Supplemental Material

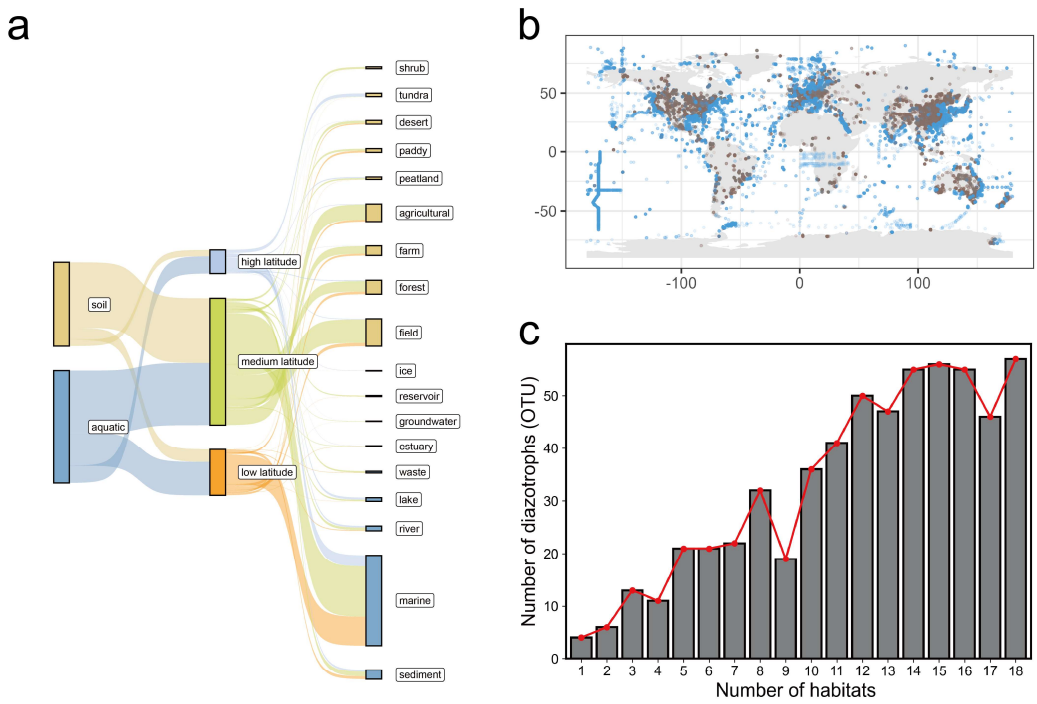
Reduction of global diazotroph diversity is driven by anthropogenic climate change

Peng Li[1], Zhuo Pan[1] *, Jingyu Sun[1], Yu Geng[1], Yiru Jiang[1], Yue-zhong Li[1], Zheng Zhang[1] *

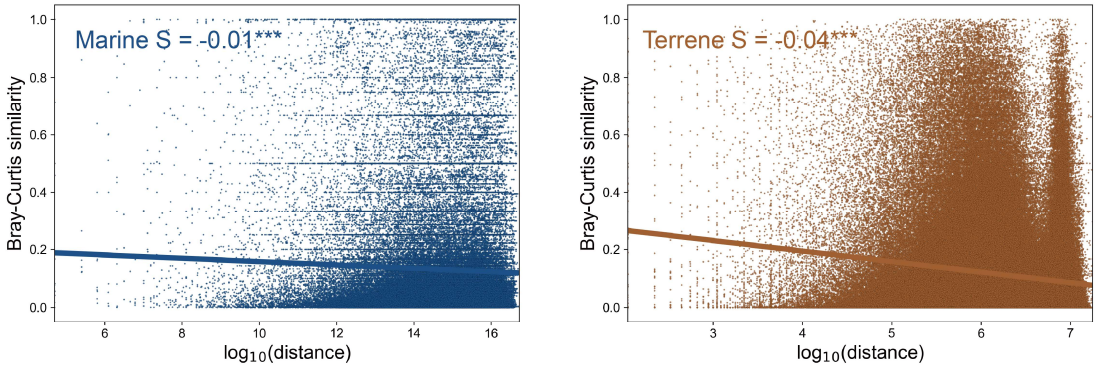
[1] State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China

*Address correspondence to Zhuo Pan (E-mail: panzhuo@sdu.edu.cn, ORCID: 0000-0003-4149-7044) or Zheng Zhang (E-mail: zhangzheng@sdu.edu.cn, ORCID: 0000-0001-9971-6006)

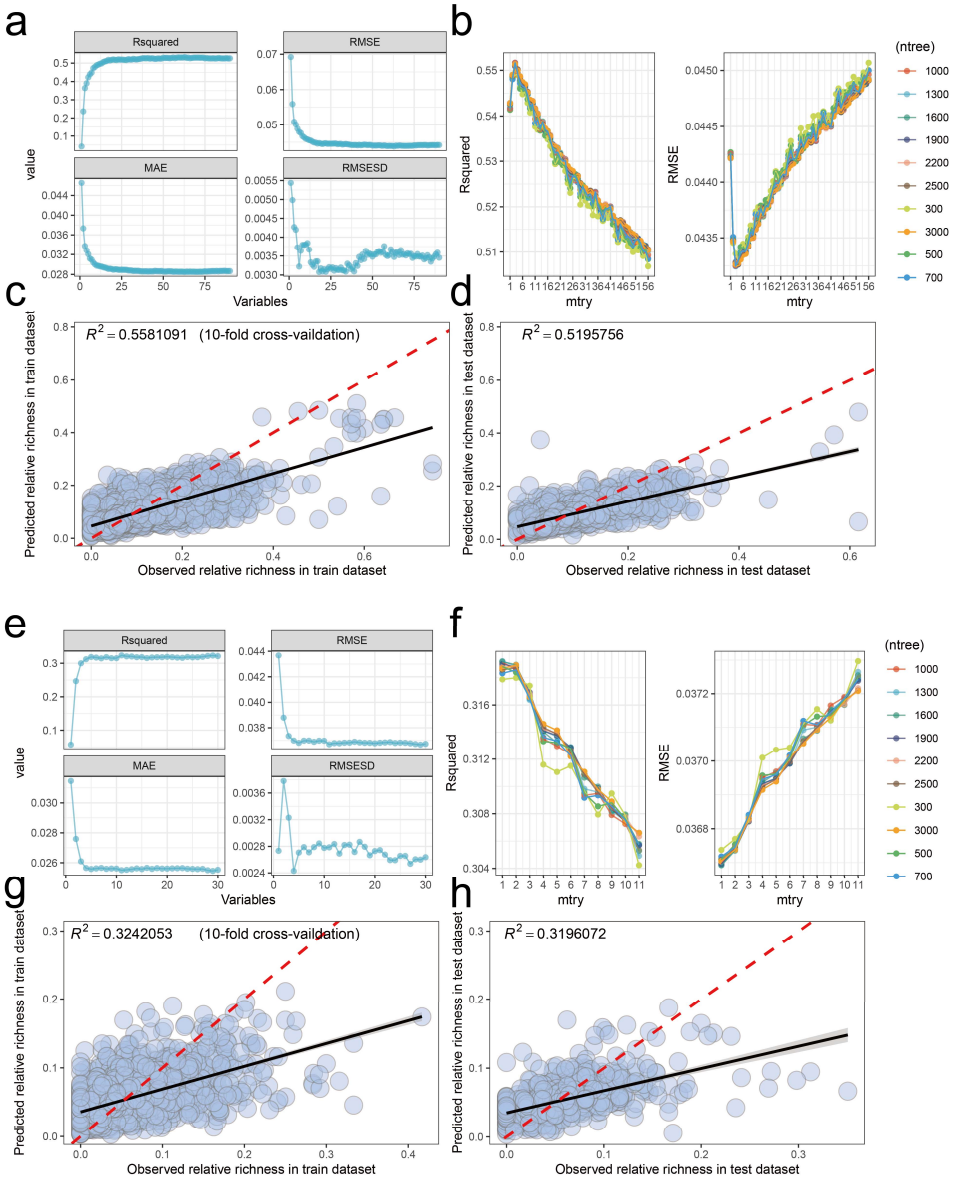
Content



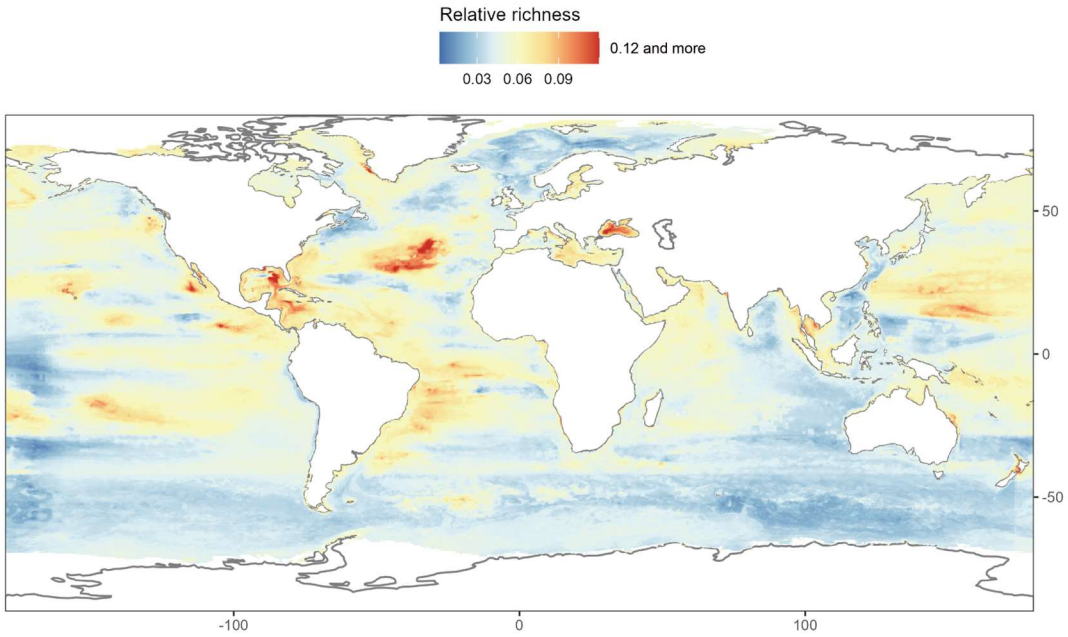
Supplementary Fig. 1 Geographic information of sequencing samples. a The geographical location of 137,672 samples from the MAP database. **b** The categorization of these samples according to their respective habitats. **c** The distribution of diazotrophic OTUs in 18 habitats. A diazotrophic OTU was considered to be distributed within a habitat if it appeared in multiple samples from that habitat ($n \geq 2$).



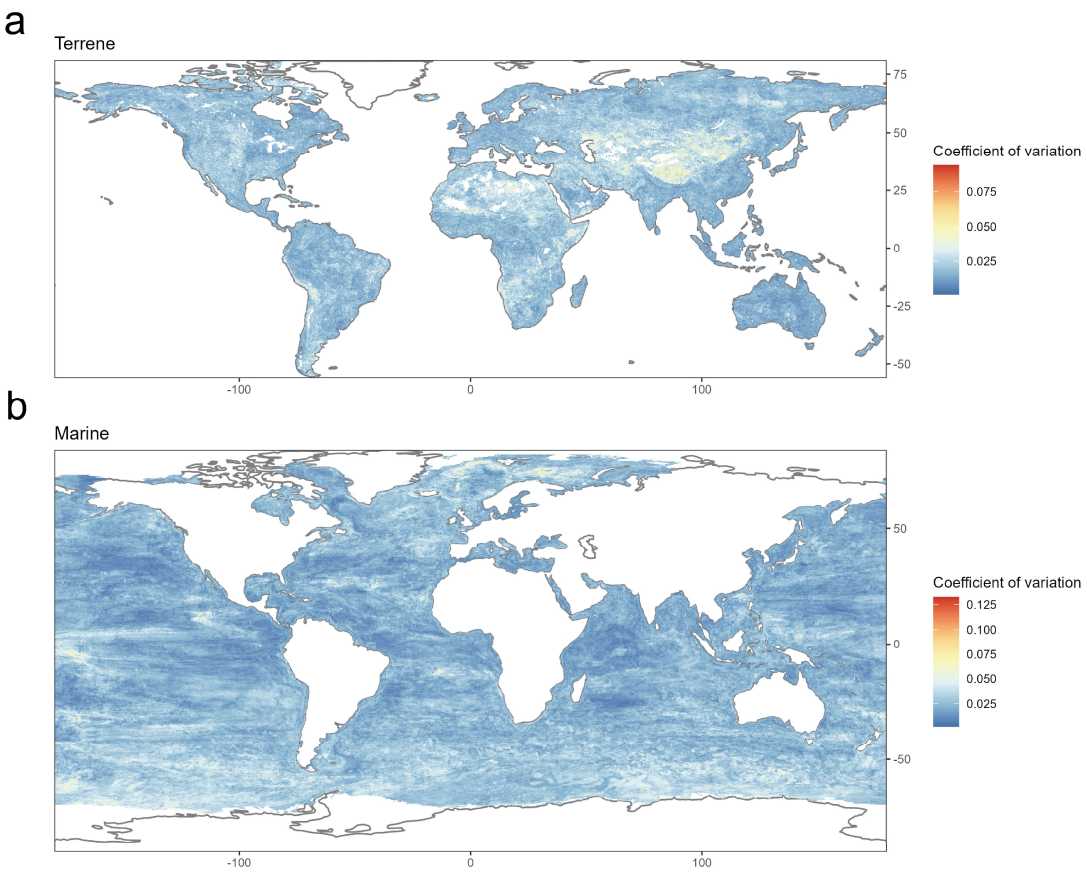
Supplementary Fig. 2 The community similarity and distance-decay relationships of diazotrophic communities across terrestrial and marine ecosystems. A total of 100 extractions were performed on the terrestrial or marine samples, with each extraction randomly selecting 100 individual samples. For each subgroup, the Bray-Curtis similarity among diazotrophic communities was calculated, and its relationship with the $\log_{10}$-transformed geographic distances (meters) was analysed.



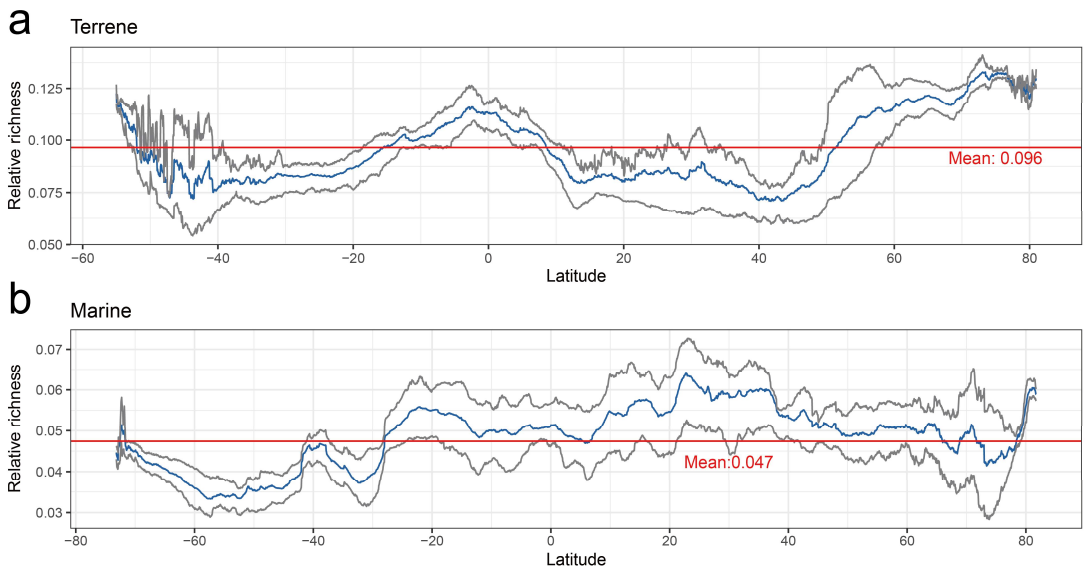
Supplementary Fig. 3 Construction and evaluation of random forest models for relative abundance prediction. **a, e** Feature selection for random forest algorithms to predict the relative richness of diazotrophs based on 10-fold cross-validation. The cross-validated R^2 and root mean square error (RMSE) were used to select the optimal feature sets with the best performance via recursive feature elimination (RFE). **b, f** Hyperparameter tuning for machine learning algorithms to predict the relative richness of diazotrophs based on 10-fold cross-validation. **c, d, g, h** Cross-validation and model assessment. Scatter plot illustrating the performance of the optimal random forest model using 10-fold cross-validation training datasets and independent test datasets.



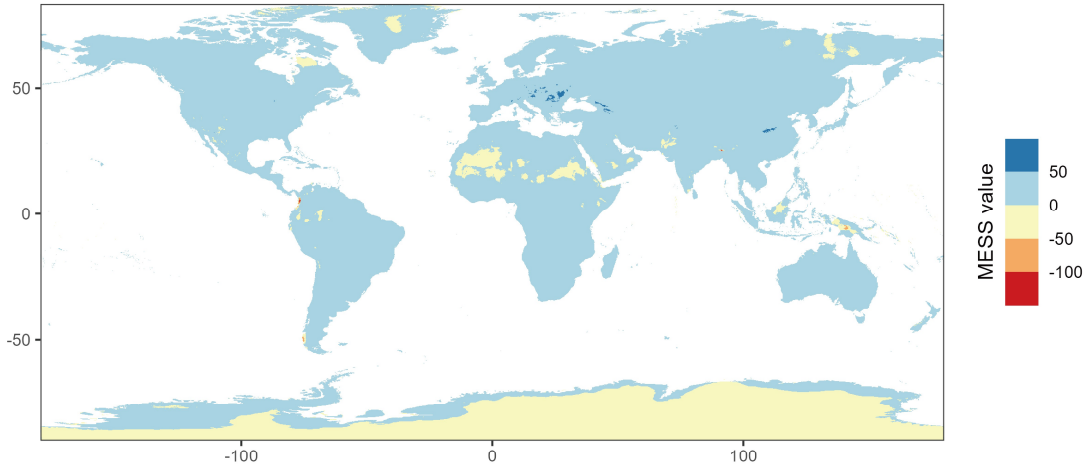
Supplementary Fig. 4 Global maps of the relative richness of marine diazotrophs. Based on spatial covariates of marine environments (Supplementary Dataset 5), we predicted the relative richness of marine diazotrophs globally via a random forest model. A total of 4/5 of the samples were regarded as the model training dataset, and 1/5 were regarded as the model testing dataset (training dataset with 10-fold cross-validation R^2 = 0.324, testing set with R^2 = 0.319; Supplementary Fig. 3).



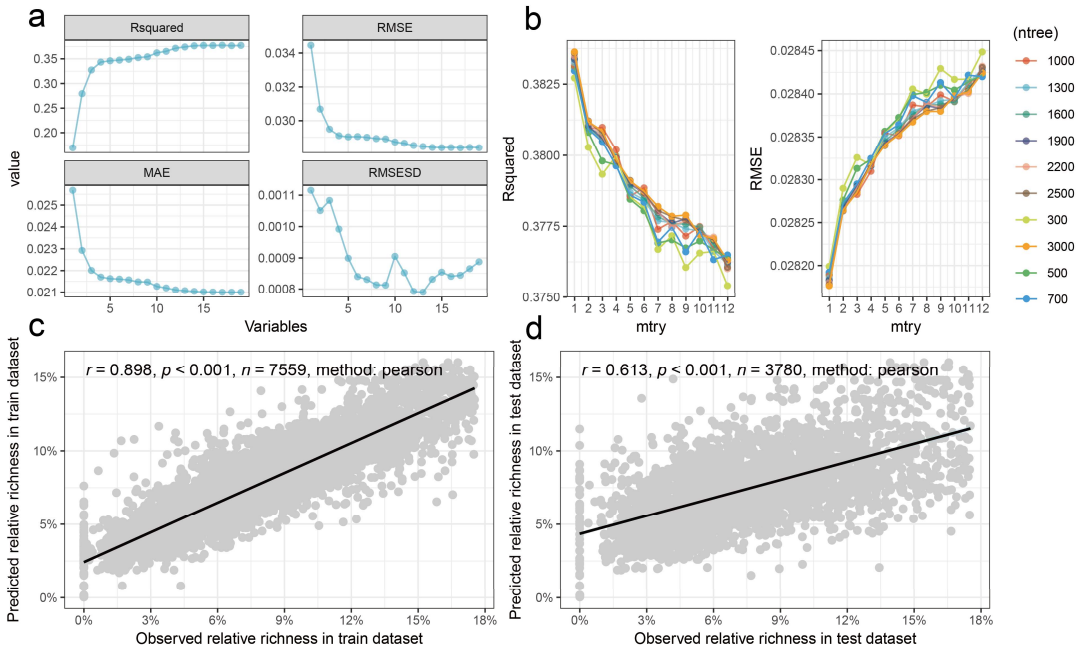
Supplementary Fig. 5 Uncertainty of global maps for predicting the relative richness of diazotrophs. a. map of terrestrial environments; **b.** map of marine environment. The color indicates the coefficient of variation (CV) of ten individual predictions, and a lower CV indicates more reliable predictions.



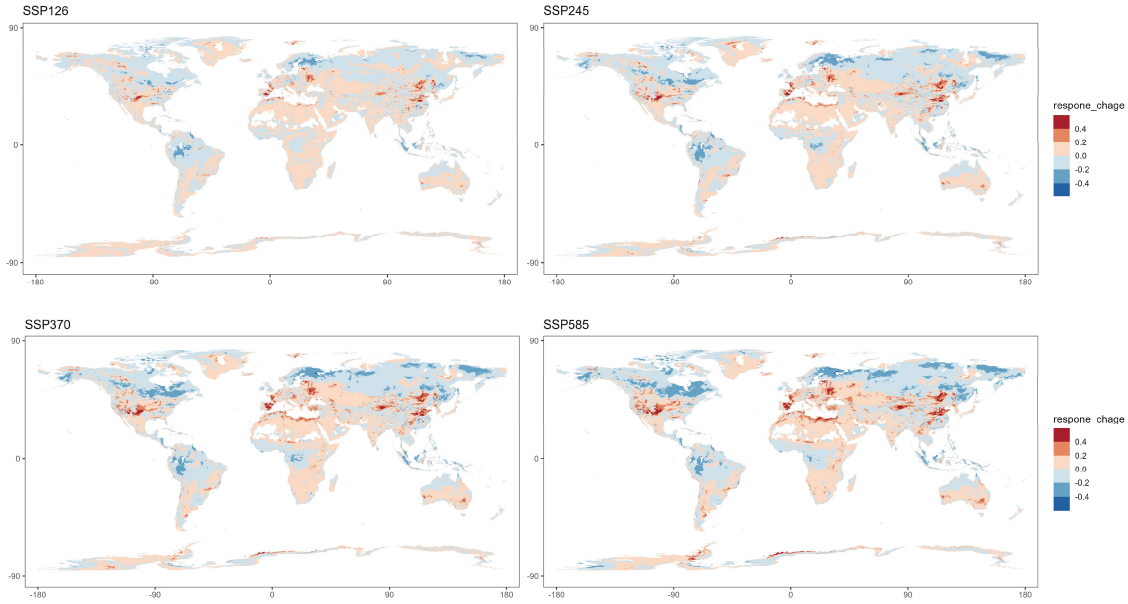
Supplementary Fig. 6 Mean relative richness of diazotrophs across latitude under different climate scenarios. a. terrestrial environments; **b.** marine environment. For any latitude, gray lines represent the upper and lower quartiles (25th and 75th), blue lines represent the medians, and red lines represent the overall means.



Supplementary Fig. 7 Multivariate environmental similarity surface (MESS) across our sampling locations. Based on the 11,339 non-redundant geographic locations collected for the MAP database sample, we used multivariate environmental similarity surface analysis to verify the reliability of climate modelling extrapolations to the global level. The blue areas (MESS values > 0) represent some extrapolation reliability, while the yellow or red areas (MESS values ≤ 0) represent environmental variables that are out of the range of the training data, so we removed these predictively unstable pixels from subsequent modelling predictions.



Supplementary Fig. 8 Construction and evaluation of the relative richness-climate model. The training data were obtained from 19 bioclimatic variables (1970-2000) provided by WorldClim and Pearson correlation test using 2/3 of the samples as a model training dataset and 1/3 as a validation dataset. Feature selection and hyperparameter tuning for the random forest model to predict the relative richness of diazotrophs under future climate change scenarios based on a grid search and 10-fold cross-validation. **a** During RFE, the best subsets of climate variables for prediction were selected based on the highest R^2. **b** After tuning, final model has the lowest cross-validated RMSE value or the highest R-squared. **c, d** The Pearson correlation test was used to examine the correlation between the observed and predicted abundance. The lines represent the least squares regression fit, and the *P*-value was subsequently calculated.



Supplementary Fig. 9 Predicted changes in the relative richness of diazotrophs under different future climate scenarios. A relative richness-climate model was constructed by random forest using the relative richness of diazotrophs and climate variables under four different climate scenarios. The predictive models were cross-validated by the common Pearson correlation test using 2/3 of the samples as the model training dataset and 1/3 of the samples as the validation dataset. All climate variables were derived from WorldClim using a 5 min (approximately 0.083°) resolution. The prediction of relative richness under future climate conditions relies on data derived from 21 different CMIP6 downscaled global change models (GCMs; see detailed information in Methods). The relative changes in the relative richness under the different GCMs compared to those under the current climate conditions were averaged.

Supplementary Table 1. CMIP6 downscaled global change models for future relative richness of diazotrophs.

GCM names	SSP126	SSP245	SSP370	SSP585
ACCESS-CM2	√	√	√	√
BCC-CSM2-MR	√	×	×	×
CMCC-ESM2	√	√	√	√
EC-Earth3-Veg	√	√	√	√
FIO-ESM-2-0	√	√	×	√
GFDL-ESM4	√	×	√	×
GISS-E2-1-G	√	√	√	√
HadGEM3-GC31-LL	√	√	×	√
INM-CM5-0	√	√	√	√
IPSL-CM6A-LR	√	√	√	√
MIROC6	√	√	√	√
MPI-ESM1-2-HR	√	√	√	√
MRI-ESM2-0	√	√	√	√
UKESM1-0-LL	√	√	√	√

Supplementary Table 2 Selected covariates and IncNodePurity (relative importance) for predicting the relative richness of diazotrophs in terrestrial environments.

Variable names	Abbreviations	IncNodePurity	Variable class
Sand content	sand	0.456963172	Soil properties
Broadleaf deciduous tree temperate	PFT7	0.272231208	Human activities
Annual Precipitation	bio_12	0.751497693	Climatic variables
Fourth principal component of the first 38 biocimatic variables	bio39	0.460078935	Climatic variables
Precipitation of Coldest Quarter	bio_19	0.562392892	Climatic variables
Total nitrogen	nitrogen	0.432777563	Soil properties
pH in H2O	phh2o	0.709111893	Soil properties
Cation exchange capacity	cec	0.443317288	Soil properties
Mean Temperature of Wettest Quarter	bio_8	0.487155927	Climatic variables
Temperature Annual Range (bio5-bio6)	bio_7	0.606827822	Climatic variables
Aridity index	ai	0.708394142	Climatic variables
Mean Temperature of Warmest Quarter	bio_10	0.559298083	Climatic variables
Mean moisture index of coldest quarter	bio35	0.490409371	Climatic variables
Third principal component of the first 37 biocimatic variables	bio38	0.558145052	Climatic variables
Highest weekly radiation	bio21	0.633339669	Climatic variables
Coarse fragments	cfvo	0.443650321	Soil properties
Annual mean moisture index	bio28	0.648054469	Climatic variables
Soil Carbon : Nitrogen (C:N) ratio	CN30cm	0.481852284	Soil properties
Coefficient of variation; Normalized dispersion of Vegetation Indices (EVI)	cv	0.480029113	Soil properties
Mean Diurnal Range	bio_2	0.585698657	Climatic variables
Silt content	silt	0.50470044	Soil properties
Phosphorus Fertilizer Application	pfertilizer	0.433531328	Human activities
Soil organic carbon content	soc	0.504392596	Soil properties
Human development index	HDI	0.366345784	Human activities
Longitude	longitude	0.590511558	Others
Temperature Seasonality	bio_4	0.661369734	Climatic variables
Organic carbon stock	ocs	0.438734746	Soil properties
Radiation seasonality	bio23	0.555250529	Climatic variables

Mean moisture index of wettest quarter	bio32	0.539337093	Climatic variables
Human Modification of Terrestrial System	HMTS	0.414920261	Human activities
Precipitation of Warmest Quarter	bio_18	0.573216306	Climatic variables
Lowest weekly radiation	bio22	0.532507863	Climatic variables
Anthropogenic Biomes of the World	anthrome	0.27614773	Human activities
Fifth principal component of the first 39 biocimatic variables	bio40	0.496013475	Climatic variables
Isothermality (bio2/bio7)	bio_3	0.626263609	Climatic variables
Max Temperature of Warmest Month	bio_5	0.519059308	Climatic variables
Annual Mean Temperature	bio_1	0.557161374	Climatic variables
Moisture index seasonality	bio31	0.524106917	Climatic variables
Human influence index	HII	0.35100332	Human activities
Mean Temperature of Driest Quarter	bio_9	0.572975142	Climatic variables
C3 Grass	PFT13	0.329418885	Human activities
Organic carbon density	ocd	0.46648097	Soil properties
Mean moisture index of warmest quarter	bio34	0.561154728	Climatic variables
Other crop: rainfed	PFT27	0.200295829	Human activities
Radiation of warmest quarter	bio26	0.695421	Climatic variables
Bulk density1	bdod	0.51642724	Soil properties
Precipitation Seasonality	bio_15	0.53084427	Climatic variables
Other crop: irrigated	PFT29	0.217281927	Human activities
Entropy; Disorderliness of EVI	Entropy_	0.370651722	Others
Nitrogen Fertilizer Application	nfertilizer	0.474661113	Human activities
Precipitation of Wettest Quarter	bio_16	0.568614536	Climatic variables
Clay content	clay	0.407078503	Soil properties
Latitude	latitude	0.693231382	Others
Soil microbial biomass carbon	SMC30cm	0.487738851	Soil properties
Soil microbial biomass nitrogen	SMN30cm	0.505307305	Soil properties
Biomass	biomass	0.407684567	Others
Highest weekly moisture index	bio29	0.507756313	Climatic variables
Phosphorus in Manure Production	pmanure	0.500328949	Human activities
Precipitation of Wettest Month	bio_13	0.601079471	Climatic variables
Bioenergy crop: irrigated	PFT30	0.096668631	Human activities
C4 Grass	PFT14	0.163783192	Human activities
Nitrogen in Manure Production	nmanure	0.495457858	Human activities
Needleleaf evergreen tree temperate	PFT1	0.181544322	Human activities

Supplementary Dataset 1 The nitrogenase-related genes from representative genomes of prokaryotes.

Supplementary Dataset 2 The genomic classification information of prokaryotes based on colocalization of nitrogenase genes.

Supplementary Dataset 3 Geographic data of the 137,672 samples collected from MAP database.

Supplementary Dataset 4 The classification and 16S sequence of 593 diazotrophic OTUs.

Supplementary Dataset 5 Covariates for predicting the relative richness of diazotrophs in terrestrial and marine environments.