Supplementary Materials

Reduction of global diazotroph diversity is driven by anthropogenic climate change

Peng Li[1], Zhuo Pan[2, 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.

[2] The Affiliated Cancer Hospital of Zhengzhou University & Henan Cancer Hospital, Zhengzhou 450008, 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)

This PDF file includes:

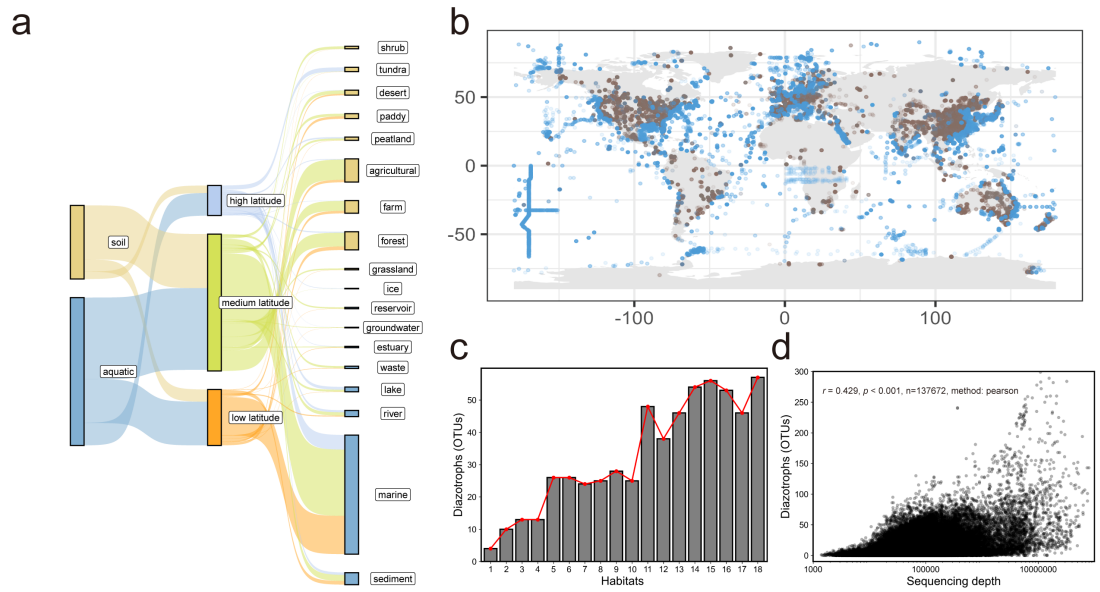
Supplementary Figures 1 to 11

Supplementary Tables 1 and 2

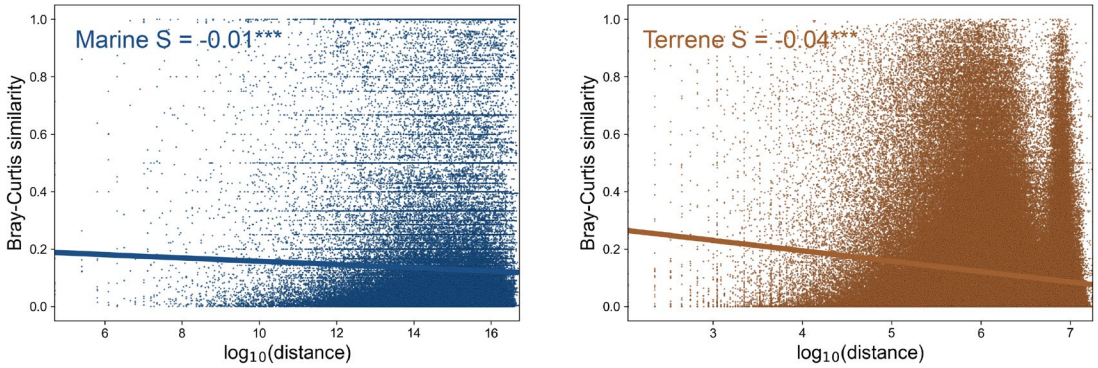
Legends for Supplementary Dataset 1 to 6

Other Supplementary Materials for this manuscript include the following:

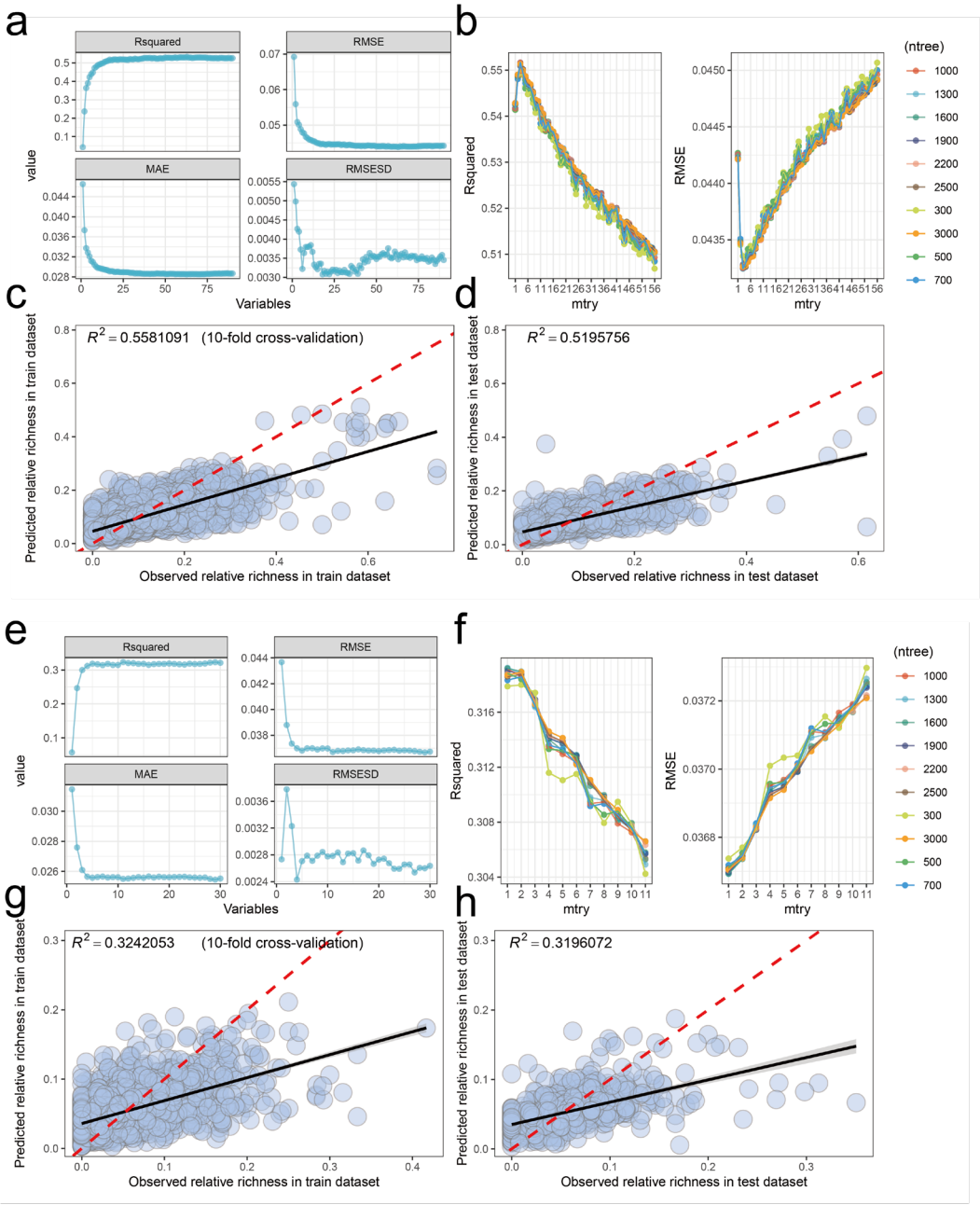
Supplementary Dataset 1 to 6



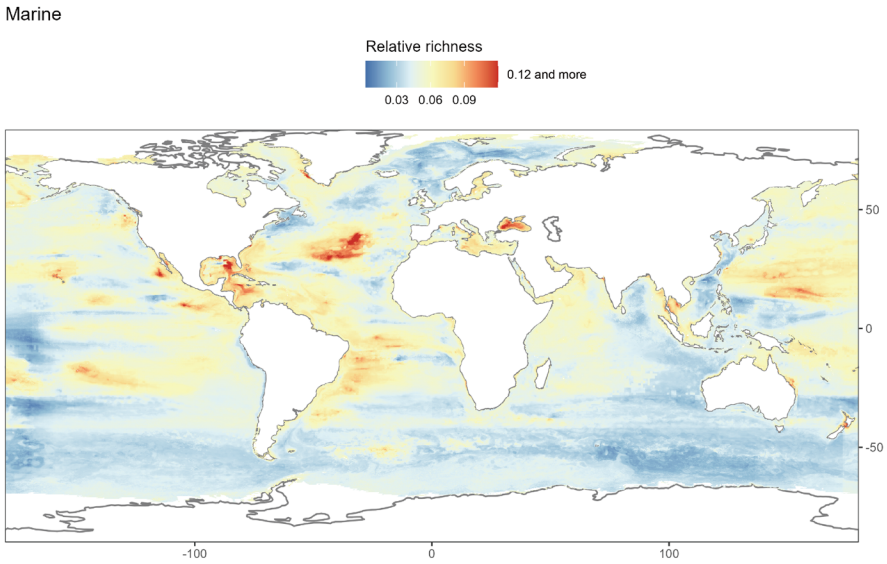
Supplementary Fig. 1 Geographic information of sequencing samples. a Geographical location of 137,672 samples from the MAP database. **b** Categorization of these samples according to their respective habitats. **c** Distribution of diazotrophic OTUs in 18 habitats. A diazotrophic OTU was considered to be distributed within a habitat if it appeared in multiple samples from that habitat ($n \geq 2$). **d** Relationships between sample depth and diazotroph diversity. The Pearson correlation test was used to examine the correlation between sample depth and absolute diazotroph richness, and the *P* value was calculated.



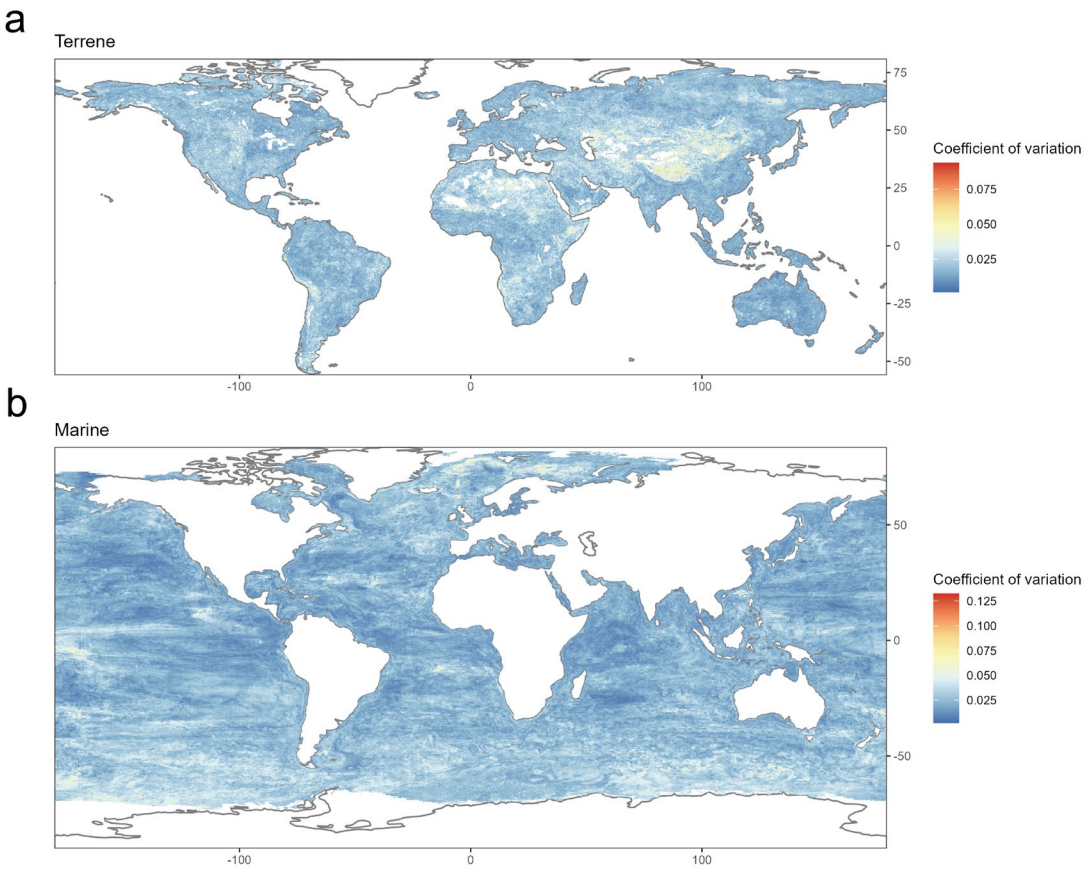
Supplementary Fig. 2 Community similarity and distance-decay relationships of diazotrophic communities across terrestrial and marine ecosystems. A total of 100 extractions were performed on the terrestrial or marine samples, with each extraction randomly selecting 100 individual samples. For each subgroup,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



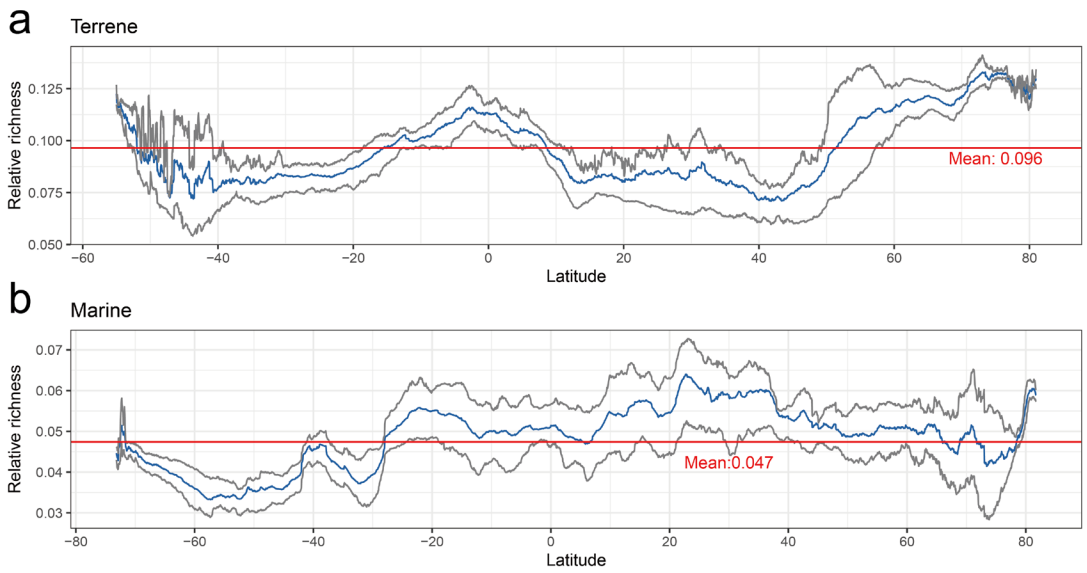
Supplementary Fig. 3 Construction and evaluation of random forest models for relative abundance prediction. **a, e** Feature selection for random forest algorithms to predict the relative richness of diazotrophs on the basis of 10-fold cross-validation. The cross-validated R^2 and root mean square error (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 on the basis of 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.



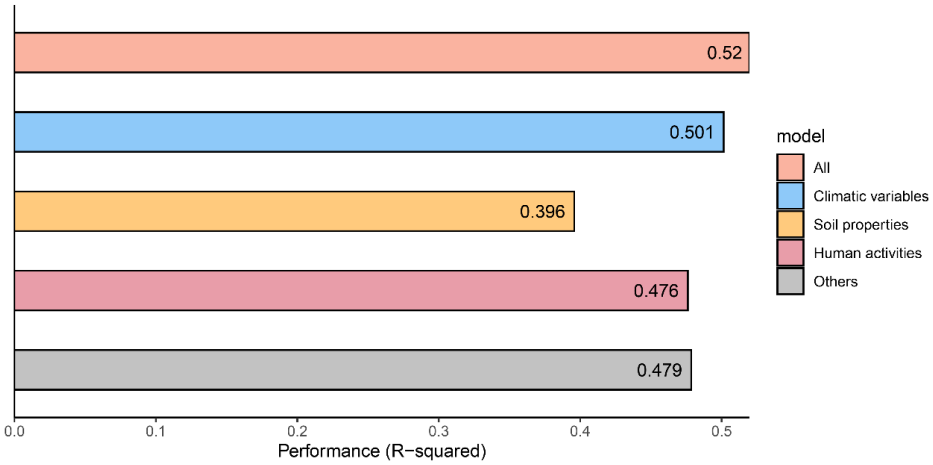
Supplementary Fig. 4 Global maps of the relative richness of marine diazotrophs. On the basis of spatial covariates of marine environments (**Supplementary Dataset 5**), we predicted the relative richness of marine diazotrophs globally via a random forest model. A total of 4/5 of the samples were regarded as the model training dataset, and 1/5 of the samples 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**).



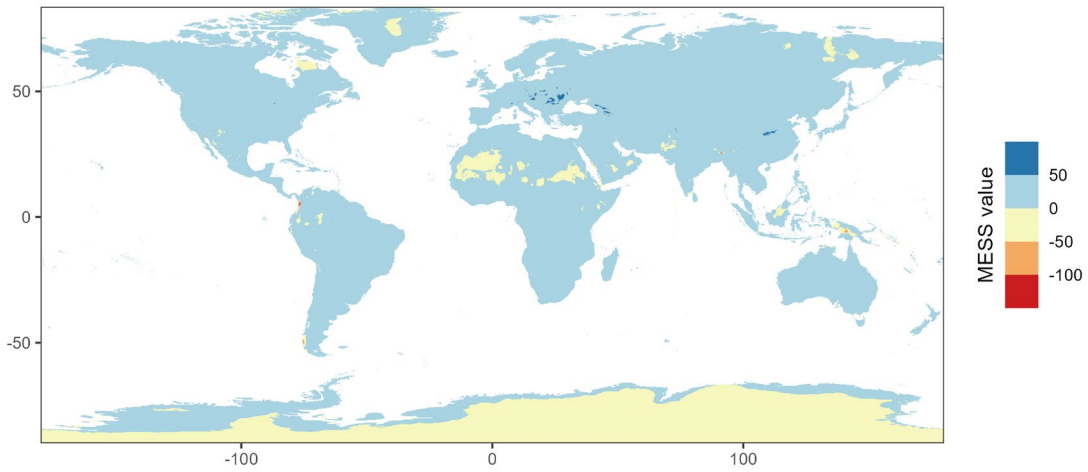
Supplementary Fig. 5 Uncertainty of global maps for predicting the relative richness of diazotrophs in terrestrial (a) and marine (b) environments. The color indicates the coefficient of variation (CV) of ten individual predictions, and a lower CV indicates more reliable predictions.



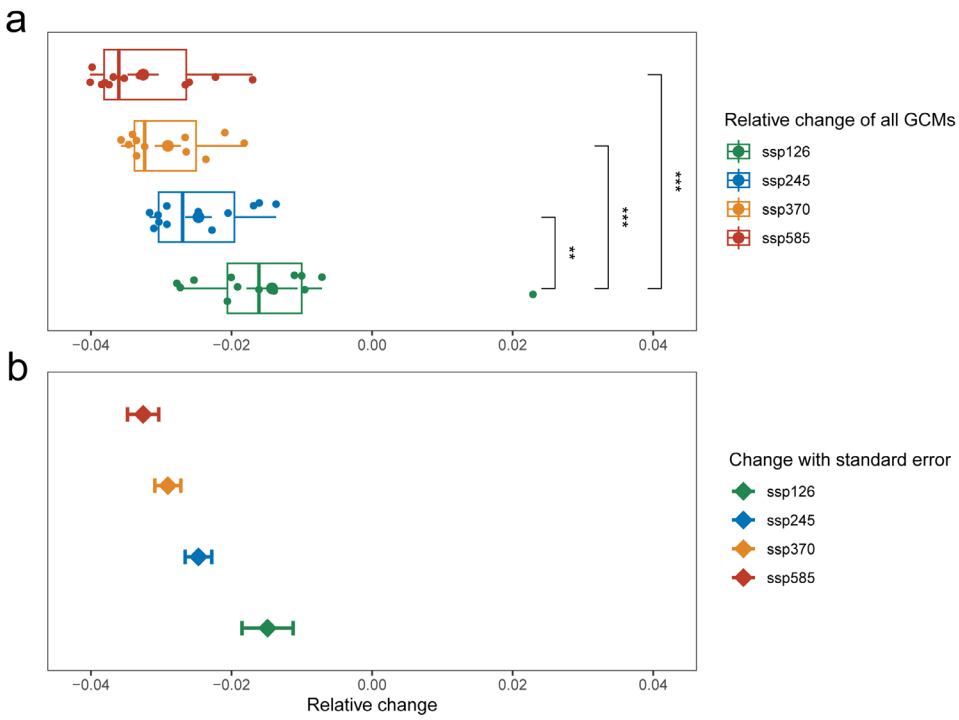
Supplementary Fig. 6 Mean relative richness of diazotrophs across latitudes. For any latitude, gray lines represent the upper and lower quartiles (25^{th} and 75^{th}), blue lines represent the medians, and red lines represent the overall means.



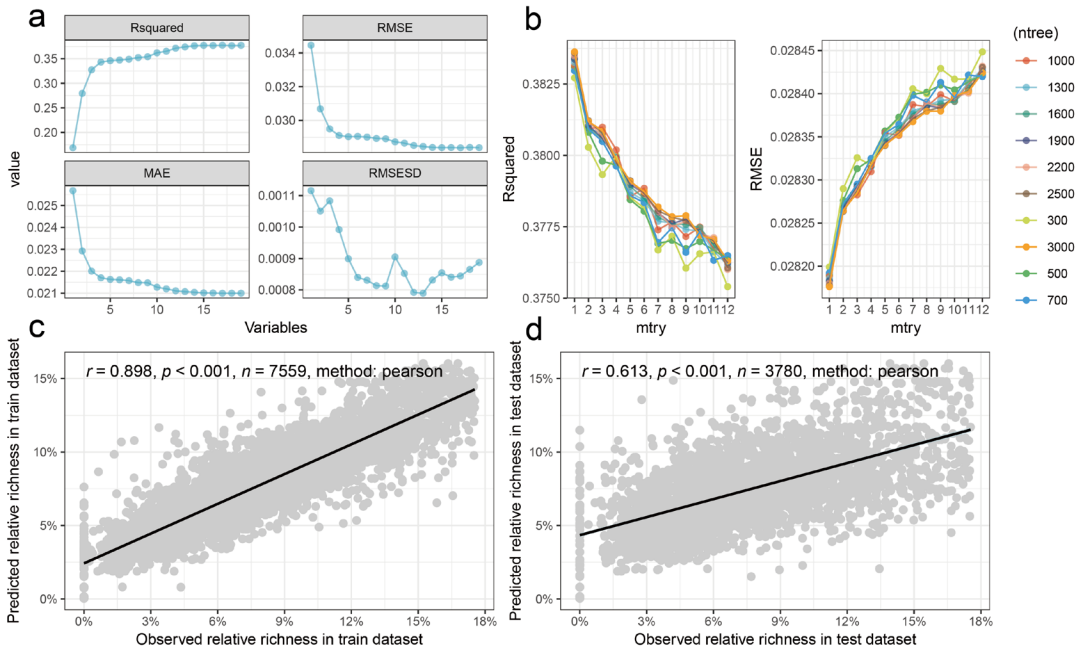
Supplementary Fig. 7 Variations in diazotroph relative richness explained by various models. Each group of environmental covariates selected by the global diazotroph relative richness map model was used to build individual random forest models. Model performance reflects the explained variation in diazotroph relative richness for different types of covariates.



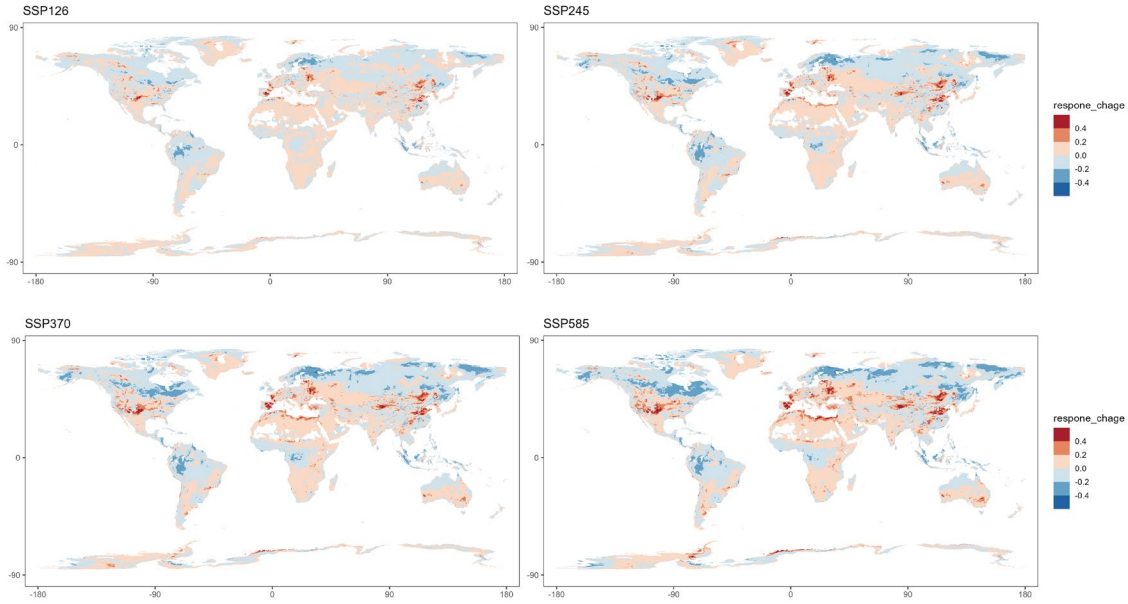
Supplementary Fig. 8 Multivariate environmental similarity surface (MESS) across our sampling locations. Based on the 11,339 nonredundant geographic locations collected for the MAP database sample, we used multivariate environmental similarity surface analysis to verify the reliability of climate modelling extrapolations to the global level. The blue areas (MESS values > 0) represent some extrapolation reliability, whereas 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. 9 Deviation range of diazotroph relative richness changes across all GCMs. On the basis of the predictions of each downscaled global change model, we calculated the variation within SSP (**a**) and the standard error for changes (**b**) of diazotroph relative richness in the future.



Supplementary Fig. 10 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 tests using 2/3 of the samples as a model training dataset and 1/3 as a test 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 square. **b** After tuning, the model has the lowest cross-validated RMSE value or the highest R-squared value when the ntree is 3000 and the mtry is 1. **c, d** The Pearson correlation test was used to examine the correlation between the observed and predicted relative richness. The lines represent the least squares regression fit, and the p value was subsequently calculated.



Supplementary Fig. 11 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 used 2/3 of the samples as the model training dataset and 1/3 of the samples as the test dataset. All the climate variables were derived from WorldClim at 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 with those under the current climate conditions were averaged.

Supplementary Table 1 CMIP6 downscaled global change models for the 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 of covariates in randomforest model) for predicting the relative richness of diazotrophs in terrestrial environments.

IncNodePurity	Var_name	Var_class
0.456963172	sand	Soil properties
0.272231208	PFT7	Human activities
0.751497693	bio_12	Climatic variables
0.460078935	bio39	Climatic variables
0.562392892	bio_19	Climatic variables
0.432777563	nitrogen	Soil properties
0.709111893	phh2o	Soil properties
0.443317288	cec	Soil properties
0.487155927	bio_8	Climatic variables
0.606827822	bio_7	Climatic variables
0.708394142	ai	Climatic variables
0.559298083	bio_10	Climatic variables
0.490409371	bio35	Climatic variables
0.558145052	bio38	Climatic variables
0.633339669	bio21	Climatic variables
0.443650321	cfvo	Soil properties
0.648054469	bio28	Climatic variables
0.481852284	CN30cm	Soil properties
0.480029113	cv	Soil properties
0.585698657	bio_2	Climatic variables
0.50470044	silt	Soil properties
0.433531328	pfertilizer	Human activities
0.504392596	soc	Soil properties
0.366345784	HDI	Human activities
0.590511558	longitude	Others
0.661369734	bio_4	Climatic variables
0.438734746	ocs	Soil properties
0.555250529	bio23	Climatic variables
0.539337093	bio32	Climatic variables
0.414920261	HMTS	Human activities
0.573216306	bio_18	Climatic variables
0.532507863	bio22	Climatic variables
0.27614773	anthrome	Human activities
0.496013475	bio40	Climatic variables
0.626263609	bio_3	Climatic variables
0.519059308	bio_5	Climatic variables
0.557161374	bio_1	Climatic variables
0.524106917	bio31	Climatic variables
0.35100332	HII	Human activities
0.572975142	bio_9	Climatic variables
0.329418885	PFT13	Human activities
0.46648097	ocd	Soil properties

0.561154728	bio34	Climatic variables
0.200295829	PFT27	Human activities
0.695421	bio26	Climatic variables
0.51642724	bdod	Soil properties
0.53084427	bio_15	Climatic variables
0.217281927	PFT29	Human activities
0.370651722	Entropy_	Others
0.474661113	nfertilizer	Human activities
0.568614536	bio_16	Climatic variables
0.407078503	clay	Soil properties
0.693231382	latitude	Others
0.487738851	SMC30cm	Soil properties
0.505307305	SMN30cm	Soil properties
0.407684567	biomass	Others
0.507756313	bio29	Climatic variables
0.500328949	pmanure	Human activities
0.601079471	bio_13	Climatic variables
0.096668631	PFT30	Human activities
0.163783192	PFT14	Human activities
0.495457858	nmanure	Human activities
0.181544322	PFT1	Human activities

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

Supplementary Dataset 2 Genomic classification information of prokaryotes based on co-occurrence nitrogenase genes.

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

Supplementary Dataset 4 Classification and 16S rRNA sequence of 593 diazotrophic OTUs.

Supplementary Dataset 5 Correlation coefficients between environmental covariates and the predicted relative abundance of diazotrophs.

Supplementary Dataset 6 Covariates for predicting the relative richness of diazotrophs in terrestrial and marine environments.