Supporting Information

Hunter-Cevera et al. 10.1073/pnas.1321421111

SI Text

Matrix Assignments. The formulation and analysis of matrix population models are extensively covered in ref. 1. Here we describe the element assignments for the projection matrix specified by Eqs. **4** and **5**.

Division. Division is dependent on cell size as described by Eq. **4**. We assume that a cell divides into two daughter cells, each half the size of the original. If cells in a size class less than or equal to k divide, where

$$k = 1 + \frac{1}{\Delta_\nu}, \qquad \text{[S1]}$$

this produces daughter cells that are less than $\nu_{\min}$ and are assigned to the first size class. Note that Δ_ν must be chosen such that $1/\Delta_\nu$ is an integer. The fraction of cells that divide in a given size class, δ, is therefore described by elements along the first row of $\mathbf{A}(t;\boldsymbol{\theta})$

$$a_{1,j}(t;\boldsymbol{\theta}) = 2\delta\left(t,\nu_j;\boldsymbol{\theta}\right), \quad \text{for} \quad j = 1,2,\ldots k, \qquad \text{[S2]}$$

and by the supradiagonal

$$a_{j+1-k,j}(t;\boldsymbol{\theta}) = 2\delta\left(t,\nu_j;\boldsymbol{\theta}\right), \quad \text{for} \quad j = k+1,\ldots m. \qquad \text{[S3]}$$

Growth. The fraction of cells in each size class that grow into the next largest class, γ, is dependent on incident radiation as described by Eq. **5**. The elements of $\mathbf{A}(t;\boldsymbol{\theta})$ that correspond to cell growth occur along the first subdiagonal

$$a_{j,j+1}(t;\boldsymbol{\theta}) = \gamma(t;\boldsymbol{\theta})\left[1-\delta\left(t,\nu_j;\boldsymbol{\theta}\right)\right], \quad \text{for} \quad j = 1,2,\cdots m-1. \qquad \text{[S4]}$$

Stasis. Cells that neither divide nor grow remain in the same size class. These transitions appear on the main diagonal of $\mathbf{A}(t;\boldsymbol{\theta})$

$$a_{jj}(t,\boldsymbol{\theta}) = \begin{cases} [1-\gamma(t,\boldsymbol{\theta})]\left[1-\delta\left(t,\nu_j;\boldsymbol{\theta}\right)\right] + 2\delta\left(t,\nu_j;\boldsymbol{\theta}\right) & \text{for} \quad j = 1 \\ [1-\gamma(t,\boldsymbol{\theta})]\left[1-\delta\left(t,\nu_j;\boldsymbol{\theta}\right)\right] & \text{for} \quad 2 \le j \le m-1, \\ \left[1-\delta\left(t,\nu_j;\boldsymbol{\theta}\right)\right] & \text{for} \quad j = m. \end{cases} \qquad \text{[S5]}$$

All matrix elements not assigned by Eqs. **S2–S5** are zero.

The use of a two-subpopulation model was important for accurate estimation of division rates of both the *Synechococcus* cultures and natural *Synechococcus* populations. Subpopulation dynamics often differed from each other in terms of parameter estimates and subpopulation division rate estimates. We also find that by eliminating the first 6 h of the day from the model fitting produces a better representation of the observed cell size distributions. This suggests that there is a feature of the *Synechococcus* growth and division cycle that we are not capturing in our equations with regard to cell size dynamics right after dawn. However, this does not appear to impact the model's ability to estimate division rates accurately.

Parameter Estimation. To estimate the parameters, we fit the model to hourly observations of the number of cells in each size class. It would be natural to assume that our observations have a multinomial distribution. Our data, however, are overdispersed relative to the multinomial. To account for this overdispersion in the data, we instead assume that the observation at hour t has a Dirichlet-multinomial distribution (2) whose probability density function is given by

$$f\left[\hat{\mathbf{n}}(t);\mathbf{w}(t,\boldsymbol{\theta})\right] = \frac{\hat{N}(t)!}{\hat{n}_1(t)!\,\hat{n}_2(t)!\cdots\hat{n}_m(t)!}\left[\frac{\Gamma(s)}{\Gamma\left(\hat{N}(t)+s\right)}\right] \times \frac{\prod_{i=1}^{m}\Gamma\left(\hat{n}_i(t)+sw_i(t,\boldsymbol{\theta})\right)}{\prod_{i=1}^{m}\Gamma\left(sw_i(t,\boldsymbol{\theta})\right)}. \qquad \text{[S6]}$$

In Eq. **S6**, Γ is the gamma function, $\hat{n}_i(t)$ is the number of cells observed in size class i at time t, and $\hat{N}(t)$ is the total number of cells observed at time t: $\hat{N}(t) = \sum_{i=1}^{m}\hat{n}_i(t)$. Finally, $\mathbf{w}(t)$ is the distribution of cells in each size class obtained from the model:

$$\mathbf{w}(t) = \frac{\mathbf{n}_1(t)+\mathbf{n}_2(t)}{N(t)}. \qquad \text{[S7]}$$

The parameter s is a precision parameter; the larger s, the less the dispersion and the closer the density (Eq. **S6**) is to the multinomial density. To use the Dirichlet-multinomial distribution, we must specify this parameter, which brings the total number of parameters to 13 with s now included in $\boldsymbol{\theta}$.

Our estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, maximizes the likelihood function

$$L\left(\boldsymbol{\theta}|\hat{\mathbf{n}}\right) = \prod_{t=s}^{24} f\left[\hat{\mathbf{n}}(t);\mathbf{w}(t,\boldsymbol{\theta})\right], \qquad \text{[S8]}$$

subject to the constraints listed in Table S1.

Confidence Interval Construction. The confidence intervals reflect the uncertainty surrounding the model estimate due to sampling error, given the assumption that the underlying model structure is a correct representation of cell dynamics. One can construct a confidence interval for any parameter in $\hat{\boldsymbol{\theta}}$ with the Fisher information matrix (FIM), $\mathcal{I}(\hat{\boldsymbol{\theta}})$, the entries of which are

$$-\mathrm{E}\left[\frac{\partial^2 \log L\left(\hat{\boldsymbol{\theta}}\right)}{\partial\theta_i\partial\theta_j}\right], \qquad \text{[S9]}$$

where E denotes the expectation. Here we calculate the *observed* FIM as the $n \times n$ matrix $\mathbf{I}(\hat{\boldsymbol{\theta}})$, with elements

$$-\frac{\partial^2 \log L\left(\hat{\boldsymbol{\theta}}\right)}{\partial\theta_i\partial\theta_j}. \qquad \text{[S10]}$$

We use the asymptotic normality of $\hat{\boldsymbol{\theta}}$ (3) to construct confidence intervals around a particular parameter in $\hat{\boldsymbol{\theta}}$ as

$$\theta(j) \pm C\sqrt{\left[\mathbf{I}_{jj}^{-1}\left(\hat{\boldsymbol{\theta}}\right)\right]}, \qquad \text{[S11]}$$

where $\mathbf{I}_{jj}^{-1}(\hat{\boldsymbol{\theta}})$ refers to the jth diagonal entry of the inverse observed FIM, and C is the desired critical value of the normal distribution (i.e., 1.96 for 95% CI). Second derivatives were estimated with finite-difference calculations. For the diagonal elements of the FIM,

we used the center difference rule when the maximum likelihood estimate of the parameter was away from the bounds and a forward difference or backward difference equation when the parameter was on the bounds. In the case of mixed partial derivatives for the rest of the elements in the FIM, we used a combination of forward, backward, and center differences to calculate the second derivative depending on whether the parameter was on the bounds or not.

We use this approach to obtain confidence intervals for the division rate by treating the division rate as a parameter. Instead of calculating it from the other parameters, we estimate it directly and calculate δ_{max} of one subpopulation instead. The relationship between δ_{max} of one of the subpopulations and the calculated division rate is monotonic. Therefore, if we specify a division rate, we should be able to solve for the corresponding δ_{max}, with all other parameters held constant. We use a root-finding solver (fzero solver offered in MATLAB) to identify the δ_{max} that produces a specified division rate. We then run the model forward with the found δ_{max}. Given some parameter combinations, certain division rate values are not feasible within the potential range of δ_{max}, [0,1]. This has the potential to cause issues when calculating the second derivatives for the observed FIM. Generally, however, we do not run into this problem often. Evaluating the FIM when parameters are at the bounds may violate some of the asymptotic normality assumptions that allow us to calculate confidence intervals for parameters in $\hat{\boldsymbol{\theta}}$. We investigated this by simulating data with parameters chosen on the bounds (with sampling from the Dirichlet multinomial distribution), fitting this simulated data with our model, and calculating the confidence intervals. We found that a majority of the resulting confidence intervals contained the true division rates and thus feel confident in this approach even if parameter values were on the bounds.

Model Differences from Sosik et al. The model differs from that presented in Sosik et al. (4) in three ways. First, two subpopulations are allowed to exist and behave according to their own growth (γ) and division (δ) functions. Second, the division function used here is different from that presented in Sosik et al.

$$\delta(t,\nu_j;\boldsymbol{\theta}) = \begin{cases} 0 & \text{for} \quad t \le 6, \\ \left(\dfrac{a\nu_j^b}{1+a\nu_j^b}\right)\delta_{max} & \text{for} \quad t > 6, \end{cases} \qquad \textbf{[S12]}$$

as our function no longer contains the a parameter. Estimated parameters of simulated multinomial sampled data demonstrated an inverse relationship between parameters δ_{max} and a, indicating an unnecessary flexibility for the division function. Division rate estimates for simulated data were nearly identical between the two division function formulations. The third difference between the models is that a starting distribution is fit according to a mixture of log normal distributions for the two subpopulations. In Sosik et al. (4), the observed distribution at hour 0 was used as the starting distribution.

Parameter estimation is also different from Sosik et al. (4), who used a nonlinear least squares approach. Here, we use a maximum likelihood approach. We also only fit the model to a partial day. These changes from the original model version are supported (and some initially inspired) by our evaluation of the model's ability to estimate division rate of both cultured and natural *Synechococcus* populations.

Dilution Series Experiments. For each individual experiment, seawater at the farthest point off the Woods Hole Oceanographic Institution Iselin dock was collected 2 h before dawn via bucket sample. Water was passed through a 233-μm mesh to exclude larger zooplankton predators, but not the protozoa that typically feed on *Synechococcus*. Water was kept in the dark in 24-L carboys until experiments were started within 1.5 h after sampling. A portion of the water was filtered through 20-μm mesh (by gravity) before filtration by peristaltic pump with a 0.2-μm inline Versapor capsule filter (Pall Corporation) to yield filtered seawater. The inline filter was acid-washed before use and in between every two dilution series experiments by pumping 10% (vol/vol) HCl through the filter, followed by flushing with 2 L of Milli-Q (Millipore Corporation) water. Before filtrate was collected, ∼1 L was passed through the inline filter and discarded. Acid-cleaned 1.25-L polycarbonate bottles were triple rinsed with whole seawater before being filled with portions of whole and filtered seawater. Water was combined to yield dilutions corresponding to fractions of either [0.1, 0.3, 0.6 and 1] or [0.2, 0.4, 0.6 and 1] of whole seawater. Each dilution level was prepared in triplicate, and the total final volume in each bottle was 1 L. Lower fractions were used for dilution in the summer because *Synechococcus* abundance was 10-fold higher than in the autumn. We aimed to have >200 cells·mL^{-1} to resolve hourly cell size distributions with FlowCytobot (FCB). All bottles were shaded with two layers of window screen. Five of the 12 dilution series experiments demonstrated no change in net growth rate across the dilution fractions. Considering the cell concentration at which a linear net growth rate response was observed in the other experiments, these results are consistent with cell concentration of *Synechococcus* being above the grazer ingestion saturation threshold even in the most diluted bottle. Therefore, no division rate estimate could be obtained for these experiments, and the data were not used. The remaining seven dilution series experiments provided division estimates that could be compared with the matrix population model (Table S2).

We let each experiment run for a total of 2 d, treating each day as a separate case to compare model and dilution-based division rate estimates. We recognize that bottle effects may influence dilution series experiments that run for more than 24 h, but in this case, our goal was not to assess precisely in situ division rate but rather to compare division rates with those from the model and to investigate the possible impact of grazing on model estimates. In most of the experiments conducted, grazing appeared to increase during the second day, which provided an effective dataset for the latter objective. For each experimental point, division rate was only calculated over a 24-h period.

To determine if a linear relationship between net growth rate and dilution factor was appropriate, we tested a one-phase regression model against a two-phase regression model for each dilution series experiment (5). The one-phase model is the standard linear regression model: $Y = \beta_0 + \beta_1 X + \varepsilon$, where Y is the observed data, X is the regressor, β_0 and β_1 are the intercept and slope parameters, respectively, and ε is the error. The two-phase model is $Y = \beta_0 + \beta_1 X + \beta_2(X - X_C)I(X) + \varepsilon$, where X_C is a change point at which the data are better represented by a different linear form and $I(X)$ is an indicator function, such that $I(X) = 1$ if $X \ge X_C$ and 0 otherwise. First we fit both the one-phase and two-phase model by minimizing the residual sum of squares (RSS). We then calculate an F statistic, which is well approximated by an F distribution with 3 degrees of freedom in the numerator and $n - 4$ degrees of freedom in the denominator (5)

$$F_{stat} = \frac{\frac{(RSS_1 - RSS_2)}{3}}{\frac{RSS_2}{n-4}} \sim F(3, n-4),$$

and n is the number of data points. We reject the null hypothesis of a single phase in favor of two phases at significance level $\alpha = 0.05$. In the case where the null hypothesis is not supported, the division rate is taken as $\hat{\beta}_0$ from the fitted two-phase model. Confidence intervals for the intercept were constructed from the profile log likelihood of β_0 and the likelihood ratio test. The confidence interval included values of β_0 such that the

likelihood ratio $(2[\log L(\hat{\beta}_0) - \log L(\beta_0)] \sim \chi^2_1$, from profile log likelihoods) would not be rejected at significance level $\alpha = 0.05$.

Martha's Vineyard Coastal Observatory Environmental Data. Core observations of incident short wave radiation come from the Martha's Vineyard Coastal Observatory (MVCO) Meteorological Mast (41°20.996′ N, 70°31.60′ W) from an Eppley pyranometer and were downloaded from ftp://mvcodata.whoi.edu/pub/mvcodata/. Water temperature was obtained using a MicroCat CTD (SeaBird Electronics) deployed next to the FCB at the MVCO offshore tower (41°19.500′ N, 70°34.0′ W). Data gaps in 2008 were filled with temperature data from the offshore node (12 m depth, 41°20.1950′ N, 70°33.3865′ W), also downloaded from ftp://mvcodata.whoi.edu/pub/mvcodata/.

Culture Identification. Colony PCR was used to amplify the diversity marker *ntcA* from culture cells. Approximately 5 μL of pelleted cells was added to a 50-μL reaction mixture from Qiagen Taq PCR Master Mix Kit. BSA was added for a final concentration of 0.2 μg/μL. Primers used were 1F and 4R as described in ref. 6 at a final concentration of 2 μM. Cycling conditions consisted of an initial denaturation period of 4 min at 94 °C, followed by 40 cycles of 1 min at 94 °C, 1 min at 45 °C, and 1 min at 72 °C, and a final extension step at 72 °C for 10 min.

The PCR product (an expected 449-bp fragment) was gel purified (Qiagen Qiaquick gel extraction kit) and cloned into sequencing vectors (TOPO TA Kit; Invitrogen). Twenty-five positive colonies (determined by blue/white selection on Kanamycin LB plates) were selected for sequencing. At least 200 ng of plasmid DNA was used in a sequencing reaction with BigDye Terminator chemistry (Applied Biosystems) and sequenced with M13 vector primer on an Applied Biosystems 3730XL capillary instrument at the Josephine Bay Paul Center Keck Facility (Marine Biological Laboratory, Woods Hole, MA). Primer and vector sequences were removed, and the resulting 404-bp *ntcA* sequences were aligned with ClustalW embedded in BioEdit software (version 7.2.0) (7). The majority of sequences were the same, with only one or two nucleotide differences, most likely due to sequencing error. To determine clade assignment, a consensus sequence was created and used in constructing a phylogenetic tree with other strain representative *ntcA* sequences obtained from National Center for Biotechnology Information GenBank. The tree was created in ARB (8) with Randomized Axelerated Maximum Likelihood [general time reversible model (GTR) with gamma-distributed rate variation among sites, rapid bootstrap analysis, 100 bootstraps]. Culture placement corresponded to clade I (9).

1. Caswell H (2001) *Matrix Population Models* (Sinauer Associates, Sunderland, MA), 2nd Ed.
2. Johnson NL, Kotz S (1972) *Distributions in Statistics: Continuous Multivariate Distributions* (Wiley, New York).
3. Azzalini A (1996) *Monographs on Statistics and Applied Probability* (Chapman and Hall, London), Vol 68.
4. Sosik HM, Olson RJ, Neubert MG, Shalapyonok A, Solow AR (2003) Growth rates of coastal phytoplankton from time-series measurements with a submersible flow cytometer. *Limnol Oceanogr* 48(5):1756–1765.
5. Hinkley DV (1969) Inference about the intersection in two-phase regression. *Biometrika* 56(3):495–504.
6. Lindell D, Padan E, Post AF (1998) Regulation of ntcA expression and nitrite uptake in the marine *Synechococcus* sp. strain WH 7803. *J Bacteriol* 180(7):1878–1886.
7. Hall TA (1999) Bioedit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/nt. *Nucleic Acids Symp Ser* 41:95–98.
8. Ludwig W, et al. (2004) ARB: A software environment for sequence data. *Nucleic Acids Res* 32(4):1363–1371.
9. Penno S, Lindell D, Post AF (2006) Diversity of *Synechococcus* and *Prochlorococcus* populations determined from DNA sequences of the N-regulatory gene *ntcA*. *Environ Microbiol* 8(7):1200–1211.

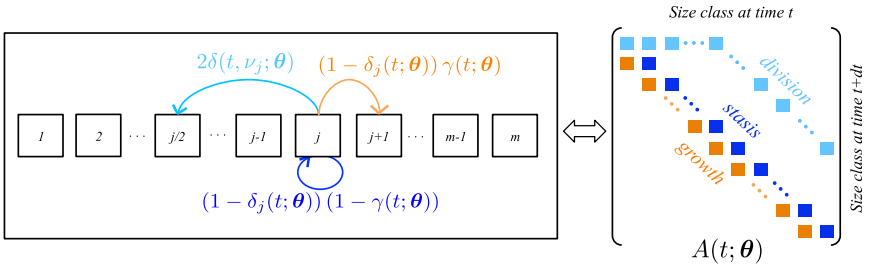

Fig. S1. Schematic representation of cell size transitions that can occur within one time step, *dt*, and how the transitions are represented in the matrix, $\mathbf{A}(t;\theta)$. The transitions are division (represented in light blue), growth (orange) and stasis (dark blue).

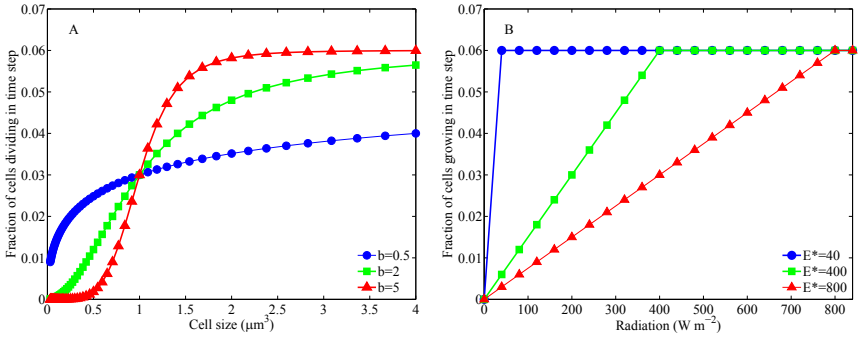

Fig. S2. (*A*) Division function, δ, with δ_{max} held constant and different values of the shape parameter, b. (*B*) Growth function, γ, with γ_{max} held constant and different values of the shape parameter, E*.

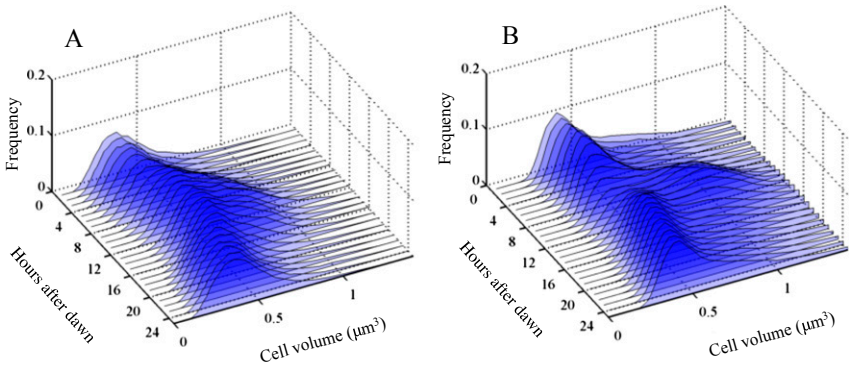

Fig. S3. (*A*) Observed hourly cell size distributions on June 16, 2008, at MVCO with estimated division rate of 1.12·d^{-1} from the model. (*B*) Cell size distributions from *Synechococcus* laboratory culture growing at division rate of 0.69·d^{-1} (note bimodal distributions for middle hours of the day).

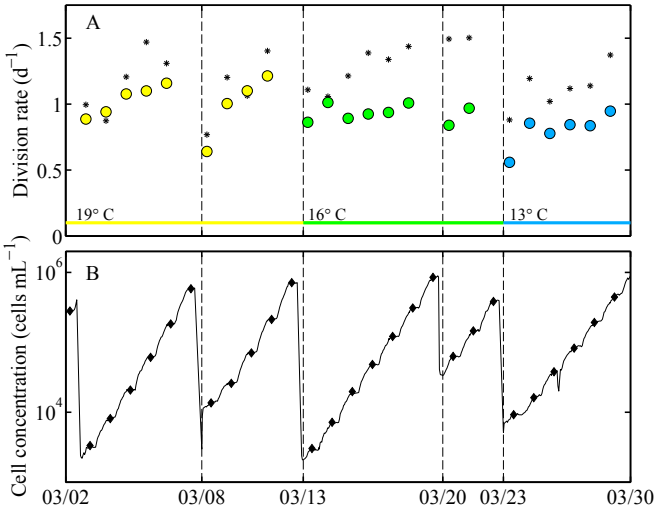

Fig. S4. (*A*) Daily division rates calculated for successive days of batch growth from the laboratory culture experiments. Colored symbols are division rates calculated from the change in cell concentration as shown in *B*. Color and shape of symbols match key in Fig. 2. Black stars are the model estimates of division rate for each day. (*B*) Cell concentration over the course of batch growth punctuated by dilution with fresh media (indicated by dashed vertical lines). Black diamonds indicate the time and cell concentration used to calculate the division rate in *A*.

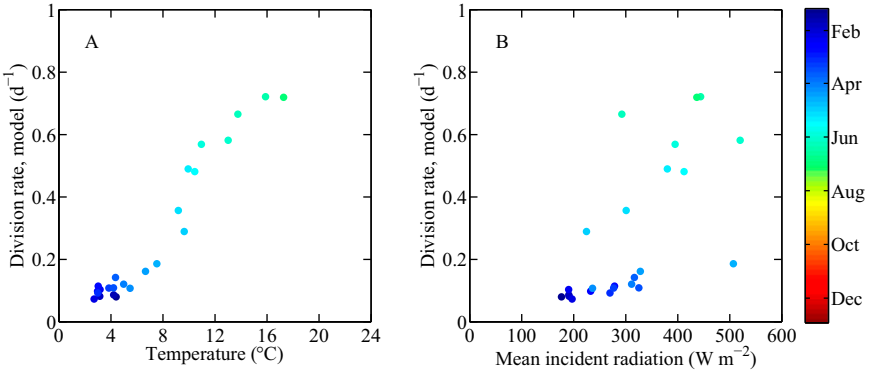

Fig. S5. (*A*) Relationship between weekly averaged division rates and temperature for the first half of 2008 (January 1–June 30). (*B*) Relationship between weekly averaged division rates and mean incident radiation for the first half of 2008. Color of symbol denotes time of year. Division rates are more strongly correlated with temperature (R = 0.97) for this period than for light (R = 0.68), suggesting that division rate is limited by temperature during this time.

Table S1. Variables, constants, and parameters for the matrix model applied to coastal *Synechococcus*

Item	Definition	Range or value	Units
$\mathbf{n}(t)$	Number of cells in each size class		
$\mathbf{w}(t)$	Proportion of cells in each size class		
$N(t)$	Total number of cells at each time step		
$\hat{\mathbf{n}}(t)$	Number of observed cells in each size class		
$\hat{N}(t)$	Total number of observed cells at each time step		
$\mathbf{A}(t; \theta)$	Projection matrix		
$E(t)$	Observed incident radiation		$W{\cdot}m^{-2}$
t^*	Data starting hour and division start time	6	Hours after dawn
ν_j	Cell size (volume)	$(2^{-5}, 4)$	μm^3
Δ_ν	Spacing between size classes	0.125	
ν_{min}	Smallest cell size	2^{-5}	μm^3
m	Number of size classes	57	
θ	Parameter vector		
μ	Division rate estimate		d^{-1}
$\delta(t, \nu_j; \theta)$	Division function		
$\gamma(t; \theta)$	Growth function		
γ_{max}	Maximum growth fraction	(0, 1]	
E^*	Growth function shape parameter	(0, max[$E(t)$]]	$W{\cdot}m^{-2}$
δ_{max}	Maximum division fraction	(0, 1]	
b	Division function shape parameter	(0, 15]	
ψ	Fraction of cells in first subpopulation at t^*	(0, 0.5]	
$\overline{\nu}_\ell$	Mean cell size of a subpopulation at t^*	[0.68, 2.38]	μm^3
σ^2	Cell size variance of both subpopulations at t^*	[0.125, 1.75]	μm^3
s	Precision parameter, Dirichlet-multinomial distribution	(0, ∞)	

Table S2. Date, symbol, dilution fractions, starting cell concentrations, division rate calculation method, extrapolated division rate, confidence intervals, and grazing rate of each dilution series experiment included in the final dataset

Date	Symbol	Day in bottle	Highest dilution	Starting concentration ($cells{\cdot}mL^{-1}$)	Estimation method	Growth rate (d^{-1})	Confidence interval ($\pm d^{-1}$)	Grazing rate (d^{-1})
June 23	•	2	0.1	278,000	Two phase	0.72	0.25	1.22
Oct 18	*	1	0.2	1,674	One phase	0.63	0.16	0.26
Oct 19	☆	2	0.2	2,409	One phase	0.53	0.09	0.33
Oct 22	○	1	0.2	893	One phase	0.69	0.14	0.08
Oct 23	▷	2	0.2	1,648	One phase	0.77	0.09	0.26
Oct 24	□	1	0.2	1,038	One phase	0.53	0.18	0.05
Oct 25	◇	2	0.2	1,667	Two phase	0.76	0.17	0.28

The first and second days of 48-h incubations are indicated separately. The symbols match those in Fig. 3.