Geophysical Research Letters

Supporting Information for

Data-driven modeling of the distribution of diazotrophs in the global ocean

Weiyi Tang[1,*] & Nicolas Cassar[1,2]

[1]Division of Earth and Ocean Sciences, Nicholas School of the Environment, Duke University, Durham, NC 27708, USA

*Now at Department of Geosciences, Princeton University, Princeton, NJ 08544, USA.

[2]Laboratoire des Sciences de l'Environnement Marin (LEMAR), UMR 6539 UBO/CNRS/IRD/IFREMER, Institut Universitaire Européen de la Mer (IUEM), Brest, France

Contents of this file

Additional Supporting Information (Files uploaded separately)

Text S1: caveats and limitations

Because of the limited number of qPCR observations, we restricted our study to four of the major diazotroph groups. In reality, there is a multitude of diazotrophic species, clades and sublineages. For example, UCYN-A1 and UCYN-A2 were combined into the group of UCYN-A in this study despite potentially exhibiting distinct niches (Farnelid et al., 2016; Thompson et al., 2014). Similarly, diazotrophs (*Richelia*+*Calothrix*) associated with diatoms are combined into the single group *Richelia*.

PCR primer bias may also limit our ability to characterize the entire diazotrophic communities. Using metagenomic sequencing, the recent study of Delmont et al. (2018) showed that non-cyanobacterial diazotrophs may have been underestimated in many oceanic regions because of primer bias. In addition, estimating diazotroph abundances based on *nifH* gene copy numbers introduces errors because of variations in cell-specific number of *nifH* gene copies and methodological issues (non-specificity of primers and probes used in qPCR) as discussed in White et al. (2018). For example, *Trichodesmium* may contain 1 to >600 copies of *nifH* gene per cell (Sargent et al., 2016). Finally, we note that the presence of *nifH* gene copies does not imply that *nifH* genes are expressed or that N_2 fixation activity is occurring.

While our updated database includes new observations in the North Pacific, Southeast Pacific, the Indian and Arctic Oceans, the majority of observations come from the North Atlantic and western Pacific (Figure 1). Extrapolating beyond the range of observations and our inability to capture the full range of environmental variables influencing the abundance of diazotrophs may introduce uncertainty. In addition, observed diazotroph abundances represent a time-integrated response to environmental conditions which may not necessarily be captured by contemporaneous observations (e.g. measured nutrient concentrations may not reflect availability). Finally, spatial and temporal mismatch may bias our estimates, such as when using climatologies of environmental factors instead of contemporaneous measurements to estimate depth-integrated diazotroph abundances.

This study provides preliminary statistical relationships between diazotroph abundances and environmental factors at the global scale. A mechanistic understanding of environmental controls on the abundance and distribution of different types of diazotrophs will require more field observations and most importantly lab culture experiments.

Text S2: data sources, data matching, data transformation and machine learning

I_s and WS were obtained from NCEP/NCAR reanalysis products (Kalnay et al., 1996). Monthly climatologies of SSS (0 m), dissolved oxygen, and nutrient concentrations (0 m) were downloaded from the World Ocean Atlas 2013 (Boyer et al., 2013). DO_{min} was defined as the minimum oxygen concentration in the upper 500 meters as in Luo et al. (2014). 8-day averaged PAR and [Chl] measured by SeaWiFS and MODIS satellites were downloaded from the NASA OceanColor website. The MLD climatology was obtained from Ifremer (de Boyer Montégut et al., 2004). Annual mean Fe was estimated based on the CESM1-BGC model under the pre-industrial control experiment as presented in CMIP5 (Moore et al., 2013; Taylor, Stouffer, & Meehl, 2012). Large discrepancies between observed and modeled Fe concentrations may introduce uncertainties in our model construction. We used monthly climatologies when contemporaneously-measured predictors were unavailable. For the machine-learning training, sampling coordinates (i.e. latitude and longitude) and time were also included. They were transformed using sine and cosine functions to preserve continuity in the data as shown in equations 1 and 2 below. Data which were close to log-normal distributed, were log10

transformed based on the previous study of Luo et al. (2014). A more detailed description of the environmental factors studied here can be found in Tang, Li, & Cassar. (2019). MATLAB machine learning (random forest), curve fitting and regression packages were used in this study.

$$coordinates = \begin{pmatrix} sin\left(latitude \cdot \frac{\pi}{180}\right) \\ sin\left(longitude \cdot \frac{\pi}{180}\right) \cdot cos\left(latitude \cdot \frac{\pi}{180}\right) \\ -cos\left(longitude \cdot \frac{\pi}{180}\right) \cdot cos\left(latitude \cdot \frac{\pi}{180}\right) \end{pmatrix} \quad (1)$$

$$time = \begin{pmatrix} cos\left(month \cdot \frac{2\pi}{12}\right) \\ sin\left(month \cdot \frac{2\pi}{12}\right) \end{pmatrix} \quad (2)$$

Feature selection in random forest: the features or explanatory variables used in model construction are critical to the success of machine-learning techniques. Some of the variables used in this study are correlated (e.g. P*=DIP-DIN/16) and some variables are less important in the model construction (Figure S16). Therefore, we tested the differences in model performance by including/excluding these variables. We found that including all the variables slightly improved model prediction without compromising the computation time. We decided to keep all the variables.

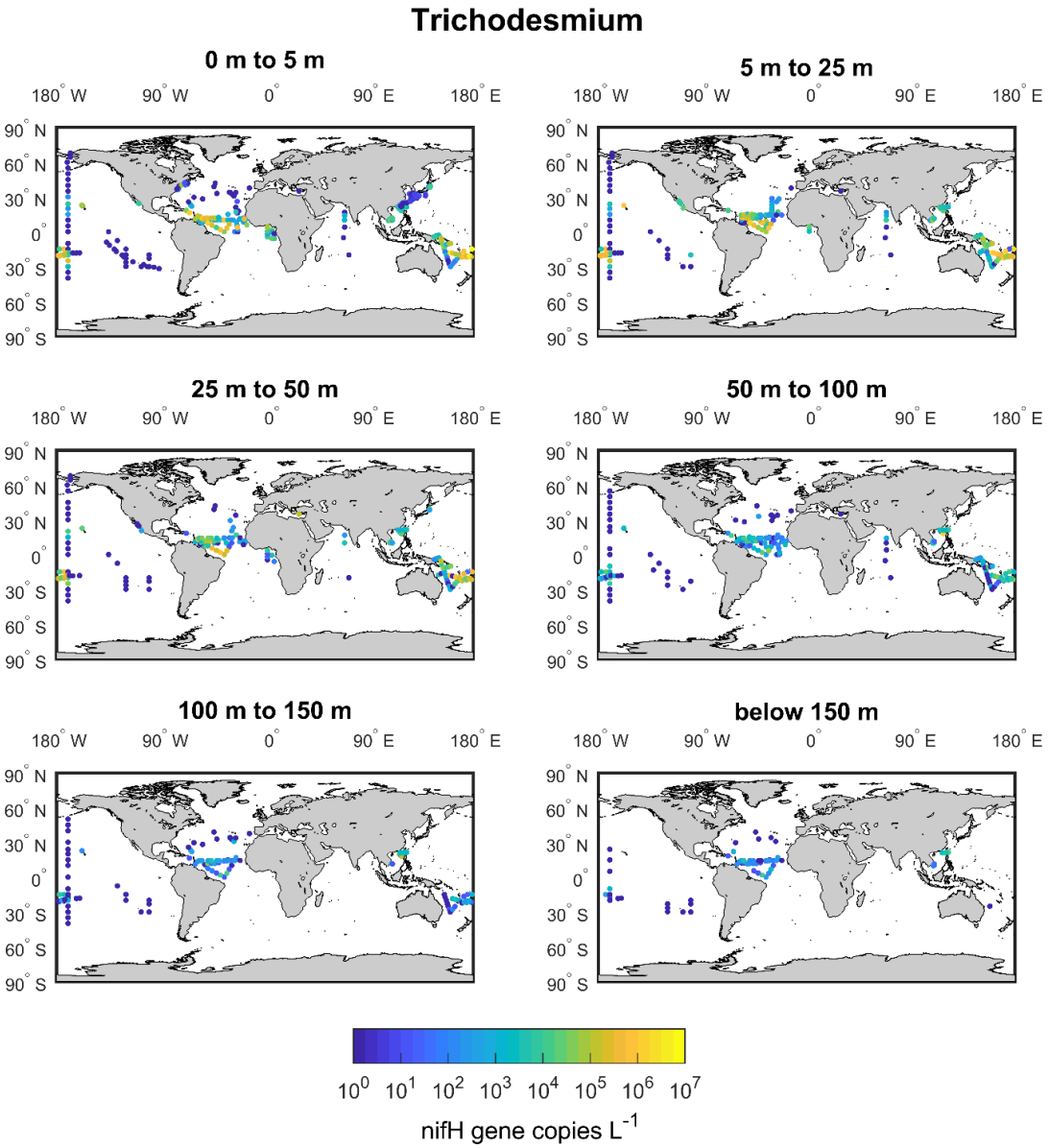


Figure S1. Volumetric abundance of *Trichodesmium* within 6 depth ranges.

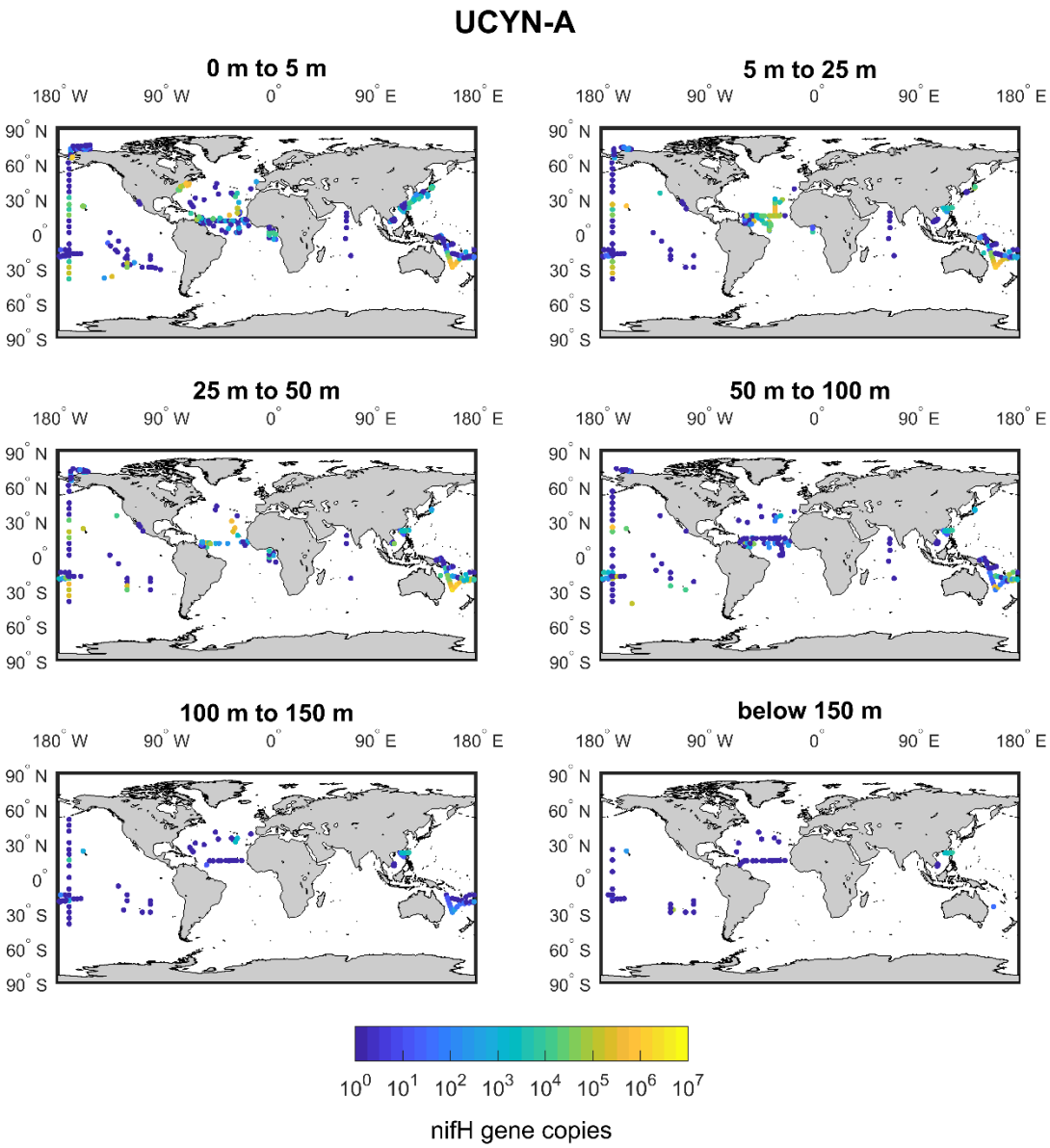


Figure S2. Volumetric abundance of UCYN-A within 6 depth ranges.

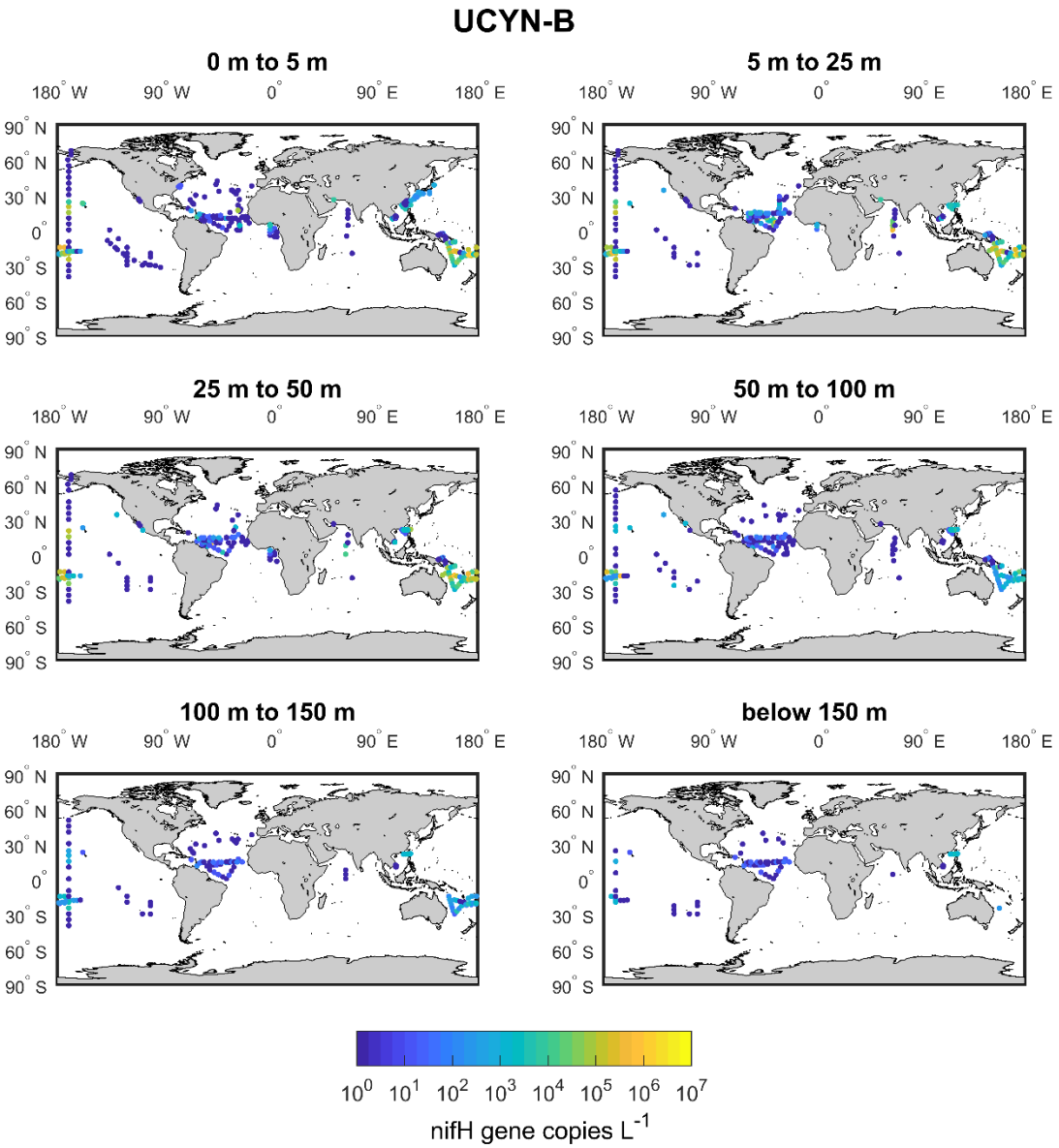


Figure S3. Volumetric abundance of UCYN-B within 6 depth ranges.

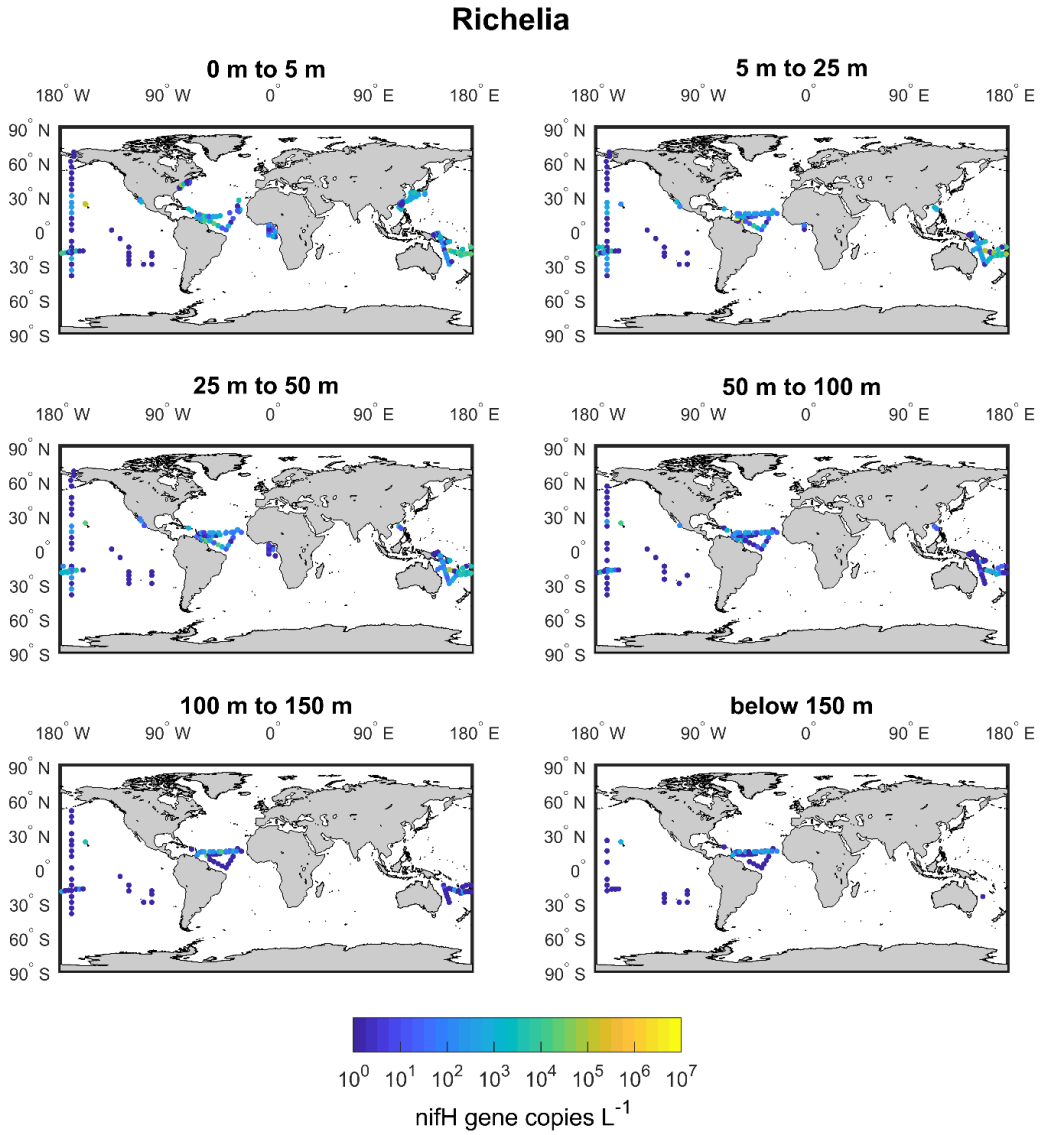


Figure S4. Volumetric abundance of *Richelia* within 6 depth ranges.

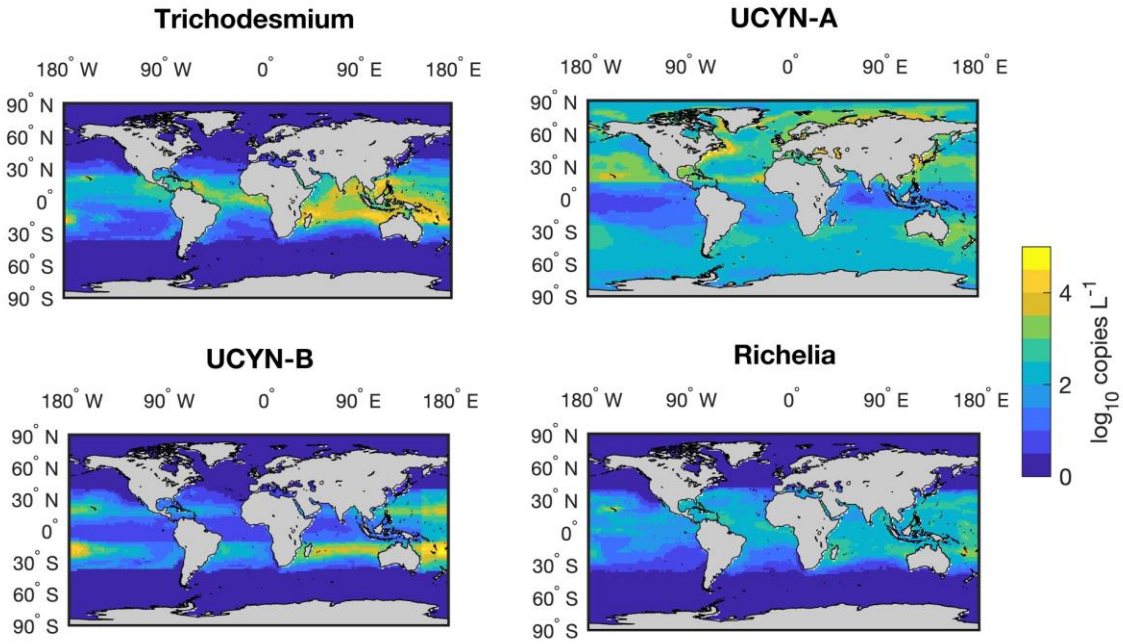


Figure S5. Simulated annual average surface volumetric abundance of *Trichodesmium*, UCYN-A, UCYN-B and *Richelia* in the global ocean.

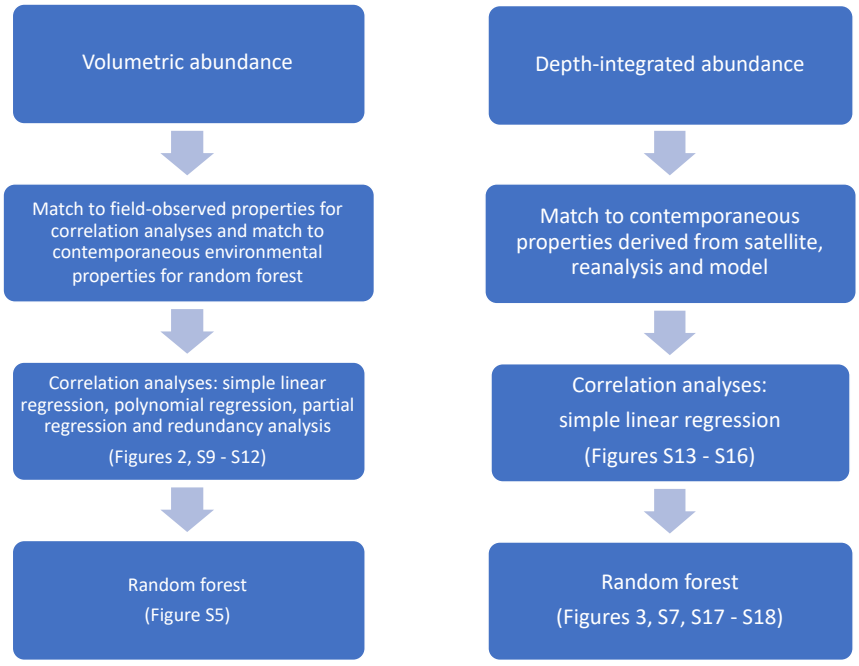


Figure S6. Workflow chart for data matching, correlation analyses and machine learning.

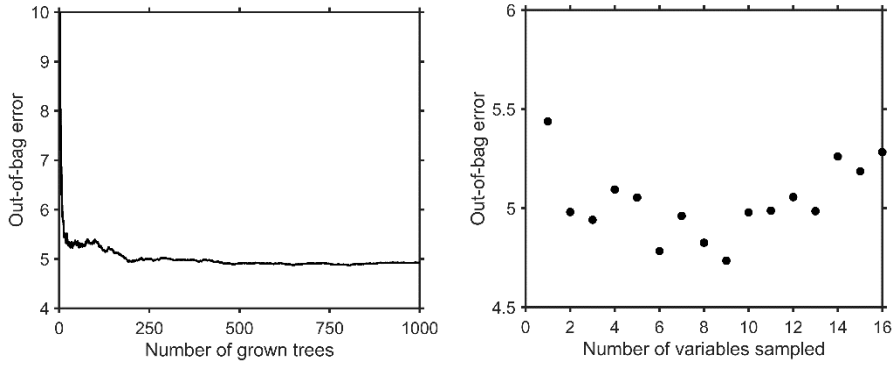


Figure S7. Out-of-bag mean square error as a function of the number of grown trees (left) and number of variables to sample (right) at each split during random forest model construction of *Trichodesmium*.

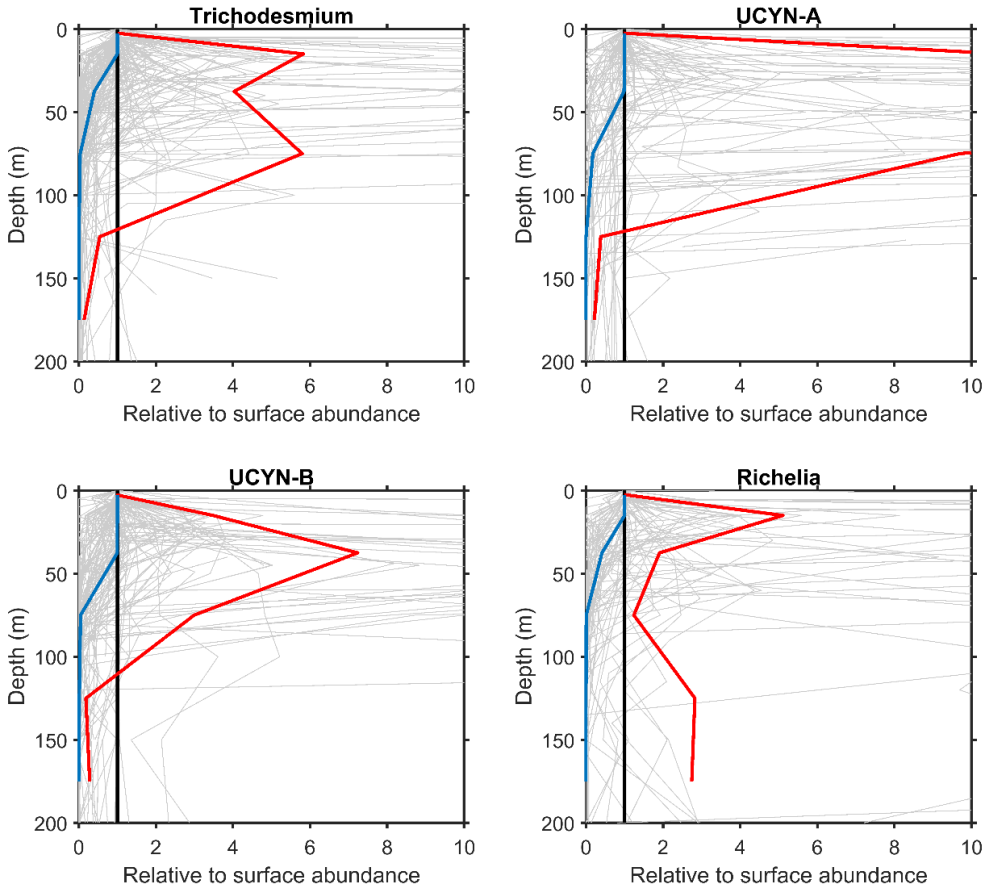


Figure S8. Volumetric abundances of each diazotrophic group within 6 depth ranges as shown in Figures S1-4 relative to surface volumetric abundances (0-5 m). Ratios of individual depth profiles are shown in light grey. The mean and median ratios are respectively shown in red and

blue. A ratio of 1 means that the diazotroph volumetric abundance at that particular depth range is equal to the surface volumetric abundance.

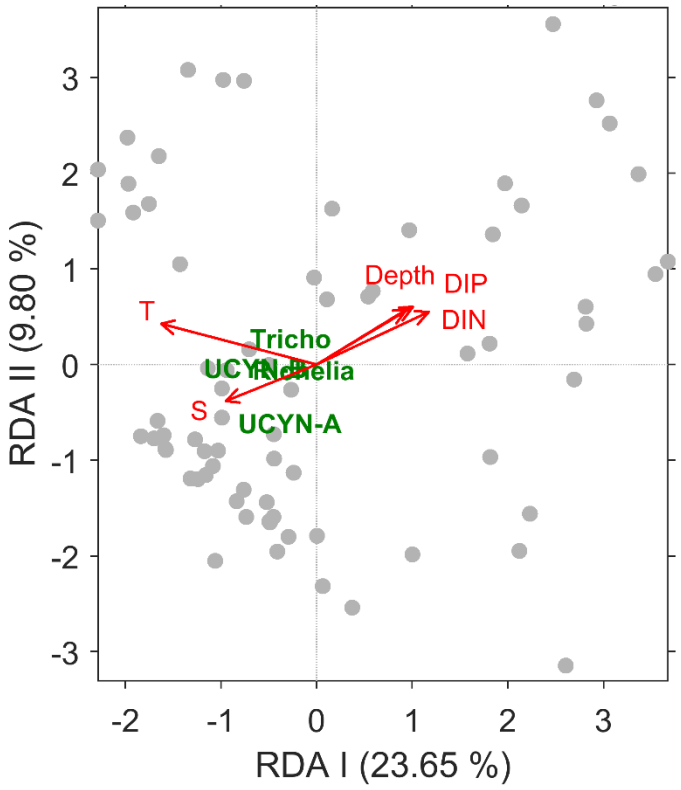
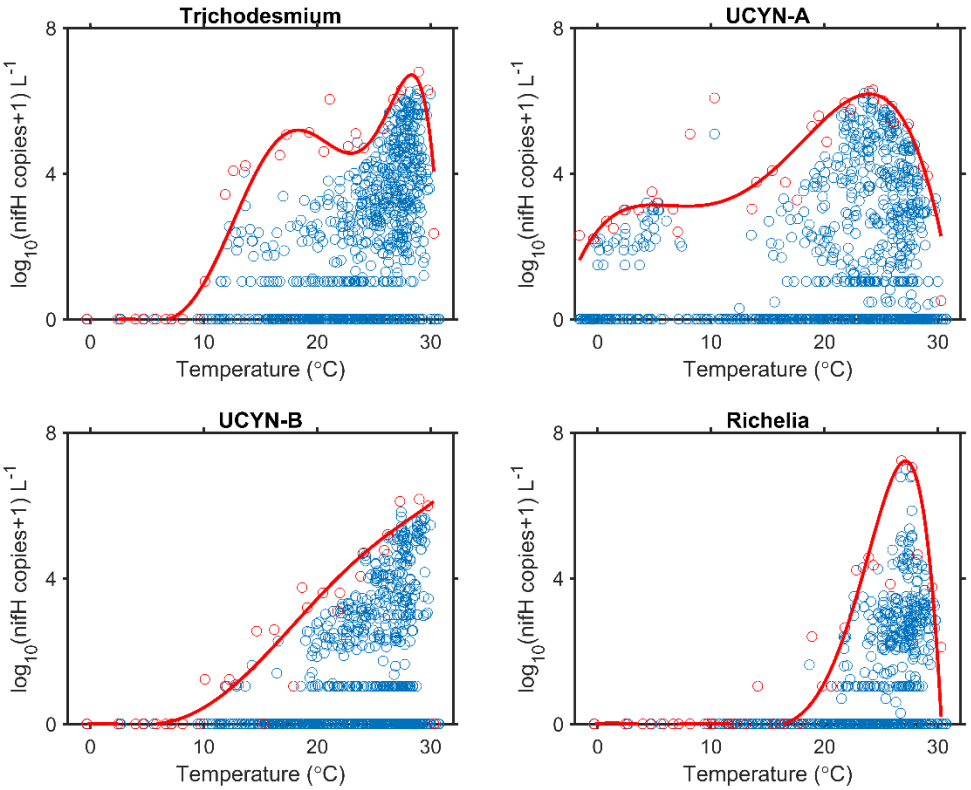


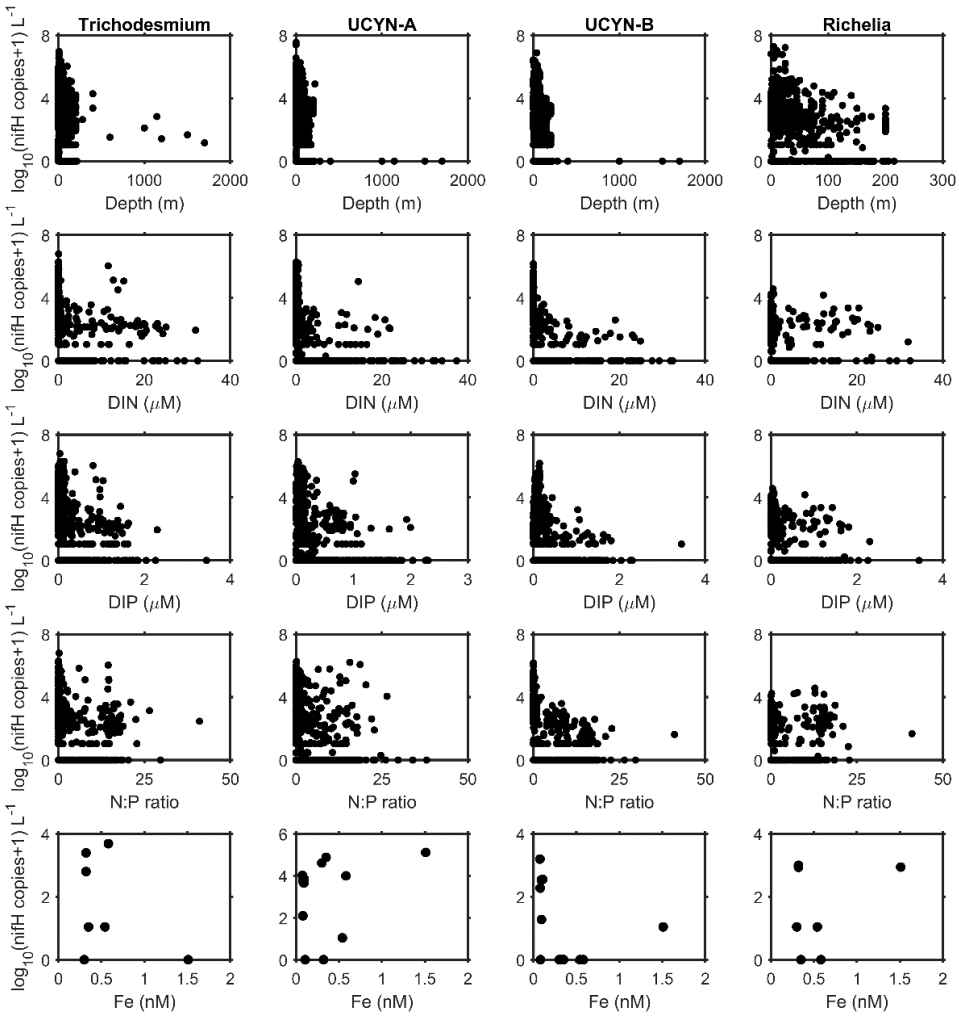
Figure S9. Redundancy analyses assessing the relations of environmental factors to volumetric abundances (log10-transformed) of four diazotrophic groups. Grey points represent the sampling sites. Red arrows show the explanatory variables including temperature (T), salinity (S), depth, dissolved inorganic nitrogen (DIN) and dissolved inorganic phosphorus (DIP). Green texts represent the diazotrophic groups. The two RDA I and RDA II can explain 23.65% and 9.8% of the variance in diazotroph abundances. *Trichodesmium*, UCYN-B and *Richelia* are clustered together, suggesting similarity in environmental controls while UCYN-A is distinct.



140

141 **Figure S10.** Upper bounds (red lines, polynomial fit) of diazotroph abundances versus
142 temperature based on the maximum *nifH* gene abundances at each temperature (red circles).
143 The polynomial fitting captures most of the variation in maximum abundances (r=0.96, 0.92,
144 0.93 and 0.96 for *Trichodesmium*, UCYN-A, UCYN-B and *Richelia*, respectively).

145



146

147 **Figure S11.** Relations between volumetric abundances of four diazotrophs and
148 contemporaneously field-observed environmental factors.

149

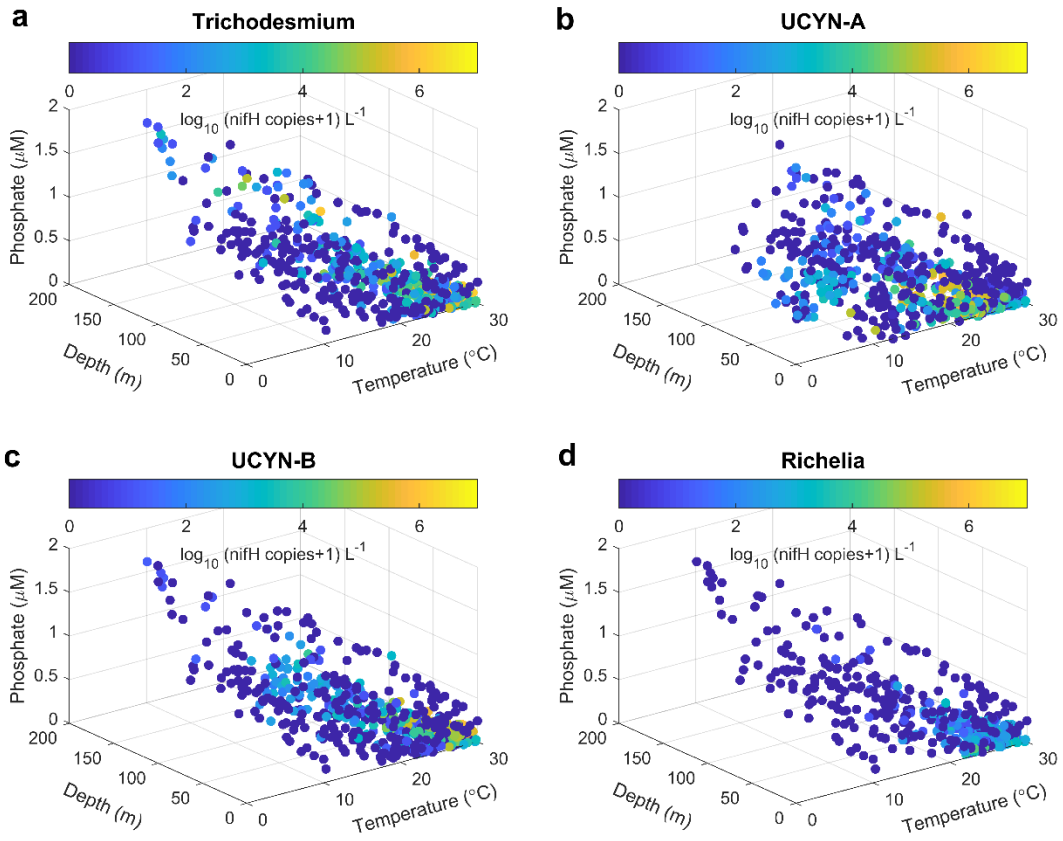
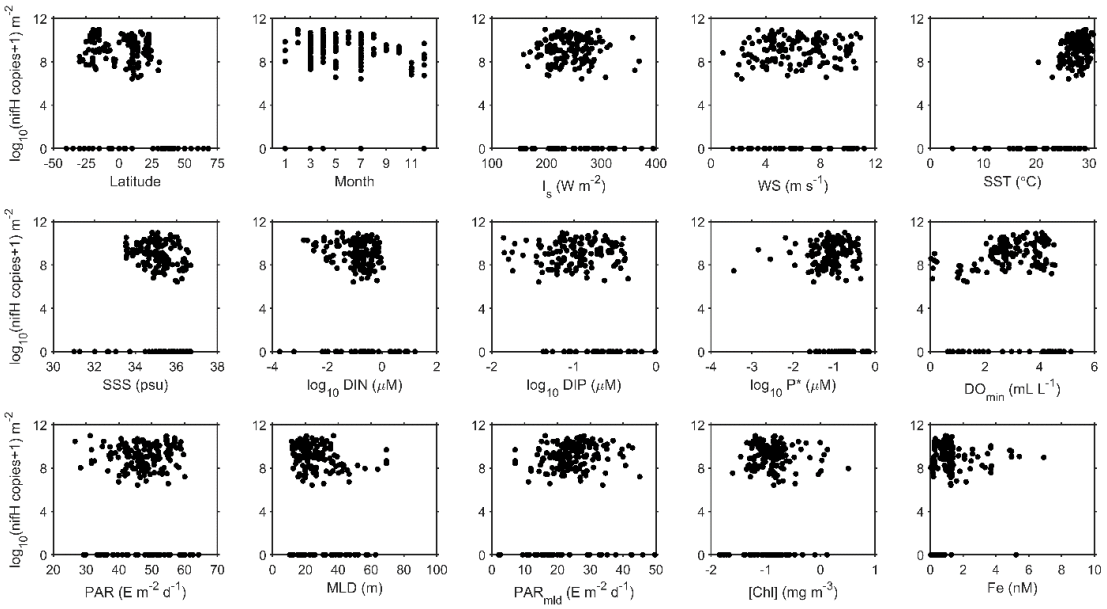


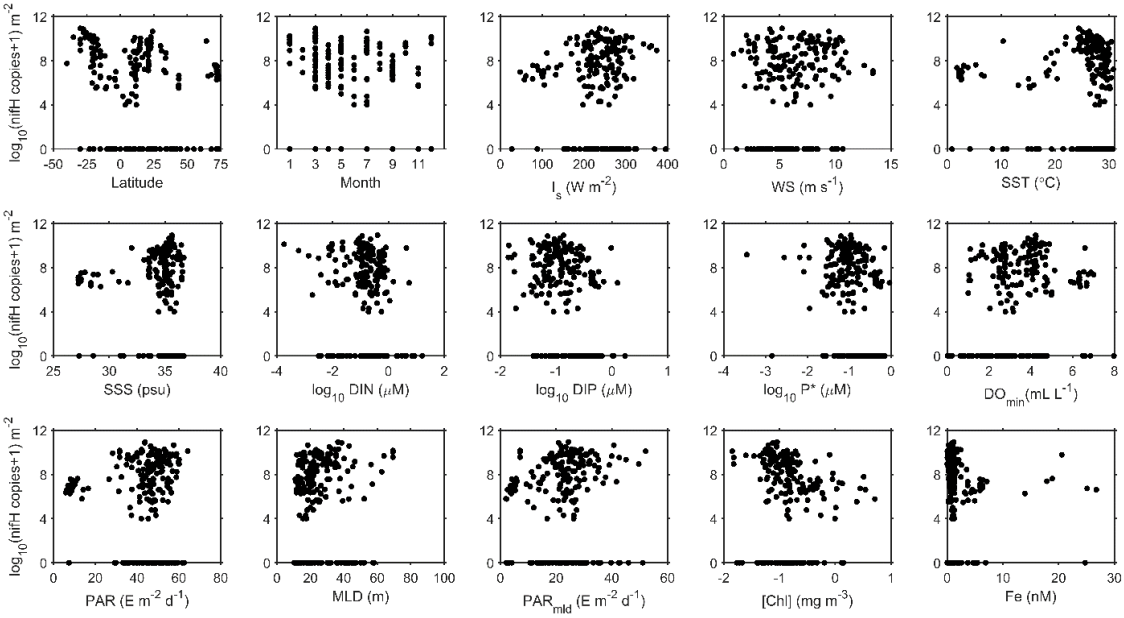
Figure S12. Relations between volumetric abundances of four diazotrophs (color-coded) and contemporaneously field-observed environmental factors including depth, temperature and phosphate concentration.



158

159 **Figure S13.** Depth-integrated abundance of *Trichodesmium* versus environmental predictors.
160 Solar radiation (I_s), wind speed (WS), sea surface temperature (SST), photosynthetically
161 available radiation (PAR) and chlorophyll-a concentration [Chl] are obtained re-analyses or
162 measured by satellites contemporaneously with diazotrophs observations while sea surface
163 salinity (SSS), nutrients, minimum oxygen in upper 500 m (DO_{min}) and mixed layer depth (MLD)
164 are obtained from monthly climatologies. Fe values are based on a modeled annual
165 climatology.

166



167

168 **Figure S14.** Depth-integrated abundance of UCYN-A versus environmental predictors.
169 Properties and data sources are as described above in the caption of Figure S10.

170

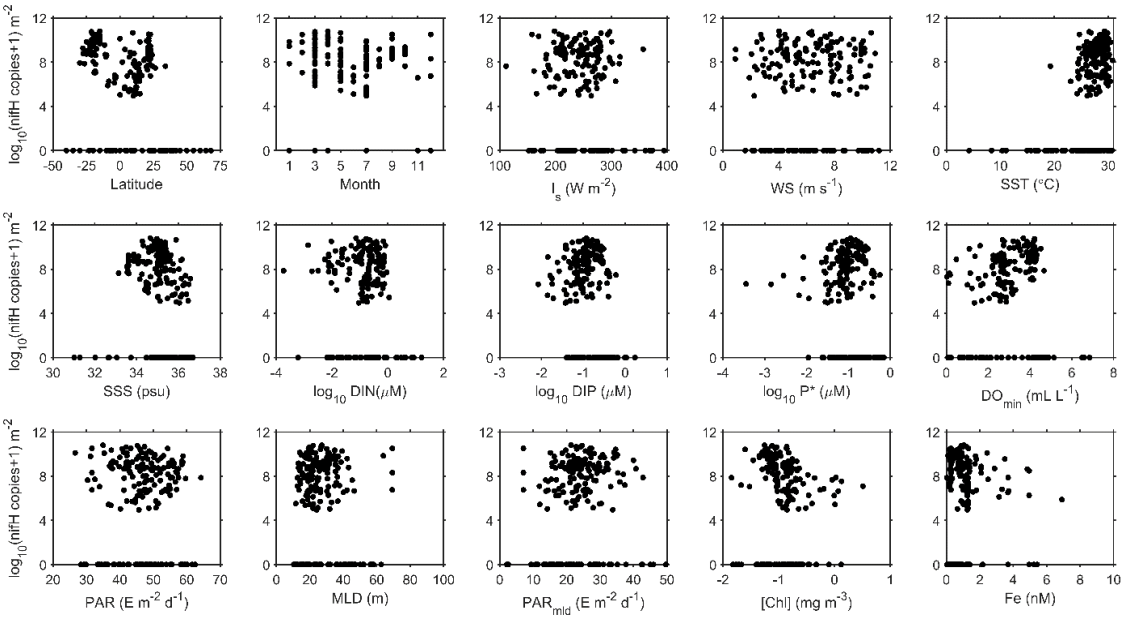


Figure S15. Depth-integrated abundance of UCYN-B versus environmental predictors. Properties and data sources are as described above in the caption of Figure S10.

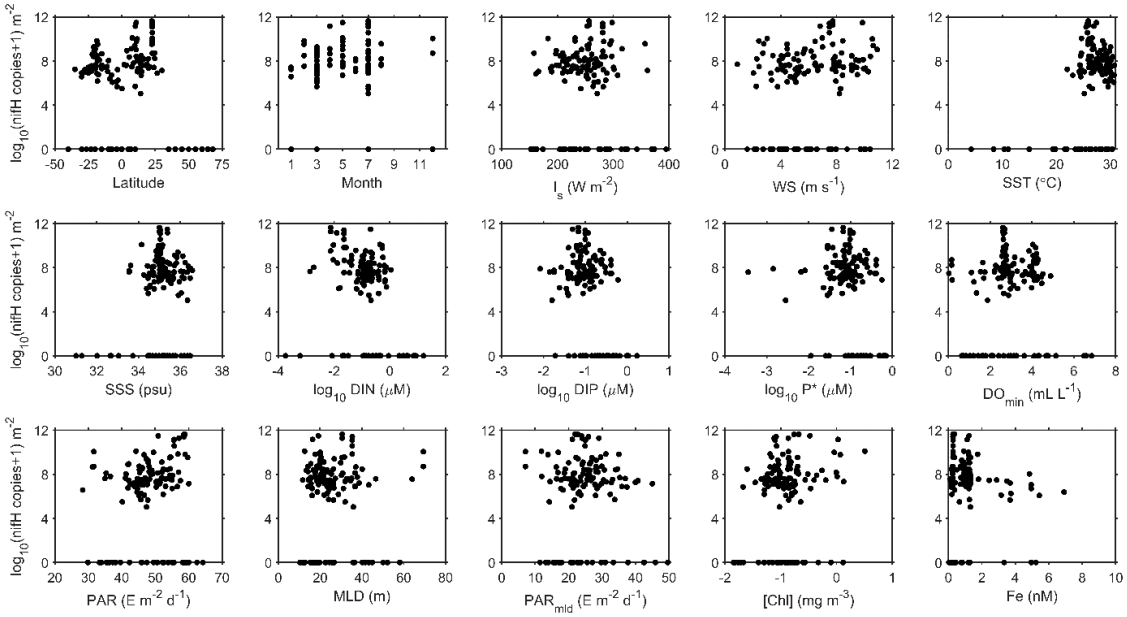


Figure S16. Depth-integrated abundance of *Richelia* versus environmental predictors. Properties and data sources are as described above in the caption of Figure S10.

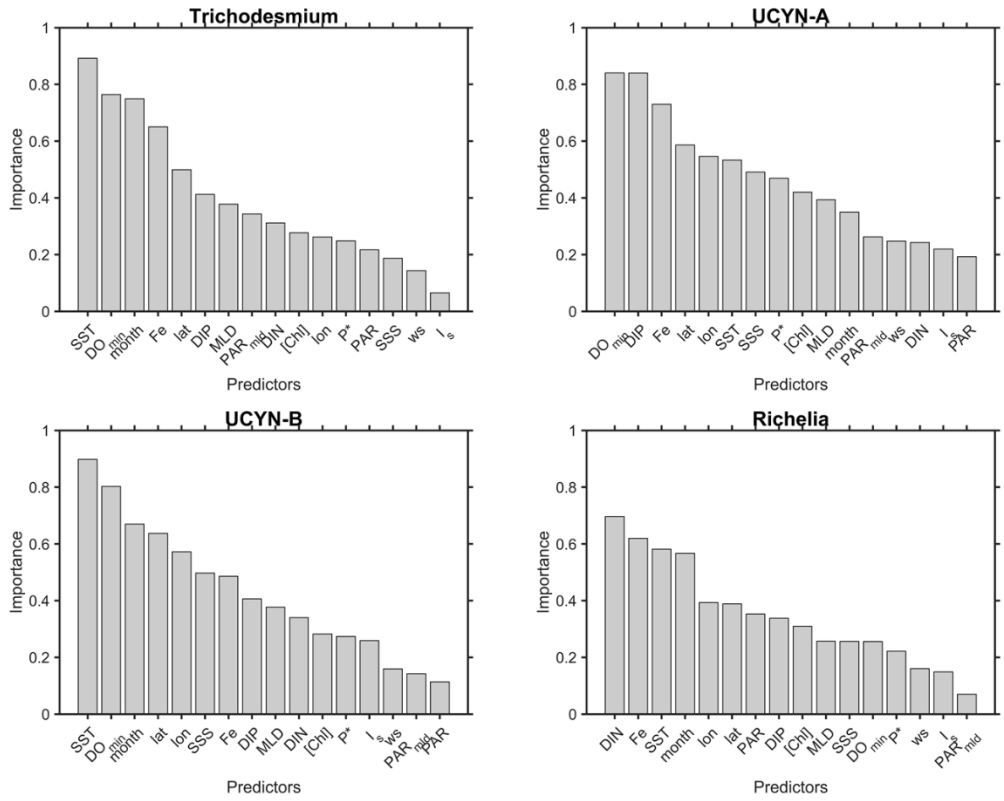


Figure S17. Predictor feature importance during model construction by random forest for four diazotrophic groups.

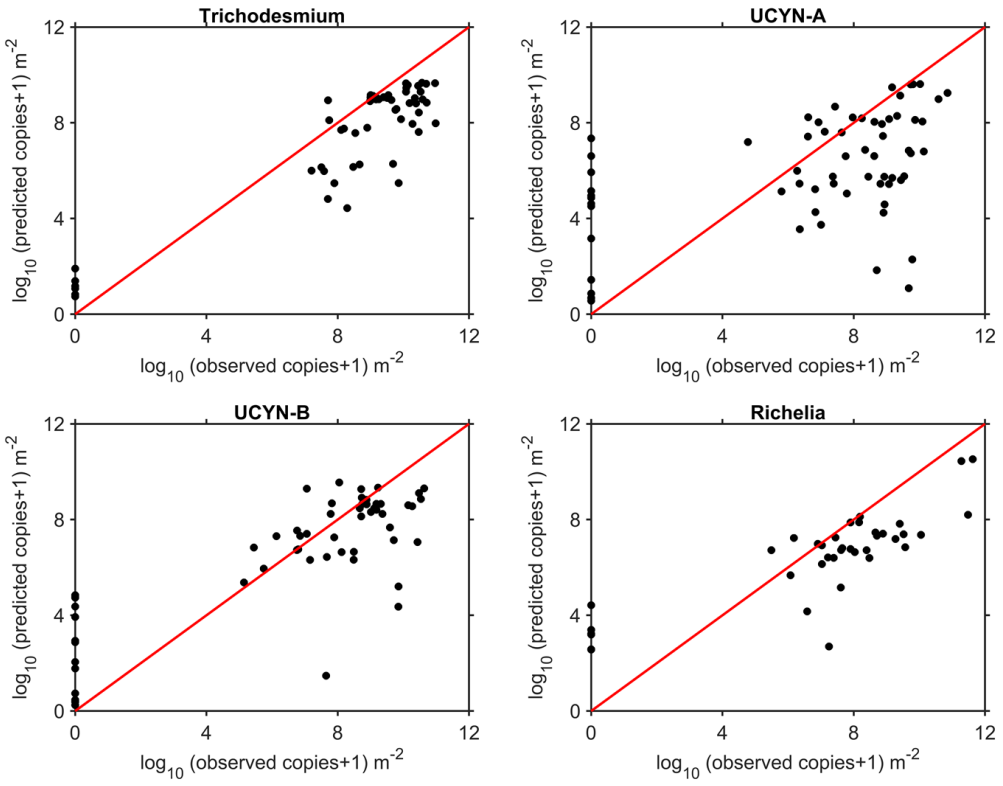


Figure S18. Observed versus predicted depth-integrated abundances of four diazotrophs in testing datasets. The average correlation coefficients are 0.87, 0.46, 0.78 and 0.79 for *Trichodesmium*, UCYN-A, UCYN-B, and *Richelia*, respectively. Grazing and other yet-to-be determined factors not included in our model may have caused false positives (i.e. incorrectly predicting the presence of a diazotroph when it is not observed in the database). Without the false positives, the correlation coefficients are 0.61, 0.22, 0.35 and 0.66 for *Trichodesmium*, UCYN-A, UCYN-B, and *Richelia*, respectively.

Table S1. Updated database of diazotrophs in the global ocean.

Table S2. Environmental properties used to model depth-integrated diazotroph abundances.

Data	Symbol	Source	Spatial resolution	Temporal resolution	Range	Log10-transformed
Surface downward solar radiation (W m^{-2})	I_s	NCEP/NCAR	2°	Daily	27.4-394.8	No
Surface wind speed (m s^{-1})	WS	reanalysis	2.5°		0.8-17.5	No
Mixed layer depth (m)	MLD	Ifremer	2°	Monthly climatology	10.4-183.8	No
Sea surface salinity (psu)	SSS		1°		27.2-36.8	No

Minimum dissolved oxygen in 0-500 m (mL L^{-1})	DO_{min}	World Ocean Atlas 2013	1°	Monthly climatology	0-8.0	No
Surface nitrate (μM)	DIN		1°		0.0002-16.2	Yes
Surface phosphate (μM)	DIP		1°		0.0008-1.7	Yes
Excess phosphate (μM)	P^*		1°		0.0004-0.9	Yes
Sea surface temperature (°C)	SST	SeaWiFS and MODIS	0.083°	8 days	0.9-31.5	No
Photosynthetically available radiation (Einstein m^{-2} d^{-1})	PAR		0.083°		5.2-64.2	No
Average PAR in mixed layer (Einstein m^{-2} d^{-1})	PAR_{mld}		0.083°		0.8-52.1	No
Chlorophyll-*a* (mg m^{-3})	[Chl]		0.083°		0.01-5.2	Yes
Modeled iron (nM)	Fe	CESM1-BGC	~0.56°× ~0.94°	Annual	0.04-28.8	No

198
199
200

201 **Movie S1.** Modeled monthly change of *Trichodesmium* abundance in the global ocean

202 **Movie S2.** Modeled monthly change of UCYN-A abundance in the global ocean

203 **Movie S3.** Modeled monthly change of UCYN-B abundance in the global ocean

204 **Movie S4.** Modeled monthly change of *Richelia* abundance in the global ocean