

The model is developed with a Bayesian approach, which requires explicit statements of prior distributions on all unknown quantities. We independently assigned on the species-specific optimal temperatures for biomass ρ_s, normal priors centered at the average temperature over the time series, 13 °C, with variance 10, and independent $Gamma(1,1)$ priors on the temperature sensitivity coefficients, δ_s. We placed exchangeable $InvGamma(u, v)$ priors on the species-specific variance parameters σ_s^2, with independent $Gamma(1,1)$ priors on the hyper-parameters u and v. To uncover the relevant environmental predictors of species biomass patterns, we assigned, as in the occurrence model, spike-and-slab priors (George & McCulloch, 1993; Owusu et al., 2016) on the linear effects $\beta_{s,k}$ of the environmental variables on the expected log-biomass. That is,

$$\beta_{s,k} \sim \mathrm{N}(0, w_{s,j}) \tag{8}$$

$$w_{s,k} = (1 - \xi_{s,k}) \times b + \xi_{s,k} \times B \tag{9}$$

where the constants $b > 0$ and $B > 0$ representing the variances of the spike and the slab parts of the mixture prior are respectively selected to be small and large to perform variable selection. Since temperature is typically one of the most informative variables in biomass models, we anticipated that it will be an important predictor in the biomass model which we parameterized in terms of species-specific optimum temperatures and temperature sensitivities to be freely estimated from the data. A zero temperature sensitivity would indicate that deviations from the optimal temperature do not induce a significant change in biomass of the focal species.

Model validation and model comparison

Model validation is a crucial part of the statistical modeling workflow. It involves the assessment of the model's adequacy at describing the data, by exploring potentially deficient aspects (if any) and finding ways of remedying them, independently of any other model. In the Bayesian framework, a standard approach to model validation is based on the notion of posterior predictive checks (Gelman et al., 1996, 2013), which relies on the ability to approximate the posterior predictive distribution, i.e., the distribution of unobserved values conditional on observed data. For a model with likelihood $L(y|\boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$, the posterior predictive distribution of an observable $\tilde{y}$ conditional on data y is defined by

$$p(\tilde{y}|y) = \int L(\tilde{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \tag{10}$$

In Eq. (10), it is assumed that $\tilde{y}$ and y are conditionally independent given $\boldsymbol{\theta}$. Since all parameters are integrated out, only information in observed data contributes to the prediction. The optimal Bayesian prediction under a quadratic loss function is the posterior predictive mean value $E[\tilde{y}|y]$.

The idea behind posterior predictive checks is to compare the observed data with replicated data simulated from the posterior predictive distribution, or alternatively to compare some test quantity $T(y, \theta)$ based on the observed data to the same statistic $T(y^{rep}, \theta)$ for replicated data from the posterior predictive distribution, and interpret systematic discrepancies between the two as evidence of model misfit. The comparison may be done visually using, for instance histograms, or formally using posterior predictive P-values or Bayesian P-values defined as

$$P_B = \Pr(T(y^{rep}, \theta) \geq T(y, \theta | y) \tag{11}$$

(Gelman et al., 1996, 2013). If the observed data are consistent with the model predictions, then P_B should be close to 0.50. Values of P_B close to 0 or 1 provide evidence for model inadequacy with P_B close to 0 indicating a lack of fit and P_B values close to 1 pointing to overfitting, which may occur when a model is needlessly too complex. A simulation-based approximation of the posterior predictive P-value is obtained as the proportion of posterior predictive data replicates for which the test quantity exceeds the one based on the original data. At the most basic level, posterior predictive checks involves the comparison of observed data to their posterior predictions. If the model fits the data, the replicated data should closely resemble the observations (Gelman et al., 1996). As a result, the difference $\epsilon = (y - y^{rep})$ between each data point, y, and its posterior predictive replicate, y^{rep}, should be distributed around zero.

When it comes to selecting among plausible models, cross-validation (e.g., Stone, 1974; Hastie et al., 2009) remains the most commonly used method for identifying the model with best out-of-sample predictive performance and presumably the model that best mimics the data-generating mechanism. The cross-validation procedure for model selection involves the following steps: (i) split the data into a training set and a validation set or test set; (ii) fit each competing model to the training data holding out the validation data; (iii) use the fitted model to predict the test data and compute a measure of predictive performance; (iv) select the model with the best overall performance. A widely used performance measure is the root mean squared predictive error (RMSPE) defined as

$$RMSPE = \left(\frac{1}{N} \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 \right)^{0.5} \tag{12}$$

where y_i and $\tilde{y}_i$ are respectively the i th omitted value and its model prediction, and N is the number of held-out values. The model with lowest RMSPE is preferable.

It is worth noting that biomass prediction under the two-part model integrates information from the occurrence and the biomass model components according to the following property: if X is a semi-continuous variable such that $E[X|X > 0] = \mu$ and $\Pr(X = 0) = \theta$, then $E[X] = (1 - \theta)\mu$. Stated otherwise,

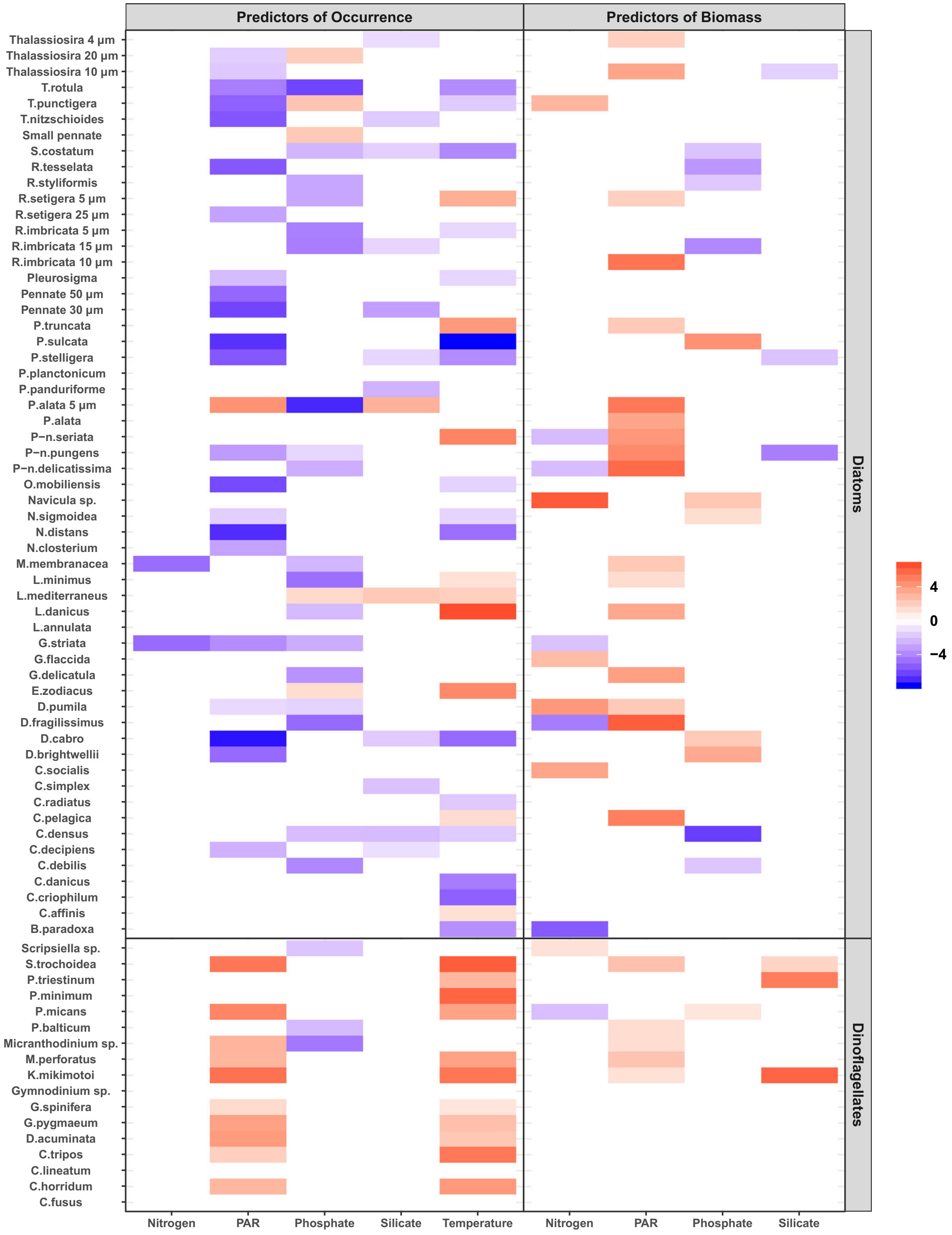
$$E[X] = E[X|X > 0]\{1 - \Pr(X = 0)\} \tag{13}$$

We assessed the validity of our two-part model hereafter identified as Model 1 through posterior predictive checks (Gelman et al., 1996, 2013). We subsequently compared our model's out-of-sample predictive performance in terms of RMSPE (Eq. 12) to two one-stage models based on prevailing zero-handling methods namely, adding a small positive number to all outcomes (Model 2) and ignoring the zeroes (Model 3), which is virtually identical to the continuous part of our two-part model, except that the latter model is conditioned on the binary response.

We used Markov chain Monte-Carlo (Gilks et al., 1996) implemented in OpenBUGS (Thomas et al., 2006) to simulate from the joint posterior distributions under both the occurrence model and the biomass model. We ran 30,000 iterations of three parallel Markov chains starting from scattered initial values and discarded the first 10,000 iterations of each Markov chain as burn-in, thinning the remainder by a factor of 10. We assessed the convergence of the Markov chains informally through visual inspection of traceplots and autocorrelation plots.

Results

The results of the occurrence model component revealed different combinations of environmental predictors for the presence of individual phytoplankton species at Station L4, with temperature, irradiance and phosphate concentration standing out as key predictors (Fig. 2A). The temperature and irradiance coefficients were consistently positive among dinoflagellates, indicating high occurrence probability at higher temperature and irradiance levels, with virtually no dependence on nutrient concentrations. In contrast, temperature and irradiance effects on occurrence probabilities of individual diatoms were broadly negative, implying higher occurrence probability at lower temperature and irradiance levels. Of all the nutrients under consideration, phosphate emerged as the most important predictor of species occurrence in


Predictors of Occurrence
Predictors of Biomass
Thalassiosira 4 µm
Thalassiosira 20 µm
Thalassiosira 10 µm
T.rotula
T.punctigera
T.nitzschioides
Small pennate
S.costatum
R.tesselata
R.styliformis
R.setigera 5 µm
R.setigera 25 µm
R.imbricata 5 µm
R.imbricata 15 µm
R.imbricata 10 µm
Pleurosigma
Pennate 50 µm
Pennate 30 µm
P.truncata
P.sulcata
P.stelligera
P.planctonicum
P.panduriforme
P.alata 5 µm
P.alata
P−n.seriata
P−n.pungens
P−n.delicatissima
O.mobiliensis
Navicula sp.
N.sigmoidea
N.distans
N.closterium
M.membranacea
L.minimus
L.mediterraneus
L.danicus
L.annulata
G.striata
G.flaccida
G.delicatula
E.zodiacus
D.pumila
D.fragilissimus
D.cabro
D.brightwellii
C.socialis
C.simplex
C.radiatus
C.pelagica
C.densus
C.decipiens
C.debilis
C.danicus
C.criophilum
C.affinis
B.paradoxa
Scripsiella sp.
S.trochoidea
P.triestinum
P.minimum
P.micans
P.balticum
Micranthodinium sp.
M.perforatus
K.mikimotoi
Gymnodinium sp.
G.spinifera
G.pygmaeum
D.acuminata
C.tripos
C.lineatum
C.horridum
C.fusus
Diatoms
Dinoflagellates
Nitrogen
PAR
Phosphate
Silicate
Temperature
Nitrogen
PAR
Phosphate
Silicate
4
0
−4

◀**Fig. 2** Heatmaps conveying the sign and magnitude of posterior mean effects of relevant covariates (posterior inclusion probability ≥ 0.75 implying Bayes factor for inclusion ≥ 3 under the assumed 0.5 prior inclusion probability) for (**A**) the occurrence model (Eqs. 2–5) and (**B**) the biomass model (Eqs. 6–9). The heat scale ranges from blue (negative effect) through white (no effect) to red (positive effect), with color intensity reflecting the effect magnitude according to the displayed color bar scale. The horizontal line separates diatoms from dinoflagellates. Nitrogen represents the combined concentration of nitrate and nitrite

diatoms and was selected in 40% of them, with negative coefficients in most cases.

One of the premises of the two-part modeling of ecological data is that different ecological mechanisms may underlie the occurrence of individual species and their biomass when present, which may result in different sets of covariates being associated with species presence and their biomass when present, as it turns out to be the case for our study species at Station L4. The results of our biomass model displayed in Fig. 2B indicate that irradiance is the most important and most frequently selected predictor of biomass when species are present, with a positive effect for roughly 30% of all species including both diatoms and dinoflagellates, implying higher biomass at high irradiance levels. Phosphate and nitrogen were each selected as important predictors of biomass for 20% of the diatoms, with a mix of positive and negative effects. Silicate was selected as an important predictor of biomass for three diatoms and three dinoflagellates with negative coefficients for the diatoms and positive coefficients for the dinoflagellates.

Our parameterization of the biomass model allowed us to estimate species-specific optimum temperatures, along with temperature sensitivities. Temperature sensitivities vary considerably among and within functional types (Fig. 3). Some of the species with high temperature sensitivity such as the diatoms *Pseudo-nitzschia delicatissima*, *Chaetoceros socialis Leptocylindrus minimus*, *Leptocylindrus danicus*, *Thalassiossira* sp.4 µm, *Skeletonema costatum* and the dinoflagellates *Prorocentrum minimum*, *Prorocentrum balticum*, *Karenia mikimotoi*, and *Prorocentrum micans* are among the species identified by Widdicombe et al. (2010) as being most responsible for the patterns of abundance observed at Station L4 over the 15-year period (1992–2007) covered by their study.

In diatoms, optimum temperatures for biomass spanned the entire range (7 °C–19 °C) of observed temperatures at Station L4 over the study period. The substantial disparities in species-specific optimum temperatures suggest the existence of distinct thermal niches. Based on the posterior distributions of species-specific optimum temperatures as displayed in Fig. 4, diatoms can be separated into low-temperature species with posterior median optimum temperatures lower than 12 °C (from *S. costatum* to *M. membranacea* in Fig. 4), mid-range temperature species with posterior median of optimum temperatures between 12 and 16 °C (from Small *pennate* to *C. affinis* in Fig. 4), and high temperature species whose posterior medians of optimum temperature range from 16 to 18 °C (from *P.-n. pungens* to *L. mediterraneous* in Fig. 4).

On the other hand, optimum temperatures for individual dinoflagellates range from intermediate (posterior medians between 12.5 °C and 15 °C) to high (posterior medians larger than 16 °C). Intermediate-temperature dinoflagellates involve species from *Scripiscella* sp. to *M. perforatus* in Fig. 4, while high temperature species go from *G. pygmaeum* to *P. minimum.* In the latter category, *Prorocentrum minimum* and *Karenia mikimotoi* stand out with exceptionally high optimum temperatures (posterior medians larger than 18 °C).

We assessed the validity of our model through posterior predictive checks by comparing simulated data y^{rep} from the posterior predictive distribution to the observed data y, and found the posterior predictive P-value $P_B = \Pr(y^{rep} \geq y)$ to be 0.48, which is close to the target value of 0.5 and far from the extremes 0 and 1, indicating that the model predictions do not systematically underestimate or overestimate the observed data.

After the model validation, e.g., through posterior predictive checks, it is often useful to compare its performance to alternative models embodying different hypotheses with regard to suitable performance measures (Conn et al., 2018). The model with best out-of-sample predictive performance is usually preferred, as a common goal is to optimize predictive ability. We compared the prediction accuracy of our two-part model (Model 1) against two single-stage models based on prevailing zero-handling approaches, namely Model 2 which involves the addition of a small

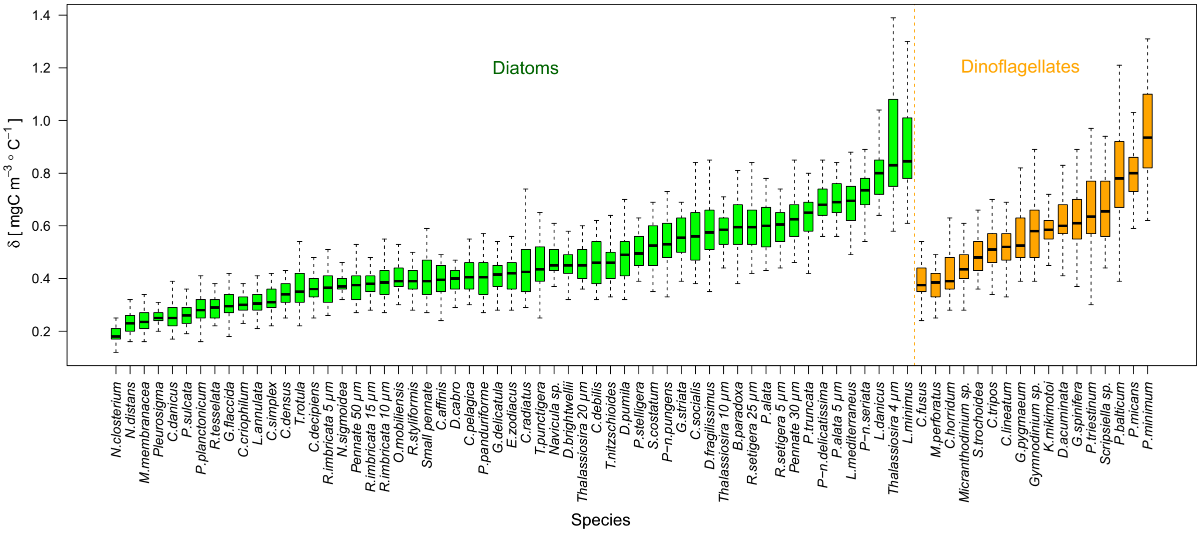


Fig. 3 Box-and-whisker plots summarizing the posterior distributions of species-specific temperature sensitivity coefficients, δ. Fill colors represent each species' functional group, green for diatoms and orange for dinoflagellates. The height of the box indicates the 25th ($Q1$) and 75th ($Q3$) per-centiles; the horizontal line inside the box is median, and the lower and upper whisker limits are defined as $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, respectively, where IRQ represents interquartile range. Within each functional group, species are sorted by increasing median of the temperature sensitivity parameter

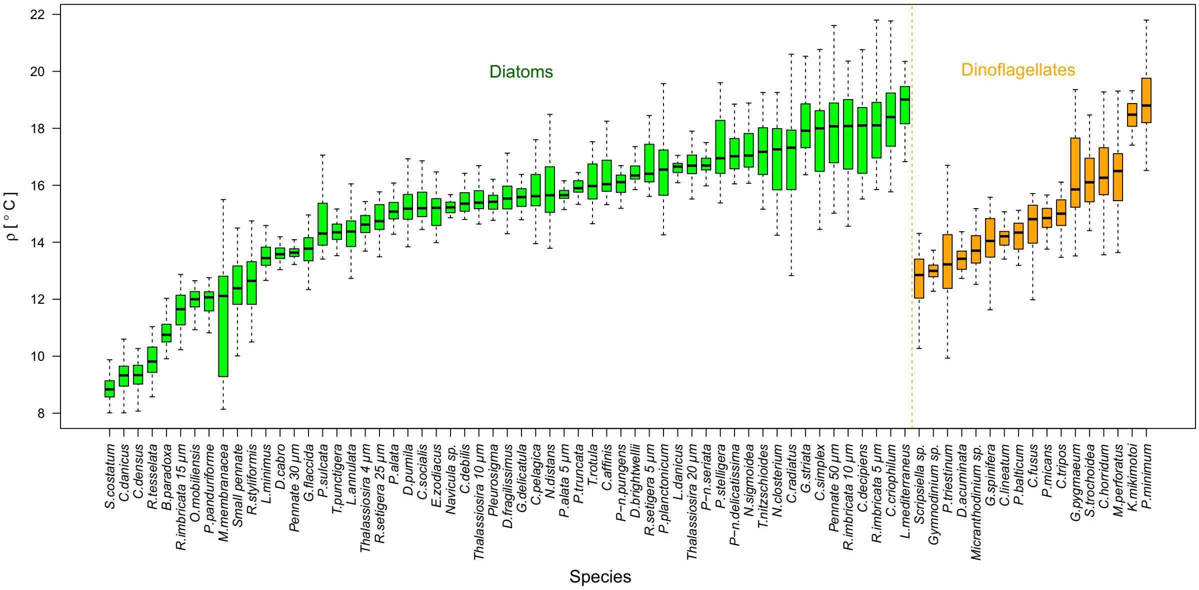


Fig. 4 Box-and-whisker plots summarizing the posterior distributions of species-specific optimum temperatures for biomass accumulation, ρ. Fill colors represent each species' functional group, green for diatoms and orange for dinoflagellates. The height of the box indicates the 25th ($Q1$) and 75th ($Q3$) per-centiles; the horizontal line inside the box is median, and the lower and upper whisker limits are defined as $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, respectively, where IQR represents interquartile range ($IQR = Q3 - Q1$). Species are sorted by increasing median of optimum temperature of biomass within each functional type

positive number ε to all response values to avoid zero outcomes (Model 2) and ignoring the zeroes (Model 3). For the purpose of the present analysis, we used $\varepsilon = 1.18 \times 10^{-5}$mg C m^{-3}), which represents half of the minimum observed biomass for either functional type. We omitted 10 randomly selected outcome values (5 positive values and 5 zeroes) and evaluated the model performance at predicting the omitted values through the root mean squared prediction error (Eq. 12). For Model 2, we compared the predicted values over zero outcomes to ε. In computing the RMSPE. The RMSPE of 2.20 mg C m^{-3} under our two-part model on the scale of the observed data was 30–50% lower than the RMSPEs 4.30 mg C m^{-3} under Model 2 considering the observed the data plus a positive offset to get rid of zero outcome and 3.06 mg C m^{-3} under Model 3 considering only the strictly positive outcomes unconditionally ignoring zero outcomes.

The two-part model (Model 1) reduces the prediction error by weighing the prediction from Model 3 by the occurrence probably estimated from the binary component of the two-part model according to Eq. 13. Model 3 is ostensibly identical to the continuous component of our two-part model, but the two models differ fundamentally in that the continuous part model is conditioned on the binary response when the response in Model 3 was wrongly assumed to be unconditioned.

Discussion

Semi-continuous data consisting of a point mass at zero followed by a right-skewed continuous distribution supported on strictly positive real numbers arise frequently across disciplines, particularly in ecology. Positive continuous distributions such as the gamma and the lognormal distributions cannot be directly fitted to semi-continuous data since their support excludes zero. Models based on ad hoc zero-handling techniques perform poorly at fitting data or predicting new data. Two-part models harness the flexibility to accommodate excess zeroes by separately modeling the probability that an outcome is non-zero and describing the non-zero outcomes by a positive continuous distribution. This two-step modeling approach allows for different ecological mechanisms to determine the occurrence of a species in a specific environment and its abundance when present, by relating species-specific occurrence probabilities and expected log-biomasses of species that are observed to potentially different sets of covariates.

In this paper, we designed a two-part Bayesian model to disentangle the abiotic predictors of the occurrence and biomass patterns of phytoplankton species from semi-continuous monitoring biomass data and coincident measurements of abiotic covariates expected to affect species' occurrence, and biomass. Our two-part model integrated a Bernoulli distribution for the indicator of whether a species has positive or zero biomass and a lognormal distribution for the strictly positive biomasses, with two equations, Eqs. (3) and (7) relating respectively the occurrence probability of each species at any time and its expected log-biomass when observed to abiotic covariates. We parameterized the biomass model in terms of species-specific optimum temperature and temperature sensitivity, two traits we expected to be important in shaping species biomass patterns given the critical role of temperature in phytoplankton ecology. We linearly related the four other biotic variables (PAR, nitrogen, silicate, phosphate) to each species' expected log-biomass and imposed spike-and-slab hierarchical priors on linear effects of environmental covariates to perform variable selection by constraining the effects of irrelevant covariates to be virtually zero while freely estimating the relevant effects from the data.

We found that at Station L4 the occurrence of individual species is governed by different combinations of environmental variables with temperature, irradiance and phosphate concentration standing as the most important predictors of species occurrence (Fig. 2A). The sign and magnitude of the coefficients of abiotic variables deemed relevant by the Bayesian variable selection mechanism included in the occurrence model imply greater occurrence probability of dinoflagellates as a group at higher temperature and irradiance levels, with virtually no dependence on nutrient concentrations. This result supports the documented tendency of dinoflagellates to thrive in warmer, stratified and nutrient-depleted waters in contrast with diatoms (Le Quéré et al., 2005), and corroborates the findings of previous analyses of the L4 phytoplankton data (Mutshinda et al., 2017, 2019). In diatoms, irradiance effects on species' occurrence probabilities are broadly negative, in line with the notion that diatoms are generally adapted to low light levels (Reynolds, 2006). On the other hand, temperature effects vary widely among the diatom species with negative effects for 15% of them, implying higher occurrence probability at lower temperatures and positive effects for about the same proportion of species indicating higher occurrence probability at higher temperatures. This result substantiates the documented diversity of diatoms (Armbrust, 2009; Mutshinda et al., 2020), with collections of species adapted to different and even contrasting conditions. Of all nutrients under consideration, phosphate emerged as the most important predictor of species occurrence in diatoms, with mostly negative coefficients when deemed relevant.

For the biomass component of our two-part model, irradiance stood out as the most important predictor among all resources under consideration, with positive effects for roughly a third of the study species, indicating higher biomass with increasing irradiance. Phosphate and nitrogen (nitrate + nitrite) appeared to be important predictors of biomass for roughly 20% of the diatoms, with a mix of positive and negative effects. In dinoflagellates, nitrogen concentration emerged as an important predictor of biomass for three species, with negative effect for one of them (*Prorocentrum micans*) and positive effects for the two others (*Prorcentrum triestinum* and *Scripiscella* sp.). Thus as a rule different predictors were selected for the presence and biomass models and the sign of the effects differed as well. Most notably, higher biomass was predicted with higher irradiance for many dinoflagellate taxa while diatom presence was more likely at lower irradiance.

Optimum temperatures varied widely, suggesting the existence of difference thermal niches within each functional type. Posterior medians of diatoms' optimum temperatures spanned the entire range of the observed temperatures over the time series. Based on species-specific optimum temperatures, diatoms can be separated into spring species characterized by low optimum temperatures (posterior medians < 12 °C), fall species with mid-range optimum temperatures (posterior medians between 12 and 16 °C) and summer species with higher optimum temperatures (posterior medians > 16 °C).

Optimum temperatures for the study dinoflagellates ranged from intermediate to high, suggesting an increase in dinoflagellate biomass with increasing sea surface temperature from low biomass during spring to high biomass during summer, with potential for intense but brief blooms of species with high optimum temperatures such as *Prorocentrum minimum* and *Karenia mikimotoi*. These are well-documented harmful algal bloom taxa (HABs). In 2003, the English Channel experienced a massive bloom of *K. mikimotoi* from the end of June to the beginning of August due to exceptionally warm conditions (Vanhoutte-Bruniera et al., 2008).

HABs may impact the marine ecosystem either directly, through its hemolytic cytotoxin, or indirectly through hypoxia, with far-reaching implications, including widespread mortality of wild fishes and benthic invertebrates. The economic losses induced by fish kills due to red tides can be enormous. Certain types of HABs are also linked to low oxygen (hypoxic) conditions. HABs are notoriously difficult to predict using mechanistic models. Although a full exploration of this issue is beyond the scope of this study, we suggest that the optimum temperatures and sensitivities documented here may be relevant for this problem and anticipate that the two-part modeling approach proposed here can be used, in conjunction with environmental forecasting models, to predict the location and magnitude of HABS.

Our two-part model achieved a reduction of about one-third to one-half in root mean squared prediction error relative to the two most common methods for describing biomass data with many zeroes, indicating a dramatic improvement in model error is possible with our two-step approach.

While some of our results confirm well-known phenomena, other findings, such as the difference between variables predicting presence vs. biomass magnitude, are hardly ever documented. This finding has significant ecological implications: the important predictors of presence/absence are relatable for niche characterization while the predictors of biomass/abundance are potentially pertinent for modeling other contributions to ecosystem services. This dichotomy illustrates the value of the two-part modeling approach for analyzing semi-continuous ecological data without wasting information or compromising the data integrity as implied by models based on commonly used zero-handling techniques. In addition, the identification of optimal temperature for biomass and the sensitivity of biomass magnitude to temperature from observational data is infrequent.

In conclusion, two-part models provide the flexibility to accommodate excess zeroes in semi-continuous data beyond the dominant ad hoc approaches. With this flexibility comes great benefits including a better fit to data, a higher predictive performance, and the ability to decouple the drivers of species occurrence and biomass patterns from observational data. In addition to these statistical findings, our model identified under-appreciated differences between the factors that promote species presence and the factors that promote higher biomass of individual species, with clear differences between diatoms and dinoflagellates.

Acknowledgements This work was supported by the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems/CBIOMES (Grant ID: 549935, AJI). C.E.W. was funded through the UK Natural Environment Research Council's National Capability Long-term Single Centre Science Programme, "Climate Linked Atlantic Sector Science", Grant No. NE/R015953/1, and is a contribution to Theme 1.3—Biological Dynamics. Phytoplankton biomass and environmental data were provided by the Plymouth Marine Laboratory's Western Channel Observatory www.westernchannelobservatory.org.uk, which was funded as part of the UK's Natural Environmental Research Council's National Capability.

Data availability Source data were obtained from and are available from the Western Channel Observatory www.westernchannelobservatory.org.uk. The L4 environmental data are available at the British Oceanographic Data Centre https://www.bodc.ac.uk/ (doi: https://doi.org/10.5285/f1968a39-26bf-55fe-e044-000b5de50f38).

Code availability The OpenBUGS code used to fit the model to data is available in the Online Supplementary Materials.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Aitchison, J., 1955. On the distribution of a positive random variable having a discrete probability mass at the origin. Journal of the American Statistical Association 50: 901–908.

Aitchison, J. & J. A. C. Brown, 1957. The Lognormal Distribution (with special reference to its uses in economics), Cambridge University Press, London, pp. 94–99.

Amemiya, T., 1974. Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. Econometrica 42: 999–1012.

Armbrust, E. V., 2009. The life of diatoms in the world's oceans. Nature 459: 185–192.

Cameron, A. & P. Trivedi, 1998. Regression analysis of count data, University Press, Cambridge.

Clarke, K. R. & R. H. Green, 1988. Statistical design and analysis for a "biological effects" study. Marine Ecology Progress Series 46: 213–226.

Cragg, J. G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39: 829–844.

Conn, P. B., D. S. Johnson, P. J. Williams, S. R. Melin & M. B. Hooten, 2018. A guide to Bayesian model checking for ecologists. Ecological Monographs 88: 526–542.

Crow, E. L. & K. Shimizu, 1988. Lognormal distributions: Theory and applications, Marcel Dekker, New York:, 47–51.

Feng, C., H. Wang, N. Lu, T. Chen, H. He, Y. Lu & X. M. Tu, 2014. Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry 26: 105–109.

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

Fletcher, D., D. MacKenzie & E. Villouta, 2005. Modelling skewed data with many zeroes: a simple approach combining ordinary and logistic regression. Environmental and Ecological Statistics 12: 45–54.

Gelman, A., X.-L. Meng & H. S. Stern, 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6: 733–807.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari & D. B. Rubin, 2013. Bayesian data analysis, 3rd ed. Chapman & Hall, London:

George, E. I. & R. E. McCulloch, 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88: 881–889.

Gilks, W. R., S. Richardson & D. J. Spiegelhalter (eds), 1996. Markov Chain Monte Carlo in practice. Chapman and Hall, London.

Harris, R., 2010. The L4 time-series: the first 20 years. Journal of Plankton Research 32(5): 577–583.

Hastie, T., R. Tibshirani & J. Friedman, 2009. The elements of statistical learning: Data mining, inference, and prediction, Springer, New York:

Jeffreys, H., 1961. Theory of probability, 3rd ed. Oxford University Press, Oxford:

Kass, R. & A. Raftery, 1995. Bayes factors. Journal of the American Statistical Association 90: 773–795.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in mnufacturing. Technometrics 34: 1.

Le Quéré, C., S. P. Harrison, I. C. Prentice, et al., 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Global Change Biology 11: 2016–2040.

Martin, T. G., B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low-Choy, A. J. Tyre & H. P. Possingham, 2005. Zero tolerance ecology: improving ecological inference by modeling of zero observations. Ecology Letters 8: 1235–1246.

Maunder, M. N. & A. E. Punt, 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research 70: 141–159.

May, R. M., 1975. Patterns of species abundance and diversity. In Cody, M. L. & J. M. Diamond (eds), Ecology and evolution of communities Harvard University Press, Cambridge, pp. 81–120.

McCarthy, M., 2007. Bayesian methods in ecology, Cambridge University Press, New York.

McCullagh, P. & J. Nelder, 1989. Generalized linear models, 2nd ed. Chapman and Hall/CRC, Boca Raton:

McGill, B. J., 2003. A test of the unified neutral theory of biodiversity. Nature 422: 881–885.

Menden-Deuer, S. & E. J. Lessard, 2000. Carbon to volume relationships for dinoflagellates, diatoms, and other protest plankton. Limnology and Oceanography 45: 569–579.

Min, Y. & A. Agresti, 2002. Modeling non-negative data with clumping at zero: a survey. Journal of the Iranian Statistical Society 1: 7–33.

Mullahy, J., 1986. Specification and testing of some modified count data models. Journal of Econometrics 33: 341–365.

Mutshinda, C. M., R. B. O'Hara & I. P. Woiwod, 2009. What drives community dynamics? Proceedings of the Royal Society London, Series B 276: 2923–2929.

Mutshinda, C. M., R. B. O'Hara & I. P. Woiwod, 2011. A multispecies perspective on ecological impacts of climatic forcing. Journal of Animal Ecology 80: 101–107.

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

Mutshinda, C. M., Z. V. Finkel & A. J. Irwin, 2013b. Which environmental factors control phytoplankton populations? A Bayesian variable selection approach. Ecological Modelling 269: 1–8.

Mutshinda, C. M., Z. V. Finkel, C. E. Widdicombe & A. J. Irwin, 2016. Ecological equivalence of species within phytoplankton functional groups. Functional Ecology 30: 1714–1722.

Mutshinda, C. M., Z. V. Finkel, C. E. Widdicombe & A. J. Irwin, 2017. Phytoplankton traits from long-term oceanographic time-series. Marine Ecology Progress Series 576: 11–25.

Mutshinda, C. M., Z. V. Finkel, C. E. Widdicombe & A. J. Irwin, 2019. Bayesian inference to partition determinants of community dynamics from observational time series. Community Ecology 20: 238–251.

Mutshinda, C. M., Z. V. Finkel, C. E. Widdicombe & A. J. Irwin, 2020. A trait-based clustering for phytoplankton biomass modeling and prediction. Diversity 12: 295.

Mwalili, S., E. Lesaffre & D. Declerck, 2008. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Statistical Methods in Medical Research 17: 123–139.

Mwanza, C., 2010. Bayesian analysis of community dynamics, University of Helsinki, Helsinki:

Neelon, B., A. O'Malley & V. Smith, 2016. Modeling zero-modified count and semicontinuous data in health services research Part 1: background and overview. Statistics in Medicine 35: 5070–5093.

Olsen, M. K. & J. L. Schafer, 2001. A two-part random-effects model for semicontinuous longitudinal data. Journal of the American Statistical Association 96: 730–745.

Owusu, R. A., C. M. Mutshinda, I. Antai, K. Q. Dadzie & E. M. Winston, 2016. Which UGC features drive web purchase intent? A spike-and-slab Bayesian Variable Selection Approach. Internet Research. 26: 22–37.

Pennington, M., 1983. Efficient estimators of abundance for fish and plankton surveys. Biometrics 39: 281–286.

Preston, F. W., 1948. The commonness, and rarity, of species. Ecology 29: 254–283.

Reynolds, C. S., 2006. Ecology of phytoplankton, Cambridge University Press, Cambridge, MA.

Rubec, P. J., R. Kiltie, E. Leone, R. O. Flamm, L. E. McEachron & C. Santi, 2016. Using delta-generalized additive models to predict spatial distributions and population abundance of Juvenile Pink Shrimp in Tampa Bay, Florida. Marine and Coastal Fisheries 8: 232–243.

Stefánsson, G., 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES Journal of Marine Science 53: 577–588.

Stone, M., 1974. Cross-validation choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B 36: 111–147.

Su, L., B. D. Tom & V. T. Farewell, 2009. Bias in 2-part mixed models for longitudinal semicontinuous data. Biostatistics 10: 374–389.

Sugihara, G., 1980. Minimal community structure: an explanation of species abundance patterns. American Naturalist 116: 770–787.

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

Vanhoutte-Bruniera, A., S. L. Lyons, F. Gohin, L. Fernand, A. Ménesguen & P. Cugier, 2008. Modelling the *Karenia mikimotoi* bloom that occurred in the western English Channel during summer 2003. Ecological Modelling 210: 351–376.

Widdicombe, C., D. Eloire, D. Harbour, R. Harris & P. Somerfield, 2010. Long-term phytoplankton community dynamics in the Western English Channel. Journal of Plankton Research 32: 643–655.

Wang, X., X. Feng & X. Song, 2020. Joint analysis of semicontinuous data with latent variables. Computational Statistics and Data Analysis 151: 107005.

Xu, L., A. D. Paterson, W. Turpin & W. Xu, 2015. Assessment and selection of competing models for zero-inflated microbiome data. PLoS ONE 10: 129606.