Supplementary Information- **High-emission socioeconomic pathways threaten *phoD*-harboring bacterial communities in cold ecosystems**

Lin Xu[1,2], Chaonan Li[3], Xiangzhen Li[4], Minjie Yao[4], Bo Tu[2], Lixia Wang[1], Li Zhang[1], Chengming You[1], Lihua Tu[1], Bo Tan[1], Yongping Kou[2*], Zhenfeng Xu[1*]

[1] *National Forestry and Grassland Administration Key Laboratory of Forest Resources Conservation and Ecological Safety on the Upper Reaches of the Yangtze River & Forestry Ecological Engineering in the Upper Reaches of the Yangtze River Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu, 611130, China;*

[2] *CAS Key Laboratory of Mountain Ecological Restoration and Bioresource Utilization& Ecological Restoration and Biodiversity Conservation Key Laboratory of Sichuan Province, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, 610041, China ;*

[3]*Ecological Security and Protection Key Laboratory of Sichuan Province, Mianyang Normal University, Mianyang, 621000, China*

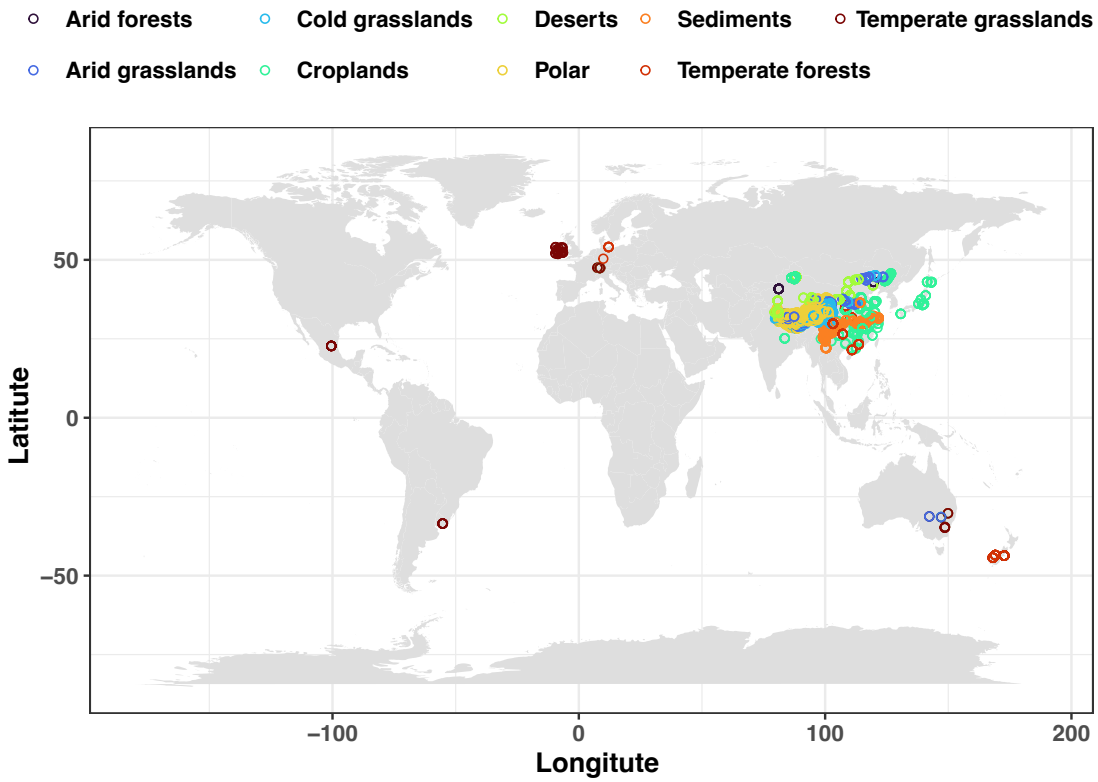
[4]*Engineering Research Center of Soil Remediation of Fujian Province University, College of Resources and Environment, Fujian Agriculture and Forestry University, Fuzhou 350002, China;*

***Correspondence:**

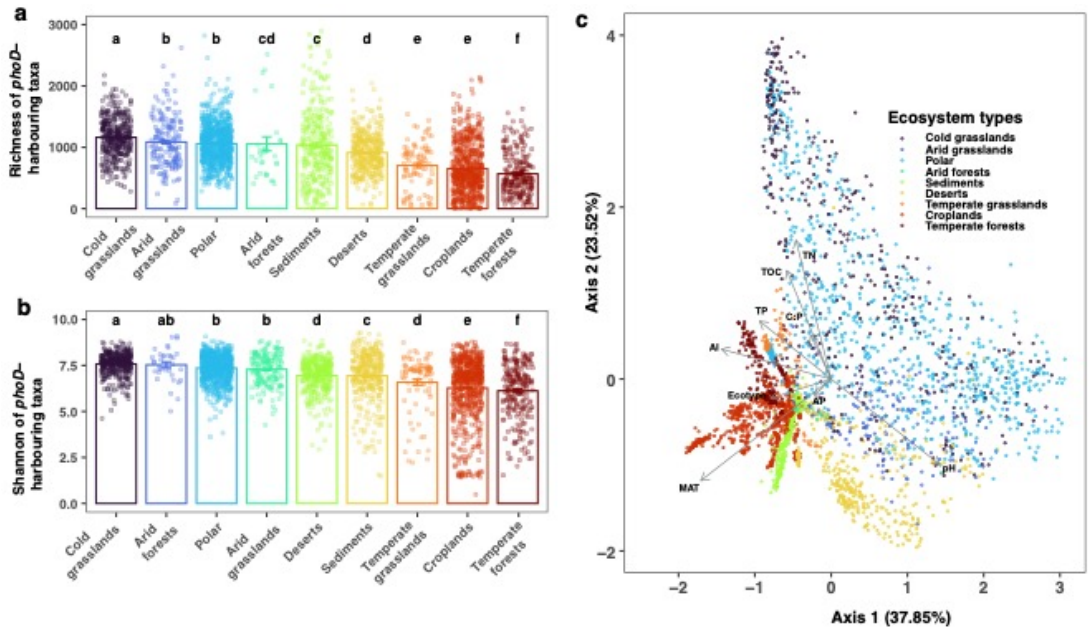
Yongping Kou, E-mail: kouyp@cib.ac.cn

Zhenfeng Xu, E-mail: xuzf@sicau.edu.cn

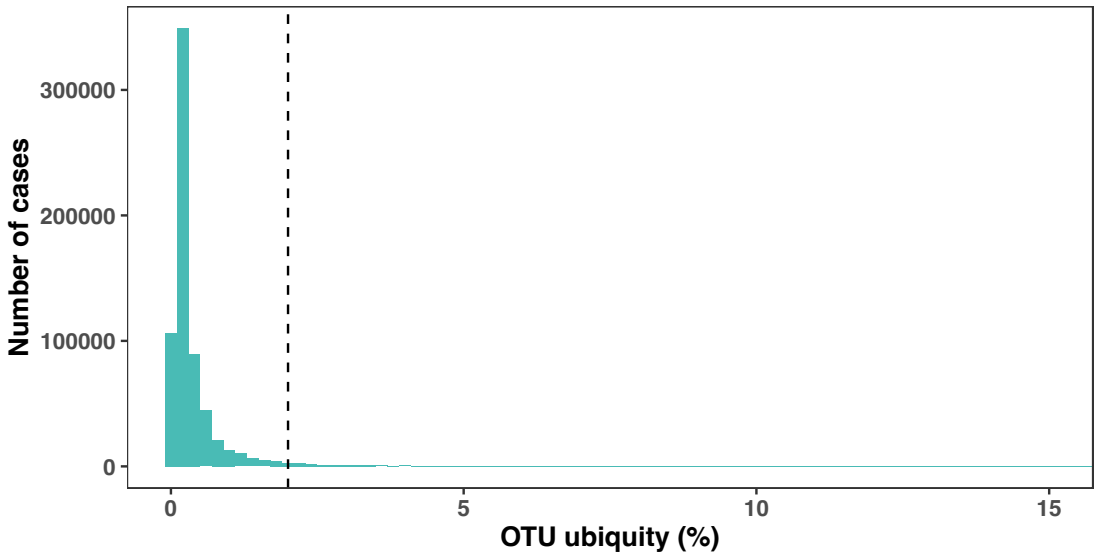
Supplementary Figures



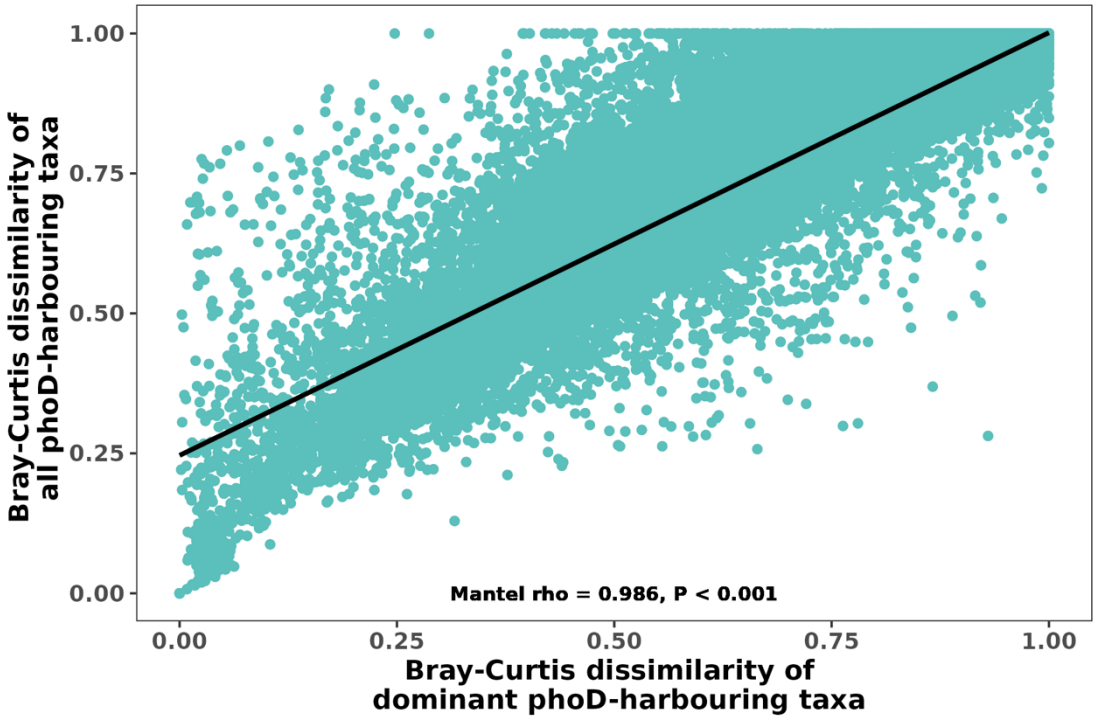
Supplementary Fig. S1 Site locations of the 3195 samples across 60 publications included in this study. The ecosystem types are classified according to Köppen climate classification and land use types that the samples were collected. Croplands include arid, cold and temperate croplands. Different ecosystem types are represented by distinct colours.



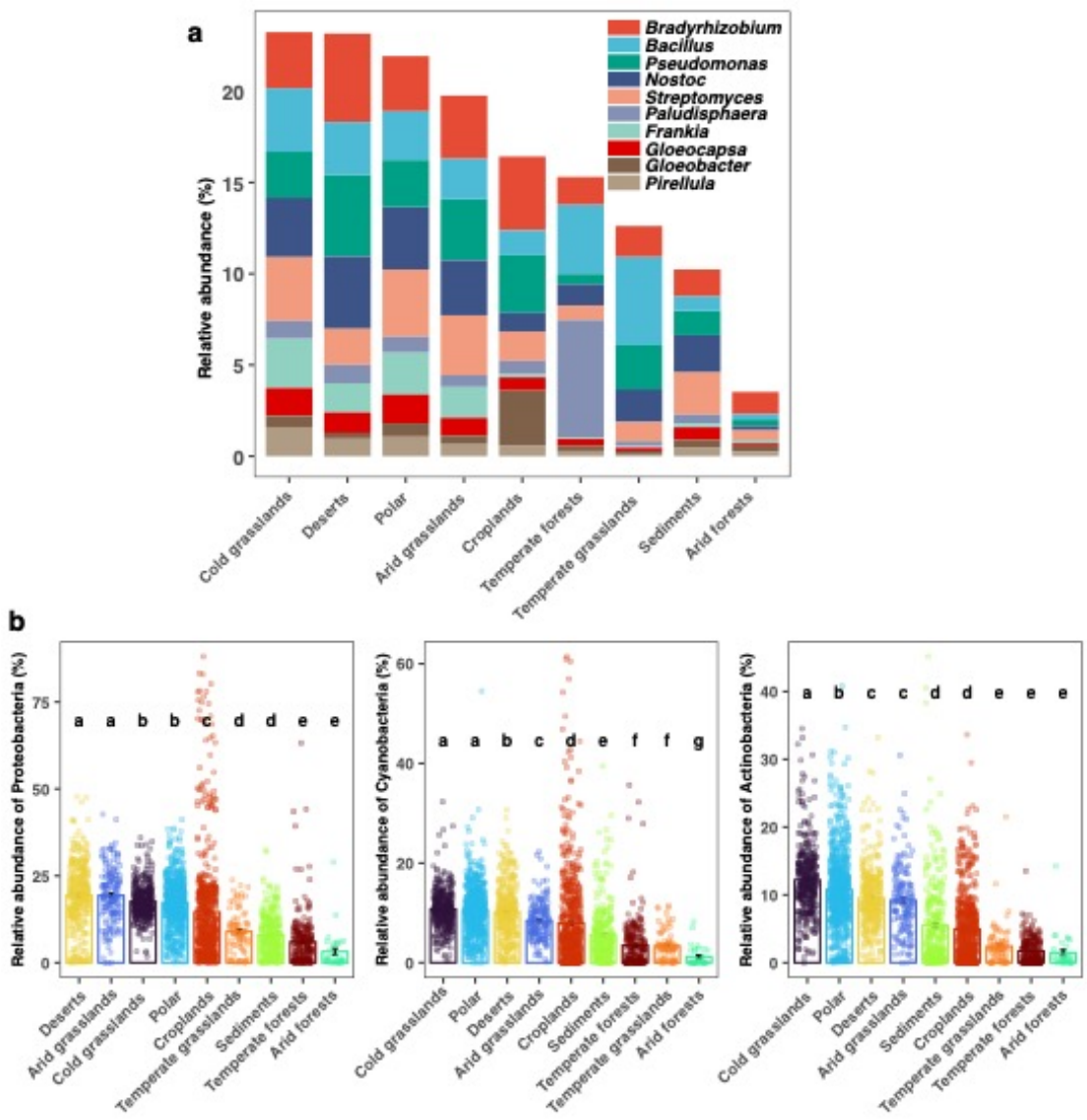
Supplementary Fig. S2 a Observed richness (indicated by chao1 index) among ecosystem types, **b** Observed Shannon diversity among ecosystem types. In panel **a** and **b**, the data are shown as mean value ± standard error (SE), and the centre line in the boxplot signifies the mean, while the lower and upper hinges denote the standard error around the mean. Each whisker corresponds to the minimum and maximum values of the data, respectively. Different lowercase letters indicate statistically significant differences according to the Kruskal-Wallis test at $P < 0.05$. **c** Distance-based redundancy analysis (dbRDA) of *phoD*-harbouring communities based on Bray-Curtis dissimilarity. Different ecosystem types are indicated by distinct colours.



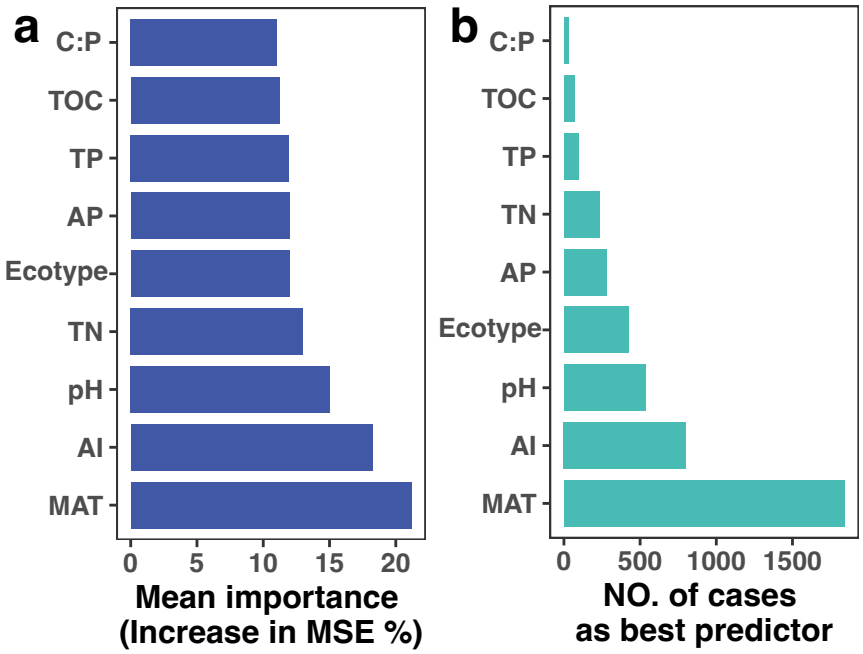
Supplementary Fig. S3 Histograms illustrating the distributions of ubiquity for phylotypes across the 3175 samples from 60 studies. A dashed line represents the threshold used to identify dominant phoD-harbouring phylotypes, defined as those present in more than 100 samples.



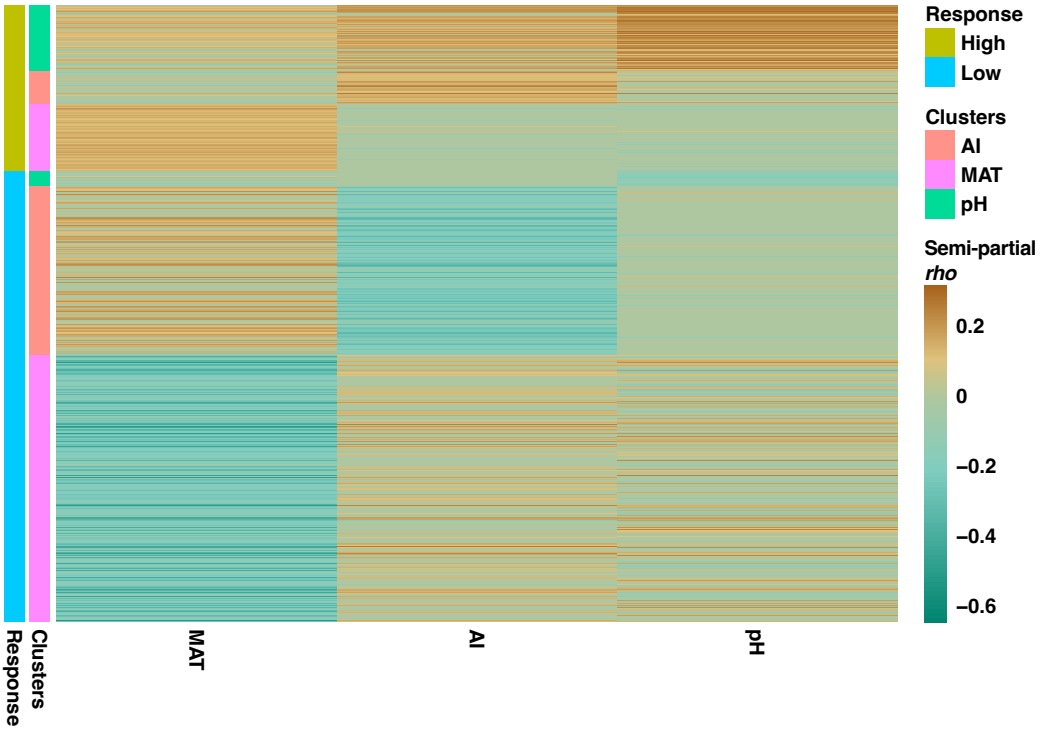
Supplementary Fig. S4 Relationships between beta diversity (differences in community compositions among samples) of the dominant (19,194 phylotypes) and the remaining 650,327 *phoD*-harbouring phylotypes, based on abundance-considering Bray-Curtis dissimilarity distance.



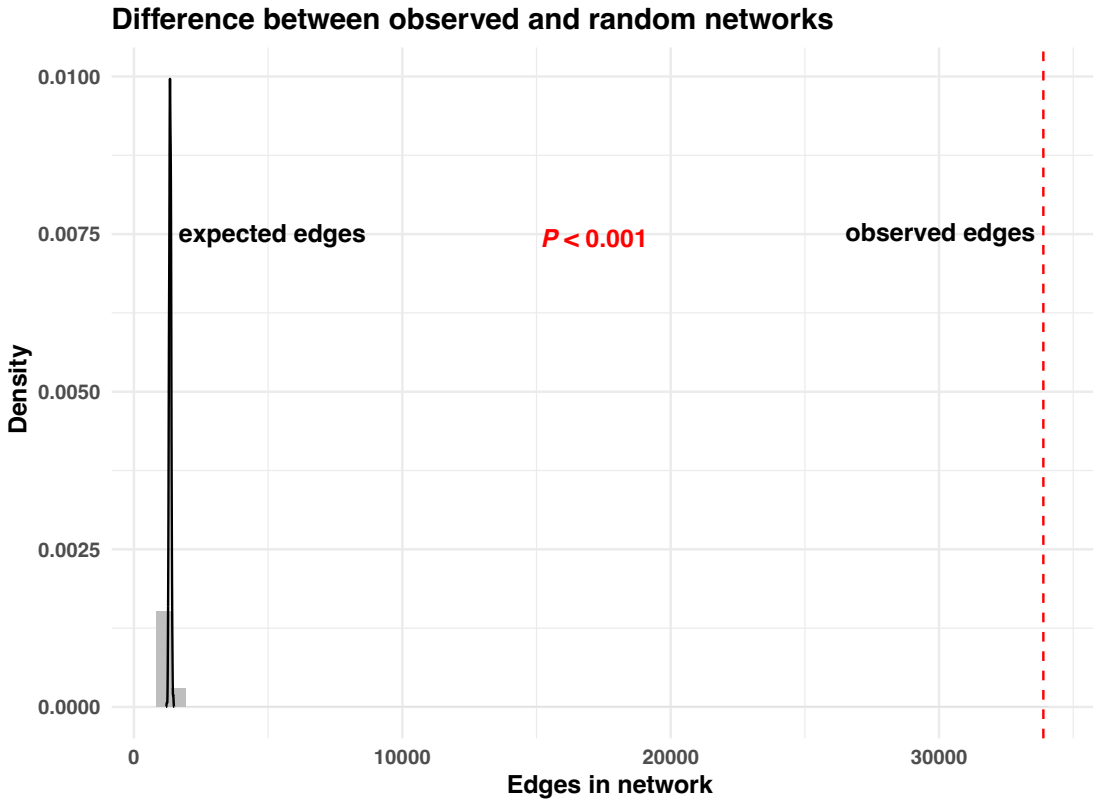
Supplementary Fig. S5 a Representative genera of dominant phoD-harboring phylotypes. **b** Differences in relative abundances of the top 3 groups of the dominant *phoD*-harbouring phylotypes among ecosystem types. The data are shown as mean value ± standard error (SE), and the centre line in the boxplot signifies the mean, while the lower and upper hinges denote the standard error around the mean. Each whisker corresponds to the minimum and maximum values of the data, respectively. Different lowercase letters indicate significant differences according to the Kruskal-Wallis test at $P < 0.05$.



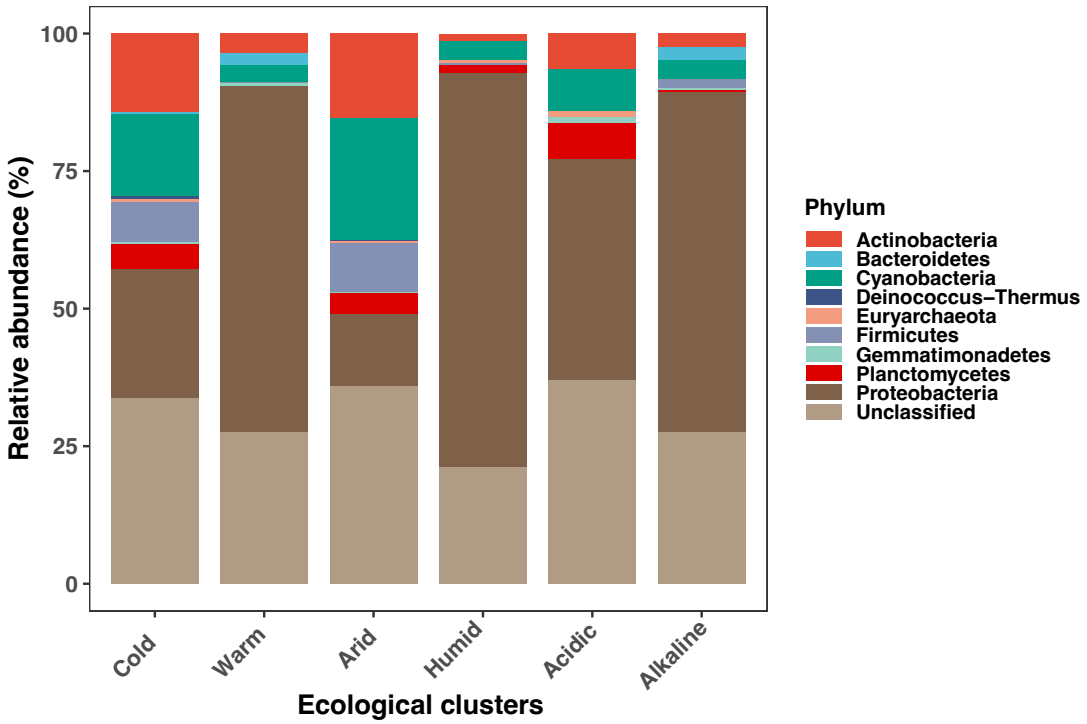
Supplementary Fig. S6 The major predictors of the distribution of dominant *phoD*-harbouring taxa across continents. **a** The averaged importance of environmental factors across 19,194 random forest models in predicting the relative abundances of *phoD*-harbouring taxa. **b** The number of cases, out of 19,194 random forest models, for which a particular environmental factor emerged as the best predictor for the dominant *phoD*-harbouring taxa. AI, aridity index; MAT, mean annual temperature; TOC, soil (or sediment) total organic carbon; TN, soil (or sediment) total nitrogen; TP, soil (or sediment) total phosphorus, AP soil (or sediment) available phosphorus, Ecotype, Ecosystem types.



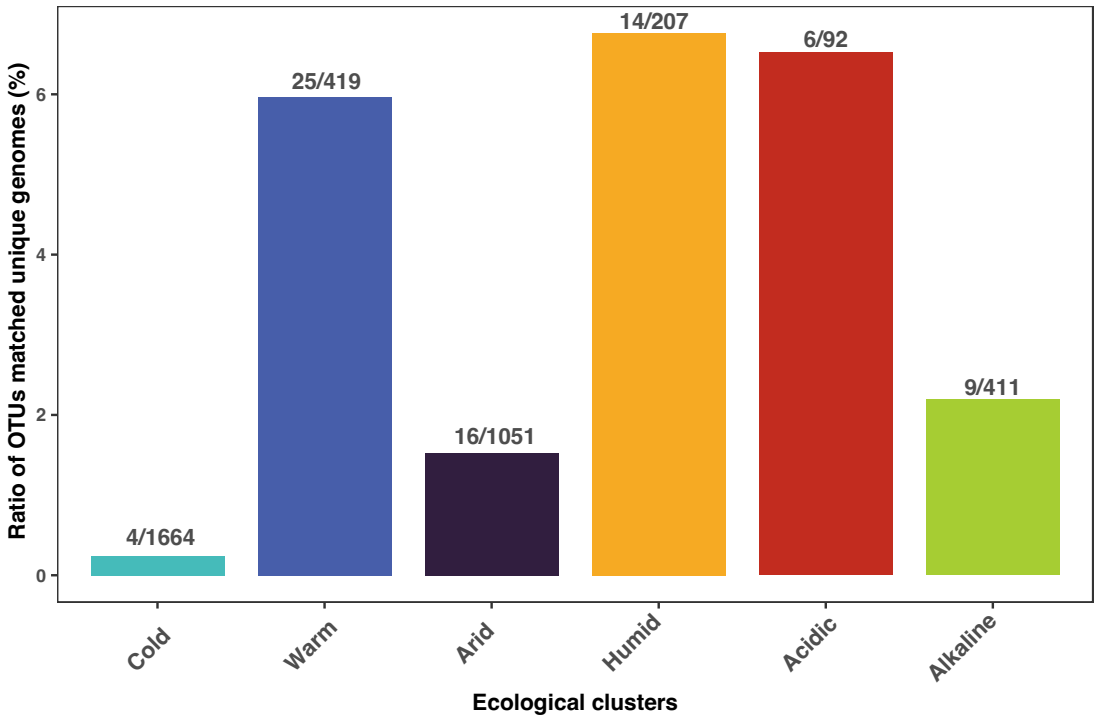
Supplementary Fig. S7 Heatmap illustrating semi-partial correlations between the relative abundances of each *phoD*-harbouring phylotype with the top 3 important environmental factors identified by random forest machine learning models. Data were sorted and coloured using the ecological cluster information provided in Supplementary Table S3. MAT, mean annual temperature, AI, aridity index.



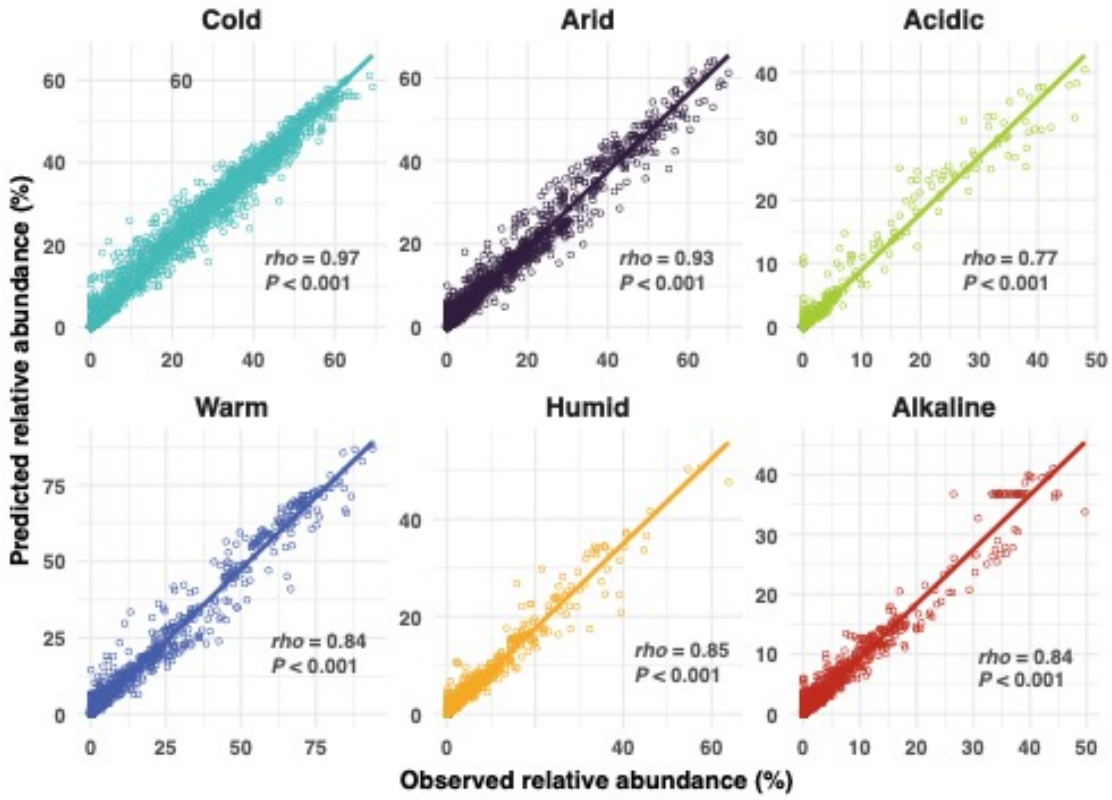
Supplementary Fig. S9 Distributions of environmental assembly across observed co-occurrence network and null models (dashed line). The histogram indicates the expected distribution of number of edges between taxa with occurrence in the networks based on 1000 random graphs under the Erdős–Rényi model if there were no structuring of co-occurrence patterns by environmental forces (e.g. the null model). The dashed line indicates the number of edges between taxa in the observed co-occurrence network, and the *P*-value signifies the significance level of differences between the observed versus expected edge numbers.



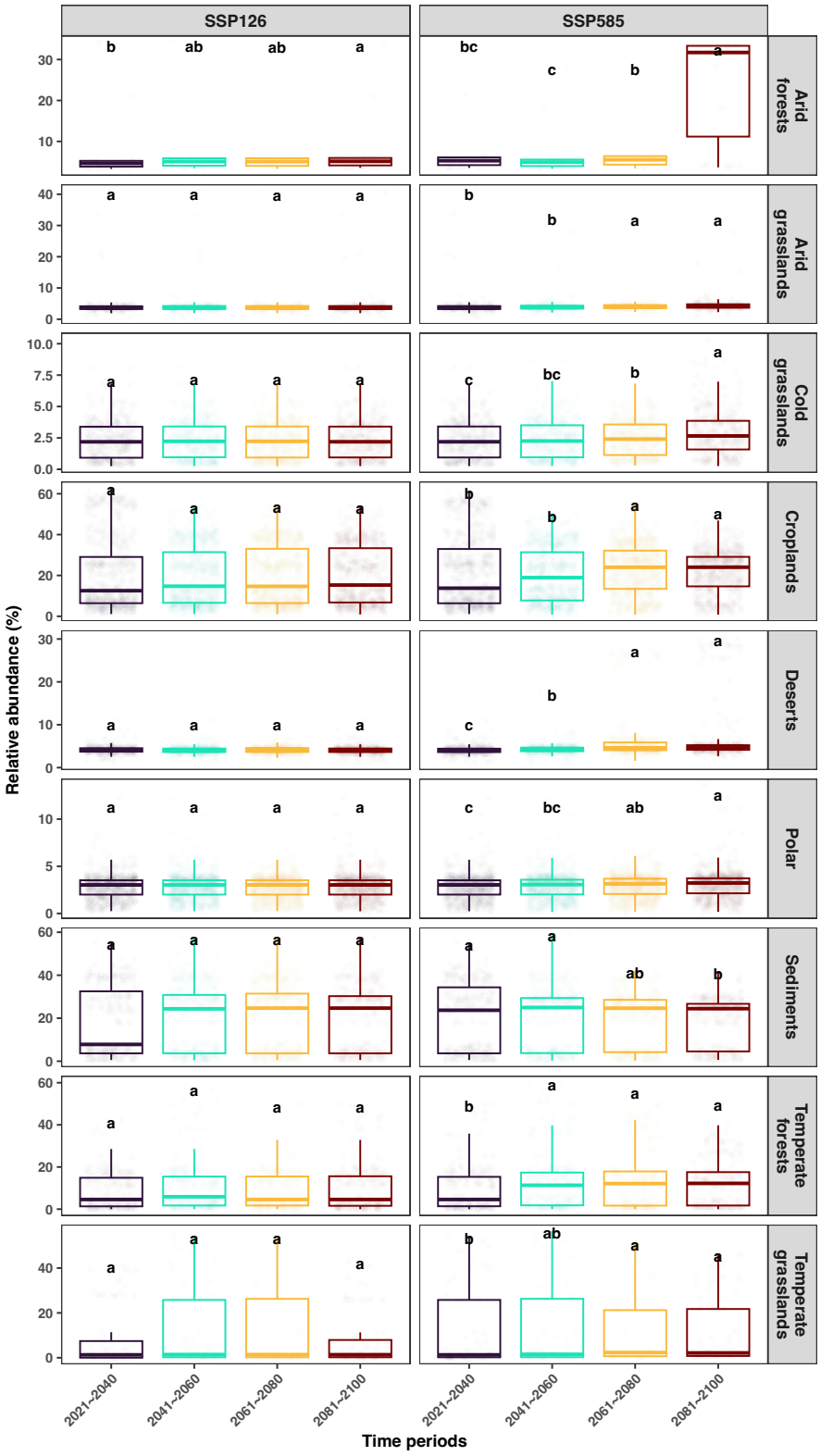
Supplementary Fig. S9 Taxonomic composition (% of OTUs within each cluster) for the six well-defined and other ecological clusters of *phoD*-harbouring phylotypes sharing habitat preferences.



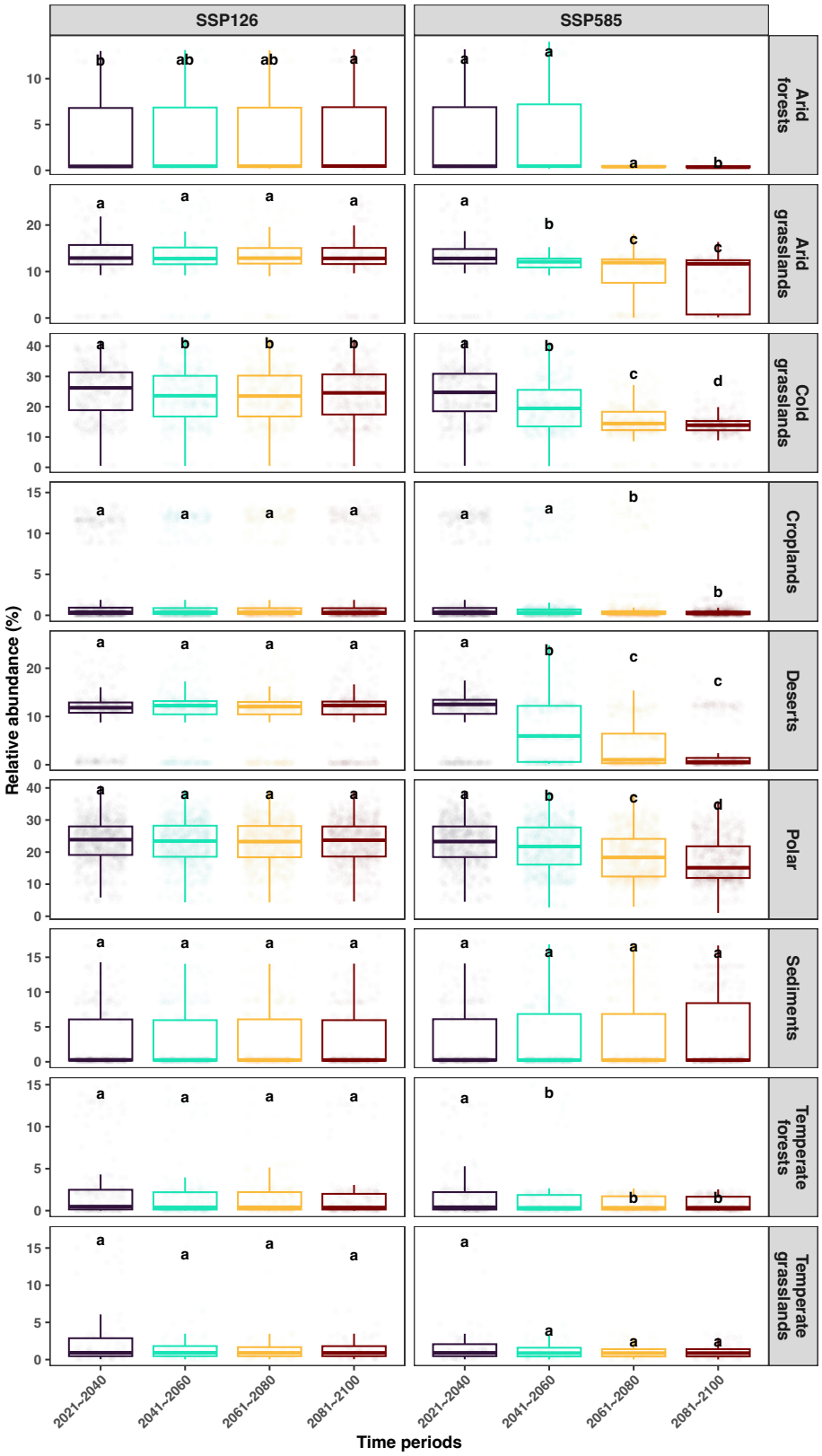
Supplementary Fig. S10 Ratio of matched genomes within each ecological cluster. The total number of phylotypes for which representative genomic data are available per cluster is shown at the top of each bar plot.



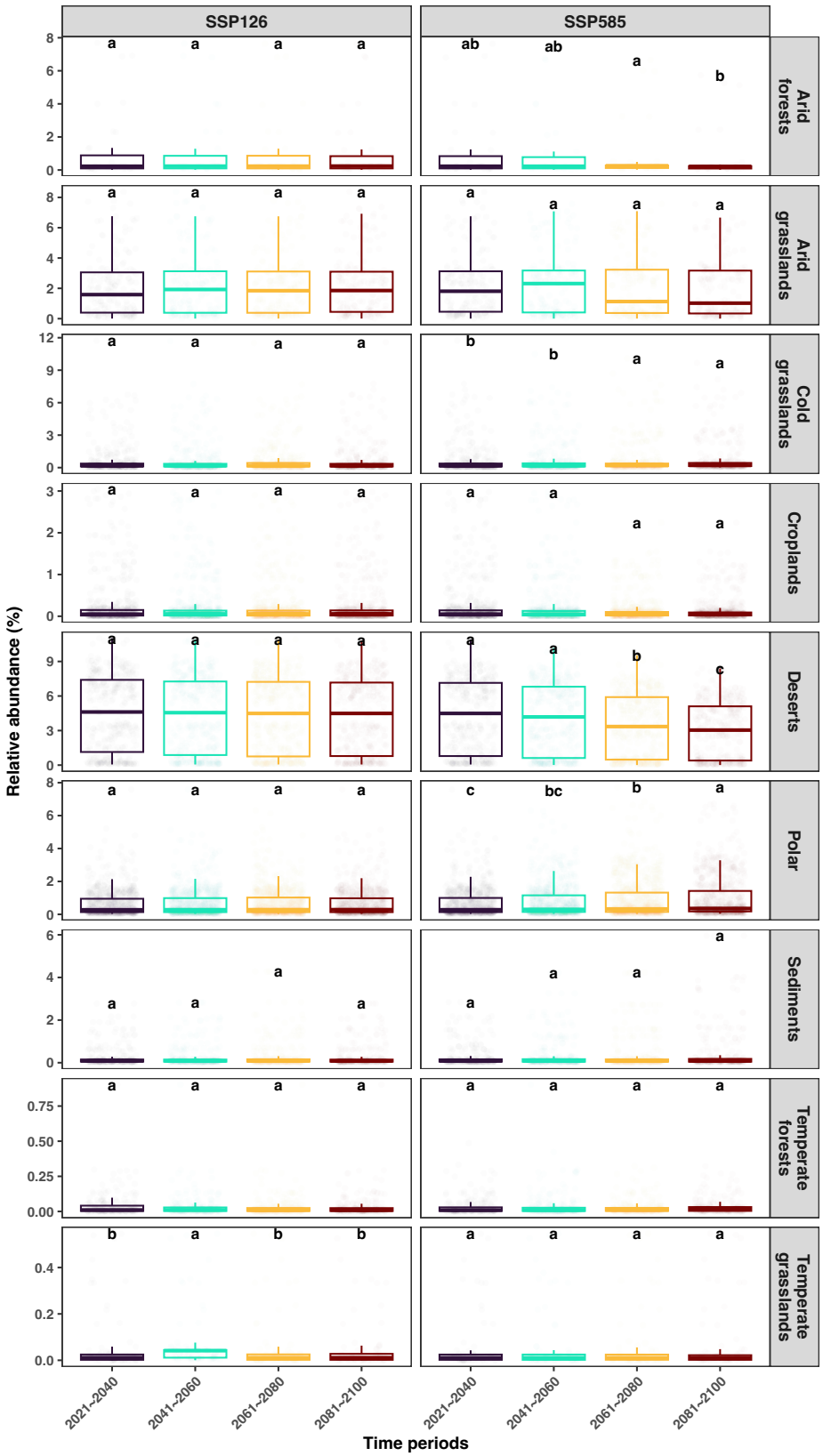
Supplementary Fig. S11 Performances of the machine learning random forest models in predicting the relative abundances of major ecological clusters. The Spearman's *rho* and *P* values indicate the strength of the correlation and its significance, respectively, between the observed and the predicted values generated by the random forest models.



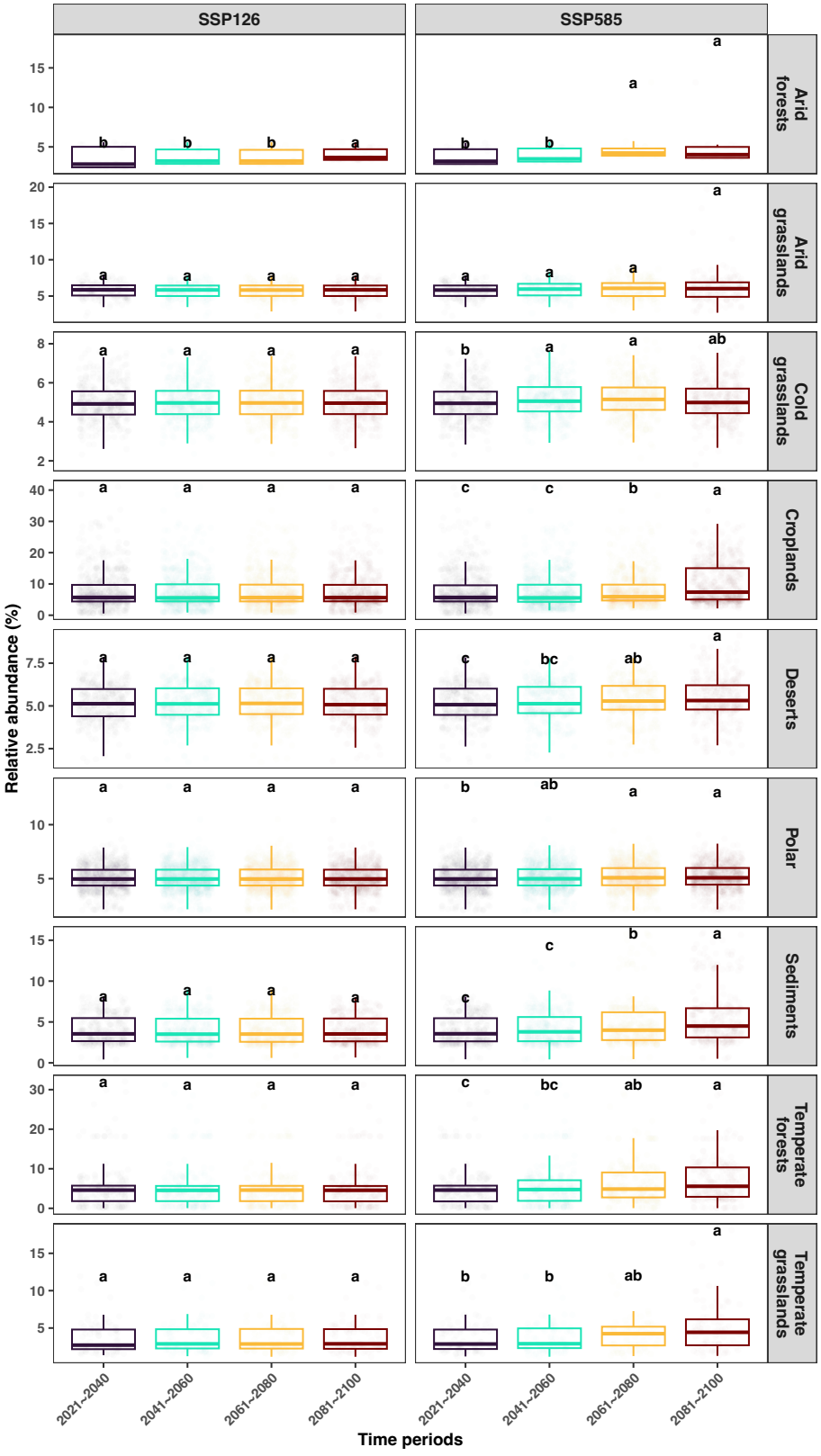
Supplementary Fig. S12 Boxplot illustrating the differences in predicted relative abundances of the warm cluster across future time periods under sustainable (SSP126) and high-emission (SSP585) socioeconomic pathways for various ecosystem types. Different lowercase letters indicate statistically significant differences according to the Kruskal-Wallis test at $P < 0.05$.



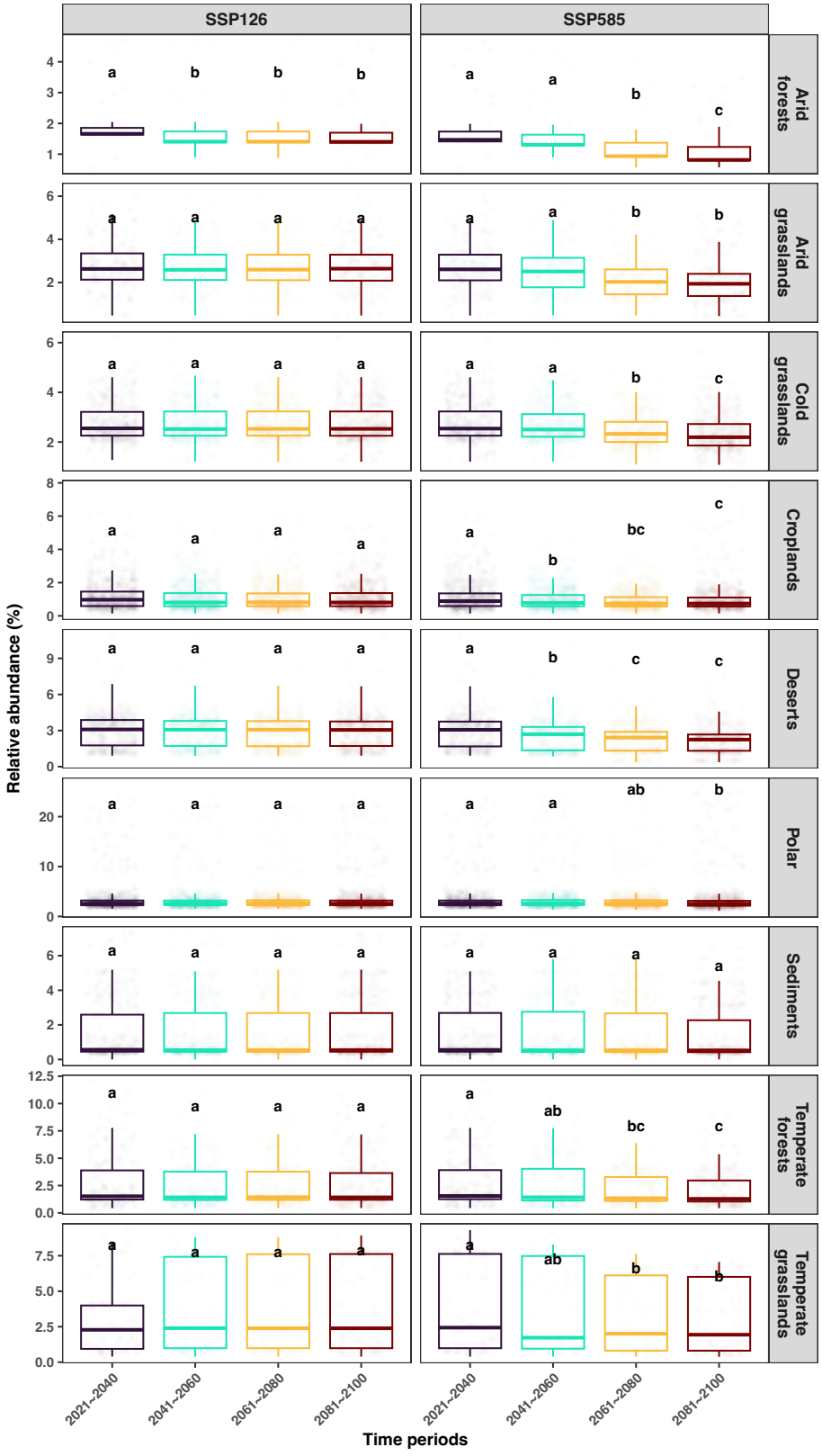
Supplementary Fig. S13 Boxplot illustrating the differences in predicted relative abundances of the cold cluster across future time periods under sustainable (SSP126) and high-emission (SSP585) socioeconomic pathways for various ecosystem types. Different lowercase letters indicate statistically significant differences according to the Kruskal-Wallis test at $P < 0.05$.



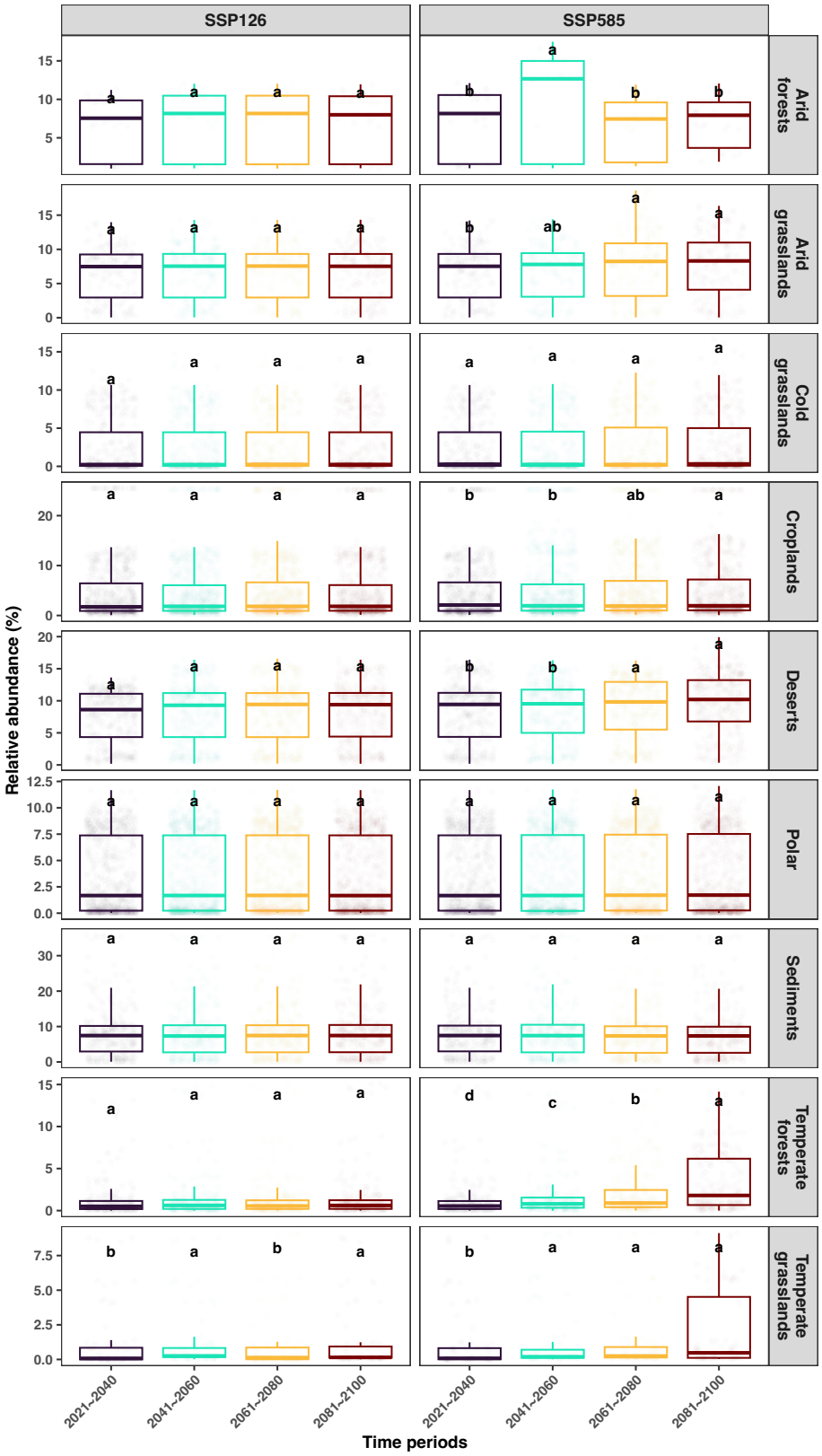
Supplementary Fig. S14 Boxplot illustrating the differences in predicted relative abundances of the arid cluster across future time periods under sustainable (SSP126) and high-emission (SSP585) socioeconomic pathways for various ecosystem types. Different lowercase letters indicate statistically significant differences according to the Kruskal-Wallis test at $P < 0.05$.



Supplementary Fig. S15 Boxplot illustrating the differences in predicted relative abundances of the humid cluster across future time periods under sustainable (SSP126) and high-emission (SSP585) socioeconomic pathways for various ecosystem types. Different lowercase letters indicate statistically significant differences according to the Kruskal-Wallis test at $P < 0.05$.



Supplementary Fig. S16 Boxplot illustrating the differences in predicted relative abundances of the acidic cluster across future time periods under sustainable (SSP126) and high-emission (SSP585) socioeconomic pathways for various ecosystem types. Different lowercase letters indicate statistically significant differences according to the Kruskal-Wallis test at $P < 0.05$.



Supplementary Fig. S17 Boxplot illustrating the differences in predicted relative abundances of the alkaline cluster across future time periods under sustainable (SSP126) and high-emission (SSP585) socioeconomic pathways for various ecosystem types. Different lowercase letters indicate statistically significant differences according to the Kruskal-Wallis test at $P < 0.05$.

Supplementary Tables

Supplementary Table S1 List of studies included is this study.

Supplementary Table. S2 Environmental information retrieved from the 60 studies included in this study.

Supplementary Table. S3 List of identified dominant *phoD*-harbouring phylotypes from the 3175 soil samples across continents. The list includes information on the taxonomic identity of each phylotype, the ecological cluster it was assigned to, the most closely related reference genome, the ubiquity best predictor and the R^2 values of the random forest models.

Supplementary Table. S4 The KO numbers, function descriptions, gene names, and metabolic processes of the investigated phosphorus cycling genes.

Supplementary Table. S5 Gene count table of 74 *phoD*-harbouring phylotypes (out of the 3,844 total) that belong to well-defined ecological clusters and were well matched to genomes in the GTDB database.

All supplementary tables are available as separate ".xlsx" file