



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole de biologie

**SPATIAL DISTRIBUTION OF ARCTIC SOIL BACTERIA UNDER
CURRENT AND FUTURE ENVIRONMENTAL CONDITIONS**

**Travail de Maîtrise universitaire ès Sciences en comportement, évolution et
conservation, spécialisation « géosciences, écologie et environnement »**
*Master Thesis of Science in Behaviour, Evolution and Conservation,
specialisation « Geosciences, Ecology and Environment »*

par

Liyo SHIGA

**Directeur : Prof. Antoine Guisan
Superviseuse : Dr. Lucie Malard
Expert·e·s : Anonymous
Département d'écologie et évolution (DEE)**

Janvier 2023

ABSTRACT

Since microorganisms dominate soil biodiversity and play key role in ecosystem functions, soil microbial biogeography and effect of environmental changes has recently gained attention. However, our knowledge about soil bacterial distribution in the Arctic region and how it may be affected in the future remains scarce. With the rapid warming of the Arctic and the resulting release of organic carbon from thawing soils, greenhouse gases emission is likely to increase by enhanced microbial activity. Considering this positive feedback and the region's sensitivity to climate change, forecasting the future of Arctic soil bacterial distribution is crucial. Here, Species Distribution Models (SDM) were applied to provide spatial distribution of soil bacterial community in the Arctic region under current environmental conditions and under three different representative concentration pathways (RCP) corresponding to low, medium and high greenhouse gas emission scenario, for 2050 and 2070. The results reveal a positive relationship between latitude and soil bacterial diversity, and that both soil bacterial diversity and community structure are expected to change with combined climate and edaphic changes under all scenarios. This predictive insight into how soil bacterial diversity may be altered by climate change is essential to improve our understanding of its influence on ecosystems worldwide.

KEYWORDS

Microbial communities, soil microbiome, biogeography, polar medium, species distribution model, spatial pattern predictions, species richness, climate change

RÉSUMÉ

Étant donné que les micro-organismes dominent la biodiversité du sol et jouent un rôle clé dans les fonctions de l'écosystème, la biogéographie microbienne du sol et les effets des changements environnementaux ont récemment retenu l'attention. Cependant, nos connaissances sur la distribution des bactéries du sol dans la région arctique et sur la manière dont elle pourrait être affectée à l'avenir restent limitées. Avec le réchauffement rapide de l'Arctique et la libération du carbone organique des sols en fonte qui en résulte, les émissions de gaz à effet de serre sont susceptibles d'augmenter en raison d'une activité microbienne accrue. Au vu de cette rétroaction positive et de la sensibilité de la région au changement climatique, il est crucial de prédire l'avenir de la distribution des bactéries du sol arctique. Ici, la modélisation de la distribution des espèces (SDM) a été appliquée pour déterminer la distribution spatiale de la communauté bactérienne du sol dans la région arctique dans les conditions environnementales actuelles et selon trois voies de concentration représentatives (RCP) correspondant aux scénarios d'émissions faibles, moyennes et élevées de gaz à effet de serre pour 2050 et 2070. Les résultats révèlent une relation positive entre la latitude et la diversité bactérienne du sol, et l'on s'attend à ce que la diversité bactérienne du sol et la structure de la communauté changent avec les changements climatiques et édaphiques combinés dans tous les scénarios. Cet aperçu des effets du changement climatique sur les communautés bactériennes est essentiel pour améliorer notre compréhension de l'influence des bactéries du sol sur les écosystèmes au niveau global.

INTRODUCTION

Climate change is causing major changes on species distribution and ecosystem function worldwide. While these climate change effects have been widely recognized on macroorganism for decades (Araújo et al., 2011; Mantyka-pringle et al., 2012), the impact on soil bacterial community has recently gained attention (Castro et al., 2010; Chu et al., 2020; Jansson & Hofmockel, 2020; Mod et al., 2021). Microorganisms dominate soil biodiversity in terms of richness and abundance (Maron et al., 2018) and are involved in a range of ecosystem functions such as biogeochemical cycling, primary production and organic matter decomposition (Delgado-Baquerizo et al., 2016). Recent advances have been made in soil microbial biogeography (Chu et al., 2020), providing for example evidence that soil microbial biomes differ from terrestrial biomes that are based on animal and plant distribution (Vasar et al., 2022). Dominant taxa, such as Proteobacteria, Actinobacteria and Acidobacteria (Delgado-Baquerizo et al., 2018) and ecological drivers of soil microbial distribution, such as pH, soil organic carbon and moisture, (Fierer, 2017) were identified. Despite these efforts, significant gaps on soil biodiversity study remain in the Arctic region (Cameron et al., 2018; Chu et al., 2020).

Low temperatures, water and nutrient availability combined with high variation of UV radiation are common abiotic stresses of the Polar regions (Jani et al., 2022). This characteristic extreme climate limits the presence of plant and animal species and, as a result, microorganisms dominate these systems. Due to the harsh environmental conditions which require specific physiological adaptations, the Arctic region exhibits unique biogeographical patterns and taxonomic composition of soil bacterial community compared to non-polar regions (Ji et al., 2022). Recently, Malard et al. (2019) have provided the broadest baseline database of Arctic soil bacterial diversity, where pH has been identified as the most explicative factor of soil bacterial community, among other drivers including soil organic carbon, moisture and conductivity. To advance our understanding of the soil bacterial biogeography, the next step is to model the spatial distribution across the Arctic region.

Furthermore, the Arctic region is known to be warming four times faster than the rest of the world (Rantanen et al., 2022). The permafrost, defined as a permanently frozen layer in the soil, is harbouring a substantial amount of organic carbon (Tarnocai et al., 2009). This ground is thawing at an increasing rate (Brown & Romanovsky, 2008) and releasing the stored organic

carbon as a food source for microorganism. This provides the potential to enhance bacterial activity which, in turn, leads to the release of the greenhouse gases as carbon dioxide and methane. This permafrost carbon-climate feedback results in accelerated global warming (Koven et al., 2011). Beyond the permafrost thaw, other warming effects such as higher precipitation (Bachar et al., 2010; She et al., 2018; Wu et al., 2020) or soil acidification (Mod et al., 2021; Zhou et al., 2021) are likely to change soil bacterial diversity and community structure. Altogether, the need to forecast the future of soil bacterial distribution is evident.

This study aims to complete the whole picture of soil bacterial distribution and is a step towards understanding the complex effects of climate change on microbial communities, especially in a region particularly sensitive to climate change. As soil microorganisms display biogeographical distribution patterns, we first hypothesize that soil bacterial diversity is not homogeneous across the Arctic region, expecting hotspots of diversity (Malard & Pearce, 2018) and a latitudinal gradient (Bahram et al., 2018; Malard et al., 2022). Considering the implications of Arctic warming on key factors shaping soil bacterial distribution, the second assumption is that changes in diversity and community structure will occur in the future. To test these hypotheses, we modelled the distribution of soil bacteria over the Arctic region (i) under current, and (ii) under future environmental conditions by applying a Species Distribution Modelling (SDM) framework. The community analysis was conducted in two parts: alpha-diversity modelling (richness and Shannon diversity) and community structure modelling. Future distributions were projected under three representative concentration pathways (RCP) corresponding to low, medium and high greenhouse gas emission scenario, for 2050 and 2070.

MATERIALS AND METHODS

1. Data acquisition and processing

The dataset comes from a previously published Pan-Arctic survey of soil bacterial communities (Malard et al., 2019), where soil was sampled in the Arctic region across 43 sites, between April and September 2017 (Figure 1). At each site, three to five soil samples were collected within 100 m² under the most common vegetation for a total of 200 soil samples. Each sample was collected at 5-15 cm depth to have a total of 150 g of soil. In addition, conductivity, moisture, pH and soil organic carbon (SOC) were measured at the sampling sites. Soil DNA was

extracted using the PowerSoil kit (Qiagen, Hilden, Germany) and the V3-V4 region of the 16S rRNA gene was amplified by polymerase chain reaction (PCR) and sequenced using Illumina Miseq (Illumina, San Diego, Ca, USA).



Figure 1: Map of the 43 sampling sites in the Arctic region at which 200 soil samples in total were collected.

After demultiplexing and removing barcodes, sequenced paired-end data was processed following the DADA2 pipeline tutorial (v1.16.0; Callahan et al., 2016) in R (RStudio Team, 2022) to produce amplicon sequence variants (ASV) with the number of reads per ASV in each sample and the associated taxonomy. In this step, the sequenced data was filtered with a maximum of expected errors of two and trimmed at 230 and 200 length for forward and reverse reads respectively. Then, the ASVs were inferred using the DADA2 algorithm to finally merge paired reads and remove chimeras. Taxonomy was assigned to the output sequences using the SILVA reference database (v138.1; Quast et al., 2012) using a naive Bayesian classifier method with a minimum bootstrap confidence of 50. Contaminants were removed based on negative control samples using a frequency- and prevalence-based contaminant identification with a probability threshold of 0.5 using the decontam package (Davis et al., 2018). From an initial set of 47639 ASVs, those present at <15 sites were filtered out using the phyloseq package (McMurdie & Holmes, 2013) to have sufficient prevalence to prevent the models from overfitting. In total, 2523 ASVs across the 200 samples were used for the spatial distribution analysis.

2. Environmental data

An ensemble of climatic, topographic, and edaphic variables was used to model soil bacterial distribution. While soil properties and climate conditions are fundamental drivers of their distribution (Fierer, 2017), topography also plays a role on nutrient and water accumulation in soil and therefore microbial activity (Liu et al., 2020). Distribution in the future was

projected according to three representative concentration pathways (RCP) used for the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) (IPCC, 2017): RCP2.6, RCP4.5, RCP8.5, corresponding to minimum, medium and maximum greenhouse gas emission scenario, respectively, for 2050 and 2070. Four climate variables and one edaphic variable with high reliance on climate change were used to predict under the different scenarios: annual precipitation over zero (rain) and under zero (snow), freezing degree days, thawing degree days and active layer thickness. Since freezing degree days and thawing degree days are related to the permafrost and its associated active layer, they were good candidates for forecasting, while variation in annual precipitation over and under zero may affects microbial distribution by changing soil water availability.

From an initial set of 25 variables, the environmental data of 21 variables could be obtained for the 200 sampling sites (Table 1), while the nearest value within a perimeter of ten kilometres was extracted when no data was available at the sampling site coordinates. Among the 21 variables, a preliminary selection based on the correlation was made to avoid collinearity. To do so, a dendrogram was built with $[1 - \text{Pearson correlation coefficient}]$ as distance to perform the hierarchical clustering (Figure 2) using the *ggdendro* package (Vries & D. Ripley, 2022). Two clusters of predictors highly correlated with each other (Pearson correlation > 0.7) were identified. The larger cluster included clay, freezing degree days (fdd), mean annual ground temperatures (magt), extent, annual precipitation under zero (prec_snow), cation exchange capacity (cec), annual precipitation (prec) and annual precipitation over zero (prec_rain). Among them, the three predictors available for future predictions (fdd, prec_snow, prec_rain) were conserved. In the same way, thawing degree days (tdd), which was also available under the future scenarios, was conserved among the second cluster including normalized difference vegetation index (ndvi) and growing degree days (gdd). In this way, all the predictors used for future predictions were present in this selection. In total, 14 predictors were used as initial variables to fit the models: elevation (elev), topographic roughness index (tri), soil pH in 5cm depth (pH), annual precipitation under zero (prec_snow), soil type (soiltype), annual precipitation over zero (prec_rain), freezing degree days (fdd), topographic wetness index (twi), thawing degree days (tdd), annual potential incoming direct solar radiation (rad), active layer thickness (active_layer), soil organic carbon stock in 5-15 cm depth (soc), land cover classification (esa) and topographic

position index (tpi). In order to assess to which extent our models extrapolate the environmental conditions outside the range of the observation data, the coverage of sampling sites along environmental gradients was evaluated.

Table 1: Description of the initial set of 21 predictors including climatic, topographic and edaphic variables. Variable name is the abbreviation used in the present study.

Variable	Description	Source
active_layer	Active layer thickness (2000-2014)	Aalto et al., 2018
cec	Cation exchange capacity of soil in centimol positive charge per kg	Poggio et al., 2021
clay	Clay content mass fraction in % in 15 cm depth	Poggio et al., 2021
elev	Altitude in meters	Amatulli et al., 2018
esa	European Space Agency - Land cover classification	ESA CCI Land Cover, 2017
fdd	Freezing degree days (2000-2014), number of days when temperatures are below 0 degrees	Fick & Hijmans, 2017
gdd	Growing degree days (2000-2014), number of days when temperatures are above 3 degrees	Fick & Hijmans, 2017
magt	Mean annual ground temperatures (2000-2014)	Aalto et al., 2018
ndvi	Normalized Difference Vegetation Index (2000-2014)	Didan, 2015
soc	Soil organic carbon stocks in the upper 200 cm	Poggio et al., 2021
extent	Permafrost extent (no permafrost, continuous, discontinuous, sporadic, isolated)	Brown et al., 2002
ph	Soil pH in H ₂ O in 15 cm depth (need to be divided by 10)	Poggio et al., 2021
rad	Potential incoming direct solar radiation in millijoule per square centimetre per year	Amatulli et al., 2018
prec	Annual precipitation in millimetre (2000-2014)	Fick & Hijmans, 2017
prec_rain	Annual precipitation over zero in millimetre (2000-2014)	Fick & Hijmans, 2017
prec_snow	Annual precipitation under zero in millimetre (2000-2014)	Fick & Hijmans, 2017
soiltype	Soil type	Poggio et al., 2021
tdd	Thawing degree days (2000-2014), number of days when temperatures are above 0 degrees	Fick & Hijmans, 2017
tpi	Topographic position index	Amatulli et al., 2018
tri	Topographic roughness index	Amatulli et al., 2018

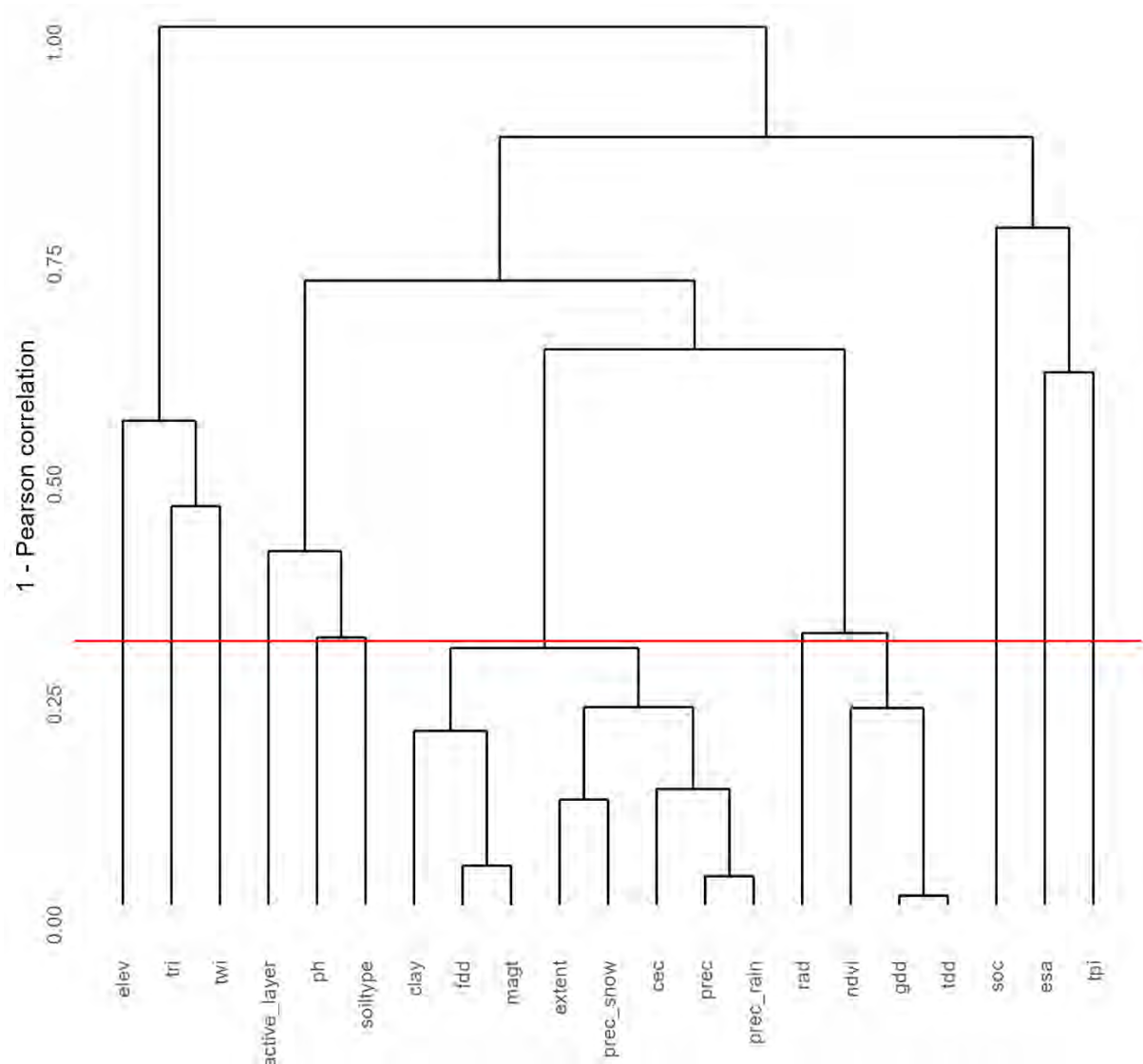


Figure 2: Dendrogram based on hierarchical clustering of the 21 environmental variables. Distances were calculated as $[1 - \text{Pearson correlation coefficient}]$. Clusters of correlated variables were defined using a threshold of a Pearson correlation value of 0.7 (horizontal red line).

3. Soil bacterial distribution modelling

To investigate soil bacterial distribution in the Arctic region, different community properties were targeted: alpha-diversity (richness and Shannon diversity) and community structure, using ASV as taxonomic unit. Each property was modelled independently as a function of environmental conditions, based on a common framework used in ecology, Species Distribution Model (SDM) (Guisan & Zimmermann, 2000). Two spatial modelling approaches were used: alpha-diversity was predicted directly, based on the principle “assemble first, predict later” (Ferrier & Guisan, 2006), while community structure was predicted by stacking

relative abundance predictions of individual ASVs (Stacked Species Distribution Modelling, S-SDM) (Guisan & Rahbek, 2011). The predictions were based on an ensemble forecasting framework, which aims to average several models to overcome the uncertainty between them (Araújo & New, 2007). Here, one regression-based approach (Generalized Linear Models, GLM) and one classification tree algorithm (Random Forests, RF) were used in the ensemble models, the two methods differing in their performance and bias (Elith et al., 2006). All analyses were conducted in R (RStudio Team, 2022) and figures and maps were produced using the ggplot2 package (Wickham, 2016).

3.1. *Richness and Shannon diversity*

Alpha-diversity measures (richness and Shannon diversity) were directly modelled as a function of environmental variables. Richness data corresponded to the number of ASV per sample and Shannon diversity index was calculated using the vegan package (Oksanen et al., 2022).

Preliminary analysis intended to identify the best regression-based method between elastic-net regularized Generalized Linear Models (GLMnet) with 100% lasso, which is able to shrink coefficient variables to zero, and Generalized Linear Models with formula defined by stepwise regression (GLMsw). The stepwise model selection was performed by Akaike information criterion (AIC) with search in both directions, and the most complex formula included all the 14 variables with quadratic terms and without interactions. Both methods involved automatic variable selection procedures. Although elastic-net is recognized to generally perform better (Lenters et al., 2018), the stepwise regression method was slightly better evaluated in our study when modelling diversity (Figure S1). However, GLMsw predicted ecologically irrelevant values of richness ($>10^{18}$) when projecting over the Arctic region, unlike GLMnet which had more shrunk predictions due to its penalization-based method. Therefore, GLMnet was chosen over GLMsw for variable selection and as modelling approach. Regarding the RF method, there was no actual variable selection since it relies on a multitude of decision trees, each node of which providing information on the importance of the variable.

GLMnet and RF models were fit using the initial set of 14 predictors to build the ensemble models. GLMnet with 100% lasso was fit using lambda.1se as the optimal lambda and glmnet package (Friedman et al., 2010). RF was fit using 1000 trees (bootstrap samples) with 4

variables per tree and randomForest package (Liaw & Wiener, 2002). GLMnet and RF predictions were ensembled by weighting according to their respective R^2 mean value obtained from the evaluation (see below). Predictive models were evaluated using resampling procedures to test their predictive power on partially independent data. The data partitioning strategy was a repeated split-sample cross-validation (100x), where 75% of the data was assigned to the training set and the remaining 25% to the evaluation set. This method is both simple to run and powerful (Guisan et al., 2017). The metric used for evaluation was R-Squared (R^2) and Root Mean Squared Error (RMSE) computed with the package caret (Kuhn et al., 2022).

Ensemble model projection over space (Arctic region) and standard deviation between the GLMnet and RF predictions were mapped. The relation between diversity and latitude was tested using linear and quadratic regression. Then, projection over time (2050 and 2070) for the three scenarios and differences between current and future (2070) diversity distribution were mapped.

3.2. Community structure

Analyses prior to modelling were conducted to have a baseline understanding of how soil bacterial community may change under the different climate change scenarios. It included a Spearman correlation heatmap based on environmental variables and the community at phylum level. The Spearman correlations and their associated P-values were calculated using the hmisc package (Harrell Jr, 2022). Based on this, 9 variables that can potentially influence community structure in the future were identified. Relationships between each phylum and the 9 predictors were tested using linear regression. Significant relationships were then illustrated in a scatter plot for each of the 9 environmental gradients.

To assess changes in soil bacterial community structure, the relative abundance of each ASV was modelled individually and predictions across the 200 sampling sites were assembled to obtain community structure under each climate-change scenario. Since the ASV distribution model predictions were restricted to the sampling sites, edaphic variables measured in the field were used for modelling. Conductivity and moisture data were added to the set of variables used while pH and soil organic carbon data replaced and were compared to those provided by global databases (Figure S2). The field-measured variables showed a greater

variation than online data and thus have the potential to be better predictors for bacterial distribution.

Following the same method as for diversity distribution modelling, GLMnet and RF models were fit using the initial set of 14 predictors to build the ensemble model. The frequency of variable selection by GLMnet among the pool of modelled ASVs was calculated for each predictor used. Model performances were estimated by repeated split-sample cross-validation (100x). Considering the poor model qualities suggested by the R^2 and RMSE values (Figure S3) and to avoid not removing a significant number of ASVs, Spearman correlation was used instead as metric for evaluation and for weighting ensembled predictions, and models with Spearman correlation <0.2 were removed from the analysis.

For each ASV, relative abundances were predicted and summed across the 200 sampling sites, to be assembled into a community under each scenario. Scenario-based community compositions were visualized in a phylum-level barplot. To evaluate the change in the community structure, a distance matrix was computed with Bray-Curtis as a dissimilarity metric using the vegan package (Oksanen et al., 2022), which has the advantage of quantifying differences between communities at the ASV level. Bray-Curtis dissimilarity is a measure of beta-diversity that indicates the degree of difference between communities in term of species abundance (Bray & Curtis, 1957). To represent the dissimilarity in a two dimensional plot, a Principal Coordinate Analysis (PCoA) scatter plot was computed using the ape package (Paradis & Schliep, 2019).

RESULTS

1. Environment variables

The coverage of sampling sites along environmental gradients showed in which extent our models extrapolate the environmental conditions outside the range of the observation data (Figure S4). Overall, environmental variation over the Arctic region was relatively well represented for the sample size and sampling sites, although the conditions sampled for some variables only partially covered the full range of values (e.g., active layer thickness, freezing degree days).

2. *Richness and Shannon diversity distribution*

2.1 *Modelling*

Soil bacterial diversity ranged from 97 to 912 ASV per sample for richness and from 3.53 to 6.14 for Shannon diversity across the 200 samples. GLMnet and RF models were fit using the initial set of 14 predictors. Environmental variables used in GLMnet with 100% lasso were determined by shrinking variable coefficients to zero. For richness modelling, 18 terms were present in the output formula including all variables in linear or quadratic form or both. The formula was implemented as follows:

$$\text{Richness} \sim \text{elev}^2 + \text{tri}^2 + \text{ph}^2 + \text{prec_snow} + \text{prec_snow}^2 + \text{soiltype}^2 + \text{prec_rain}^2 + \text{fdd}^2 + \text{twi}^2 + \text{tdd} + \text{rad} + \text{active_layer} + \text{active_layer}^2 + \text{soc} + \text{esa} + \text{esa}^2 + \text{tpi} + \text{tpi}^2$$

Unlike for richness modelling, GLMnet for Shannon diversity modelling has not selected all the variables. The formula was composed of 17 terms and has not included `prec_rain` as predictor:

$$\text{Shannon diversity} \sim \text{elev} + \text{elev}^2 + \text{tri}^2 + \text{ph}^2 + \text{prec_snow} + \text{soiltype} + \text{fdd}^2 + \text{twi}^2 + \text{tdd} + \text{rad}^2 + \text{active_layer} + \text{active_layer}^2 + \text{soc} + \text{esa} + \text{esa}^2 + \text{tpi} + \text{tpi}^2$$

The five variables used for future predictions (`active_layer`, `fdd`, `prec_rain`, `prec_snow`, `tdd`) were chosen as richness predictors whereas one of them (`prec_rain`) was not present in the Shannon diversity model.

2.2 *Evaluation*

When evaluating the transferability of the fitted models, the repeated split-sample cross-validation (100x) on richness data indicated a R^2 mean value of 0.59 and 0.68 and a RMSE mean value of 128.48 and 122.91 for GLMnet and RF algorithm, respectively, between the observed and predicted values (Figure 3.A). Although there was no substantial difference in quality between GLMnet and RF models, RF had slightly higher R^2 and lower RMSE, indicative of a better model, and GLMnet showed greater variation through cross-validations. The same procedure on Shannon diversity data indicated a R^2 mean value of 0.36 and 0.50 and a RMSE mean value of 0.37 and 0.33 for GLMnet and RF algorithm, respectively (Figure 3.B). The two models presented a greater difference in predictive performance and justified the R^2 -weighted ensembling method. More generally, the RF model consistently performed better.

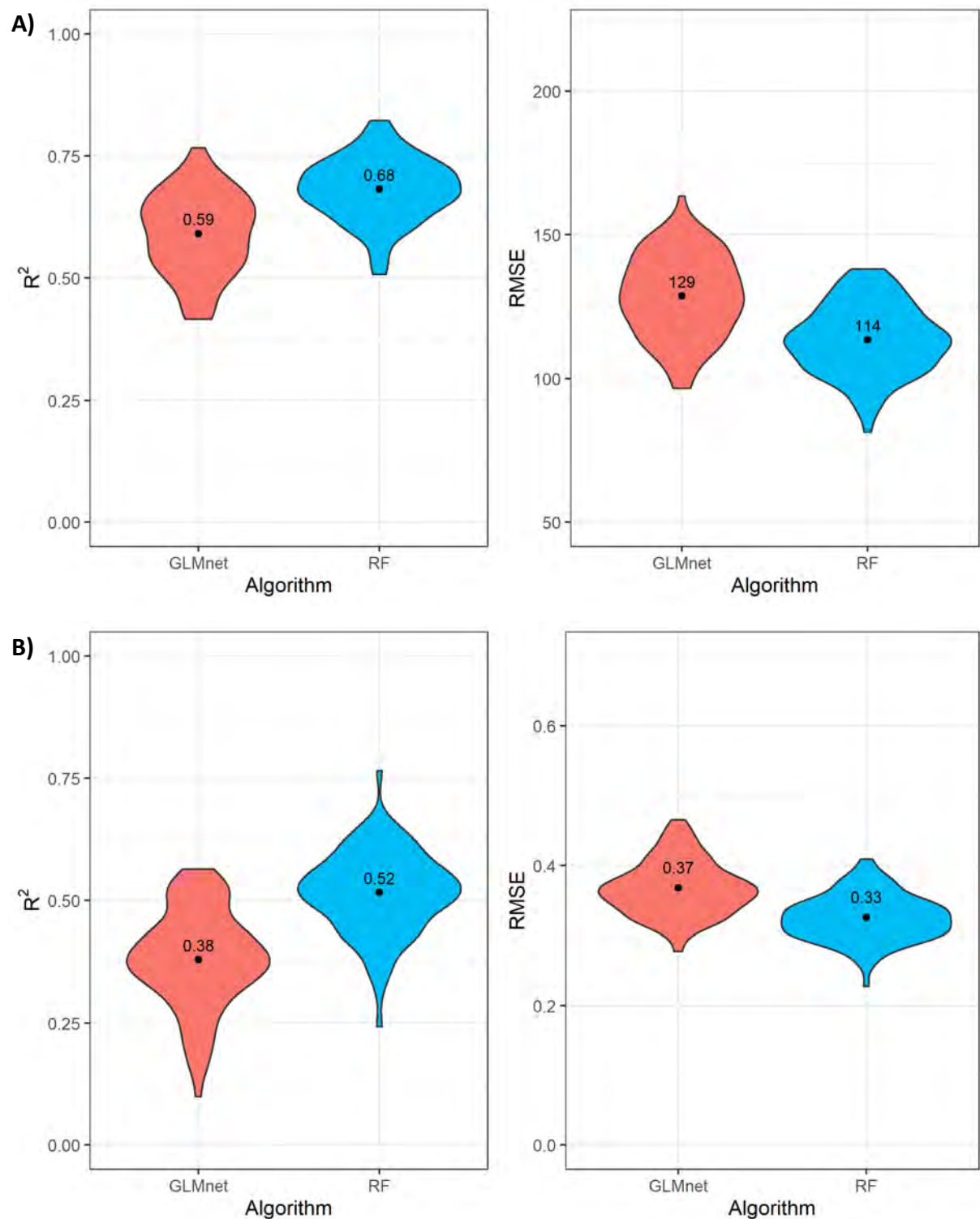


Figure 3: Model evaluation by repeated split-sample cross-validation (100x), where 75% of the data was assigned to calibration and the remaining 25% to evaluation. The two algorithms GLMnet and RF were evaluated with R^2 and RMSE. Mean point and its associated value is indicated. R^2 mean was used to weight predictions for ensembling **(A)** Evaluation for soil bacterial richness model. **(B)** Evaluation for Shannon diversity model.

2.3 *Projection over space*

Diversity distribution was projected over space as a function of environmental conditions throughout the Arctic region. Richness predictions ranged from 123 to 782 ASVs with a median of 344 ASVs, the higher richness located in the upper part of Russia and lower richness in Northern Quebec (Figure 4). Shannon diversity predictions ranged from 2.6 to 5.7 (Figure 5). Overall, higher richness was identified in higher latitude, whereas lower richness was associated with lower latitude relative to the Arctic region, especially in north America (Figure 4.A). The same gradient was observed for Shannon diversity (Figure 5.A). Linear and quadratic regression confirmed the positive effect of latitude on diversity (Figure 6), both on richness ($P<0.001$, linear model $R^2=0.23$, quadratic model $R^2=0.31$) and Shannon diversity ($P<0.001$, linear model $R^2=0.17$, quadratic model $R^2=0.24$).

Richness and Shannon diversity distribution had mostly the same pattern, only a few isolated areas showed opposite trend. For example, the southwestern part of Russia displayed high richness and low Shannon diversity. On the other hand, Iceland and the most southern part of Russia harbored low richness and high Shannon diversity. Standard deviation maps showed higher variation between GLMsw and RF models when low diversity was predicted (Figure 4.B, 5.B). This inconsistency occurred differentially across space. Predictions ranged from 0 to 1148 ASVs for GLM and from 150 to 801 ASVs for RF, whereas Shannon diversity predictions ranged from 0 to 6.2 for GLM and from 4.2 to 5.9 for RF. This implies that GLM predictions extrapolated beyond the range of observed diversity values (97 to 912 for richness and 3.5 to 6.1 for Shannon diversity), whereas RF predicted within the observed range.

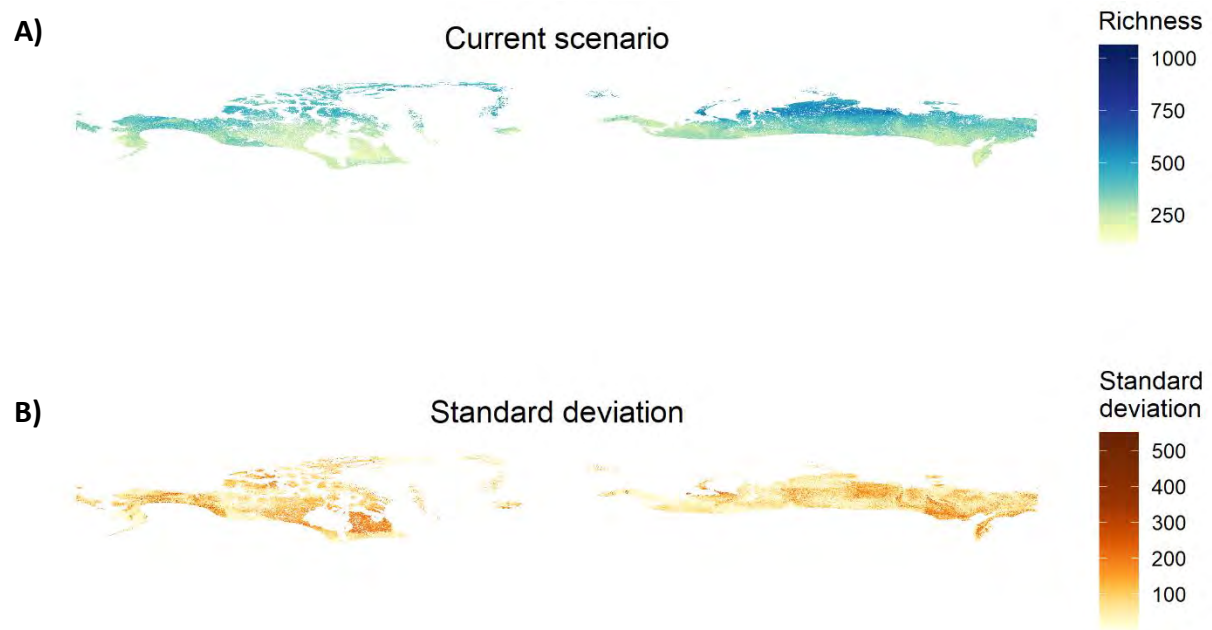


Figure 4: Soil bacterial richness in Arctic region under current environmental conditions. **(A)** Ensemble model projection calculated as the R^2 -weighted mean of GLM and RF predictions at 1km of resolution. **(B)** Standard deviation of the two model predictions.

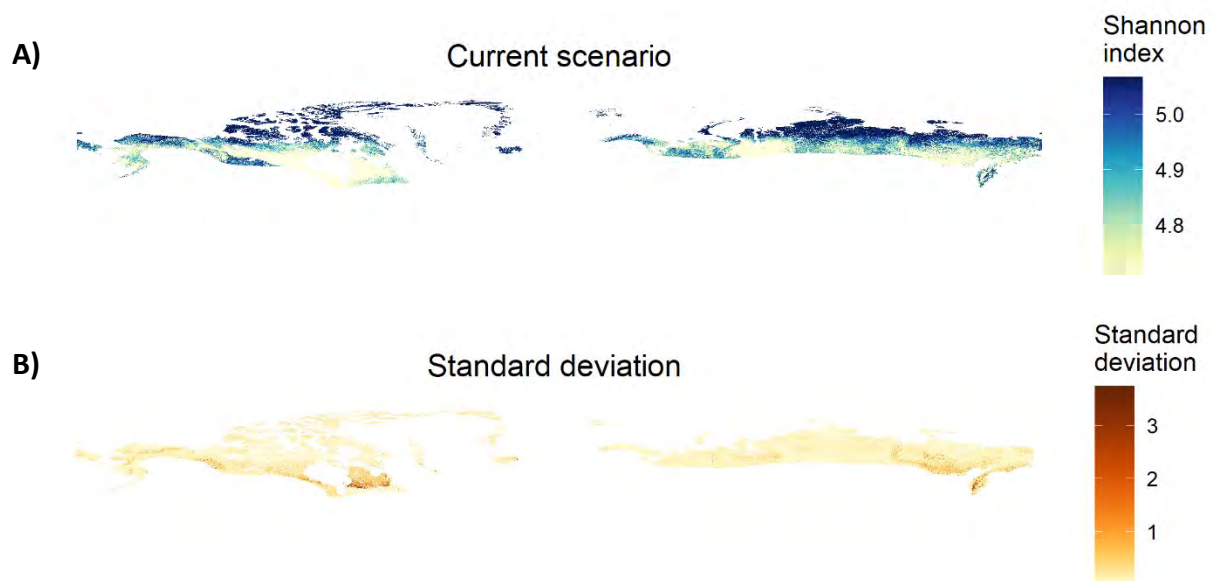


Figure 5: Shannon diversity in Arctic region under current environmental conditions. **(A)** Ensemble model projection calculated as the R^2 -weighted mean of GLM and RF predictions at 1km of resolution. **(B)** Standard deviation of the two model predictions.

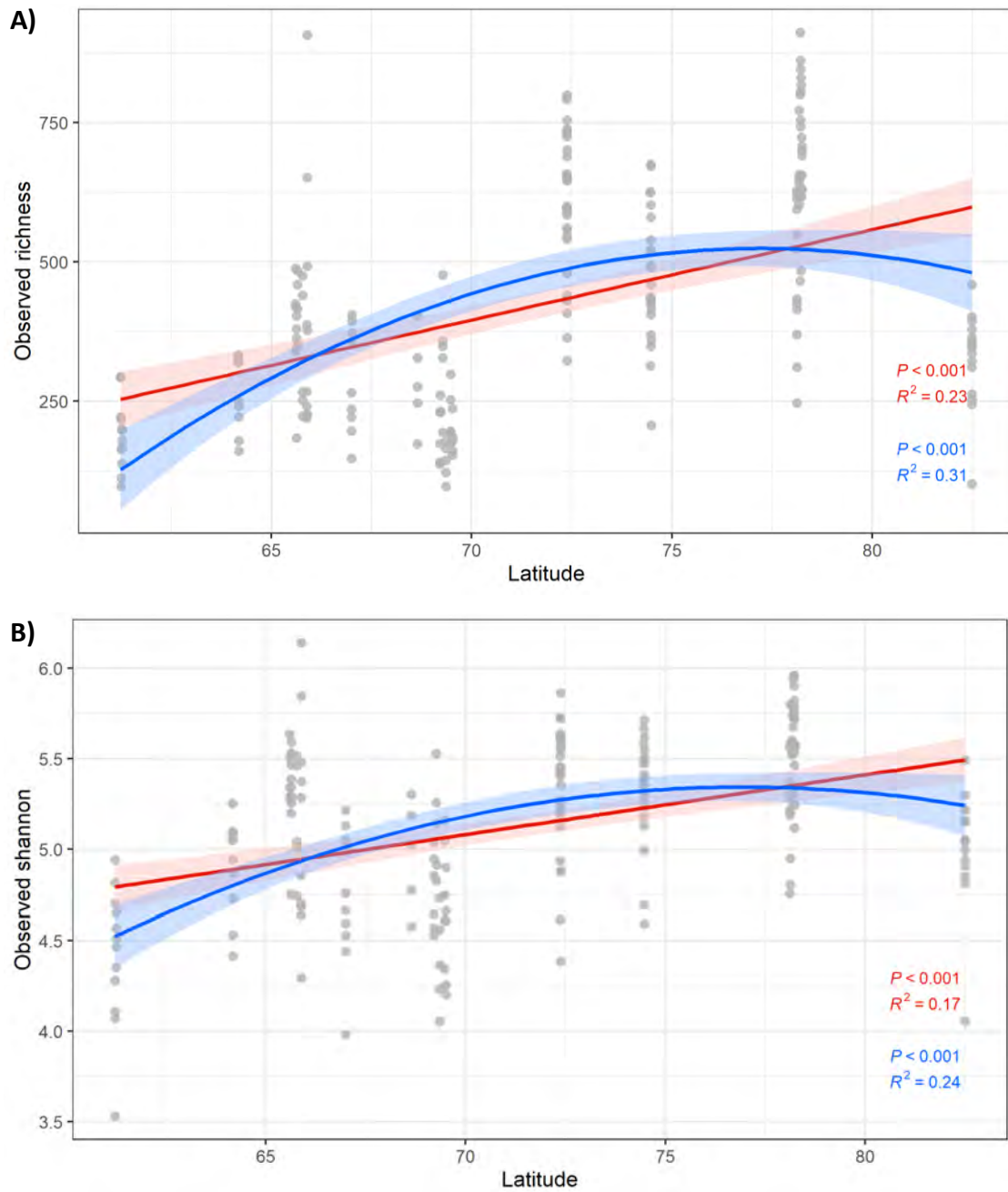


Figure 6: (A) Soil bacterial richness and **(B)** Shannon diversity at sampling sites along latitudinal gradient. Linear (red line) and quadratic (blue line) regression were performed and their respective P and R^2 values were calculated.

2.4 *Projection over time*

Diversity was then projected over time under the three RCP scenarios, in 2050 and 2070. Richness predictions generally increased, both through the RCP scenarios with increasing greenhouse gases emission and through time, gradually (Figure 7). The richness median shifted from 344 to 388 ASVs and the maximum richness from 782 to 1063 ASVs. Shannon diversity predictions however did not change its range between the different scenarios, with values between 2.6 and 5.7 and a median of 4.9 (Figure 8).

To visualize the spatial changes between current and future scenarios, diversity differences were mapped (Figure 7.D, 8.D). Under the RCP2.6 scenario (best emission-based scenario), richness differences were relatively small (yellow), highlighting some wide region with differences up to 271 ASV (orange) and negative differences (blue). Jumping to the RCP 4.5 and 8.5 scenarios (medium and worst emission-based scenarios), the same clusters subjected to diversity changes increased in size and intensity. Although Shannon diversity showed mostly no changes (yellow), more localized changes gained in size and intensity with scenario severity but in a smaller extent, with a maximum difference of 2.5.

Regions of positive, neutral and negative changes differed between richness and Shannon diversity map. In summary, soil bacterial diversity generally increased across the Arctic under all scenarios of climate change considered although some clusters of loss of diversity were identified for both, richness and Shannon diversity.

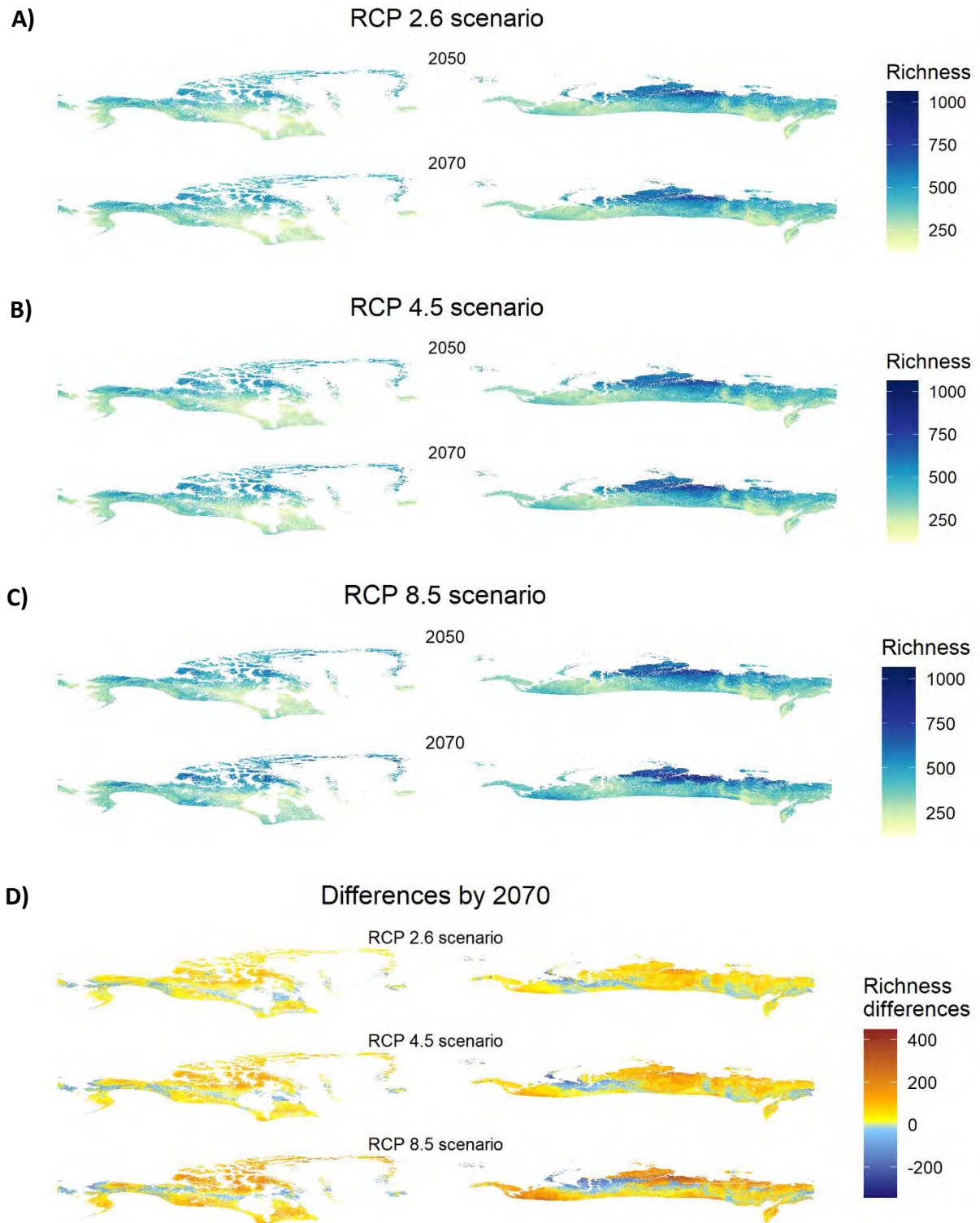


Figure 7: Soil bacterial richness in Arctic region under RCP 2.6, 4.5 and 8.5 scenarios. **(A, B, C)** Forecasts for 2050 and 2070 have been computed for each scenario at 1km of resolution. Ensemble model projection was calculated as the R^2 -weighted mean of GLM and RF predictions. **(D)** Soil bacterial richness differences between current and future conditions in 2070 for each scenario.

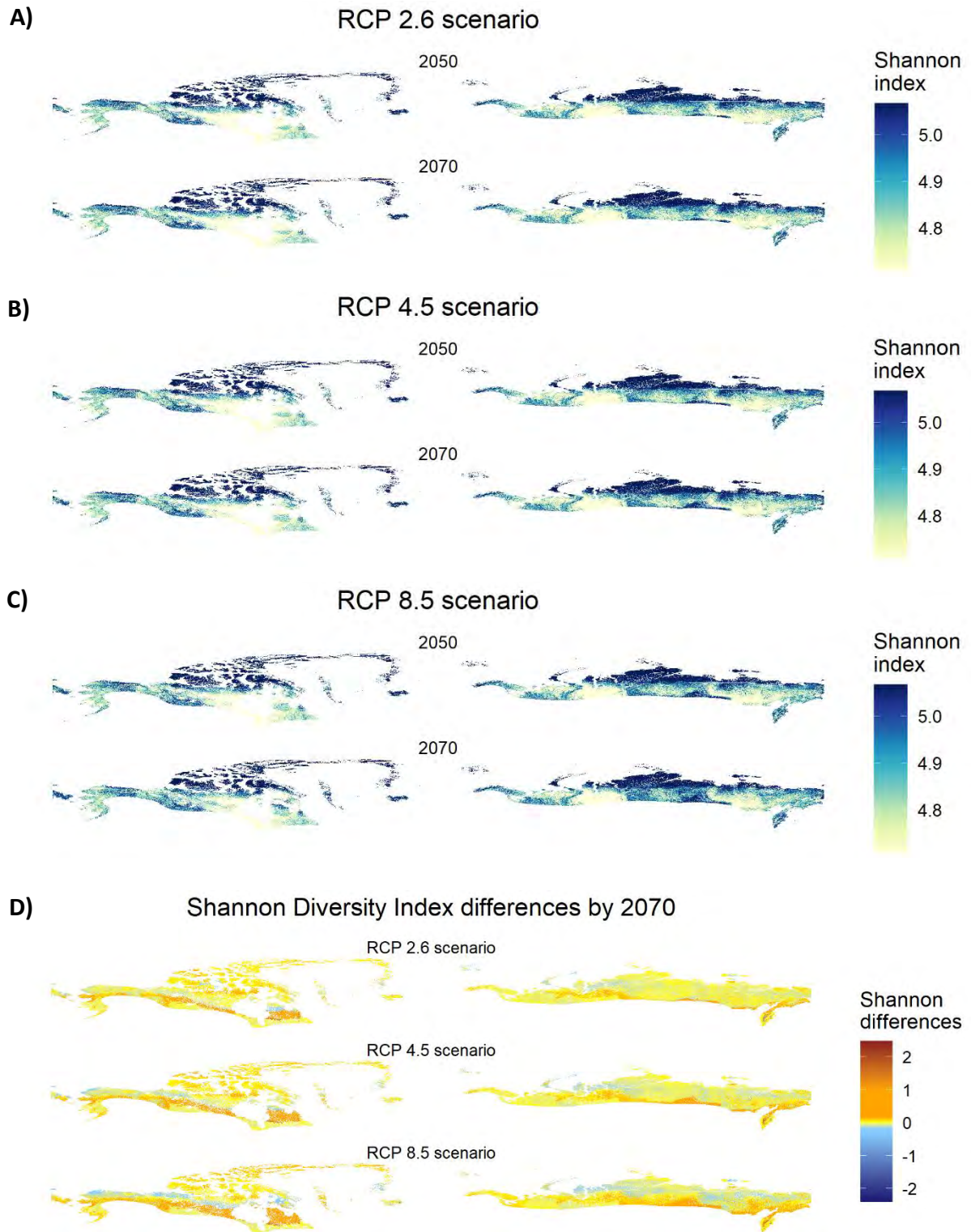


Figure 8: Shannon diversity in Arctic region under RCP 2.6, 4.5 and 8.5 scenarios. **(A, B, C)** Forecasts for 2050 and 2070 have been computed for each scenario at 1km of resolution. Ensemble model projection was calculated as the R^2 -weighted mean of GLM and RF predictions. **(D)** Shannon diversity differences between current and future conditions in 2070 for each scenario.

3 Community structure

3.1 Environment-phyla correlation

To have indication of how soil bacterial community may change in the future, we first computed a Spearman correlation heatmap based on environmental variables and bacterial communities to determine how bacterial phyla correlate with current environmental conditions (Figure 9). It revealed that phyla responded differently to the 14 environment gradients. For example, pH was correlated with a total of 20 phyla, where 12 phyla had strong positive correlation (red) and 4 phyla had strong negative (blue) correlation (***) indicates $P \leq 0.001$).

Figure 10 focused on 9 environmental variables expected to be affected in the future, including those used for future predictions: active layer thickness, freezing degree days, moisture, pH, precipitation over and under zero, radiation, soil organic carbon and thawing degree days. They were tested as predictor of relative abundance of each phylum and significant relations were shown in Figure 10 showing differential sensitivity within soil bacterial community to environmental gradient. Most of the relationships were negative for these 9 variables, except pH. Variables affecting the highest and lowest number phyla were pH and active layer thickness, respectively. Overall, these results highlight some high correlations between phyla and their environment and therefore, the likely strong impacts of changing environmental conditions on soil bacterial community structure.

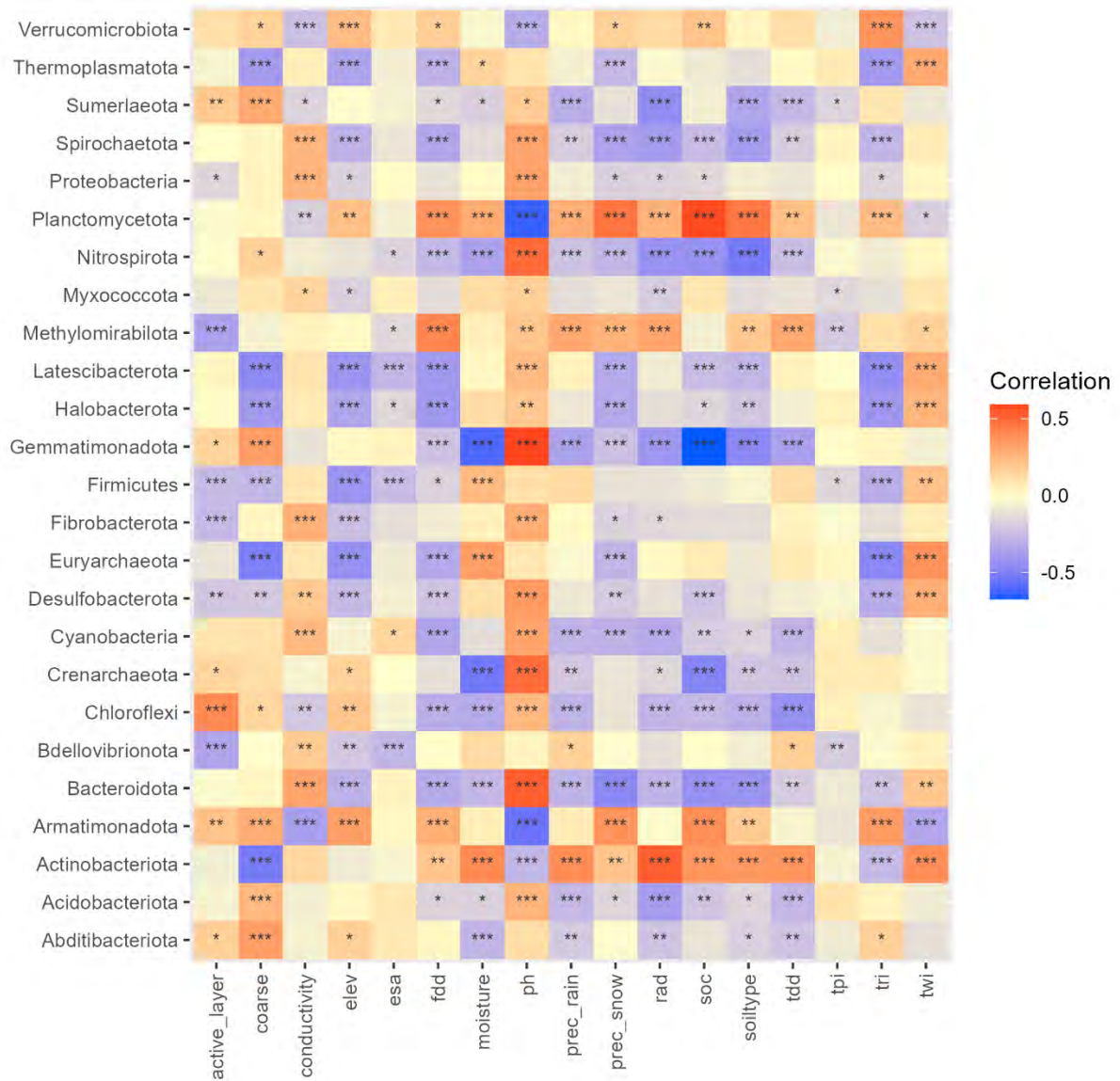


Figure 9: Spearman correlation heatmap based on environmental variables (X axis) and the bacterial community at phylum level (Y axis). The colour gradient indicates the correlation value, blue and red for positive and negative correlation, respectively. The stars indicate the significance level (* when $0.01 < P \leq 0.05$, ** when $0.001 < P \leq 0.01$, *** when $P \leq 0.001$).

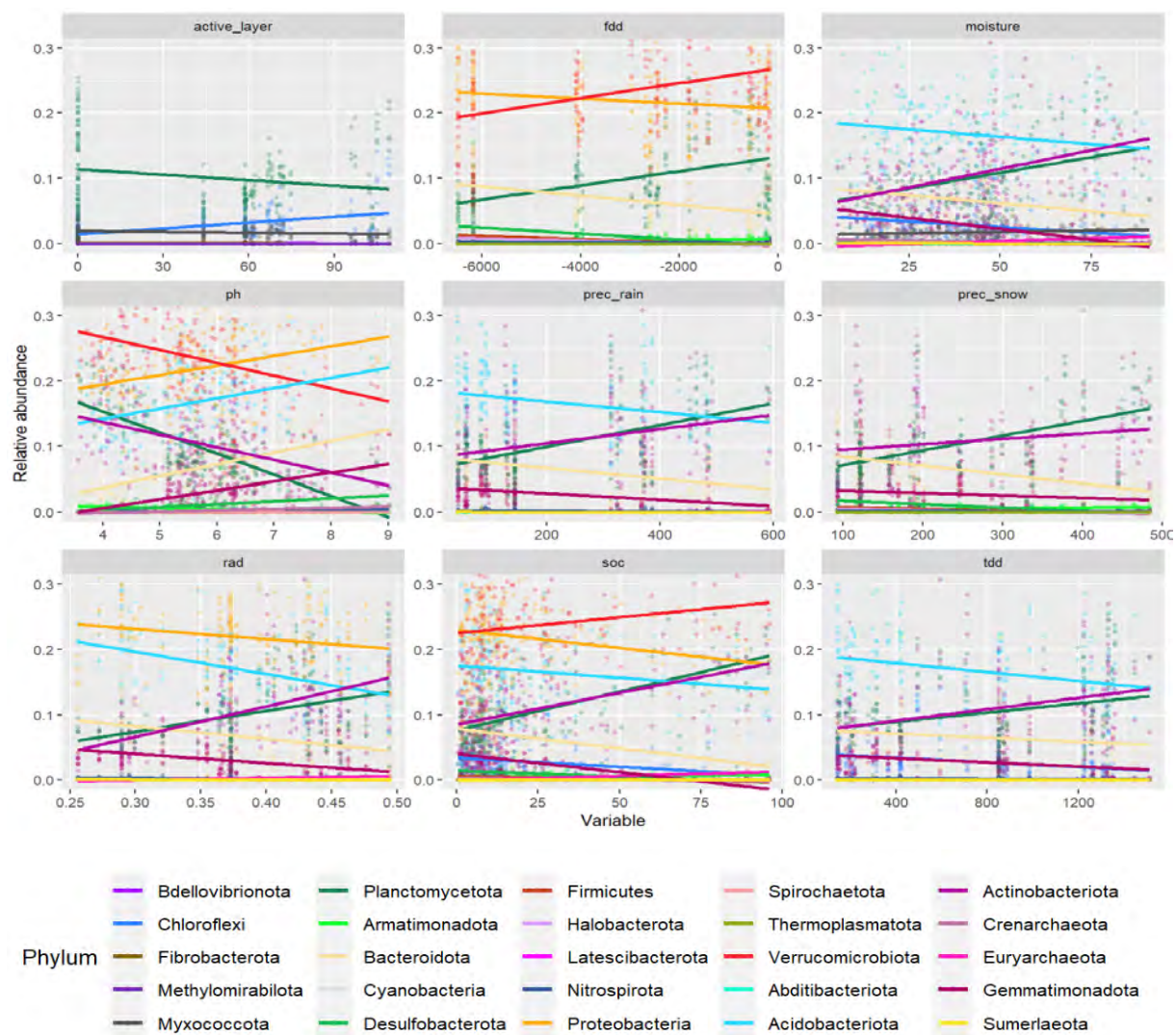


Figure 10: Soil bacterial communities along environmental gradients. Phylum relative abundance were plotted as a function of 9 predictors: active layer thickness (active_layer), freezing degree days (fdd), moisture, pH, annual precipitation over zero (prec_rain) and under zero (prec_snow), radiation (rad), soil organic carbon (soc) and thawing degree days (tdd). The relationships were tested using linear regression and only significant relations ($P \leq 0.05$) are shown in this figure.

3.2 Modelling

From the initial set of 2523 ASV, the distribution of 90 ASVs could not be modelled using GLMnet. Distribution models of the remaining 2433 ASVs were fit using GLMnet and 2523 models were fit using RF, using the initial set of 14 predictors. The variable selection by GLMnet when modelling the distribution of the 2337 ASVs was reported (Figure 11). Each variable was selected and included in a minimum of 20% of the 2433 ASV distribution models. With the most frequent selection, pH was chosen by GLMnet as a predictor to explain the

distribution of 59% of the ASVs. The ranking was followed by freezing degree days, moisture, conductivity, esa (land cover classification) and radiation which have been selected between 40% and 55% of the times. Some predictors used for future predictions were selected less frequently, as precipitation over zero (20%) and active layer thickness (30%).

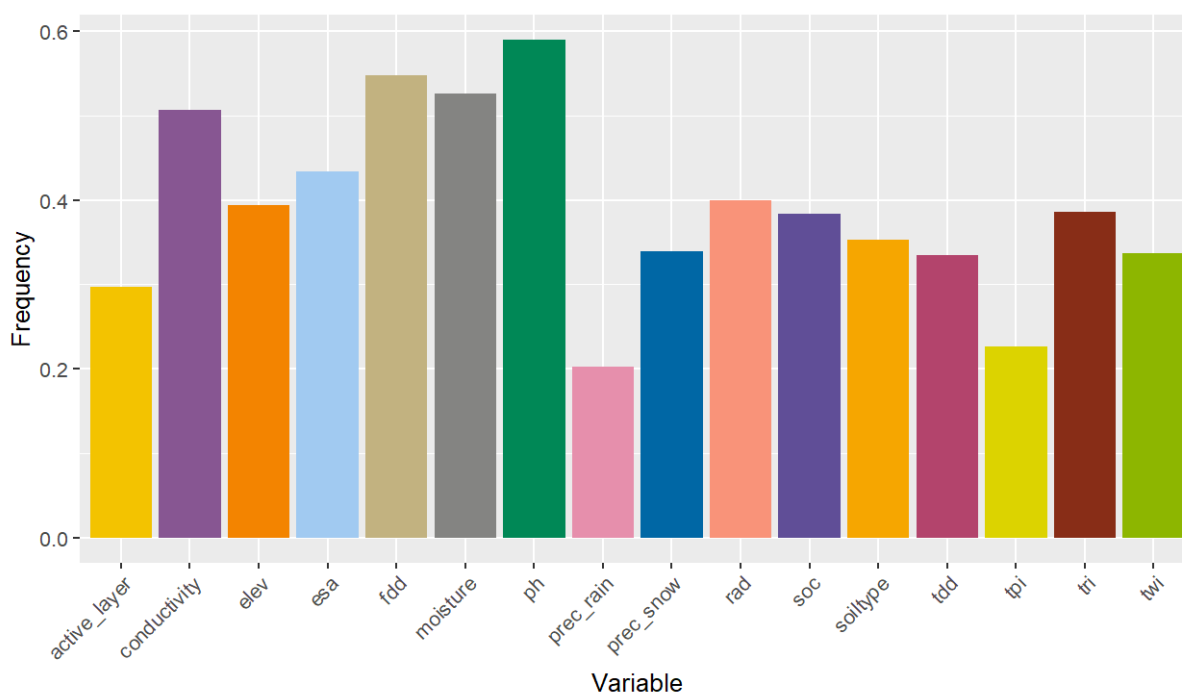


Figure 11: Environmental variables selection by GLMnet. The frequency is the number of times the variable was selected out of the 2337 ASVs.

3.3 Evaluation

Distribution model of the 2433 ASVs for GLM and 2523 ASVs for RF were evaluated by repeated split-sample cross-validation (100x), using Spearman correlation (Figure 12). Based on this metric, models with a mean correlation <0.2 were filtered out, corresponding to 96 models for GLM and 126 models for RF removed from the set of evaluated models. A total of 80 ASVs were filtered for both algorithms. The remaining 2443 ASVs were used for community analysis, of which 106 were modelled only with RF and 46 only with GLM. The mean Spearman correlation was also used as weight for ensembling GLM and RF predictions. In general, there was a high variation in model performance.

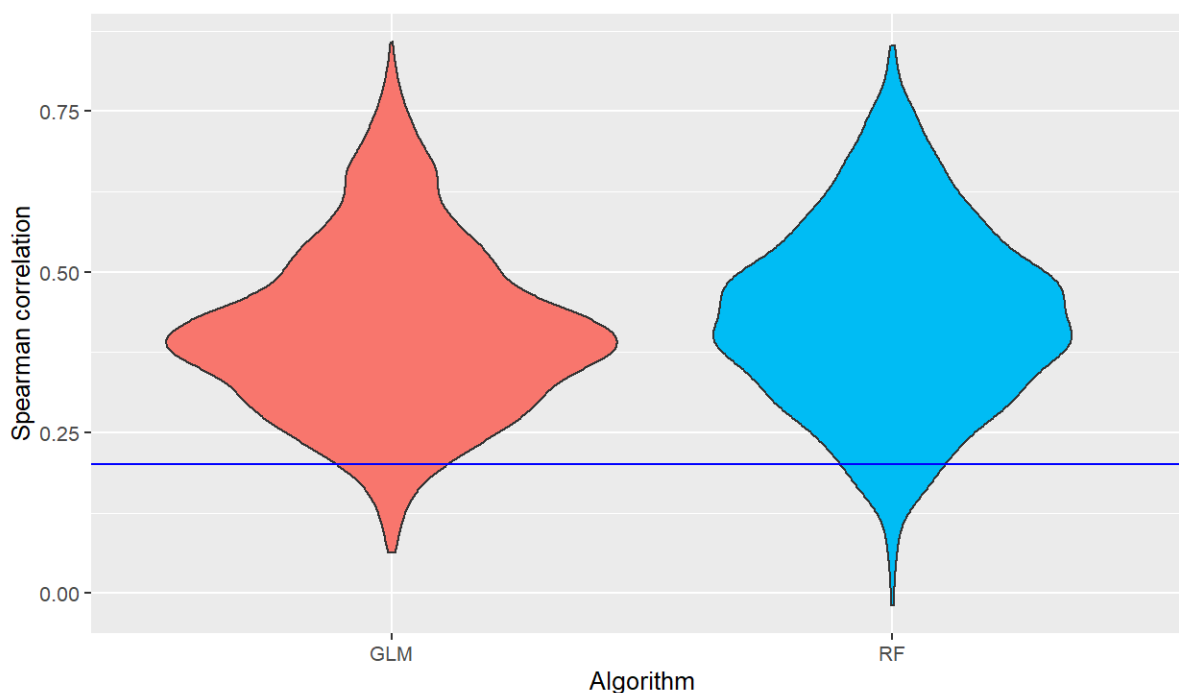


Figure 12: Model evaluation by repeated split-sample cross-validation (100x), where 80% of the data was assigned to calibration and the remaining 20% to evaluation. Each ASV distribution model was individually evaluated with Spearman correlation for both GLMnet and RF algorithms and their associated mean Spearman correlation values were plotted in this violin plot. The horizontal blue line indicates a correlation of 0.2 which was used as a threshold to filter out models with poor performance.

3.4 Community structure changes

Distribution of each ASV were projected under current and future conditions across the 200 sampling sites and assembled to build the community at phylum level. The relative abundance of the 25 phyla was calculated for each climate scenarios. One ASV (ASV1584) was identified as having considerable changes across future scenarios, becoming largely dominant in each future community (Figure S5). Its assigned taxonomy corresponded to the genus *Steroidobacter* from the phyla of Proteobacteria. Its dominance made the visualization of community composition difficult and was therefore removed from the community analysis. The resulting community revealed no major changes between the different scenarios at phylum level (Figure 13). Low abundance phyla (less than 1%) were grouped into one category (Phylum <1%), which allows the visualization of the 10 most abundant phyla (from the most to the less abundant under current conditions): Verrucomicrobiota (22.1%), Proteobacteria (21.5%), Acidobacteriota (16.1%), Planctomycetota (10.5%), Actinobacteriota (9.5%), Bacteroidota (8.1%), Chloroflexi (3.2%), Gemmatimonadota (2.9%), Myxococcota (2.1%) and

Desulfobacterota (1.3%). The less abundant phyla accounted for 2.6% of total abundance under current conditions and decreased to 2.3% under future conditions (RCP 8.5 in 2070). Except for Myxococcota, the relative abundance of other phyla did not vary by more than 1.5% between scenarios. Indeed, Myxococcota was the only phylum showing progressive changes, tending to increase over the scenarios (relative abundance from 2.1% to 7.3%), at the expense of all other phyla.

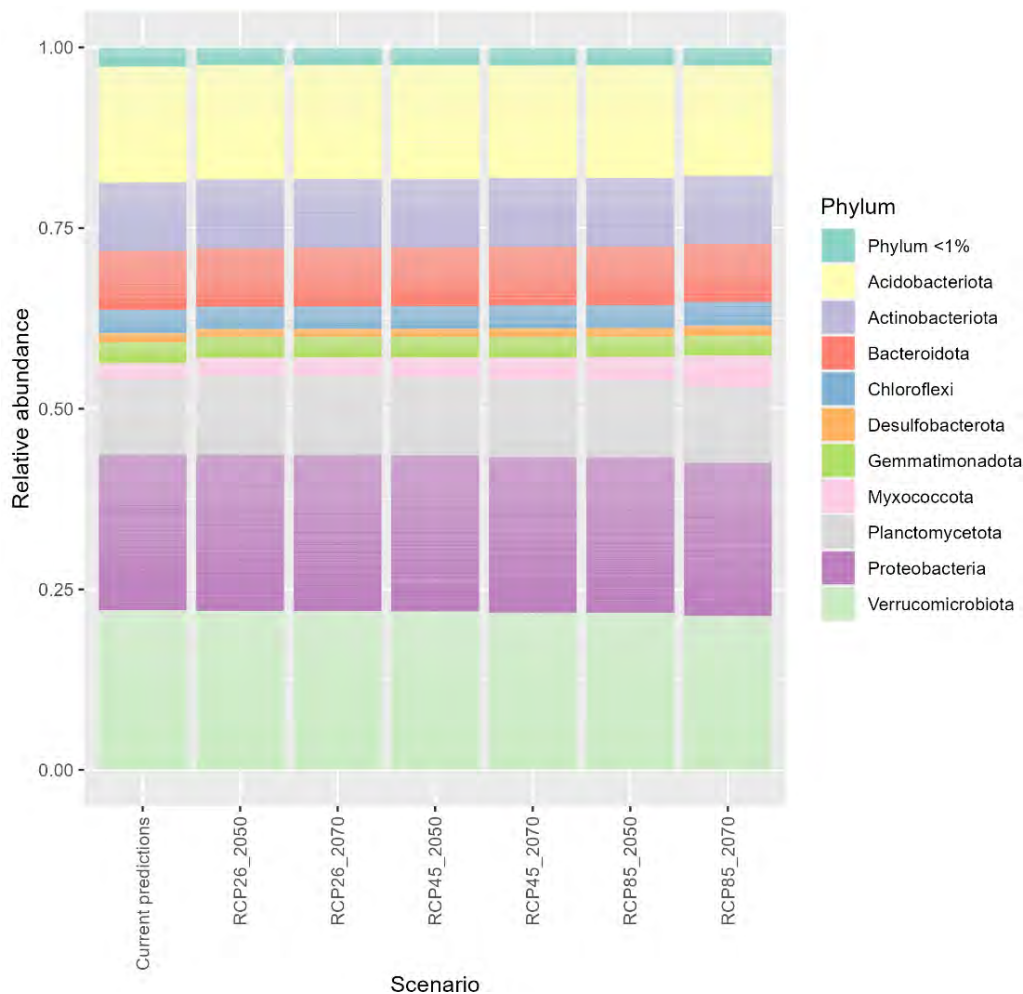


Figure 13: Structure of the soil bacterial communities at phylum level under current and future scenarios. The ASV1584 was omitted in the analysis.

To investigate community changes at a finer resolution, Bray-Curtis dissimilarity index was calculated between the scenario-based communities, where a high value indicates high dissimilarity (Table 2). Given that this index is bounded between 0 and 1, the values obtained were relatively small (<0.11). The dissimilarity between the current and the different future scenarios tended however to increase with the severity of the scenario and the time step. Principal coordinate analysis (PCoA), which spatially reflects these dissimilarities, highlighted

the community changes between the current and all other future predictions (Figure 14). Among the RCP scenarios, RCP 8.5 (2070) was the most differentiated from the others in term of community composition. The remaining scenarios, RCP 2.6 (2050 and 2070), RCP 4.5 (2050 and 2070) and RCP 8.5 (2050) were more closely related.

Table 2: Distance matrix based on Bray-Curtis dissimilarity metric. High value indicates a high dissimilarity.

	<i>Current predictions</i>	<i>RCP26_2050</i>	<i>RCP26_2070</i>	<i>RCP45_2050</i>	<i>RCP45_2070</i>	<i>RCP85_2050</i>
<i>RCP26_2050</i>	0.043					
<i>RCP26_2070</i>	0.045	0.003				
<i>RCP45_2050</i>	0.047	0.007	0.006			
<i>RCP45_2070</i>	0.051	0.015	0.015	0.014		
<i>RCP85_2050</i>	0.052	0.015	0.014	0.011	0.008	
<i>RCP85_2070</i>	0.109	0.075	0.074	0.071	0.070	0.066

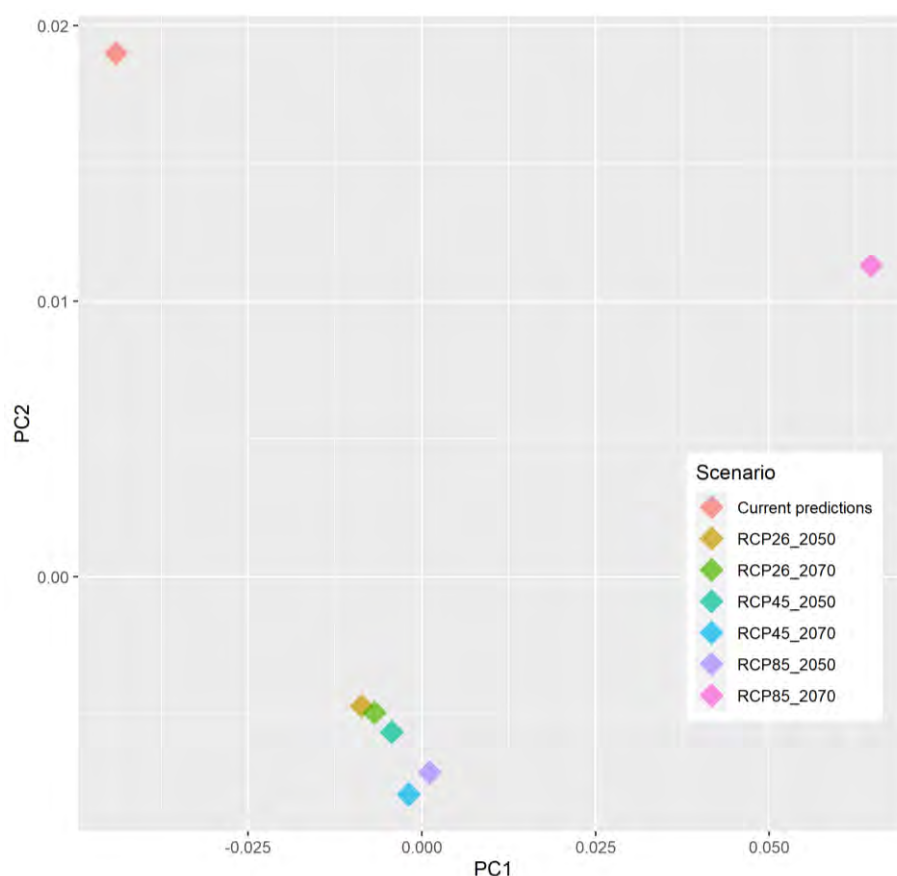


Figure 14 Principal Coordinate Analysis (PCoA) scatter plot based on the Bray-Curtis distance matrix, illustrating dissimilarity in ASV-level structure between the predicted community under current conditions and the different scenario-based communities.

DISCUSSION

To fill gaps in soil bacterial biogeography, SDM-based models were developed to provide spatial distribution of soil bacterial diversity and community in the Arctic region under current and future environmental conditions.

1. Richness and Shannon diversity distribution

All or most of the 14 environmental variables used were able to partially explain the distribution of soil bacterial richness and Shannon diversity and were integrated in the models. This confirmed previous studies that climate, edaphic and topographic factors influence soil microbial diversity (Delgado-Baquerizo & Eldridge, 2019; Fierer, 2017), including in the Arctic (Dimitriu & Grayston, 2010; Malard & Pearce, 2018; Siciliano et al., 2014).

Model evaluation identified Random Forests (RF) as the most powerful algorithm, supporting its wide use in microbial distribution studies (Delgado-Baquerizo & Eldridge, 2019; Pajunen et al., 2016; Větrovský et al., 2019). The poorest performance of the Shannon diversity model could be explained by the narrower range of the index or, more probably, by the complexity of the metric, accounting for both ASV richness and evenness of their abundance. This implies that a lower Shannon diversity could be the result of fewer ASV or the dominance of certain ASV, leading to less predictable values. Compared to individual ASV models, community-level models measured by richness and Shannon diversity performed better. This difference may be due to the fact that the weak information given by rarer ASVs were pooled into more general information and it is thus more efficient to link predictors with the whole community.

Model projections suggested that the Arctic region harbors some hotspots of diversity, as expected by Malard & Pearce (2018). Surprisingly, the highest diversity was concentrated in the northern part of the region. This observation, supported by the significant positive relationship between diversity and latitude, were in contradiction with the latitudinal diversity gradient hypothesis and previous studies (Bahram et al., 2018; Malard et al., 2022). However, the same findings were demonstrated for fungal diversity (Větrovský et al., 2019). This inconsistency may be caused by the underrepresentation of high latitudinal regions which introduces bias at the global scale. Indeed, the present analysis focused on a truncated latitudinal gradient with the potential to reveal different trends, unlike global studies. As microbial distribution can change rapidly under warming (Cavicchioli et al., 2019), another

possible explanation was that their suitable habitat has recently shifted under climate change and soil bacteria had already migrated when sampled, resulting in a temporal shift between the climatic conditions used and the actual conditions when sampling the communities.

Some regions appeared to harbour high richness but were characterized by high to low Shannon diversity. This indicated that for the same ASV richness, the evenness can vary. A lower Shannon diversity with fixed richness suggested the dominance of a few ASVs among the community. In contrary, fixed richness with higher Shannon diversity indicated a good evenness in ASV relative abundance.

The prediction differences between GLM and RF algorithms illustrated by the standard deviation maps reflected the uncertainty among the ensemble predictions. Some deviations were expected as the two models translated different information and GLM predicted more extreme values than RF. Since GLM was a regression-based method and the environmental gradient was not fully covered, extrapolation outside the range of observed diversity was difficult to avoid.

Assuming the stability of the ecological niche through time, soil bacterial diversity was expected to change under all the considered scenarios. Although some areas experienced a loss of diversity, the changes were mostly positive, suggesting an increased in bacterial activity and thus possible interference in ecosystem functioning. These results were in line with the hypothesis that permafrost thaw releases more organic carbon for microbial activity (Koven et al., 2011; Tarnocai et al., 2009). Moreover, the bidirectionality of diversity changes were in good agreement with previous studies showing that microbial diversity responds differently to warming depending on the habitat types (Lladó et al., 2017; Větrovský et al., 2019). While Yang et al. (2021) found negative effects of global change on soil microbial diversity, abundance is expected to be enhanced in cold regions, stronger than in temperate regions (Chen et al., 2015). These inconsistencies highlighted the need to consider the direction of the impact of climate change specific to the ecosystem types.

Shannon diversity changes occurred at little amplitude. This could be due to the Shannon diversity distribution model which did not include all the variables used for future scenarios, compared to the richness distribution model. This implies that richness distribution changes were the result of all the environmental future variations compared to Shannon diversity

changes which did not depend on changes in annual precipitation over zero. Another reason was that the different component of the metric had reverse effect on the index. For example, gain in richness would result in higher Shannon diversity which can be lowered if some ASV become dominant, the information would be therefore blurred.

2. Community structure

The direction and intensity of the relationship between environmental variables and phyla were identified using heatmap and linear regressions and could be used as an indicator of how community may vary with changing environments. The differential responses to the environmental gradient among the phylum-level community points to potential changes in community structure under future climate. Note that impacts on bacterial communities can be species specific and the signal can thus be weakened if opposite relationships occurred within a phylum. First of all, pH positively correlated with most of the phyla but decreased with some of the most abundant phyla (Verrucomicrobiota, Actinobacteriota and Planctomycetota). Soil acidification caused by higher precipitation would therefore alter soil bacterial community as reported previously (Mod et al., 2021; Zhou et al., 2021) and benefit the few most abundant at the expense of most phyla. Nevertheless, the association between precipitation and pH has been found to be weaker in high latitudes since pH is expected to be more dependent on the mineralogy of the bedrock than precipitation (Vasar et al., 2022). Then, soil organic carbon, which is a main driver of soil bacterial community (Fierer, 2017), had negative correlations with most phyla. Its availability is expected to increase in the future, and presumably may lead to a decreased abundance of most phyla for the benefit of Verrucomicrobiota, Actinobacteriota and Planctomycetota, again. Increasing radiation, associated to the increasing temperature, had negative correlations with most phyla and positive for Planctomycetota, Methyloirabilota (less abundant phyla) and Actinobacteriota. Moisture and annual precipitation over and under zero, all together relate to soil water availability, can alter soil nutrient availability indirectly for bacterial growth (Wu et al., 2020). The directions of their relationship with community were more diverse, some corroborated and some did not with previous studies (Bachar et al., 2010; She et al., 2018). However, the inconsistent responses were not particularly surprising given the fact that they may be specific to the considered ecosystem. Finally, thawing degree day and active layer thickness are likely to increase with warming and inversely for freezing degree days. As anticipated, most phyla

were significantly affected by thawing and freezing degree days. On the other hand, active layer thickness was surprisingly correlated with less phyla. Based on these hypothetical sensitivities, climate change effects would mostly enhance the abundance of the most abundant phyla and rarely advantage the less abundant phyla.

All the 14 environmental variables were selected and integrated in at least 20% of the models to explain individual ASV distribution. The more frequently selected variables concurred well with those identified as highly correlated with the soil bacterial community discussed in the previous analysis (e.g., pH, freezing degree days, moisture and radiation). This result confirmed that pH was the most important factor found in the literature (Chu et al., 2020; Fierer, 2017; Malard et al., 2019; Yashiro et al., 2016). The different choices of predictors between the ASVs implied that future changes were driven by different set of predictors, and possible including no or a few of the future variable available for our analyses.

Although some ASV distribution model were removed from the community analysis due to bad performance, model evaluations indicated Spearman correlation of up to 0.8. Inevitably, there was high variation in model performance due to the variation in information given by ASV distribution to fit the models. Linking environmental variables with rare or little abundant ASVs resulted in underpowered models.

Individual ASV distribution predictions were stacked to provide an overall community structure across the sampling sites under current and future environmental conditions. It was probable that the considerable changes of ASV1584 was an artefact, due to the poor model qualities, and was not considered in the community analysis. Contrary to expectations, stacked ASV distribution models suggested no noteworthy changes in community structure at phylum level, although a slight increase in relative abundance of Myxococcota. This change was also not expected since this phylum was poorly correlated with a few environmental variables. Based on the PCoA and Bray-Curtis dissimilarity index, all communities under future scenarios appeared to be different from the current community at ASV level. Among the future scenarios, the worst-case scenario in 2070 is likely to induce stronger changes in community structure compared to other scenarios, which is consistent with the characteristics of this scenario.

While modelling microbial communities in the future is essential to provide a better understanding of the impacts of climate change and the associated consequences on ecosystems (Chu et al., 2020; Malard & Pearce, 2018), we should note that model performances were relatively poor. Therefore, the results from such analyses should be treated with caution. There are several possible explanations impacting model performance. To begin with, the sample size, unevenly distributed across the region, was not able to capture all variable gradients. This understandable limitation underlined the difficulty of collecting data in such an inaccessible region. Another possible reason why the models may fail to explain soil bacterial distribution would be a lack of precision in spatial or temporal scale. Indeed, soil bacterial community may respond to smaller-scale changes. For example, abundance can be affected by micro-climate, micro-topography or micro-edaphic factors (Hill et al., 2016), whereas cell grid resolution in the present study is one kilometre. In terms of temporal scale, some climate data were averaged over a large period (2000-2014). However, bacterial diversity and community structure may be sensitive to meteorological data (Malard et al., 2022) or seasonal conditions (Hill et al., 2016), which suggests that much smaller temporal scale would be required to improve prediction accuracy.

CONCLUSION

Taken together, these results showed that hotspots of soil bacterial diversity are present and mainly located in the highest latitude of the Arctic region, and that both soil bacterial diversity and community structure were likely to change with combined climate and edaphic changes under the three different RCP scenarios. This study contributed to improve our knowledge about global biogeography of soil microbial community and points towards the idea that climate change affects the different part of the globe in different ways. It demonstrated that species distribution modelling approaches can be used to provide predictive insight into how soil bacterial diversity may vary across spatio-temporal scales but predicting changes in community structure remains challenging. Moreover, identifying environmental drivers specific to the region is crucial to accurately anticipate consequences of environmental changes. The natural next step for future research is to determine how bacterial communities are linked to their ecosystem functions and what their changes therefore imply by including soil bacterial data in climate models. Although the relationship between bacterial distribution and ecological function is well supported by observational studies, very little is known to

assign functions to taxa (Chu et al., 2020). Therefore, metagenomic and metatranscriptomic analyses associated to greenhouse gas flux measurement would be critical to improve our understanding of soil bacterial influence on ecosystems worldwide.

ACKNOWLEDGEMENTS

I thank the following people for helping with this project: Lucie Malard, who provided the data I used, for her guidance throughout my work and feedback, Valentin Verdon for his help with modelling, Antoine Guisan for his feedback on my thesis and the rest of the Spatial Ecology Group when I had questions.

REFERENCES

- Aalto, J., Karjalainen, O., Hjort, J., & Luoto, M. (2018). Statistical Forecasting of Current and Future Circum-Arctic Ground Temperatures and Active Layer Thickness. *Geophysical Research Letters*, 45(10), 4889–4898. <https://doi.org/10.1029/2018GL078007>
- Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data*, 5(1), Article 1. <https://doi.org/10.1038/sdata.2018.40>
- Araújo, M. B., Alagador, D., Cabeza, M., Nogués-Bravo, D., & Thuiller, W. (2011). Climate change threatens European conservation areas. *Ecology Letters*, 14(5), 484–492. <https://doi.org/10.1111/j.1461-0248.2011.01610.x>
- Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22(1), 42–47. <https://doi.org/10.1016/j.tree.2006.09.010>
- Bachar, A., Al-Ashhab, A., Soares, M. I. M., Sklarz, M. Y., Angel, R., Ungar, E. D., & Gillor, O. (2010). Soil Microbial Abundance and Diversity Along a Low Precipitation Gradient. *Microbial Ecology*, 60(2), 453–461. <https://doi.org/10.1007/s00248-010-9727-1>
- Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., Bengtsson-Palme, J., Anslan, S., Coelho, L. P., Harend, H., Huerta-Cepas, J., Medema, M. H., Maltz, M. R., Mundra, S., Olsson, P. A., Pent, M., Pölme, S., Sunagawa, S., Ryberg, M., ... Bork, P. (2018). Structure and function of the global topsoil microbiome. *Nature*, 560(7717), 233–237. <https://doi.org/10.1038/s41586-018-0386-6>
- Bray, J. R., & Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27(4), 325–349. <https://doi.org/10.2307/1942268>
- Brown, J., Ferrians, O., Heginbottom, J. A., & Melnikov, E. (2002). *Circum-Arctic Map of Permafrost and Ground-Ice Conditions, Version 2*. National Snow and Ice Data Center. <https://doi.org/10.7265/skbk-kf16>

- Brown, J., & Romanovsky, V. E. (2008). Report from the International Permafrost Association: State of permafrost in the first decade of the 21st century. *Permafrost and Periglacial Processes*, 19(2): 255-260. <https://doi.org/10.1002/ppp.618>
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), Article 7. <https://doi.org/10.1038/nmeth.3869>
- Cameron, E. K., Martins, I. S., Lavelle, P., Mathieu, J., Tedersoo, L., Gottschall, F., Guerra, C. A., Hines, J., Patoine, G., Siebert, J., Winter, M., Cesarz, S., Delgado-Baquerizo, M., Ferlian, O., Fierer, N., Kreft, H., Lovejoy, T. E., Montanarella, L., Orgiazzi, A., ... Eisenhauer, N. (2018). Global gaps in soil biodiversity data. *Nature Ecology & Evolution*, 2(7), Article 7. <https://doi.org/10.1038/s41559-018-0573-8>
- Castro, H. F., Classen, A. T., Austin, E. E., Norby, R. J., & Schadt, C. W. (2010). Soil Microbial Community Responses to Multiple Experimental Climate Change Drivers. *Applied and Environmental Microbiology*, 76(4), 999–1007. <https://doi.org/10.1128/AEM.02874-09>
- Cavicchioli, R., Ripple, W. J., Timmis, K. N., Azam, F., Bakken, L. R., Baylis, M., Behrenfeld, M. J., Boetius, A., Boyd, P. W., Classen, A. T., Crowther, T. W., Danovaro, R., Foreman, C. M., Huisman, J., Hutchins, D. A., Jansson, J. K., Karl, D. M., Koskella, B., Mark Welch, D. B., ... Webster, N. S. (2019). Scientists' warning to humanity: Microorganisms and climate change. *Nature Reviews Microbiology*, 17(9), Article 9. <https://doi.org/10.1038/s41579-019-0222-5>
- Chen, J., Luo, Y., Xia, J., Jiang, L., Zhou, X., Lu, M., Liang, J., Shi, Z., Shelton, S., & Cao, J. (2015). Stronger warming effects on microbial abundances in colder regions. *Scientific Reports*, 5(1), Article 1. <https://doi.org/10.1038/srep18032>
- Chu, H., Gao, G.-F., Ma, Y., Fan, K., & Delgado-Baquerizo, M. (2020). Soil Microbial Biogeography in a Changing World: Recent Advances and Future Perspectives. *MSystems*, 5(2), e00803-19. <https://doi.org/10.1128/mSystems.00803-19>

- Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). *Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data* (p. 221499). bioRxiv. <https://doi.org/10.1101/221499>
- de Vries, A., & D. Ripley, B. (2022). *ggdendro: Create Dendrograms and Tree Diagrams Using “ggplot2”* (R package version 0.1.23). <https://CRAN.R-project.org/package=ggdendro>
- Delgado-Baquerizo, M., & Eldridge, D. J. (2019). Cross-Biome Drivers of Soil Bacterial Alpha Diversity on a Worldwide Scale. *Ecosystems*, 22(6), 1220–1231. <https://doi.org/10.1007/s10021-018-0333-2>
- Delgado-Baquerizo, M., Maestre, F. T., Reich, P. B., Jeffries, T. C., Gaitan, J. J., Encinar, D., Berdugo, M., Campbell, C. D., & Singh, B. K. (2016). Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nature Communications*, 7(1), Article 1. <https://doi.org/10.1038/ncomms10541>
- Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., Maestre, F. T., Singh, B. K., & Fierer, N. (2018). A global atlas of the dominant bacteria found in soil. *Science*, 359(6373), 320–325. <https://doi.org/10.1126/science.aap9516>
- Didan, K. (2015). *MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006*. NASA EOSDIS Land Processes DAAC. <https://doi.org/10.5067/MODIS/MOD13A2.006>
- Dimitriu, P. A., & Grayston, S. J. (2010). Relationship Between Soil Properties and Patterns of Bacterial β -diversity Across Reclaimed and Natural Boreal Forest Soils. *Microbial Ecology*, 59(3), 563–573. <https://doi.org/10.1007/s00248-009-9590-0>
- Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., ... E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2), 129–151. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>

- ESA. *Land Cover CCI Product User Guide Version 2*. (2017). [Tech. Rep.].
maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf
- Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, 43(3), 393–404. <https://doi.org/10.1111/j.1365-2664.2006.01149.x>
- Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12), 4302–4315.
<https://doi.org/10.1002/joc.5086>
- Fierer, N. (2017). Embracing the unknown: Disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology*, 15(10), Article 10. <https://doi.org/10.1038/nrmicro.2017.87>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.
- Guisan, A., & Rahbek, C. (2011). SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, 38(8), 1433–1444. <https://doi.org/10.1111/j.1365-2699.2011.02550.x>
- Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models. With applications in R*. <https://doi.org/10.1017/9781139028271>
- Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2), 147–186. [https://doi.org/10.1016/S0304-3800\(00\)00354-9](https://doi.org/10.1016/S0304-3800(00)00354-9)
- Harrell Jr, F. E. (2022). *Hmisc: Harrell Miscellaneous* (R package version 4.7-0). <https://CRAN.R-project.org/package=Hmisc>
- Hill, R., Saetnan, E. R., Scullion, J., Gwynn-Jones, D., Ostle, N., & Edwards, A. (2016). Temporal and spatial influences incur reconfiguration of Arctic heathland soil bacterial community structure. *Environmental Microbiology*, 18(6), 1942–1953. <https://doi.org/10.1111/1462-2920.13017>

- Intergovernmental Panel On Climate Change (IPCC). (2017). *IPCC Fifth Assessment Report (AR5) Observed Climate Change Impacts Database, Version 2.01* [Data set]. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). <https://doi.org/10.7927/H4FT8J0X>
- Jani, K., Mahajan, A., Kajale, S., Ashar, A., & Sharma, A. (2022). Soil bacterial community structure, metabolic adaptations and their functional interactions to abiotic factors in Antarctica. *Polish Polar Research*, 21–36. <https://doi.org/10.24425/ppr.2021.138588>
- Jansson, J. K., & Hofmockel, K. S. (2020). Soil microbiomes and climate change. *Nature Reviews Microbiology*, 18(1), Article 1. <https://doi.org/10.1038/s41579-019-0265-7>
- Ji, M., Kong, W., Jia, H., Delgado-Baquerizo, M., Zhou, T., Liu, X., Ferrari, B. C., Malard, L., Liang, C., Xue, K., Makhalanyane, T. P., Zhu, Y.-G., Wang, Y., Pearce, D. A., & Cowan, D. (2022). Polar soils exhibit distinct patterns in microbial diversity and dominant phylotypes. *Soil Biology and Biochemistry*, 166, 108550. <https://doi.org/10.1016/j.soilbio.2022.108550>
- Koven, C. D., Ringeval, B., Friedlingstein, P., Ciais, P., Cadule, P., Khvorostyanov, D., Krinner, G., & Tarnocai, C. (2011). Permafrost carbon-climate feedbacks accelerate global warming. *Proceedings of the National Academy of Sciences*, 108(36), 14769–14774. <https://doi.org/10.1073/pnas.1103910108>
- Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2022). *caret: Classification and Regression Training* (R package version 6.0-92). <https://CRAN.R-project.org/package=caret>
- Lenters, V., Vermeulen, R., & Portengen, L. (2018). Performance of variable selection methods for assessing the health effects of correlated exposures in case–control studies. *Occupational and Environmental Medicine*, 75(7), 522–529.
- Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3), 18–22.

- Liu, Y., Zhang, L., Lu, J., Chen, W., Wei, G., & Lin, Y. (2020). Topography affects the soil conditions and bacterial communities along a restoration gradient on Loess-Plateau. *Applied Soil Ecology*, 150, 103471. <https://doi.org/10.1016/j.apsoil.2019.103471>
- Lladó, S., López-Mondéjar, R., & Baldrian, P. (2017). Forest Soil Bacteria: Diversity, Involvement in Ecosystem Processes, and Response to Global Change. *Microbiology and Molecular Biology Reviews*, 81(2), e00063-16. <https://doi.org/10.1128/MMBR.00063-16>
- Malard, L. A., Anwar, M. Z., Jacobsen, C. S., & Pearce, D. A. (2019). Biogeographical patterns in soil bacterial communities across the Arctic region. *FEMS Microbiology Ecology*, 95(9), fiz128. <https://doi.org/10.1093/femsec/fiz128>
- Malard, L. A., Avila-Jimenez, M.-L., Schmale, J., Cuthbertson, L., Cockerton, L., & Pearce, D. A. (2022). Aerobiology over the Southern Ocean – Implications for bacterial colonization of Antarctica. *Environment International*, 169, 107492. <https://doi.org/10.1016/j.envint.2022.107492>
- Malard, L. A., & Pearce, D. A. (2018). Microbial diversity and biogeography in Arctic soils: Microbial diversity and biogeography. *Environmental Microbiology Reports*, 10(6), 611–625. <https://doi.org/10.1111/1758-2229.12680>
- Mantyka-pringle, C. S., Martin, T. G., & Rhodes, J. R. (2012). Interactions between climate and habitat loss effects on biodiversity: A systematic review and meta-analysis. *Global Change Biology*, 18(4), 1239–1252. <https://doi.org/10.1111/j.1365-2486.2011.02593.x>
- Maron, P.-A., Sarr, A., Kaisermann, A., Lévêque, J., Mathieu, O., Guigue, J., Karimi, B., Bernard, L., Dequiedt, S., Terrat, S., Chabbi, A., & Ranjard, L. (2018). High Microbial Diversity Promotes Soil Ecosystem Functioning. *Applied and Environmental Microbiology*, 84(9), e02738-17. <https://doi.org/10.1128/AEM.02738-17>
- McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>

- Mod, H. K., Buri, A., Yashiro, E., Guex, N., Malard, L., Pinto-Figueroa, E., Pagni, M., Niculita-Hirzel, H., van der Meer, J. R., & Guisan, A. (2021). Predicting spatial patterns of soil bacteria under current and future environmental conditions. *Isme Journal*, 15(9), 2547–2560.
<https://doi.org/10.1038/s41396-021-00947-5>
- Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., Caceres, M. D., Durand, S., ... Weedon, J. (2022). *vegan: Community Ecology Package* (R package version 2.6-2). <https://CRAN.R-project.org/package=vegan>
- Pajunen, V., Luoto, M., & Soininen, J. (2016). Climate is an important driver for stream diatom distributions. *Global Ecology and Biogeography*, 25(2), 198–206.
<https://doi.org/10.1111/geb.12399>
- Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3), 526–528. <https://doi.org/10.1093/bioinformatics/bty633>
- Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1), 217–240. <https://doi.org/10.5194/soil-7-217-2021>
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1), D590–D596.
<https://doi.org/10.1093/nar/gks1219>
- Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., Vihma, T., & Laaksonen, A. (2022). The Arctic has warmed nearly four times faster than the globe since 1979. *Communications Earth & Environment*, 3(1), Article 1. <https://doi.org/10.1038/s43247-022-00498-3>

- RStudio Team. (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC.
<http://www.rstudio.com/>
- She, W., Bai, Y., Zhang, Y., Qin, S., Feng, W., Sun, Y., Zheng, J., & Wu, B. (2018). Resource Availability Drives Responses of Soil Microbial Communities to Short-term Precipitation and Nitrogen Addition in a Desert Shrubland. *Frontiers in Microbiology*, 9.
<https://www.frontiersin.org/articles/10.3389/fmicb.2018.00186>
- Siciliano, S. D., Palmer, A. S., Winsley, T., Lamb, E., Bissett, A., Brown, M. V., van Dorst, J., Ji, M., Ferrari, B. C., Grogan, P., Chu, H., & Snape, I. (2014). Soil fertility is associated with fungal and bacterial richness, whereas pH is associated with community composition in polar soil microbial communities. *Soil Biology and Biochemistry*, 78, 10–20.
<https://doi.org/10.1016/j.soilbio.2014.07.005>
- Tarnocai, C., Canadell, J. G., Schuur, E. a. G., Kuhry, P., Mazhitova, G., & Zimov, S. (2009). Soil organic carbon pools in the northern circumpolar permafrost region. *Global Biogeochemical Cycles*, 23(2). <https://doi.org/10.1029/2008GB003327>
- Vasar, M., Davison, J., Sepp, S.-K., Mucina, L., Oja, J., Al-Quraishy, S., Anslan, S., Bahram, M., Bueno, C. G., Cantero, J. J., Decocq, G., Fraser, L., Hiiesalu, I., Hozzein, W. N., Koorem, K., Meng, Y., Moora, M., Onipchenko, V., Öpik, M., ... Zobel, M. (2022). Global soil microbiomes: A new frontline of biome-ecology research. *Global Ecology and Biogeography*, 31(6), 1120–1132.
<https://doi.org/10.1111/geb.13487>
- Větrovský, T., Kohout, P., Kopecký, M., Machac, A., Man, M., Bahnmann, B. D., Brabcová, V., Choi, J., Meszárošová, L., Human, Z. R., Lepinay, C., Lladó, S., López-Mondéjar, R., Martinović, T., Mašíňová, T., Morais, D., Navrátilová, D., Odriozola, I., Štursová, M., ... Baldrian, P. (2019). A meta-analysis of global fungal distribution reveals climate-driven patterns. *Nature Communications*, 10(1), Article 1. <https://doi.org/10.1038/s41467-019-13164-8>
- Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
<https://ggplot2.tidyverse.org>

- Wu, K., Xu, W., & Yang, W. (2020). Effects of precipitation changes on soil bacterial community composition and diversity in the Junggar desert of Xinjiang, China. *PeerJ*, 8, e8433. <https://doi.org/10.7717/peerj.8433>
- Yang, Y., Li, T., Wang, Y., Cheng, H., Chang, S. X., Liang, C., & An, S. (2021). Negative effects of multiple global change factors on soil microbial diversity. *Soil Biology and Biochemistry*, 156, 108229. <https://doi.org/10.1016/j.soilbio.2021.108229>
- Yashiro, E., Pinto-Figueroa, E., Buri, A., Spangenberg, J. E., Adatte, T., Niculita-Hirzel, H., Guisan, A., & van der Meer, J. R. (2016). Local Environmental Factors Drive Divergent Grassland Soil Bacterial Communities in the Western Swiss Alps. *Applied and Environmental Microbiology*, 82(21), 6303–6316. <https://doi.org/10.1128/AEM.01170-16>
- Zhou, X., Khashi u Rahman, M., Liu, J., & Wu, F. (2021). Soil acidification mediates changes in soil bacterial community assembly processes in response to agricultural intensification. *Environmental Microbiology*, 23(8), 4741–4755. <https://doi.org/10.1111/1462-2920.15675>

SUPPLEMENTARY

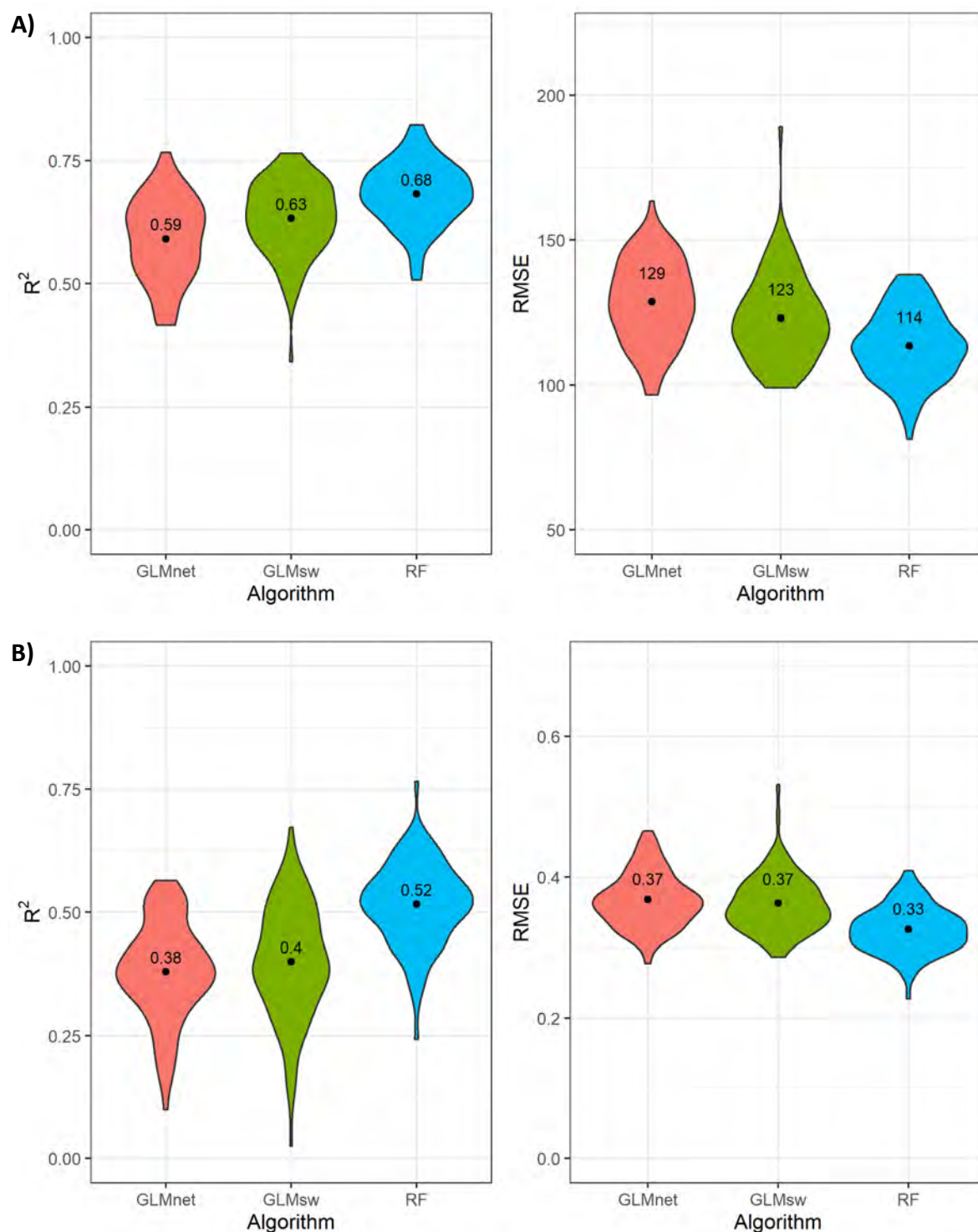


Figure S1: Model evaluation by repeated split-sample cross-validation (100x), where 75% of the data was assigned to calibration and the remaining 25% to evaluation. The three algorithms GLMnet, GLMsw and RF were evaluated with R^2 and RMSE. Mean point and its associated value is displayed. **(A)** Evaluation for soil bacterial richness model. **(B)** Evaluation for Shannon diversity model.

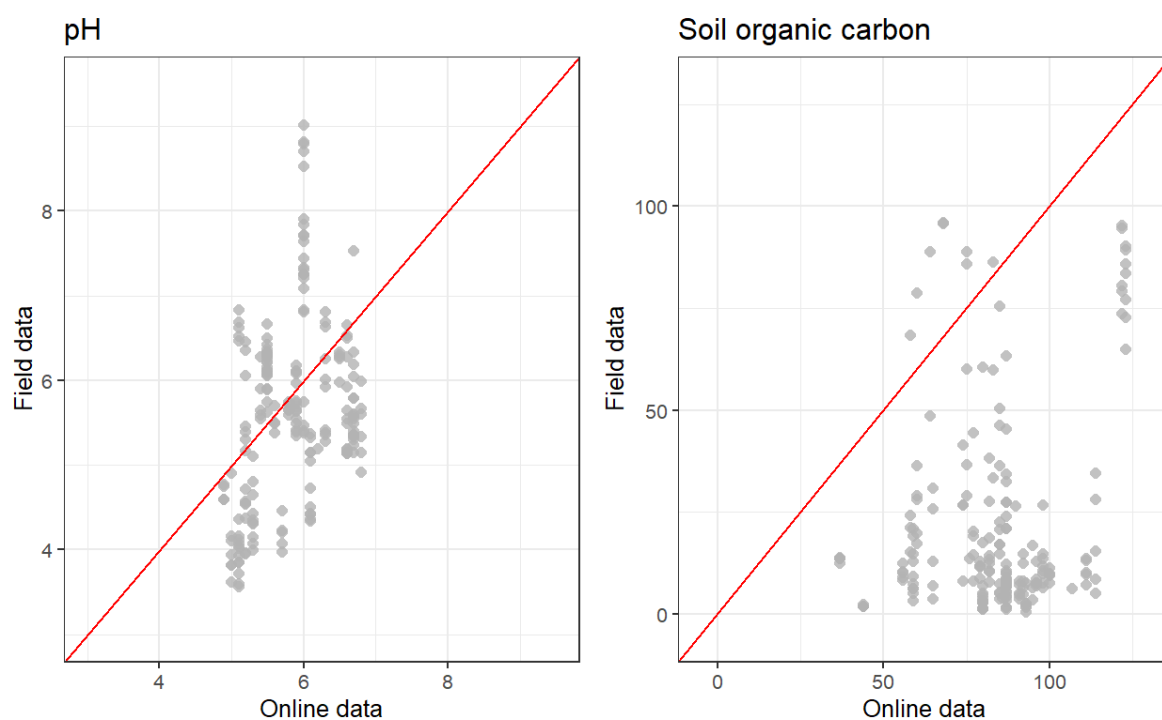


Figure S2: Comparison of data measured in the field and data from Soilgrids 2.0 for pH (left panel) and soil organic carbon (right panel). The red line corresponds to a slope of one.

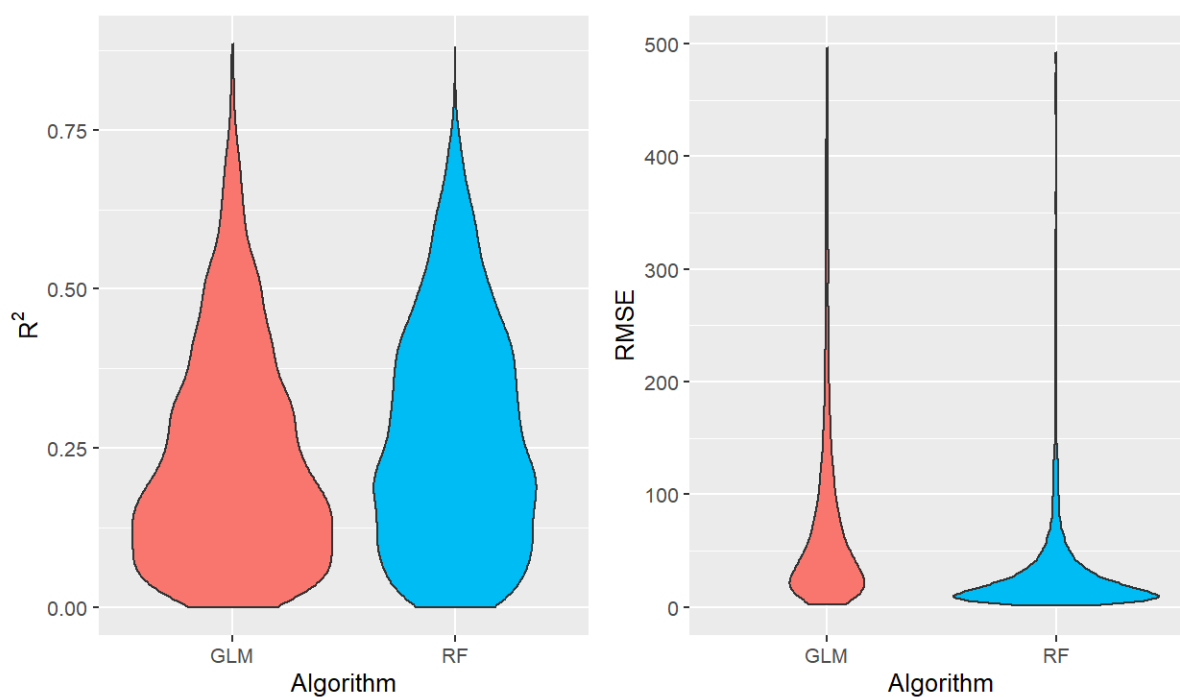


Figure S3: ASV distribution model evaluations by repeated split-sample cross-validation (100x), where 75% of the data was assigned to calibration and the remaining 25% to evaluation. Mean R^2 (right panel) and mean RMSE (left panel) values of all ASV distribution model evaluations. The two metrics used (R^2 and RMSE) indicate bad model qualities.

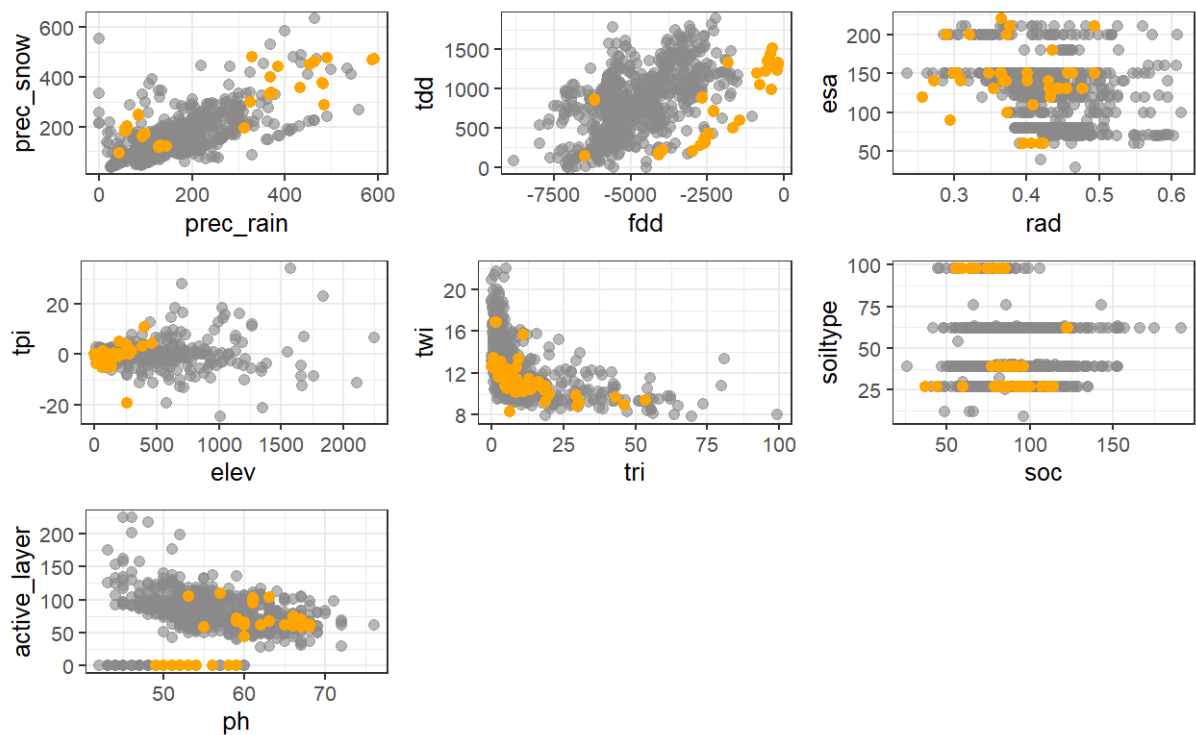


Figure S4: Coverage of sampling sites on Arctic environmental gradients for the 14 variables. Each plot illustrates the coverage for two predictors arbitrarily paired. The orange dots are the 200 sampling sites, and the grey dots are 1000 randomly selected values in the Arctic region.

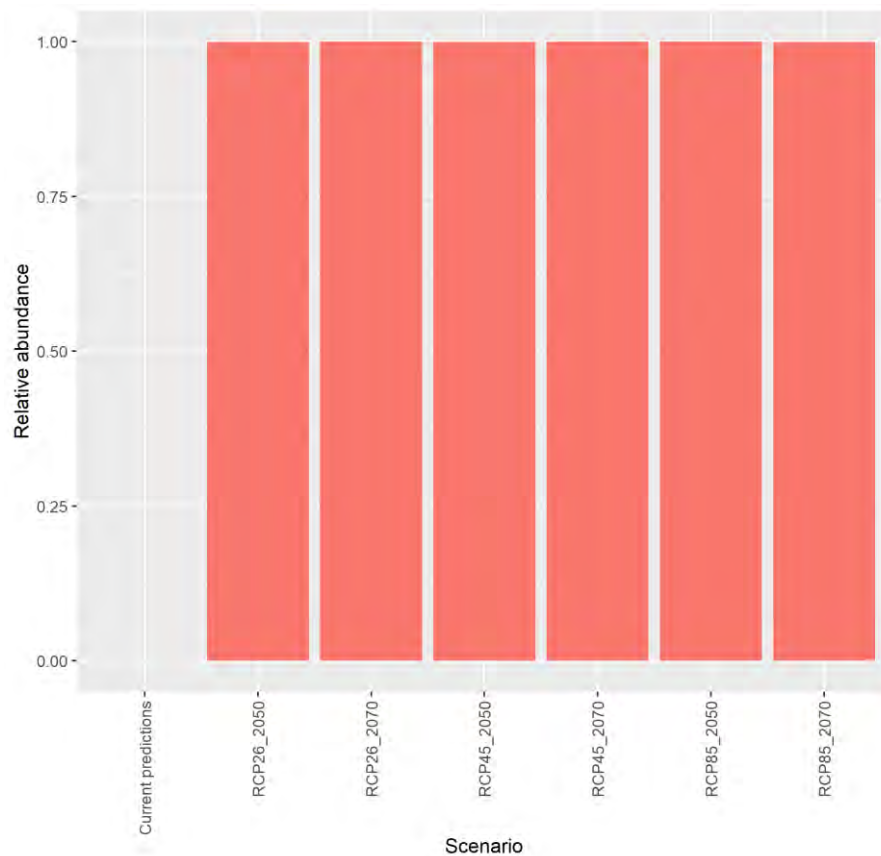


Figure S5: ASV1584 relative abundance under current conditions and the different RCP scenarios.