Can we accurately predict the distribution of soil microorganism presence and relative abundance?

Valentin Verdon[1], Lucie Malard[2], Flavien Collart[1], Antoine Adde[1], Nicolas Guex[1], Heidi Mod[3], Erika Yashiro[1], Enrique Lara[4], David Singer[5], Hélène Niculita-Hirzel[1], and Antoine Guisan[1]

[1]University of Lausanne
[2]Université de Lausanne
[3]University of Helsinki
[4]Université de Neuchâtel
[5]University of Sciences and Art Western Switzerland

September 4, 2023

Abstract

Soil microbes play a key role in shaping terrestrial ecosystems. It is therefore essential to understand what drives their distributions. While multivariate analyses have been used to characterise microbial communities and drivers of their spatial patterns, few studies focused on modelling the distribution of Operational Taxonomic Units (OTUs). Here, we evaluate the potential of species distribution models (SDMs), to predict the presence-absence and relative abundance distribution of bacteria, archaea, fungi and protist OTUs from the Swiss Alps. Advanced automated selection of abiotic covariates was used to circumvent the lack of knowledge on the ecology of each OTU. 'Presence-absence' SDMs were successfully applied to most OTUs, yielding better predictions than null models. 'Relative-abundance' SDMs were less successful, yet, they were able to correctly rank sites according to their relative abundance values. Archaea and bacteria SDMs displayed better predictive power than fungi and protist ones, indicating a closer link of the latter with the abiotic covariates used. Microorganism distributions were mostly related to edaphic covariates. In particular, pH was the most selected covariate across models. The study shows the potential of using SDM frameworks to predict the distribution of OTUs obtained from environmental DNA (eDNA) data. It underscores the importance of edaphic covariates and the need for further development of precise edaphic mapping and scenario modelling to enhance prediction of microorganism distributions in the future.

Can we accurately predict the distribution of soil microorganism presence and relative abundance from environmental DNA?

Abstract

Soil microbes play a key role in shaping terrestrial ecosystems. It is therefore essential to understand what drives their distributions. While multivariate analyses have been used to characterise microbial communities and drivers of their spatial patterns, few studies focused on modelling the distribution of Operational Taxonomic Units (OTUs). Here, we evaluate the potential of species distribution models (SDMs), to predict the presence-absence and relative abundance distribution of bacteria, archaea, fungi and protist OTUs from the Swiss Alps. Advanced automated selection of abiotic covariates was used to circumvent the lack of knowledge on the ecology of each OTU. 'Presence-absence' SDMs were successfully applied to most OTUs, yielding better predictions than null models. 'Relative-abundance' SDMs were less successful, yet, they were able

to correctly rank sites according to their relative abundance values. Archaea and bacteria SDMs displayed better predictive power than fungi and protist ones, indicating a closer link of the latter with the abiotic covariates used. Microorganism distributions were mostly related to edaphic covariates. In particular, pH was the most selected covariate across models. The study shows the potential of using SDM frameworks to predict the distribution of OTUs obtained from environmental DNA (eDNA) data. It underscores the importance of edaphic covariates and the need for further development of precise edaphic mapping and scenario modelling to enhance prediction of microorganism distributions in the future.

Keywords : amplicon sequencing, species distribution model, topsoil, eDNA, bacteria, fungi, archaea, protist, cross-validation, environmental niche.

Introduction

Soil microbes play a key role in shaping terrestrial ecosystems and their responses to climate change and land degradation (Karhu et al., 2014; Cavicchioli et al., 2019) by driving soil functions such as cycling of nutrients and carbon (Philippot et al., 2013; Bardgett & Van Der Putten, 2014; Jiao et al., 2021). For example, rising temperatures could enhance microbial activity, leading to increased carbon release from soil to the atmosphere (Crowther et al., 2016; Ballantyne et al., 2017; Rocci et al., 2021) thereby further amplifying global temperature rise. However, the mechanisms and rates of carbon and nutrient releases depend on the composition and spatial patterns of the soil microbial communities present in the environment (Nottingham et al., 2015, 2019). Variations of these patterns have been observed from micro (Nunan et al., 2003) to regional (Yashiro et al., 2018; Pinto-figueroa et al., 2019; Mazel et al., 2021; Seppey et al., 2023) and global scales (Birkhofer et al., 2012; Bahram et al., 2018). These distribution patterns could, in turn, be retroactively affected by future land-use and climatic changes (Guo et al., 2018; Cavicchioli et al., 2019; Mod et al., 2021).

To better spatially characterise and quantify soil functions, there is a need to improve knowledge on the spatial patterns of soil microbial communities and their components (Bardgett & Van Der Putten, 2014; Mod et al., 2020). Spatial patterns are commonly studied for macro-organisms using species distribution models (SDMs; Franklin, 2010; Peterson et al., 2011; Guisan et al., 2017). SDMs were first developed to relate 'presence-absence' of a taxonomic unit (e.g. species for most macro-organisms studies) to environmental conditions, relying on Hutchinson's realised environmental niche theory (Hutchinson, 1957; Baquero et al., 2021). The realised environmental niche refers to a multidimensional volume in environmental space, representing the different characteristics (e.g. soil, climate, topography, land cover/use) where the taxonomic unit can be found. SDMs characterise this hyper-volume and project it onto geographical space based on environmental conditions, in order to predict the spatial distribution of the modelled unit (Guisan et al., 2017; Araújo et al., 2019). Most of the time, these models predict probabilities of occurrence, but other properties can also be predicted, such as abundance (Waldock et al., 2022), or species and population characteristics such as fitness and genetic diversity (Lee-Yaw et al., 2022).

For microbial diversity, spatial distribution had been previously studied mostly at the community level by linking soil microbial community characteristics such as diversity metrics (Fierer & Jackson, 2006; Griffiths et al., 2016; Ren et al., 2018; Seppey et al., 2020), abundance patterns (Pinto-figueroa et al., 2019), dominance patterns (de Vries et al., 2012), and total biomass (Serna-Chavez et al., 2013; Horrigue et al., 2016) to environmental abiotic predictors such as climate and soil edaphic conditions. Community-level characteristics are commonly obtained by summarising information about Operational Taxonomic Units (OTU) within and across sampling sites, in which OTUs are defined as operationally relevant clusters of environmental DNA (eDNA) amplicon sequenced reads grouped at specific taxon levels. However, individual OTUs may have different responses to environmental factors, which may not be revealed by community-level analyses. This problem, already discussed for macro-organisms (Guisan & Rahbek, 2011), leads to the challenge of modelling individual OTUs' distributions.

To our knowledge, few microbial studies attempted to model the distribution of microorganisms at the OTU level (Mod et al., 2021). The main reasons could be methodological limitations regarding the lack of

appropriate environmental descriptors for soils (Lembrechts et al., 2020), the tremendous number of models to calibrate for a whole library of OTUs, and the difficulty to compare OTUs derived from different databases because of the lack of standardisation of algorithms that cluster sequences into OTUs (Deiner et al., 2015). Some studies used SDM frameworks below the community level, by working with sequences clustered into clades at coarse taxonomic levels (King et al., 2010). Moreover, it has been shown in plants and animals that SDM frameworks can be applied with success to taxonomic units above and below the species level, as long as there was a link between the biological unit and the environmental covariates used (Hadly et al., 2009; Smith et al., 2019).

Relating geographically-referenced eDNA sampling points to information on local environmental condition provides a possibility to model the presence vs. absence of an OTU by characterising the OTU's environmental niche (i.e. "OTU-environment relationships"). Since read counts per OTU can be somehow related to species abundance (Giner et al., 2016; Galazzo et al., 2020), abundance models can also be fitted for OTUs (Mod et al. 2021). However, the compositionality of eDNA data (Gloor et al., 2017) also means that the direct modelling of absolute abundance (i.e. read counts) might not be meaningful, leaving only the possibility to model 'presence-absence' or 'relative-abundance'.

In this study, we take advantage of recent advances in computing facilities, bioinformatics, and modelling frameworks to fit individual SDMs for all OTUs from a comprehensive mountain soil eDNA database (Yashiro et al., 2016; Pinto-figueroa et al., 2019; Seppey et al., 2020). Our study aims more specifically to test the predictive power of the SDM approach applied to a high number of OTUs using both 'presence-absence' and 'relative-abundance' data, and comparing their predictive performance. To achieve this, we first generated SDMs for more than 60,000 bacterial, archaeal, fungal and protist OTUs across a wide elevational gradient in the Western Swiss Alps. We then evaluated the predictive power of the models and explored differences between the 'presence-absence' and 'relative-abundance' based SDMs across the four microbial groups and their constitutive phyla.

Methods

Study area and data collection

Soil samples were collected at a subset of sites from a larger set of grassland plots (Dubuis et al., 2011, 2013) of the Western Swiss Alps (46°10-46°30'N; 6deg50'-7deg10'E, http://rechalp.unil.ch; Von Daniken et al., 2014; see Figure 1). It is a mountainous region with an elevation range from 425 to 3120 metres above sea level, and very heterogeneous climatic and edaphic conditions. To relate the soil microbiota to environmental values in the area, we used data from 250 sampling sites for bacteria and archaea (Yashiro et al., 2016; Mod et al., 2020), 217 for fungi (Pinto-figueroa et al., 2019; Mod et al., 2020) and 166 for protists (Seppey et al., 2020) . Details on the sampling and DNA sequencing for the three respective groups can be found in the references above, whereas information on assignment of sequenced reads to OTUs have been published in Malard et al. (2022). In brief, soil sampling was conducted from June to September (growing season) during the summers 2012 and 2013. At each selected sampling site, a 2x2 m quadrat was used to sample the top 5 cm of soil with sterilised tools at each corner and at the middle of the quadrat. The five subsamples were then pooled and homogenised into a sample of 500 g representing the site. DNA extraction was done within 36h after collection. Amplification was done targeting the V5 region of the 16S rRNA gene for bacteria and archaea (Lazarevic et al., 2009), the ITS1 rRNA gene operon region for fungi (Schmidt et al., 2013), and the V4 region of the 18S rRNA gene for protists (Stoeck et al., 2010). PCR products were sequenced on the Illumina HiSeq 2500 for 16S and ITS1 amplicons, and on the Illumina MiSeq for 18S amplicons (see Supporting information). Demultiplexing, trimming and merging of sequences, as well as clustering of sequences to obtain zero-radius OTUs (Edgar, 2018), were performed using a custom made pipeline (see details in Mod et al., (2021; Malard et al., 2022). The raw sequence data is available on NCBI bioproject number PRJNA810480 and PRJEB30010.

Proportional abundances (hereafter 'relative-abundance') were obtained by dividing each OTU read count by sequencing depth. 'Presence-absence' data were obtained for each OTU using counts superior or equal

to one as presence, and lack of detection as absence. Taxonomic assignment of OTUs was performed using the IDtaxa classifier (Murali et al., 2018) against the Silva v138 database for bacteria and archaea (Quast et al., 2012), the UNITE+INSD v9.0 database for fungi (Abarenkov et al., 2022), and the PR2 4.5 database for protists (Guillou et al., 2012). OTUs not corresponding to bacteria, archaea, fungi or protists for the corresponding markers were discarded (Supporting Information).

For each site, values representing a wide range of 78 covariates covering climatic, edaphic, topographic, landuse-landcover and remote-sensing conditions were obtained (Supporting Information). For edaphic covariates (i.e. soil characteristics), data were measured from samples collected *in situ* as described in Buri et al. (2020) and Yashiro et al. (2016). For other covariates, data were extracted from spatial layers available at 25 m resolution (Broennimann & Guisan, In review; Kulling et al., In review).

Modelling framework

'Presence-absence' modelling was performed for all OTUs present in more than 5% and less than 95% of sites, leading to 47,520, 163, 17,318 and 2,147 OTUs for bacteria, archaea, fungi and protists, respectively. 'Relative-abundance' modelling was performed for all OTUs present in more than 5% of sites, resulting in modelling of 48,316 bacterial, 163 archaeal, 17,345 fungal and 2155 protist OTUs. This selection was done in order to have enough data points to model each OTU (see suppl. Table 1 for data about OTUs removed).

The following framework was applied to each selected OTU in R v4.3.0 (R Core Team, 2023). All 'presence-absence' models were fitted using a Binomial probability distribution. 'Relative-abundance' models were fitted as read counts models using a Poisson probability distribution and the sequencing depth as an offset (i.e. equivalent to fitting a 'relative-abundance' model; Mod et al., 2021).

Covariates were selected by applying a two-step procedure (Adde et al., 2023). This procedure circumvents the lack of a priori knowledge about most OTUs ecology, and optimises model predictive performances, thereby bringing information on the ecology of the modelled organisms (Adde et al., 2023).

The first step consists of a "data snooping" approach (Dormann et al., 2013). In other words, for each of the 78 candidate covariates, univariate Generalised Linear Models (GLM) with quadratic effect were fitted (Guisan et al., 2002). The models' predictive power, estimated using the difference between the null deviance and the residual deviance, was used to select the 15 best non-correlated covariates recursively, while excluding covariates having a Pearson correlation greater than 0.7 with already selected covariates (Dormann et al., 2013). The number of preselected covariates was capped at 15 in order to limit the computing power needed for the subsequent step of the analysis which further reduced the number of selected covariates using model-embedded regularisation techniques. Four algorithms were used: Generalised linear models (GLM; Guisan et al., 2002), Generalised additive models (GAM; Guisan et al., 2002), Random Forest (RF; Cutler et al., 2007) and Gradient Boosting Machine (GBM; Elith et al., 2008). GLMs had quadratic terms, and a lasso secondary covariate selection and regularisation ("glmnet" package v4.1-7; Tay et al., 2023). GAMs used null-space penalization for covariate selection ("mgcv" package v1.8-42; Wood, 2017). A regularised form of Random Forests (RF) was built using the "RRF" package v1.9.4 (Deng, 2013; Deng & Runger, 2013). For each OTU RF model, the number of trees was determined through hyper-parametrization as in Elith et al. (2008) in order to minimise the error rate of the model (Hastie et al., 2009). We tested four different values ("ntree"=10,100,1000 or 10000). The "mtry" option was left as default value (3 for 'presence-absence' models, 5 for 'relative-abundance' models). Gradient Boosting Machine models were built with the "gbm" package v2.1.8.1 (Greenwell et al., 2022), in which hyper-parametrization procedure was performed on the number of trees (10, 100, 1000 or 10000 trees) and the shrinkage value (0.001, 0.01, or 0.1). We only tested a limited number of hyper-parameters to reduce computing costs.

Cross-validation of models' predictive power

For each of the four algorithms' best models, predictive power was assessed using the "bootstrap .632+" cross-validation procedure (Efron & Tibshirani, 1997) with 100 iterations per model. Unlike classical SDMs

cross-validations that sample data without replacement (e.g. split-sampling, k-fold, see Guisan et al., 2017), this approach uses a bootstrap sample (i.e. with replacement) and generates better estimations of model error rates (Efron, 1983; Efron & Tibshirani, 1997). For each bootstrap iteration of the 'presence-absence' model, the difference between predictions and validation data was computed using AUC (Swets, 1988) and maximised values of TSS (maxTSS) and of Kappa (Allouche et al., 2006). The values were then averaged across the 100 iterations. Results were compared to those obtained by using null models fitted with randomised data (Collart & Guisan, 2023). For each prevalence value existing in the dataset, 100 randomised models were fitted (i.e. 100 models fitting "randomly distributed OTUs" with fixed prevalence). These randomised data were used to rescale real OTUs' maxTSS values for each OTU following this formula:

$\text{maxTSS}_{\text{adj}} = (\text{maxTSS}_{\text{obs}}\text{-maxTSS}_{\text{null}})/(1\text{-maxTSS}_{\text{null}})$

with $\text{maxTSS}_{\text{adj}}$ being the adjusted maxTSS for the considered OTU's model; $\text{maxTSS}_{\text{obs}}$ being the raw maxTSS value obtained for that OTU's model; and $\text{maxTSS}_{\text{null}}$being the 95th percentile of the distribution of models fitted on random data with the same prevalence as the considered OTU. Models having a positive $\text{maxTSS}_{\text{adj}}$ were considered as having higher predictive power than expected by chance given the environmental dataset specificities. By applying the reasoning of Swets (1988) for AUC to our metric, models with $\text{maxTSS}_{\text{adj}}>0.5$ were considered as having a high predictive power.

The same procedure was applied to 'relative-abundance' models using Spearman correlation (rho) to check whether or not models were accurately ranking sites by their 'relative-abundance' values, and Coefficient of Variation (CoV) to assess the difference between predicted 'relative-abundance' values and validation values. The null model procedure could not be applied to 'relative-abundance' models due to the high computing burden required to process the thousands of microbial OTUs in the dataset, as each OTU model would necessitate its own random distribution, whereas 'presence-absence' models with the same prevalence could use the same random distribution. Hence, models were classified as "useful" when rho>0.2 (Mod et al., 2021) and as "good" when rho>0.4 (Landis & Koch, 1977).

After computing the predictive power for all OTU models, differences between organism groups and phyla within each group were explored. Model predictive powers were compared among the four main organism groups using ANOVA followed by Tukey HSD tests with Bonferroni correction, and Cohen's D effect size metrics. To check for the potential effect of the number and locations of sampling sites on model predictive power, additional bacterial models were constructed using only the exact same 166 sites as used for the protist models. Then, performances of models using 250 sites were compared performances of models using 166 sites for each OTU with a paired student's t-test.

Covariate selection and importance

To have an insight on which covariates drive OTUs presence and relative abundance, an analysis per organism group was performed. For each group and each environmental covariate, the proportion of models selecting that covariate was computed by clustering the covariates by their correlation values to match clusters formed during covariates selection (data snooping, Dormann et. al. 2013). For GLMs and GAMs, the importance of covariates was assessed using the coefficients of each covariate. For GBM and RF, covariate importance was assessed using Gini coefficients (Deng, 2013; Greenwell et al., 2022). The obtained values were scaled so that the best covariate from each model had an importance of 1, and the other covariates were linearly rescaled from 1 to 0.

Results

Data on model performance and covariate importance for each OTU can be downloaded in FigShare:*https://figshare.com/s/825799db5d4fdc9a2f87*(private link, which will be published upon acceptance).

1.Evaluation of 'presence-absence' models

Out of the 67,148 preselected OTUs in the dataset, all OTUs were successfully fitted by at least one 'presence-absence' model algorithm and 65,554 OTUs were successfully fitted by all four algorithms (details in Supporting Information). Overall, 91% of bacteria, 98% of archaea, 81% of fungi, and 60% of protists had higher predictive power than null models when modelling their distribution using GLMs (Figure 2a-d; Supporting Information). However, the proportion of "high predictive power" models (i.e. $maxTSS_{adj}>0.4$) was rather low in all groups, with 15% of bacteria, 15% of archaea, 6% of fungi, and 0.1% of protists presenting a $maxTSS_{adj}>0.4$ for GLM (see Supporting information for the other algorithms). 'Presence-absence' models for bacteria and archaea had the best predictive power followed by fungi models and protist models across all 4 modelling algorithms (Figure 2a-d). When the number of sites for bacteria and archaea was reduced from 250 to the corresponding 166 sampled of protists, we observed a small, yet significant, decrease in $maxTSS_{adj}$ (Student's paired tests: bacteria: $p<0.001$, df=45349, Cohen's D=0.15+-0.01; archaea: $p<0.001$, df=157, Cohen's D=0.33+-0.22, Supporting Information). However, these models still had higher $maxTSS_{adj}$ values than protists models (bacteria: $p_{adj}<0.001$, Cohen's D=0.96+-0.05; archaea: $p_{adj}<0.001$, Cohen's D=1.55+-0.17).

Differences in performances were observed among phyla. Within bacteria, phyla such as Chloroflexi, Acidobacteriota, and Planctomycetota displayed a higher proportion of high predictive power models (see Figure 3 for GLMs, Supporting Information for other algorithms), with 528/1537, 1171/5163, and 576/2765, respectively. Within archaea, only some OTUs from Crenarchaeota had high predictive power models (20/131), while most phyla had few assigned OTUs and no high predictive power models. Few protists and fungi had good models, with Ascomycota and Mortierellomycota phyla having some high predictive power models (589/8631 and 89/1117, respectively).

2.Evaluation of 'relative-abundance' models

'Relative-abundance' models of all 67,979 preselected OTUs were successfully fitted by at least one algorithm, while 64,732 OTUs were fitted by four algorithms (details in Supporting Information). Spearman correlation, which evaluates the ability of the models to discriminate sites by their 'relative-abundance' values, demonstrated consistent results with the 'presence-absence' models results (Figure 2). We obtained numerous "useful" models (i.e. rho>0.2; e.g. for GLMs: 85% of bacteria, 83% or archaea, 63% of fungi, 49% of protists OTUs) and some "good" models (Spearman's rho>0.4; e.g. for GLMs: 40% of bacteria, 40% of archaea, 20% of fungi, 9% of protists OTUs). Bacteria and archaea models had higher predictive power than fungi and protist ones (Figure 2; Supporting Information). Some phyla had a higher proportion of 'relative-abundance' models that correctly ranked sites of OTUs (i.e. rho>0.4), such as for the bacterial phyla Acidobacteriota (1887/5163), Chloroflexi (489/1537) and Planctomycetota (756/2755; Figure 4). Among archaea, Crenarchaeota was the only phylum with good predictive models (28/131). Among fungi, Mortierellomycota had a higher proportion of good performing models (194/1117). For some phyla such as Nanoarchaeota and all protist phyla, the modelling pipeline could not produce 'relative-abundance' models with rho>0.4. Prediction of exact 'relative-abundance' values was less successful, because their coefficient of variation between predicted and validation values were high, with median prediction error between 10% and 100% of mean observed 'relative-abundance' among the 4 studied model algorithms (Supporting Information).

3. Covariate selection and importance

Edaphic covariates were the most selected across all groups for both 'presence-absence' and 'relative-abundance' models (Figure 5). For example in GLMs, at least one edaphic covariate was selected in models for 97% of archaea OTUs, 95% of bacteria OTUs, 87% of fungi OTUs and 80% of protist OTUs. In particular, pH was the most selected covariate in models for all groups (Figure 5; Supporting Information). Climatic covariates were also highly selected in GLMs, with 67% of archaea, 56% of bacteria, 67% of fungi and 69% of protist best models including at least one climatic covariate (Figure 5). For bacteria and archaea, winter temperature was the most selected climatic covariate, while fungi and protist models selected the set of covariates corresponding to yearly average temperature, precipitation, and elevation covariates (Supporting Information). Surprisingly, for fungi, distance to roads appeared within the list of the most-selected covariates alongside the edaphic and some topographic covariates. For protists, the most selected covariates

were found among distance to roads, climate, edaphic, topographic and land-use covariates (Figure 5).

Discussion

In this study, we estimated the ability of species distribution model (SDM) frameworks to predict the 'presence-absence' and 'relative-abundance' of 67,148 soil microbial OTUs (i.e. eDNA based operational taxa) based on associated environmental conditions along an elevation gradient. For most of these OTUs, prior knowledge on their ecology had been very sparse. Nevertheless, SDM frameworks allowed better 'presence-absence' prediction than null models for more than 85% of OTUs, and 23% had models that displayed a high predictive power. Our results confirm, in line with previous studies, that eDNA sequences can be used in models of microbial environmental niches (Schroder, 2008; King et al., 2010; Mod et al., 2020, 2021; Lembrechts et al., 2020; Malard et al., 2022). For 'relative-abundance' models, 33% of predictions had good rank correlation with on-site values. However, the prediction of the exact value of OTUs' 'relative-abundance' per site yielded large errors. A potential explanation could be the biases associated with relating proportion of reads to environment due, for instance, to intraspecific variations of the number of copies of the small ribosomal subunit, as observed for some microbial organisms (Stoddard et al., 2015; Lavrinienko et al., 2021), or even biases due to primers (Vaulot et al., 2022). This result is similar to SDM studies in macro-organisms that also showed low predictive power of abundance models (Pearce & Ferrier, 2001; Torres et al., 2012). We therefore highlight the difficulty in using SDM frameworks to predict abundance or other quantitative species characteristics (Van Couwenberghe et al., 2013; Lee-Yaw et al., 2022; Waldock et al., 2022). In regard to our results, we consider likely that a 'presence-absence' approach better depicts the situation *in situ* than a quantitative approach. However, a better and more fine-tuned quantitative approach than the one presented in our study may yield better results. For instance, zero-inflated models of semi-quantitative data such as the frameworks proposed in Guisan et. al. (1998) and in Irvine et al. (2016) for plant coverages, could be adapted to model ordinal classes of microbial OTUs abundance.

Model performances might also depend on how the environmental covariates that are used reflect the OTU ecological drivers (Guisan et al., 2017). We observed edaphic covariates as being the most selected covariates across all groups, emphasising the importance of soil properties in the spatial distribution of soil microorganisms, as previously reported (Birkhofer et al., 2012; de Vries et al., 2012; Terrat et al., 2017; Malard et al., 2022). Notably, our results continue bacteria and archaea to be highly dependent on soil pH (Yashiro et al., 2016, 2018; Malard et al., 2022). However, these performances were not consistent across phyla. For example, better performances were shown in Chloroflexi which are mostly heterotrophic phototrophs (Bryant, 2019), and Acidobacteriota, known to be strongly driven by pH, as well as other edaphic properties that were directly measured on site (Jones et al., 2009; Lauber et al., 2009; Navarrete et al., 2013). The strong relationship between organisms and the abiotic conditions that were directly measured at the field sites or from the collected soil samples (as opposed to covariates indirectly-derived from models or covariates not available like biotic interactions) may explain better performances obtained for these groups.

Conversely, poor model performances obtained for several OTUs could indicate a lack of environmental covariates relevant for these OTUs. Dependence on biotic interaction may be one of the main drivers of some microorganism distribution, further impacting model performances. For example, Patescibacetria, known for their strong dependence on biotic interactions with the surrounding community (Tian et al., 2007; Herrmann et al., 2019), had a low proportion of phyla with good modelling results compared to other phyla within bacteria. In contrast, Planctomycetota, which is also documented to contain many OTUs with highly dependent biotic interactions (Kabore et al., 2020), had a higher proportion of OTUs presenting good model performance. More work focussing on these two phyla is needed to find to which extent biotic interaction and environmental conditions determine their OTUs distributions.

Spatial and/or the temporal resolution of our covariate could also be irrelevant to accurately model OTU spatial patterns (Nunan et al., 2003). For example, landscape type and structure covariates were present in the initial covariate dataset, but at a very coarse resolution, compared to the size and generation time of microorganisms. The selection and importance of these covariates was low in all models, despite reports of microorganisms influenced by this kind of covariate (e.g. for protists Seppey et al., 2023). Moreover, it has

been shown in macro-organisms that information about micro-scale environmental conditions could improve model predictive power (Pradervand et al., 2014; Carter et al., 2016; Lembrechts et al., 2019, 2020).

Modelled organism characteristics can influence model performances (Guisan, Graham, et al., 2007; McCune et al., 2020; Collart et al., 2023). For instance, niche breadth is often reported as impacting the predictive power of SDMs, with species presenting large niches (i.e. generalists) tending to be harder to model than species presenting small ones (i.e. specialists) (Guisan & Hofer, 2003; Guisan, Zimmermann, et al., 2007; Marshall et al., 2015; Regos et al., 2019; Hallman & Robinson, 2020; Tessarolo et al., 2021). Malard et al. (2022) showed in the same study area that bacteria and protists have larger niche breadth than archaea and fungi. Our results combined with these findings tend to show that niche breadth may not be the main driving factor of microbial models' predictive power. Whether niche breadth has or not an impact on these microorganisms distribution still needs further investigation.

Another explanation for differences in model performances can be the number of sampling sites (Thuiller et al., 2004; Hernandez et al., 2006; Wisz et al., 2008; Tessarolo et al., 2021; Chevalier et al., 2022). However, while we showed that the number of sites has an impact on performance, this explanation cannot stand alone to explain model performance differences among taxa. Moreover, when we fitted models for bacteria and archaea with the exact same sites used for protists, the loss of predictive power was only marginal, and the archaea/bacteria models' predictive power remained much higher than protist models' predictive power. Moreover, the generated null distributions showed little effect of prevalence on the performance metrics of the null models, with only some effect for extreme prevalence values. To improve model performances for these OTUs with very low counts of presence or absences, ensembles of small models could be used instead (Breiner et al., 2015, 2018).

Species distribution modelling frameworks were first designed to model distributions of macro-organism species (Franklin, 2010; Peterson et al., 2011; Guisan et al., 2017). Using eDNA barcoding OTUs as modelled entities implies that organisms with different ecological requirements are potentially clustered into the same OTU, thereby potentially resulting in misleading predictions and poor modelling results (Qiao et al., 2017; Smith et al., 2019). However, with macro-organisms, other taxonomic levels than Linnaean species have been modelled successfully (Hadly et al., 2009; Smith et al., 2019). Moreover, in an exploratory study, Mod et al. (2021) tested the effect of different clustering levels of OTUs on their models' mean predictive power, and did not find any strong effect.

A usual application of species distribution models is the prediction of the modelled entity's presence outside of sampling locations and time (i.e. 'projections'; Guisan & Thuiller, 2005). Our results showed a large dependence of our soil-borne OTUs' presence-absence patterns on edaphic conditions in the soil. Consequently, projections in time and space of alpine soil-borne microorganisms would necessitate the development of edaphic maps and associated scenarios of change (Mod et al., 2021). Yet, mapping soil properties isn't an easy task, even under current conditions (Cianfrani et al., 2018). SoilGrids maps (Hengl et al., 2017) represent a possibility, but their resolution (250m) is currently not precise enough for local study areas, especially in rugged mountain landscapes as in the western Swiss Alps. Moreover, deriving future predictions with models including soil covariates will not be possible until scenarios of soil changes are also simultaneously developed (as can be found currently for climate and land-use; Mod et al., 2021). Yet, soil evolution under global change is still rather uncertain. While some studies predict an acidification of mountain soils through pollution (Hedl et al., 2011), others predict more mitigated responses of soil pH, carbon and nitrogen content (Davidson & Janssens, 2006; Trumbore & Czimczik, 2008; Rocci et al., 2021), with a lag between climatic changes and edaphic changes (Ladau et al., 2018; Mod et al., 2021). Taken together, to make full use of soil microorganism SDMs, we need to develop an ecologically relevant representation of covariates and their future scenarios.

To conclude, we showed that SDMs can be used to accurately predict presence-absence and relative abundance of microbial OTUs. Both 'presence-absence' and 'relative-abundance' approaches explore different aspects of microbial OTU distributions that can be helpful in ecological research on soil function and management. Particularly, relative abundance models as developed in this study could be used to discriminate

areas with higher prevalence of some OTUs from areas with lower relative abundance of these OTUs. However, we advise future authors to pay special attention to the coefficient of variation of such models before giving too much credit to the actual predicted value of relative abundance on sites. These models open the way to develop spatial maps to predict soil OTU compositional changes and their spatial distributions in future soil and landscape scenarios. In this context, we urge that fine-scale maps be generated, as well as future scenarios for soil edaphic covariates because of their importance as the main drivers of soil microbial OTU distribution.

List of figures:

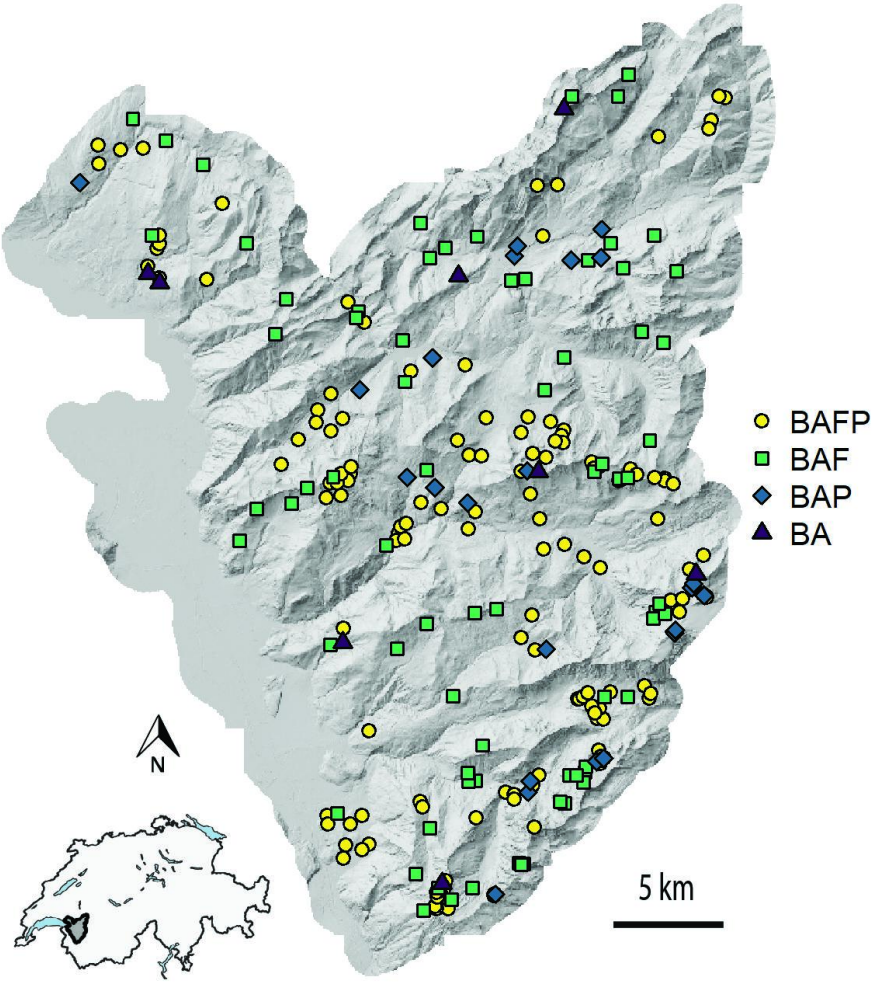

Figure 1. Distribution of the sampling sites in the Western Swiss Alps. DNA extraction was performed on samples from 250 sites, and amplification and sequencing were done on samples from all 250 sites for the 16S rRNA gene (Bacteria: B + Archaea: A), from 217 sites for ITS1 rRNA gene operon (Fungi: F), and from 166 sites for 18S rRNA gene (Protist: P). Sites sharing data from different microbial communities are referred to as a combination of the respective abbreviations.

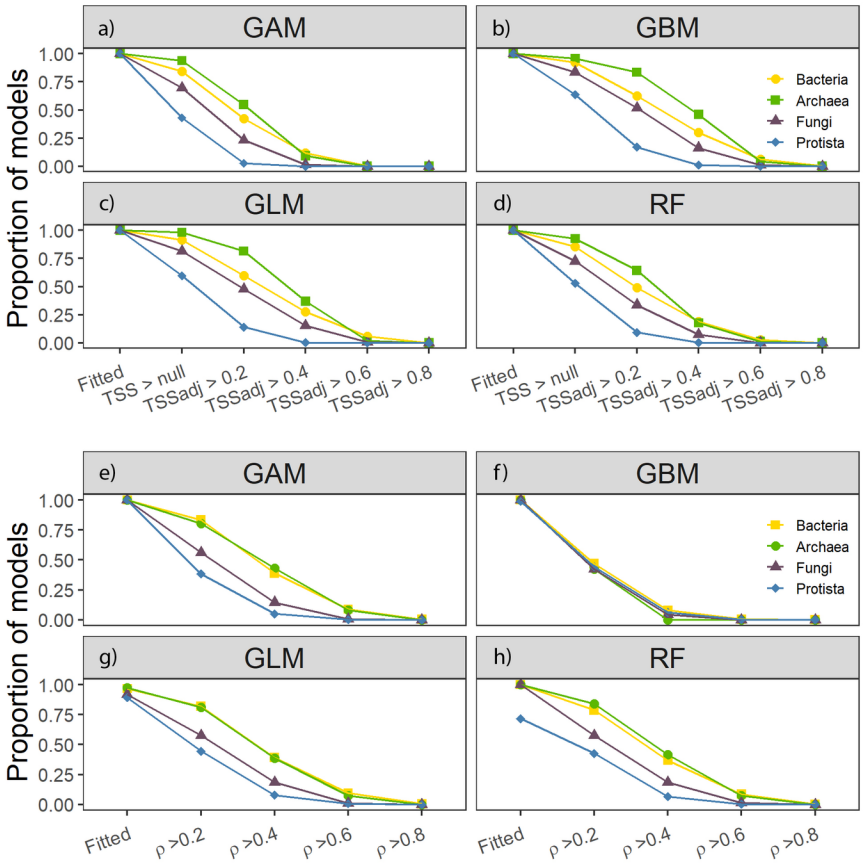

Figure 2. Predictive power obtained for Bacteria, Archaea, Fungi, and Protist and evaluated using the adjusted maxTSS for 'presence-absence' models (a,b,c,d), and Spearman's rho ([?]) for 'relative-abundance' models (e,f,g,h). For each threshold, the proportion of individual OTU models that obtained a greater value is shown. The x-axis labels are explained as follows: Fitted: the algorithm was able to fit a model; $TSS > TSS_{null}$: the predictive power of a model evaluated against null models; TSS_{adj}: predictive power metric rescaled so that $TSS_{adj}=0$ corresponds to a model with a predictive power equal to the 5% best null models. Bacteria and archaea models were fitted with 250 sites, while fungi were fitted with 217 and protists with 166 sites.

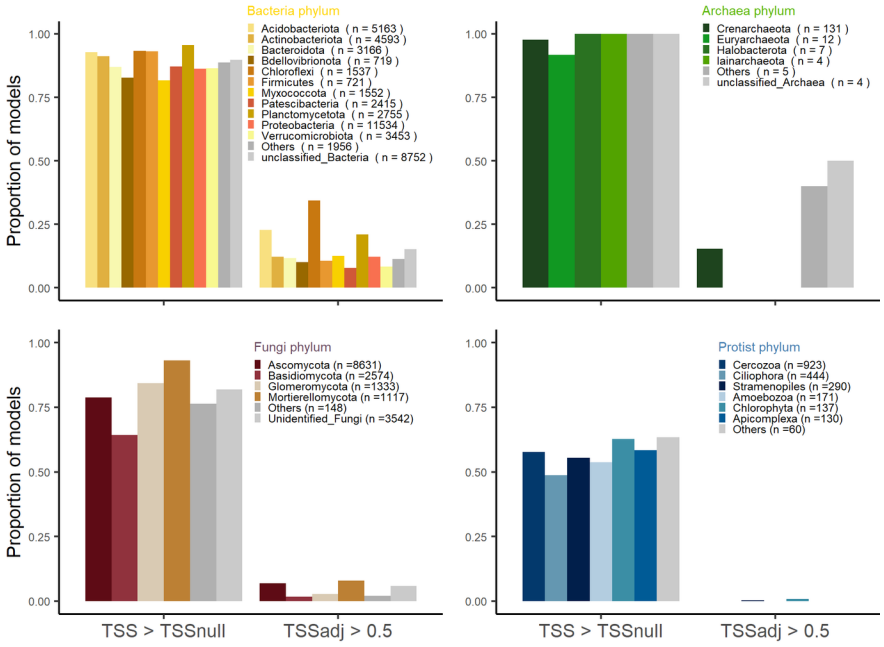

Figure 3. Performances of 'presence-absence' Generalised Linear Models across phyla. The proportion of the OTUs' models with better predictive power than null models ($TSS > TSS_{null}$), and the proportion of these models with good predictive capacities ($TSS_{adj} > 0.5$) are shown. Corresponding figures for other modelling algorithms are available in Supporting Information.

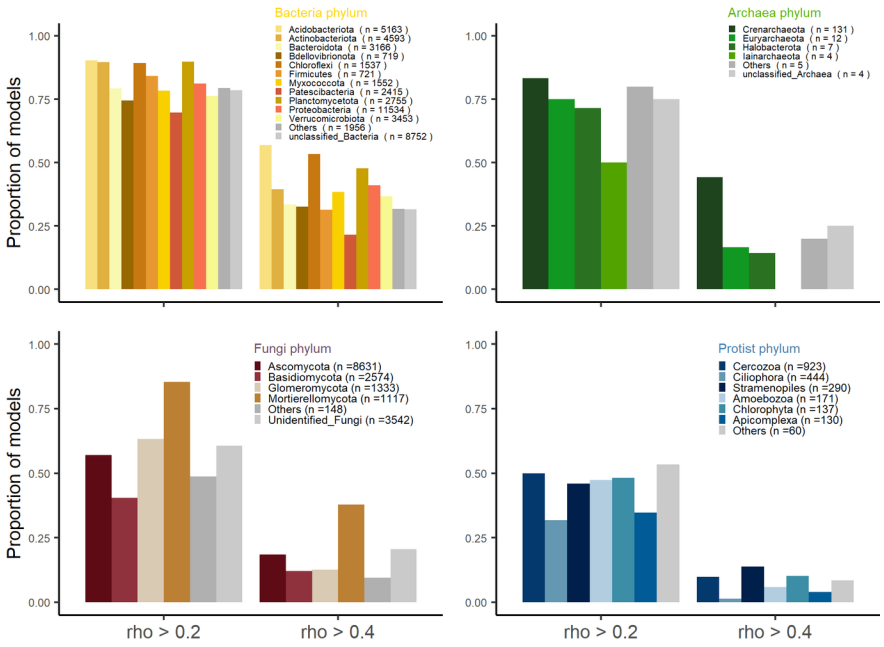

Figure 4. Performances of 'Relative-abundance' Generalised Linear Models across phyla. The proportion of OTUs with models having a spearman correlation coefficient (rho) between predictions and validation

data above 0.2 and 0.4 are shown. Corresponding figures for other modelling algorithms are available in Supporting Information.

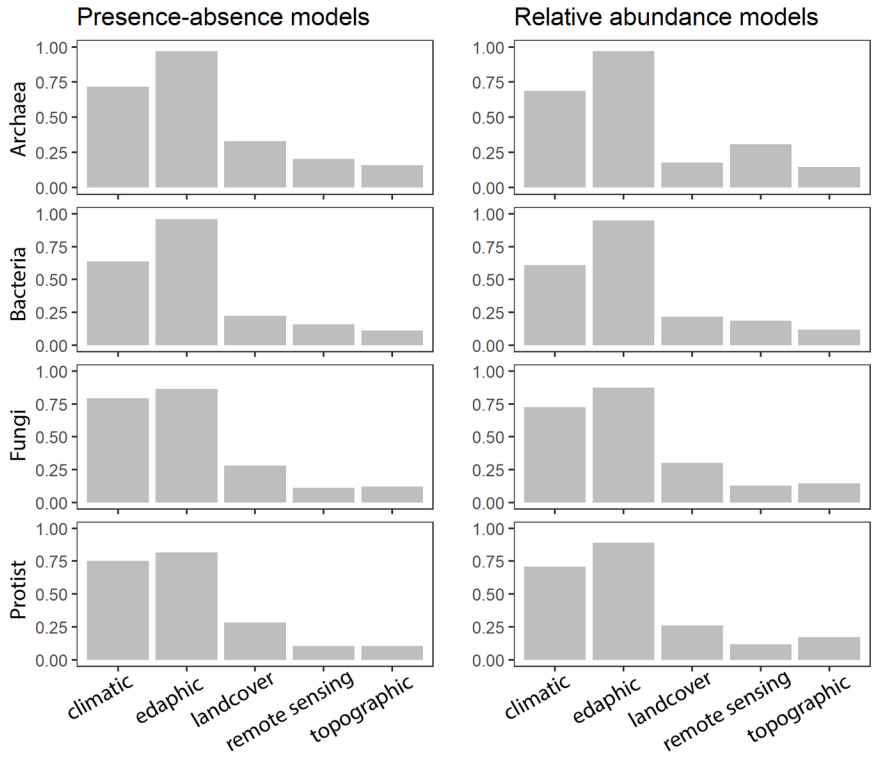

Figure 5. Proportion of GLMs selecting covariates within edaphic, climatic, land cover, remote-sensing and topographic groups of covariates for Bacteria, Archaea, Fungi, and Protist. Corresponding figures for other modelling algorithms are available in Supporting Information.

Data availability statement

The raw sequence data is available on NCBI bioproject number PRJNA810480 and PRJEB30010. All codes are available on github (to respect double blind reviewing, link will only be given upon acceptance). Modelling results for each individual OTU are available on Figshare:*https://figshare.com/s/825799db5d4fdc9a2f87*(private link, which will be published upon acceptance).

References

Abarenkov, K., Zirk, A., Piirmann, T., Pöhönen, R., Ivanov, F., Nilsson, R. H., & Kõljalg, U. (2022). *Full UNITE+INSD dataset for Fungi*[Application/gzip]. UNITE Community. https://doi.org/10.15156/BIO/2483925

Adde, A., Rey, P.-L., Fopp, F., Petitpierre, B., Schweiger, A. K., Broennimann, O., Lehmann, A., Zimmermann, N. E., Altermatt, F., Pellissier, L., & Guisan, A. (2023). Too many candidates: Embedded covariate selection procedure for species distribution modelling with the covsel R package.*Ecological Informatics*, *75*, 102080. https://doi.org/10.1016/j.ecoinf.2023.102080

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS).*Journal of Applied Ecology*, *43*(6), 1223–1232.

https://doi.org/10.1111/j.1365-2664.2006.01214.x

Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*,*5*(1), 1–12. https://doi.org/10.1126/sciadv.aat4858

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., Põlme, 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

Ballantyne, A., Smith, W., Anderegg, W., Kauppi, P., Sarmiento, J., Tans, P., Shevliakova, E., Pan, Y., Poulter, B., Anav, A., Friedlingstein, P., Houghton, R., & Running, S. (2017). Accelerating net terrestrial carbon uptake during the warming hiatus due to reduced respiration. *Nature Climate Change*, *7*(2), 148–152. https://doi.org/10.1038/nclimate3204

Baquero, F., Coque, T. M., Galán, J. C., & Martinez, J. L. (2021). The Origin of Niches and Species in the Bacterial World. *Frontiers in Microbiology*,*12*, 657986. https://doi.org/10.3389/fmicb.2021.657986

Bardgett, R. D., & Van Der Putten, W. H. (2014). Belowground biodiversity and ecosystem functioning. *Nature*, *515*(7528), 505–511. https://doi.org/10.1038/nature13855

Birkhofer, K., Schöning, I., Alt, F., Herold, N., Klarner, B., Maraun, M., Marhan, S., Oelmann, Y., Wubet, T., Yurkov, A., Begerow, D., Berner, D., Buscot, F., Daniel, R., Diekötter, T., Ehnes, R. B., Erdmann, G., Fischer, C., Foesel, B., ... Schrumpf, M. (2012). General relationships between abiotic soil properties and soil biota across spatial scales and different land-use types. *PLoS ONE*, *7*(8). https://doi.org/10.1371/journal.pone.0043292

Breiner, F. T., Guisan, A., Bergamini, A., & Nobis, M. P. (2015). Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution*, *6*(10), 1210–1218. https://doi.org/10.1111/2041-210X.12403

Breiner, F. T., Nobis, M. P., Bergamini, A., & Guisan, A. (2018). Optimizing ensembles of small models for predicting the distribution of species with few occurrences. *Methods in Ecology and Evolution*, *9*(4), 802–808. https://doi.org/10.1111/2041-210X.12957

Broennimann, O., & Guisan, A. (In review). CHclim25—A spatially and temporally very high resolution climatic dataset for Switzerland. 2023. *Under Review in Scientific Data*.

Bryant, D. A. (2019). Phototrophy and Phototrophs. In *Reference Module in Life Sciences*(p. B9780128096338207000). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20672-9

Buri, A., Grand, S., Yashiro, E., Adatte, T., Spangenberg, J. E., Pinto-Figueroa, E., Verrecchia, E., & Guisan, A. (2020). What are the most crucial soil variables for predicting the distribution of mountain plant species? A comprehensive study in the Swiss Alps. *Journal of Biogeography*,*47*(5), 1143–1153. https://doi.org/10.1111/jbi.13803

Carter, A., Kearney, M., Mitchell, N., Hartley, S., Porter, W., & Nelson, N. (2016). Modelling the soil microclimate: Does the spatial or temporal resolution of input parameters matter? *Frontiers of Biogeography*,*7*(4). https://doi.org/10.21425/F5FBG27849

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), 569–586. https://doi.org/10.1038/s41579-019-0222-5

Chevalier, M., Zarzo-Arias, A., Guélat, J., Mateo, R. G., & Guisan, A. (2022). Accounting for niche truncation to improve spatial and temporal predictions of species distributions. *Frontiers in Ecology and Evolution*, *10*, 944116. https://doi.org/10.3389/fevo.2022.944116

Cianfrani, C., Buri, A., Verrecchia, E., & Guisan, A. (2018). Generalizing soil properties in geographic space: Approaches used and ways forward. *PLoS ONE*,*13*(12), 1–17. https://doi.org/10.1371/journal.pone.0208823

Collart, F., Broennimann, O., Guisan, A., & Vanderpoorten, A. (2023). Ecological and biological indicators of the accuracy of species distribution models: Lessons from European bryophytes. *Ecography*, e06721. https://doi.org/10.1111/ecog.06721

Collart, F., & Guisan, A. (2023). Small to train, small to test: Dealing with low sample size in model evaluation. *Ecological Informatics*,*75*, 102106. https://doi.org/10.1016/j.ecoinf.2023.102106

Crowther, T. W., Todd-Brown, K. E. O., Rowe, C. W., Wieder, W. R., Carey, J. C., Machmuller, M. B., Snoek, B. L., Fang, S., Zhou, G., Allison, S. D., Blair, J. M., Bridgham, S. D., Burton, A. J., Carrillo, Y., Reich, P. B., Clark, J. S., Classen, A. T., Dijkstra, F. A., Elberling, B., ... Bradford, M. A. (2016). Quantifying global soil carbon losses in response to warming. *Nature*, *540*(7631), 104–108. https://doi.org/10.1038/nature20150

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY.*Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Davidson, E. A., & Janssens, I. A. (2006). Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature*,*440*(7081), 165–173. https://doi.org/10.1038/nature04514

de Vries, F. T., Manning, P., Tallowin, J. R. B., Mortimer, S. R., Pilgrim, E. S., Harrison, K. A., Hobbs, P. J., Quirk, H., Shipley, B., Cornelissen, J. H. C., Kattge, J., & Bardgett, R. D. (2012). Abiotic drivers and plant traits explain landscape-scale patterns in soil microbial communities.*Ecology Letters*, *15*(11), 1230–1239. https://doi.org/10.1111/j.1461-0248.2012.01844.x

Deiner, K., Walser, J.-C., Mächler, E., & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, *183*, 53–63. https://doi.org/10.1016/j.biocon.2014.11.018

Deng, H. (2013).*Guided Random Forest in the RRF Package*. https://doi.org/10.48550/ARXIV.1306.0237

Deng, H., & Runger, G. (2013). Gene selection with guided regularized random forest.*Pattern Recognition*, *46*(12), 3483–3489. https://doi.org/10.1016/j.patcog.2013.05.018

Descombes, P., Walthert, L., Baltensweiler, A., Meuli, R. G., Karger, D. N., Ginzler, C., Zurell, D., & Zimmermann, N. E. (2020). Spatial modelling of ecological indicator values improves predictions of plant distributions in complex landscapes. *Ecography*, *43*(10), 1448–1463. https://doi.org/10.1111/ecog.05117

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance.*Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dubuis, A., Giovanettina, S., Pellissier, L., Pottier, J., Vittoz, P., & Guisan, A. (2013). Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables.*Journal of Vegetation Science*, *24*(4), 593–606. https://doi.org/10.1111/jvs.12002

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.-P., & Guisan, A. (2011). Predicting spatial patterns of plant species richness: A comparison of direct macroecological and species stacking

modelling approaches: Predicting plant species richness. *Diversity and Distributions*,*17*(6), 1122–1131. https://doi.org/10.1111/j.1472-4642.2011.00792.x

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, *78*(382), 316–331. https://doi.org/10.1080/01621459.1983.10477973

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548–560. https://doi.org/10.1080/01621459.1997.10474007

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(3), 626–631. https://doi.org/10.1073/pnas.0507535103

Franklin, J. (2010).*Mapping Species Distributions: Spatial Inference and Prediction*(1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511810602

Galazzo, G., Van Best, N., Benedikter, B. J., Janssen, K., Bervoets, L., Driessen, C., Oomen, M., Lucchesi, M., Van Eijck, P. H., Becker, H. E. F., Hornef, M. W., Savelkoul, P. H., Stassen, F. R. M., Wolffs, P. F., & Penders, J. (2020). How to Count Our Microbes? The Effect of Different Quantitative Microbiome Profiling Approaches. *Frontiers in Cellular and Infection Microbiology*, *10*, 403. https://doi.org/10.3389/fcimb.2020.00403

Giner, C. R., Forn, I., Romac, S., Logares, R., De Vargas, C., & Massana, R. (2016). Environmental Sequencing Provides Reasonable Estimates of the Relative Abundance of Specific Picoeukaryotes. *Applied and Environmental Microbiology*, *82*(15), 4757–4766. https://doi.org/10.1128/AEM.00560-16

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional.*Frontiers in Microbiology*, *8*(NOV), 1–6. https://doi.org/10.3389/fmicb.2017.02224

Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers. (2022). *gbm: Generalized Boosted Regression Models* (2.1.8.1) [Computer software].

Griffiths, R. I., Thomson, B. C., Plassart, P., Gweon, H. S., Stone, D., Creamer, R. E., Lemanceau, P., & Bailey, M. J. (2016). Mapping and validating predictions of soil bacterial biodiversity using European and national scale datasets. *Applied Soil Ecology*, *97*, 61–68. https://doi.org/10.1016/j.apsoil.2015.06.018

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., ... Christen, R. (2012). The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604. https://doi.org/10.1093/nar/gks1160

Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene.*Ecological Modelling*, *157*(2–3), 89–100. https://doi.org/10.1016/S0304-3800(02)00204-1

Guisan, A., Graham, C. H., Elith, J., Huettmann, F., & the NCEAS Species Distribution Modelling Group. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*,*13*(3), 332–340. https://doi.org/10.1111/j.1472-4642.2007.00342.x

Guisan, A., & Hofer, U. (2003). Predicting reptile distributions at the mesoscale: Relation to climate and topography: Predicting reptile distributions at the mesoscale. *Journal of Biogeography*, *30*(8), 1233–1243. https://doi.org/10.1046/j.1365-2699.2003.00914.x

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., Theurillat, J.-P., & Kienast, F. (1998). Predicting the potential distribution of plant species in an alpine environment. *Journal of Vegetation Science*, *9*(1), 65–74. https://doi.org/10.2307/3237224

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat Suitability and Distribution Models: With Applications in R*. Cambridge University Press. https://doi.org/10.1017/9781139028271

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S., & Peterson, A. T. (2007). WHAT MATTERS FOR PREDICTING THE OCCURRENCES OF TREES: TECHNIQUES, DATA, OR SPECIES' CHARACTERISTICS? *Ecological Monographs*, *77*(4), 615–630. https://doi.org/10.1890/06-1060.1

Guo, F., Lenoir, J., & Bonebrake, T. C. (2018). Land-use change interacts with climate to determine elevational species redistribution. *Nature Communications*, *9*(1), 1–7. https://doi.org/10.1038/s41467-018-03786-9

Hadly, E. A., Spaeth, P. A., & Li, C. (2009). Niche conservatism above the species level.*Proceedings of the National Academy of Sciences of the United States of America*, *106*(SUPPL. 2), 19707–19714. https://doi.org/10.1073/pnas.0901648106

Hallman, T. A., & Robinson, W. D. (2020). Deciphering ecology from statistical artefacts: Competing influence of sample size, prevalence and habitat specialization on species distribution models and how small evaluation datasets can inflate metrics of performance. *Diversity and Distributions*, *26*(3), 315–328. https://doi.org/10.1111/ddi.13030

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hédl, R., Petřík, P., & Boublík, K. (2011). Long-term patterns in soil acidification due to pollution in forests of the Eastern Sudetes Mountains.*Environmental Pollution*, *159*(10), 2586–2593. https://doi.org/10.1016/j.envpol.2011.06.014

Hengl, T., Mendes De Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, *29*(5), 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

Herrmann, M., Wegner, C.-E., Taubert, M., Geesink, P., Lehmann, K., Yan, L., Lehmann, R., Totsche, K. U., & Küsel, K. (2019). Predominance of Cand. Patescibacteria in Groundwater Is Caused by Their Preferential Mobilization From Soils and Flourishing Under Oligotrophic Conditions.*Frontiers in Microbiology*, *10*, 1407. https://doi.org/10.3389/fmicb.2019.01407

Horrigue, W., Dequiedt, S., Prévost-bouré, N. C., Jolivet, C., Saby, N. P. A., Arrouays, D., Bispo, A., Maron, P., & Ranjard, L. (2016). Predictive model of soil molecular microbial biomass. *Ecological Indicators*,*64*, 203–211. https://doi.org/10.1016/j.ecolind.2015.12.004

Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, *22*(0), 415–427. https://doi.org/10.1101/SQB.1957.022.01.039

Irvine, K. M., Rodhouse, T. J., & Keren, I. N. (2016). Extending Ordinal Regression with a Latent Zero-Augmented Beta Distribution. *Journal of Agricultural, Biological and Environmental Statistics*, *21*(4), 619–640. https://doi.org/10.1007/s13253-016-0265-2

Jiao, S., Peng, Z., Qi, J., Gao, J., & Wei, G. (2021). Linking Bacterial-Fungal Relationships to Microbial Diversity and Soil Nutrient Cycling. *MSystems*, *6*(2), e01052-20. https://doi.org/10.1128/mSystems.01052-20

Jones, R. T., Robeson, M. S., Lauber, C. L., Hamady, M., Knight, R., & Fierer, N. (2009). A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *The ISME Journal*,*3*(4), 442–453. https://doi.org/10.1038/ismej.2008.127

Kaboré, O. D., Godreuil, S., & Drancourt, M. (2020). Planctomycetes as Host-Associated Bacteria: A Perspective That Holds Promise for Their Future Isolations, by Mimicking Their Native Environmental Niches in Clinical Microbiology Laboratories. *Frontiers in Cellular and Infection Microbiology*,*10*, 519301. https://doi.org/10.3389/fcimb.2020.519301

Karhu, K., Auffret, M. D., Dungait, J. A. J., Hopkins, D. W., Prosser, J. I., Singh, B. K., Subke, J.-A., Wookey, P. A., Ågren, G. I., Sebastià, M.-T., Gouriveau, F., Bergkvist, G., Meir, P., Nottingham, A. T., Salinas, N., & Hartley, I. P. (2014). Temperature sensitivity of soil respiration rates enhanced by microbial community response. *Nature*, *513*(7516), 81–84. https://doi.org/10.1038/nature13604

King, A. J., Freeman, K. R., McCormick, K. F., Lynch, R. C., Lozupone, C., Knight, R., & Schmidt, S. K. (2010). Biogeography and habitat modelling of high-alpine bacteria. *Nature Communications*, *1*(5). https://doi.org/10.1038/ncomms1055

Külling, N., Adde, A., & Fopp, F., Schweiger, A. K., Broennimann, O., Rey, P.L., Giuliani, G., Goicolea, T., Petitpierre, B., Zimmermann, N.E., Pellissier, L., Altermatt, F., Lehmann,A., and Guisan, A. (In review). SWECO25: A cross-thematic raster database for ecological research in Switzerland. *Under Review in Scientific Data*.

Ladau, J., Shi, Y., Jing, X., He, J.-S., Chen, L., Lin, X., Fierer, N., Gilbert, J. A., Pollard, K. S., & Chu, H. (2018). Existing Climate Change Will Lead to Pronounced Shifts in the Diversity of Soil Prokaryotes. *MSystems*,*3*(5). https://doi.org/10.1128/msystems.00167-18

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lauber, C. L., Hamady, M., Knight, R., & Fierer, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology*,*75*(15), 5111–5120. https://doi.org/10.1128/AEM.00335-09

Lavrinienko, A., Jernfors, T., Koskimäki, J. J., Pirttilä, A. M., & Watts, P. C. (2021). Does Intraspecific Variation in rDNA Copy Number Affect Analysis of Microbial Communities? *Trends in Microbiology*, *29*(1), 19–27. https://doi.org/10.1016/j.tim.2020.05.019

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Østerås, M., Schrenzel, J., & François, P. (2009). Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *Journal of Microbiological Methods*, *79*(3), 266–271. https://doi.org/10.1016/j.mimet.2009.09.012

Lee-Yaw, J., L. McCune, J., Pironon, S., & N. Sheth, S. (2022). Species distribution models rarely predict the biology of real populations. *Ecography*,*2022*(6), 1–16. https://doi.org/10.1111/ecog.05877

Lembrechts, J. J., Broeders, L., de Gruyter, J., Radujković, D., Ramirez-Rojas, I., Lenoir, J., & Verbruggen, E. (2020). A framework to bridge scales in distribution modeling of soil microbiota. *FEMS Microbiology Ecology*, *96*(5), 1–9. https://doi.org/10.1093/FEMSEC/FIAA051

Lembrechts, J. J., Nijs, I., & Lenoir, J. (2019). Incorporating microclimate into species distribution models. *Ecography*, *42*(7), 1267–1279. https://doi.org/10.1111/ecog.03947

Malard, L. A., Mod, H. K., Guex, N., Broennimann, O., Yashiro, E., Lara, E., Mitchell, E. A. D., Niculita-Hirzel, H., & Guisan, A. (2022). Comparative analysis of diversity and environmental niches of soil bacterial,

archaeal, fungal and protist communities reveal niche divergences along environmental gradients in the Alps. *Soil Biology and Biochemistry,169*(April), 108674. https://doi.org/10.1016/j.soilbio.2022.108674

Marshall, L., Carvalheiro, L. G., Aguirre-Gutiérrez, J., Bos, M., Groot, G. A., Kleijn, D., Potts, S. G., Reemer, M., Roberts, S., Scheper, J., & Biesmeijer, J. C. (2015). Testing projected wild bee distributions in agricultural habitats: Predictive power depends on species traits and habitat type. *Ecology and Evolution*, *5*(19), 4426–4436. https://doi.org/10.1002/ece3.1579

Mazel, F., Malard, L., Niculita-Hirzel, H., Yashiro, E., Mod, H. K., Mitchell, E. A. D., Singer, D., Buri, A., Pinto, E., Guex, N., Lara, E., & Guisan, A. (2021). Soil protist function varies with elevation in the Swiss Alps.*Environmental Microbiology*, *00*. https://doi.org/10.1111/1462-2920.15686

McCune, J. L., Rosner-Katz, H., Bennett, J. R., Schuster, R., & Kharouba, H. M. (2020). Do traits of plant species predict the efficacy of species distribution models for finding new occurrences? *Ecology and Evolution*, *10*(11), 5001–5014. https://doi.org/10.1002/ece3.6254

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*. https://doi.org/10.1038/s41396-021-00947-5

Mod, H. K., Scherrer, D., Di Cola, V., Broennimann, O., Blandenier, Q., Breiner, F. T., Buri, A., Goudet, J., Guex, N., Lara, E., Mitchell, E. A. D., Niculita-Hirzel, H., Pagni, M., Pellissier, L., Pinto-Figueroa, E., Sanders, I. R., Schmidt, B. R., Seppey, C. V. W., Singer, D., ... Guisan, A. (2020). Greater topoclimatic control of above- versus below-ground communities. *Global Change Biology*, *26*(12), 6715–6728. https://doi.org/10.1111/gcb.15330

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome,6*(1), 140. https://doi.org/10.1186/s40168-018-0521-5

Navarrete, A. A., Kuramae, E. E., De Hollander, M., Pijl, A. S., Van Veen, J. A., & Tsai, S. M. (2013). Acidobacterial community responses to agricultural management of soybean in Amazon forest soils. *FEMS Microbiology Ecology*, *83*(3), 607–621. https://doi.org/10.1111/1574-6941.12018

Nottingham, A. T., Baath, E., Reischke, S., Salinas, N., & Meir, P. (2019). Adaptation of soil microbial growth to temperature: Using a tropical elevation gradient to predict future changes. *Global Change Biology,25*(3), 827–838. https://doi.org/10.1111/gcb.14502

Nottingham, A. T., Whitaker, J., Turner, B. L., Salinas, N., Zimmermann, M., Malhi, Y., & Meir, P. (2015). Climate Warming and Soil Carbon in Tropical Forests: Insights from an Elevation Gradient in the Peruvian Andes.*BioScience*, *65*(9), 906–921. https://doi.org/10.1093/biosci/biv109

Nunan, N., Wu, K., Young, I. M., Crawford, J. W., & Ritz, K. (2003). Spatial distribution of bacterial communities and their relationships with the micro-architecture of soil. *FEMS Microbiology Ecology,44*(2), 203–215. https://doi.org/10.1016/S0168-6496(03)00027-8

Pearce, J., & Ferrier, S. (2001). The practical value of modelling relative abundance of species for regional conservation planning: A case study.*Biological Conservation*, *98*(1), 33–43. https://doi.org/10.1016/S0006-3207(00)00139-7

Peterson, A. T., Soberon, J., Pearson, R. G., Anderson, R. P., Martinez-Meyer, E., Nakamura, M., & Araujo, M. B. (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press. https://doi.org/10.1515/9781400840670

Philippot, L., Spor, A., Henault, C., Bru, D., Bizouard, F., Jones, C. M., Sarr, A., & Maron, P.-A. (2013). Loss in microbial diversity affects nitrogen cycling in soil. *The ISME Journal*, *7*(8), 1609–1619. https://doi.org/10.1038/ismej.2013.34

Pinto-figueroa, E. A., Seddon, E., Yashiro, E., Buri, A., Niculita-hirzel, H., Meer, J. R. V. D., & Guisan, A. (2019). *Archaeorhizomycetes Spatial Distribution in Soils Along Wide Elevational and Environmental Gradients Reveal Co-abundance Patterns With Other Fungal Saprobes and Potential Weathering Capacities*. *10*(April), 1–13. https://doi.org/10.3389/fmicb.2019.00656

Pradervand, J.-N., Dubuis, A., Pellissier, L., Guisan, A., & Randin, C. (2014). Very high resolution environmental predictors in species distribution models: Moving beyond topography? *Progress in Physical Geography: Earth and Environment*, *38*(1), 79–96. https://doi.org/10.1177/0309133313512667

Qiao, H., Peterson, A. T., Ji, L., & Hu, J. (2017). Using data from related species to overcome spatial sampling bias and associated limitations in ecological niche modelling. *Methods in Ecology and Evolution*, *8*(12), 1804–1812. https://doi.org/10.1111/2041-210X.12832

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glockner, 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

R Core Team. (2023).*R: A language and environment for statistical computing*. (4.3.0) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Regos, A., Gagne, L., Alcaraz-Segura, D., Honrado, J. P., & Dominguez, J. (2019). Effects of species traits and environmental predictors on performance and transferability of ecological niche models. *Scientific Reports*,*9*(1), 1–14. https://doi.org/10.1038/s41598-019-40766-5

Ren, B., Hu, Y., Chen, B., Zhang, Y., Thiele, J., Shi, R., Liu, M., & Bu, R. (2018). Soil pH and plant diversity shape soil bacterial community structure in the active layer across the latitudinal gradients in continuous permafrost region of Northeastern China. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-24040-8

Rocci, K. S., Lavallee, J. M., Stewart, C. E., & Cotrufo, M. F. (2021). Soil organic carbon response to global environmental change depends on its distribution between mineral-associated and particulate organic matter: A meta-analysis. *Science of The Total Environment*, *793*, 148569. https://doi.org/10.1016/j.scitotenv.2021.148569

Schmidt, P.-A., Balint, M., Greshake, B., Bandow, C., Rombke, J., & Schmitt, I. (2013). Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry*, *65*, 128–132. https://doi.org/10.1016/j.soilbio.2013.05.014

Schroder, B. (2008). Challenges of species distribution modeling belowground. *Journal of Plant Nutrition and Soil Science*, *171*(3), 325–337. https://doi.org/10.1002/jpln.200700027

Seppey, C. V. W., Broennimann, O., Buri, A., Singer, D., Mitchell, E. A. D., Hirzel, H. N., & Guisan, A. (2020). Soil protist diversity in the Swiss western Alps is better predicted by topo - climatic than by edaphic variables.*Journal of Biogeography*, *September 2019*, 866–878. https://doi.org/10.1111/jbi.13755

Seppey, C. V. W., Lara, E., Broennimann, O., Guisan, A., Malard, L., Singer, D., Yashiro, E., & Fournier, B. (2023). Landscape structure is a key driver of soil protist diversity in meadows in the Swiss Alps. *Landscape Ecology*, *38*(4), 949–965. https://doi.org/10.1007/s10980-022-01572-z

Serna-Chavez, H. M., Fierer, N., & Van Bodegom, P. M. (2013). Global drivers and patterns of microbial abundance in soil. *Global Ecology and Biogeography*,*22*(10), 1162–1172. https://doi.org/10.1111/geb.12070

Smith, A. B., Godsoe, W., Rodriguez-Sanchez, F., Wang, H. H., & Warren, D. (2019). Niche Estimation Above and Below the Species Level. *Trends in Ecology and Evolution*, *34*(3), 260–273. https://doi.org/10.1016/j.tree.2018.10.012

Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. K., & Schmidt, T. M. (2015). rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development.

Nucleic Acids Research, *43*(D1), D593–D598. https://doi.org/10.1093/nar/gku1201

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science,240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software*, *106*(1). https://doi.org/10.18637/jss.v106.i01

Terrat, S., Horrigue, W., Dequietd, S., Saby, N. P. A., Lelievre, M., Nowak, V., Tripied, J., Regnier, T., Jolivet, C., Arrouays, D., Wincker, P., Cruaud, C., Karimi, B., Bispo, A., Maron, P. A., Prevost-Boure, N. C., & Ranjard, L. (2017). Mapping and predictive variations of soil bacterial richness across France. *PLoS ONE*, *12*(10), 5–8. https://doi.org/10.1371/journal.pone.0186766

Tessarolo, G., Lobo, J. M., Rangel, T. F., & Hortal, J. (2021). High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecological Indicators*, *121*, 107147. https://doi.org/10.1016/j.ecolind.2020.107147

Thuiller, W., Brotons, L., Araujo, M. B., & Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, *27*(2), 165–172. https://doi.org/10.1111/j.0906-7590.2004.03673.x

Tian, L., Cai, T., Goetghebeur, E., & Wei, L. J. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error.*Biometrika*, *94*(2), 297–311. https://doi.org/10.1093/biomet/asm036

Torres, N. M., De Marco, P., Santos, T., Silveira, L., De Almeida Jacomo, A. T., & Diniz-Filho, J. A. F. (2012). Can species distribution modelling provide estimates of population densities? A case study with jaguars in the Neotropics: Distribution models and population density. *Diversity and Distributions*, *18*(6), 615–627. https://doi.org/10.1111/j.1472-4642.2012.00892.x

Trumbore, S. E., & Czimczik, C. I. (2008). An Uncertain Future for Soil Carbon.*Science*, *321*(5895), 1455–1456. https://doi.org/10.1126/science.1160232

Van Couwenberghe, R., Collet, C., Pierrat, J.-C., Verheyen, K., & Gegout, J.-C. (2013). Can species distribution models be used to describe plant abundance patterns? *Ecography*, *36*(6), 665–674. https://doi.org/10.1111/j.1600-0587.2012.07362.x

Vaulot, D., Geisen, S., Mahe, F., & Bass, D. (2022). pr2-primers: An 18S rRNA primer database for protists. *Molecular Ecology Resources*, *22*(1), 168–179. https://doi.org/10.1111/1755-0998.13465

Von Daniken, I., Guisan, A., & Lane, S. (2014). *RechAlp.vd: Une nouvelle plateforme UNIL de support pour la recherche transdisciplinaire dans les Alpes vaudoises* [Text/html,application/pdf,text/html]. https://doi.org/10.5169/SEALS-513645

Waldock, C., Stuart-Smith, R. D., Albouy, C., Cheung, W. W. L., Edgar, G. J., Mouillot, D., Tjiputra, J., & Pellissier, L. (2022). A quantitative review of abundance-based species distribution models. *Ecography,2022*(1), 1–18. https://doi.org/10.1111/ecog.05694

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & NCEAS Predicting Species Distributions Working Group+. (2008). Effects of sample size on the performance of species distribution models.*Diversity and Distributions*, *14*(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Wood, S. N. (2017).*Generalized additive models: An introduction with R* (Second edition). CRC Press/Taylor & Francis Group.

Yashiro, E., Pinto-figueroa, E., Buri, A., Spangenberg, J. E., & Adatte, T. (2016).*Local Environmental Factors Drive Divergent Grassland Soil Bacterial Communities in the Western Swiss Alps*. *82*(21), 6303–6316. https://doi.org/10.1128/AEM.01170-16.Editor

Yashiro, E., Pinto-figueroa, E., Buri, A., Spangenberg, J. E., Adatte, T., Niculita-hirzel, H., Guisan, A., & Meer, J. R. V. D. (2018). Meta-scale mountain grassland observatories uncover commonalities as well as specific interactions among plant and non-rhizosphere soil bacterial communities. *Scientific Reports*, *January*, 1–12. https://doi.org/10.1038/s41598-018-24253-x