ECOGRAPHY

Research article

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

Valentin Verdon ✉[1], Lucie Malard[1], Flavien Collart[1], Antoine Adde[2], Erika Yashiro[1,3], Enrique Lara Pandi[4], Heidi Mod[5], David Singer[6], Hélène Niculita-Hirzel[7], Nicolas Guex[8] and Antoine Guisan[1,2,9]

[1]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
[2]Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland
[3]Department of Fundamental Microbiology, University of Lausanne, Lausanne, Switzerland
[4]Real Jardín Botánico-CSIC, Madrid, Spain
[5]Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland
[6]Changins College for Viticulture and Enology, University of Sciences and Art Western Switzerland, Nyon, Switzerland
[7]Department of Occupational Health and Environment, Centre for Primary Care and Public Health, University of Lausanne, Lausanne, Switzerland
[8]Bioinformatics Competence Center, University of Lausanne, Lausanne, Switzerland
[9]Centre Interdisciplinaire de Recherche sur la Montagne, University of Lausanne, Lausanne, Switzerland

Correspondence: Valentin Verdon (valentin.verdon@unil.ch)

Soil microbes play a key role in shaping terrestrial ecosystems. It is therefore essential to understand what drives their distribution. While multivariate analyses have been used to characterise microbial communities and drivers of their spatial patterns, few studies have focused on predicting the distribution of amplicon sequence variants (ASVs). 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 ASVs in the western Swiss Alps. Advanced automated selection of abiotic covariates was used to circumvent the lack of knowledge on the ecology of each ASV. Presence–absence SDMs could be fitted for most ASVs, yielding better predictions than null models. Relative abundance SDMs performed less well, with low fit and predictive power overall, but displayed a good capacity to differentiate between sites with high and low relative abundance of the modelled ASV. SDMs for bacteria and archaea displayed better predictive power than for fungi and protists, suggesting a closer link of the former 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 ASVs obtained from topsoil DNA. It also highlights the need for further development of precise edaphic mapping and scenario modelling to enhances prediction of microorganism distributions in the future.

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

www.ecography.org

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 carbon and nutrient cycling (Philippot et al. 2013, Bardgett and Van Der Putten 2014, Jiao et al. 2021). For example, rising temperatures could enhance microbial activity, leading to increased carbon release from the 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 release depend on the composition and spatial distribution of the soil microbial communities present in the environment (Nottingham et al. 2015, 2019). Variations of these communities 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 (Birkhofer et al. 2012, Bahram et al. 2018) scales. 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 spatially characterise and quantify soil functions better, there is a need to improve knowledge of the distribution patterns of soil microbial communities and their components (Bardgett and Van Der Putten 2014, Mod et al. 2020). Ferrier and Guisan (2006) identified three different modelling approaches to predict community characteristics in space and time. The first is 'assemble first, predict later', where one computes summary metrics that characterise community-level properties, such as species richness or evenness, then model these properties against the environment. This approach is commonly used in microbial ecology, for example, to link soil microbial community characteristics, such as diversity metrics (Fierer and 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), proportion of functional groups (Mazel et al. 2021), or total biomass (Serna-Chavez et al. 2013, Horrigue et al. 2016) to environmental abiotic predictors, such as climatic and edaphic conditions. The second is 'assemble and predict together', which consists of an environmentally constrained ordination approach, sometimes used in microbial ecology (Pellissier et al. 2014, Hugerth and Andersson 2017, Yashiro et al. 2018). The third is 'predict first, assemble later'. It first models each component of the community individually against the environment using species distribution models (SDMs) frameworks (Franklin 2010, Peterson 2011, Guisan et al. 2017) and is, so far, rarely used in microbial ecology. Because each species tends to respond individually to environmental changes (Williams and Jackson 2007), this approach is more meaningful when the modelling goal is to predict future changes in species assemblages (Guisan and Rahbek 2011). SDMs were first developed to relate presence–absence of species to environmental conditions (Austin 1971, Guisan et al. 2017). When trained with presence–absence, SDMs predict the probabilities of occurrence, but

abundance models can also be developed (Guisan and Harrell 2000, Waldock et al. 2022).

In most microbial ecology studies, the base components of communities are clusters of sequenced gene reads grouped in operational taxonomic units according to a set similarity threshold, e.g. amplicon sequence variants (ASVs) cluster sequences reads at 100% similarity after a denoising step (Callahan et al. 2017). Relating geographically referenced observations of such ASVs to local environmental conditions allows quantifying their environmental niche (i.e. the ASV–environment relationships), from which the probability of presence of the ASVs can be predicted. Read counts per ASV are sometimes considered as an estimate of taxa abundance (Giner et al. 2016, Galazzo et al. 2020). However, the compositionality of sequencing data (Gloor et al. 2017, Greenacre 2021) also means that the direct modelling of absolute abundances (i.e. read counts) is not meaningful, leaving only the possibility to model relative abundance in addition to presence–absence (Mod et al. 2021). Such models could, in turn, provide predictive information in space and time about the status of microbial communities, with potential applications in microbial biodiversity conservation and land management (Averill et al. 2022, Redford 2023). To our knowledge, while some microbial studies fitted models of operational taxonomic unit level to answer ecological questions (Merges et al. 2018, Bay et al. 2020), very few studies have attempted to evaluate the quality of their models' predictions on independent data (as in Alzarhani et al. 2019, Mod et al. 2021). Hence, there is a need to develop SDM frameworks and test their predictions on microbial data to check: 1) if some taxa are better predicted than others, as observed for macro-organisms, 2) how presence–absence and relative abundance modelling compare, and 3) if predictive performances can be linked to ASV properties (Guo et al. 2015, Collart et al. 2023) such as taxonomic assignment and niche breadth, i.e. the amplitude of values where an ASV is observed in environmental space (Thuiller et al. 2004), which is often reported to impact the predictive power of SDMs (Guisan and Hofer 2003, Guisan et al. 2007a, Marshall et al. 2015, Regos et al. 2019, Hallman and Robinson 2020, Tessarolo et al. 2021). Indeed, as with macro-organism species, microbial taxa may have different responses to environmental factors, which may not be captured by community-level analyses (Ferrier and Guisan 2006, Mod et al. 2021). If we want to successfully model and anticipate changes in the composition of future soil communities from changes in environmental conditions (e.g. to inform management and conservation of soils at large scales), we need to tackle the challenge of modelling the distribution of individual ASVs (Schröder 2008).

In this study, we take advantage of recent advances in computing facilities, bioinformatics, and modelling frameworks to fit individual SDMs for every ASV from a comprehensive mountain soil DNA database (Yashiro et al. 2016, Pinto-Figueroa et al. 2019, Seppey et al. 2020, Mazel et al. 2021, Malard et al. 2022). Our study specifically aims to test the predictive power of the SDM approach applied to a large number of ASVs using both presence–absence and relative abundance

data and to compare their predictive performance by cross-validation. To achieve this, we first generated SDMs for more than 60 000 bacterial, archaeal, fungal, and protist ASVs 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 SDMs among four microbial target groups (bacteria, fungi, archaea, and protists) and their constitutive phyla.

Material and 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) in the western Swiss Alps (46°11′20″–46°32′38″N, 6°52′05″–7°14′54″E, http://rechalp.unil.ch (Von Däniken et al. 2014), Fig. 1). It is a mountainous region with an elevation range from 425 to 3120 m a.s.l. 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
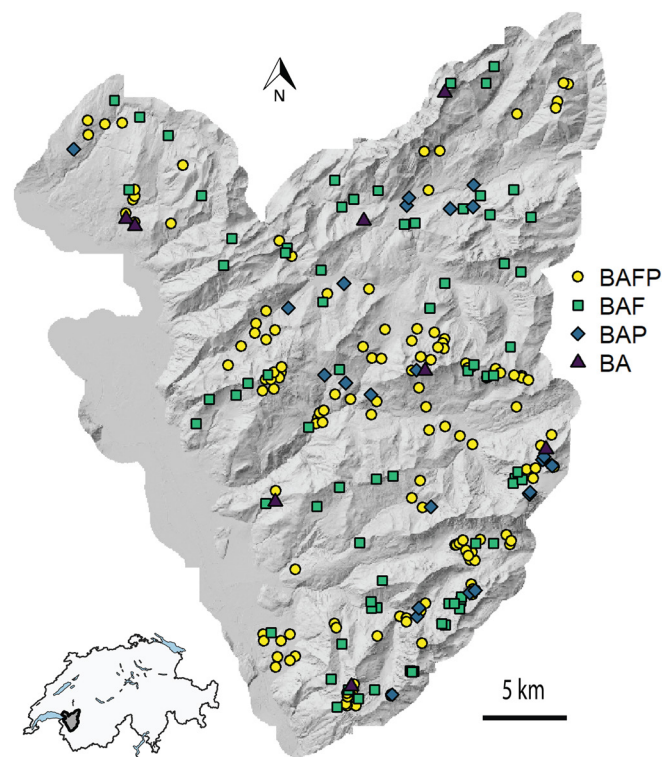

Figure 1. Distribution of the sampling sites in the western Swiss Alps. DNA extraction was performed on samples from 250 sites, and amplification and sequencing were done on samples from all 250 sites for the 16S rRNA gene (Bacteria: B + Archaea: A), from 217 sites for ITS1 rRNA gene operon (Fungi: F), 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; BAFD: all three markers were amplified; BAF: 16S and ITS1 were amplified; BAP: 16S and 18S were amplified; BA: 16S was amplified.

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, Mazel et al. 2021). Details on the sampling and DNA sequencing for the three respective groups can be found in the references above; information on assignment of sequenced reads to ASVs has been published in Malard et al. (2022). In brief, soil sampling was conducted from June to September (growing season) during the summers of 2012 and 2013. At each selected sampling site, a 2 × 2 m quadrat was used to sample the top 5 cm of soil at each corner and at the middle of the quadrat, using sterilised tools. The five subsamples were then pooled and homogenised into a sample of 500 g representing the site. DNA extraction was done within 36 h 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 (Supporting information). Demultiplexing, trimming, and merging of the sequences, as well as clustering of the sequences to obtain zero-radius ASVs (Edgar 2018), were performed using a custom-made pipeline (details in Mod et al. 2021, Malard et al. 2022).

Proportional abundances (hereafter relative abundance) were obtained by dividing each ASV read count by sequencing depth. The presence–absence data were obtained for each ASV using counts superior or equal to one as presence and lack of detection as absence. Taxonomic assignment of ASVs 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). After conversion to proportional abundance, ASVs 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 Yashiro et al. (2016) and Buri et al. (2020). For other covariates, data were extracted from spatial layers available at 25 m resolution (Broennimann unpubl. in Külling et al. 2024).

Modelling framework

Presence–absence modelling was performed for every ASV present in more than 5% and less than 95% of the sites, leading to 47 520, 163, 17 318, and 2147 ASVs for bacteria, archaea, fungi, and protists, respectively. Relative abundance modelling was performed for every ASV present in more than 5% of the sites, resulting in the modelling of 48 316 bacterial, 163 archaeal, 17 345 fungal, and 2155 protist ASVs.

This selection was applied to have enough data points for model fitting (see the Supporting information for data about ASVs removed).

The following framework was applied to each selected ASV in R ver. 4.3.0 (www.r-project.org). Covariates used in model fitting were selected by applying a two-step procedure (Adde et al. 2023). This procedure circumvents the issue regarding the lack of a priori knowledge about the ecology of most of the taxa belonging to these ASVs. Hence, the procedure treats all the selected ASVs equally relative to each other and optimises model predictive performances, after which variations observed among ASVs in the models can be associated with the underlying ecology.

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' goodness-of-fit, 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 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. We used two parametric methods: GLM (Guisan et al. 2002), generalised additive models (GAM; Guisan et al. 2002), and two machine learning methods, 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 ver. 4.1-7; Tay et al. 2023). GAMs used null-space penalization for covariate selection ('mgcv' package ver. 1.8-42; Wood 2017). For presence–absence models, parametric methods used a binomial probability distribution and a logit link function to model the probability of presence of each ASV (parameter 'family=binomial'). For relative abundance models, we used Poisson distributions with log link functions to model the ratio of the number of read per sequencing depth (parameter 'family=poisson'). For both presence–absence and relative abundance models, a regularised form of RF was built using the 'RRF' package ver. 1.9.4 (Deng and Runger 2013). For RF models, 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 10 000). GBM models were built with the 'gbm' package ver. 2.1.8.1 (Greenwell et al. 2022), in which a hyper-parametrization procedure was performed on the number of trees (10, 100, 1000, or 10 000 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 evaluation of models' predictive power

For each of the four algorithms' best models, the predictive power was assessed using the 'bootstrap .632+' cross-validation procedure (Efron and Tibshirani 1997) with 100 iterations per model. Unlike classical cross-validations used to evaluate SDMs, which 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) allowing to obtain better estimates of model error rates (Efron 1983, Efron and Tibshirani 1997). For each bootstrap iteration of the presence–absence models, the difference between predictions and validation data was computed using the area under the ROC curve (AUC; Swets 1988), and maximised values of the true skill statistics and of Kappa (Allouche et al. 2006), i.e. maxTSS and maxKappa (Guisan et al. 2017). The values were then averaged across the 100 iterations. The difference between predictions and training data (model fit) was computed using the same metrics and compared to the averaged predictive power of the model. Results were compared to those obtained by using null models fitted with randomised data (Collart and Guisan 2023); for each number of presence points (i.e. prevalence) existing in the dataset, 100 randomised models were fitted (i.e. 100 models fitting 'virtual randomly distributed ASVs'). These randomised data were used to rescale the maxTSS values for each of the real ASVs as follows:

$$maxTSS_{adj} = (maxTSS_{obs} - maxTSS_{null})/(1 - maxTSS_{null})$$

with $maxTSS_{adj}$ being the adjusted maxTSS for the considered ASV model; $maxTSS_{obs}$ being the raw maxTSS value obtained for that ASV model; and $maxTSS_{null}$ being the 95th percentile of the distribution of models fitted on random data with the same prevalence as the considered ASV. Models having a positive $maxTSS_{adj}$ were considered as having higher predictive power than expected by chance given the environmental dataset specificities. By applying the reasoning of Thuiller et al. (2004) for AUC to our metric, models with $maxTSS_{adj} > 0.5$ were considered as having a high predictive power.

The same procedure was applied to relative abundance models using Spearman correlation (rho) to check whether models were accurately ranking sites by their relative abundance values, and coefficient of variation (CoV) to assess the difference between the 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 ASVs in the dataset, as each ASV model would necessitate its own random distribution; in contrast, presence–absence models with the same prevalence could use the same random distribution. Hence, model qualities were classified as 'fair' when rho > 0.2 and as 'moderate' when rho > 0.4 (Landis and Koch 1977).

After computing the predictive power of all ASV models, the differences among the four main organism groups and within each groups' phyla were tested using ANOVA followed by Tukey tests with Bonferroni correction, and Cohen's D effect size metrics. Additionally, to assess whether modelling frameworks that specifically account for compositionality in relative abundance data would yield better predictions, models of read counts' centered-log ratios with zero replacements (Lubbe et al. 2021, Bastiaanssen et al. 2023)

were constructed and evaluated for 100 random bacteria, fungi, and protist ASVs (Supporting information).

Covariate selection and importance

To get an insight on which covariates drive the presence–absence and relative abundance of ASVs, an analysis was performed on each microorganism group. For each group and each environmental covariate, the proportion of models selecting that covariate was computed. 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 and Runger 2013, Greenwell et al. 2022). The values obtained 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

The goodness-of-fit metrics were consistently higher than the predictive power metrics (see the Supporting information).

Predictive power of presence–absence models

Out of the 67 148 ASVs identified in the dataset, at least one 'presence–absence' model algorithm could be fitted for all of the ASVs, and for 65 554 ASVs (98%), all four algorithms could be fitted (see the Supporting information). Overall, 91% of the bacteria, 98% of the archaea, 81% of the fungi, and 60% of the protists had higher predictive power than null models (Fig. 2a–d; Supporting information). However, the proportion of 'high predictive power' models (i.e. $maxTSS_{adj}$ > 0.5) was relatively low in all groups (Fig. 2a–d); e.g. for GLMs, 15% of the bacteria, 15% of the archaea, 6% of the fungi, and 0.1% of the protists presented a $maxTSS_{adj}$ > 0.5. The presence–absence models for bacteria and archaea had the best predictive power, followed by the fungi and protist models, across all four modelling algorithms (Fig. 2a–d).

Differences in performance were observed among phyla. Within bacteria, phyla such as Chloroflexi, Acidobacteriota, and Planctomycetota displayed a higher proportion of high predictive power models (Fig. 3 for GLMs; see the Supporting information for other algorithms), with 34% (528/1537), 23% (1171/5163), and 21% (576/2765), respectively. Within archaea, only some ASVs from Crenarchaeota had high predictive power models (20/131), while most phyla had only a few assigned ASVs and no high predictive power models. Few fungi and protists had good models, with Ascomycota and Mortierellomycota phyla having some high predictive power models (7%; 589/8631 and 8%; 89/1117, respectively).

Evaluation of relative abundance models' predictive power

The relative abundance models of all 67 979 preselected ASVs could be fitted by at least one algorithm, while 64 732

ASVs were fitted by all four algorithms (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 model results (Fig. 2). We obtained numerous 'fair quality' models (Spearman's rho > 0.2; e.g. for GLMs: 85% of the bacteria, 83% of the archaea, 63% of the fungi, and 49% of the protist ASVs) and some 'moderate quality' models (Spearman's rho > 0.4; e.g. for GLMs: 40% of the bacteria, 40% of the archaea, 20% of the fungi, and 9% of the protist ASVs). The bacteria and archaea models had higher predictive power than the fungi and protist models (Fig. 2, Supporting information). Some phyla had a higher proportion of relative abundance models with moderate and higher quality (i.e. rho > 0.4), such as for the bacterial phyla Acidobacteriota (1887/5163), Chloroflexi (489/1537), and Planctomycetota (756/2755; Fig. 4). Among archaea, Crenarchaeota was the only phylum with moderate and higher predictive models (28/131). Among fungi, Mortierellomycota had a higher proportion of moderately and higher 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. Their coefficients of variation between predicted relative abundance and validation values were high, with median prediction error between 10 and 100% of the mean observed relative abundance among the four studied model algorithms (Supporting information).

Covariate selection and importance

Edaphic covariates were the most selected across all groups for both the presence–absence and relative abundance models (Fig. 5). For example, in GLMs, at least one edaphic covariate was selected in models for 95% of the bacteria ASVs, 97% of the archaea ASVs, 87% of the fungi ASVs, and 80% of the protist ASVs. In particular, pH was the most selected covariate in the models for all groups (Fig. 5, Supporting information). Climatic covariates were also highly selected in GLMs, with 56% of the bacteria, 67% of the archaea, 67% of the fungi, and 69% of the protist best models including at least one climatic covariate (Fig. 5). For bacteria and archaea, winter temperature was the most selected climatic covariate, while the 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 (Supporting information). For protists, the most selected covariates were found among distance to roads, climate, edaphic, topographic and land-use covariates (Fig. 5).

Discussion

In this study, we estimated the ability of SDMs to predict the presence–absence and relative abundance of 67 148 soil
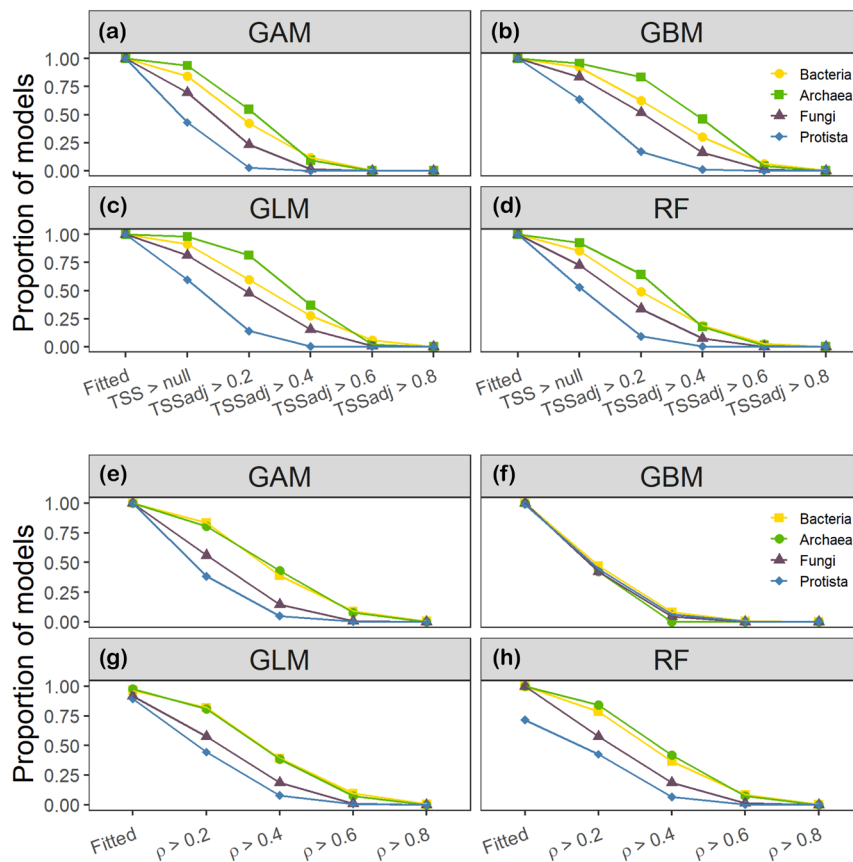
Figure 2. Predictive power obtained for bacteria, archaea, fungi, and protists and evaluated using the adjusted maxTSS for presence–absence models (a–d) and Spearman's rho (ρ) for relative abundance models (e–h). For each threshold, the proportion of individual amplicon sequence variant (ASV) models that obtained a greater value (x-axis) 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. ρ: Spearman's rho. Bacteria and archaea models were fitted with 250 sites, while fungi were fitted with 217 and protists with 166 sites. GAM: generalised additive models; GBM: gradient boosting machine; GLM: generalised linear models; RF: random forests.

microbial ASVs (i.e. soil DNA-based operational taxa) based on associated environmental conditions in the western Swiss Alps. For almost all of these ASVs, prior knowledge of their ecology had been very sparse. Nevertheless, our SDM framework allowed better presence–absence predictions than null models for more than 85% of the ASVs, and 23% had models that displayed a high predictive power. Our results confirm, in line with previous studies, that soil DNA sequences can be used in models of environmental niches of microbial taxa (Schröder 2008, King et al. 2010, Lembrechts et al. 2020, Mod et al. 2020, 2021, Malard et al. 2022). For relative abundance models, 33% had moderate and higher predictive success of on-site values ranking. However, the prediction of the exact value of ASVs' relative abundance per site yielded large errors. This result is consistent with SDM studies in macro-organisms that also showed lower predictive power in abundance models (Pearce and Ferrier 2001, Tôrres et al. 2012) and a limited correlation of presence–absence SDM predictions with observed abundances (Lee-Yaw et al. 2022). Potential explanations could be spatial processes at the population level (e.g. 'mass effects'; Kunin 1998) or biases associated with relating the proportion of reads to the environment, such as inadequacies in the handling of the compositional nature of relative abundances during model construction (Greenacre 2021), intraspecific variations in 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 primer biases (Vaulot et al. 2022). We therefore confirm here the difficulty in using SDM frameworks to predict abundance-based data (Van Couwenberghe et al. 2013, Lee-Yaw et al. 2022, Waldock et al. 2022). Regarding our results, and given the complexity of soil DNA data, it is likely that a presence–absence approach depicts the situation in situ better than a relative abundance approach.

Model performance might depend on how the environmental covariates that are used reflect the true causal ecological drivers of ASV distributions (Austin 2002, Mod et al. 2016, Guisan et al. 2017, Scherrer and Guisan 2019). As previously reported, we observed edaphic covariates as being the most selected covariates across all groups, thereby emphasising the importance of soil properties in the spatial distribution of soil microorganisms(Birkhofer et al. 2012,
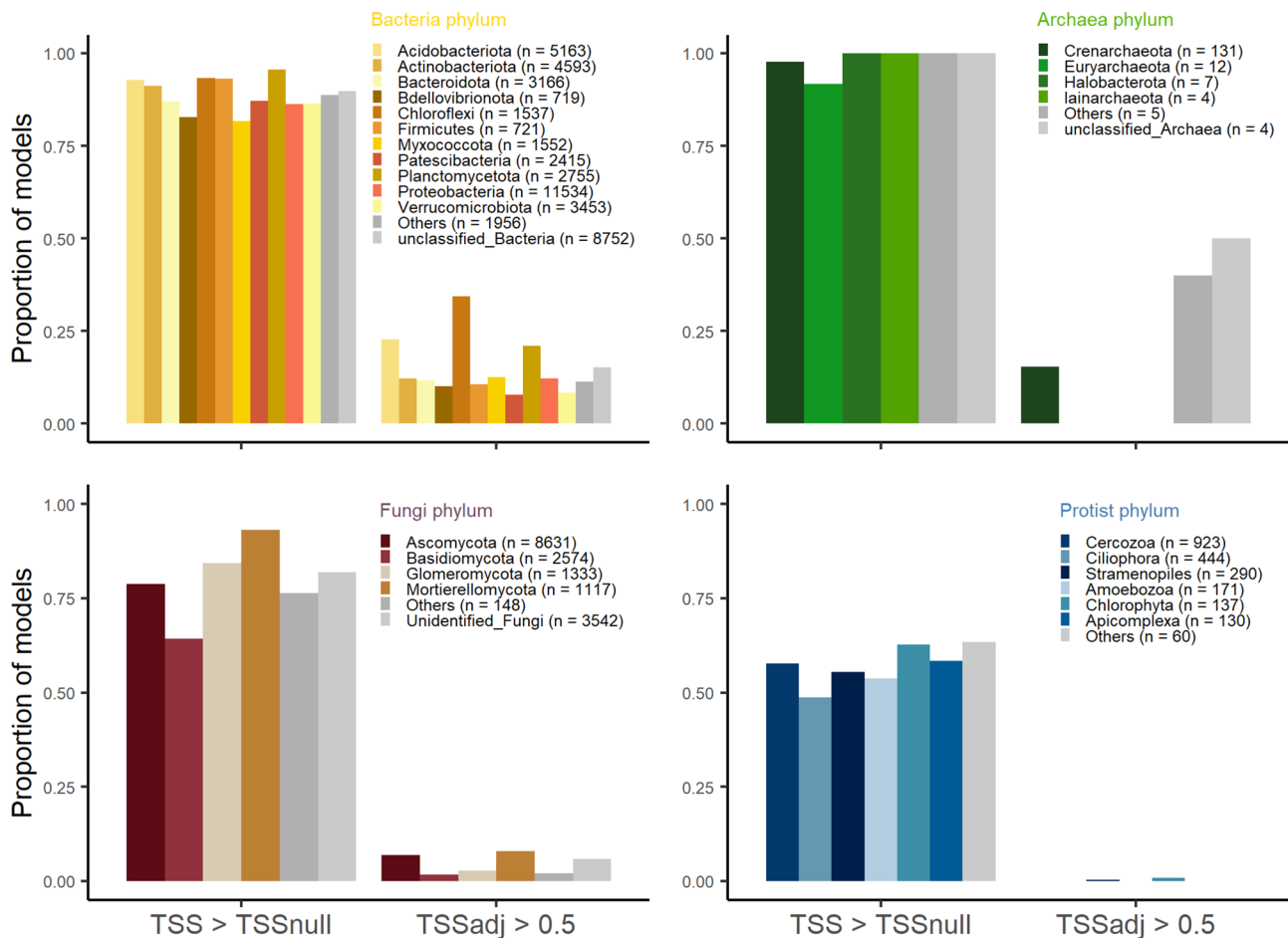
Figure 3. Performance of presence–absence generalised linear models across phyla. The proportion of the amplicon sequence variant (ASV) models with better predictive power than null models (TSS > TSS_{null}) and the proportion of these models with high predictive capacities (TSS_{adj} > 0.5) are shown. Corresponding figures for other modelling algorithms are available in the Supporting information.

de Vries et al. 2012, Terrat et al. 2017, Malard et al. 2022). Notably, our results confirm that bacteria and archaea are highly dependent on soil pH (Yashiro et al. 2016, 2018, Malard et al. 2022, Liang et al. 2023). However, their model performance was not consistent across phyla. For example, Chloroflexi, which are mostly heterotrophic phototrophs (Bryant 2019), and Acidobacteriota, known to be strongly driven by pH and other edaphic properties (Jones et al. 2009, Lauber et al. 2009, Navarrete et al. 2013), performed better than other phyla. 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 unavailable covariates such as biotic interactions, may explain the better performance obtained for these groups. The spatial and/or temporal resolution of available covariates could also lack relevance for modelling the spatial patterns of soil microbes (Nunan et al. 2003). Even for macro-organisms, information about micro-scale environmental conditions already improved the predictive power of models in several studies (Pradervand et al. 2014, Carter et al. 2016, Lembrechts et al. 2019, 2020). In soil microbial communities, because

important compositional changes can be observed at small spatial and temporal scales (Lauber et al. 2013, Degrune et al. 2017), model quality could in theory also be improved by including covariates measured at finer spatial and temporal scales. For example, covariates representing landscape type and structure were present in the initial dataset of covariates but only at a very coarse resolution, compared to the size and generation time of microorganisms. The selection and importance of these covariates were low in all models, despite reports of microorganisms being influenced by such landscape covariates (e.g. for protists, Seppey et al. 2023). Moreover, climatic covariates that summarise the state of the environment closer to the time of sampling than yearly averages might improve model quality (e.g. dynamics of edaphic conditions, Lipson et al. 1999). In addition, our modelling procedure did not include the possibility to consider interactions between covariates in parametric models. This could potentially be implemented by testing all potential interactions among the preselected covariates after the 'data snooping' step. Note, however, that tree-based approaches that automatically include covariate interactions (Guisan et al. 2006) did not yield better modelling results.
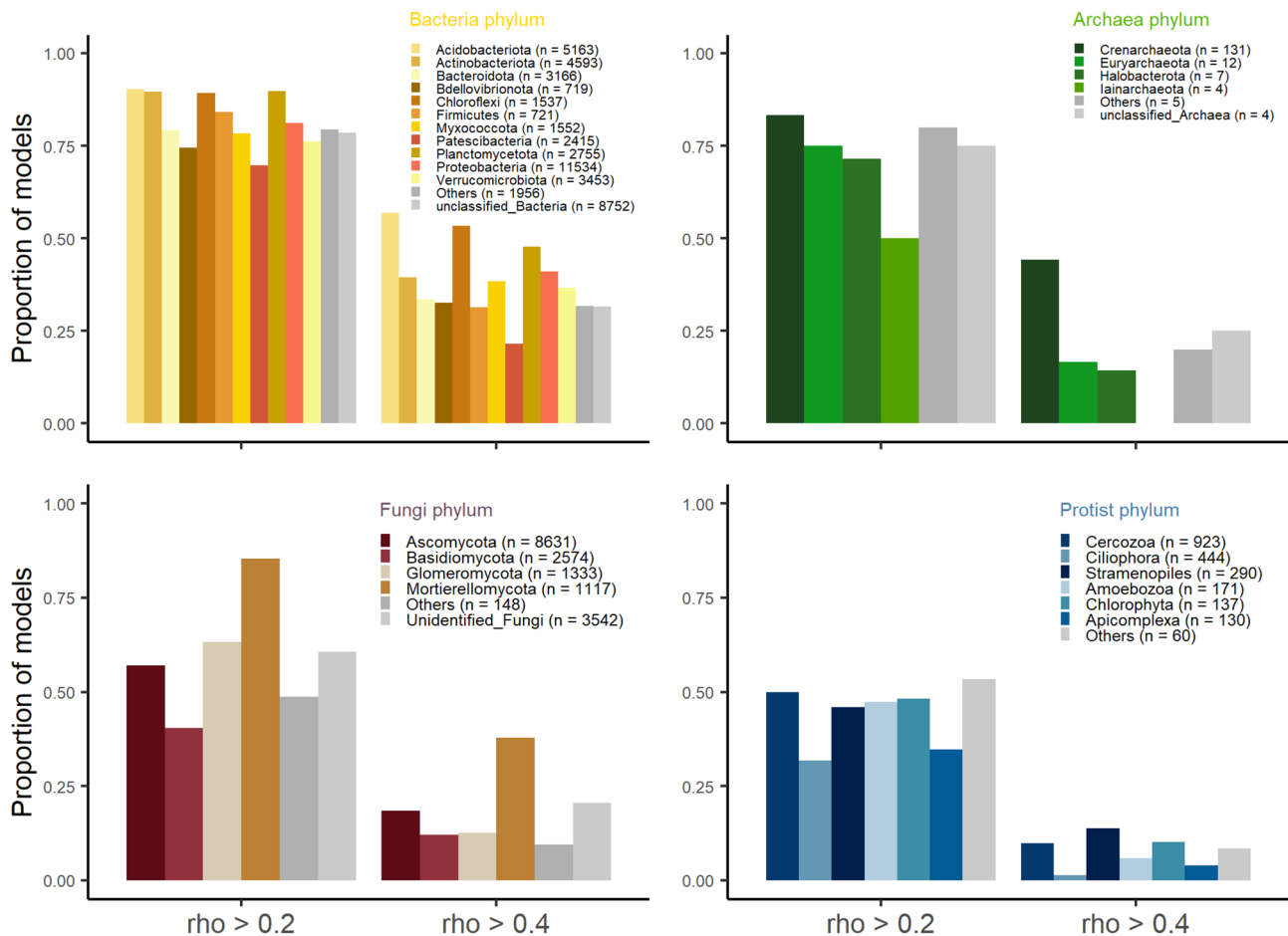
Figure 4. Performance of relative abundance generalised linear models across phyla. The proportion of amplicon sequence variants (ASVs) with models having a Spearman correlation coefficient (rho) between predictions and validation data above 0.2 and 0.4 are shown. See Supporting information for corresponding figures for other modelling algorithms.

Modelled organism characteristics can influence model performance (Guisan et al. 2007a, b, McCune et al. 2020, Collart et al. 2023). Species presenting large niches (i.e. generalists) tend to be harder to model than species presenting small ones (i.e. specialists; (Guisan and Hofer 2003, Regos et al. 2019, Hallman and Robinson 2020, Tessarolo et al. 2021). Our results tend to suggest that niche breadth for the most important covariates (e.g. pH) may be a factor driving the predictive power of microbial models.

Furthermore, the modelled distribution may be driven by biotic interactions that are not explicitly taken into account in our models, and potentially impacting the model's performance (Wisz et al. 2013). For example, the Patescibacteria superphylum had a low proportion of ASVs with high-performing models compared to other bacterial phyla, and taxa within this superphylum were reported to have potential associations with autotrophic taxa (Tian et al. 2007, Herrmann et al. 2019). These results suggest that Patescibacteria distribution is highly dependent on the resident bacterial community composition. In contrast, Planctomycetota, which is also documented to contain many ASVs with highly dependent biotic associations

(Kaboré et al. 2020), had a higher proportion of high-performing models. However, given that biotic interactions are implicitly accounted for in the observed distributions of taxa, their further inclusion in correlative SDMs does not necessarily improve the models (Dormann et al. 2018). Before using such interactions in predictive model frameworks, a formal theoretical setup would be needed to determine which biotic interactions are expected to contribute to the observed microbial distributions (Wisz et al. 2013, Dormann et al. 2018).

In our presence–absence models, any lack of detection was considered as absence, which can result in the introduction of potential biases in predictions (Benoit et al. 2018). The inclusion of sampling effort differences among sites (e.g. as in Botella et al. 2021) within modelling frameworks may improve the predictive power for some ASVs. Alternatively presence-only modelling frameworks accounting for limited detection (Dorazio 2014) could yield better predictions. For relative abundance models, improved and finer-tuned approaches, such as negative binomials that better account for overdispersion in data (Gardner et al. 1995), or centered-log ratio transformations (Aitchison 1982, Greenacre 2021) that accommodate for the compositionality of ASV data, may
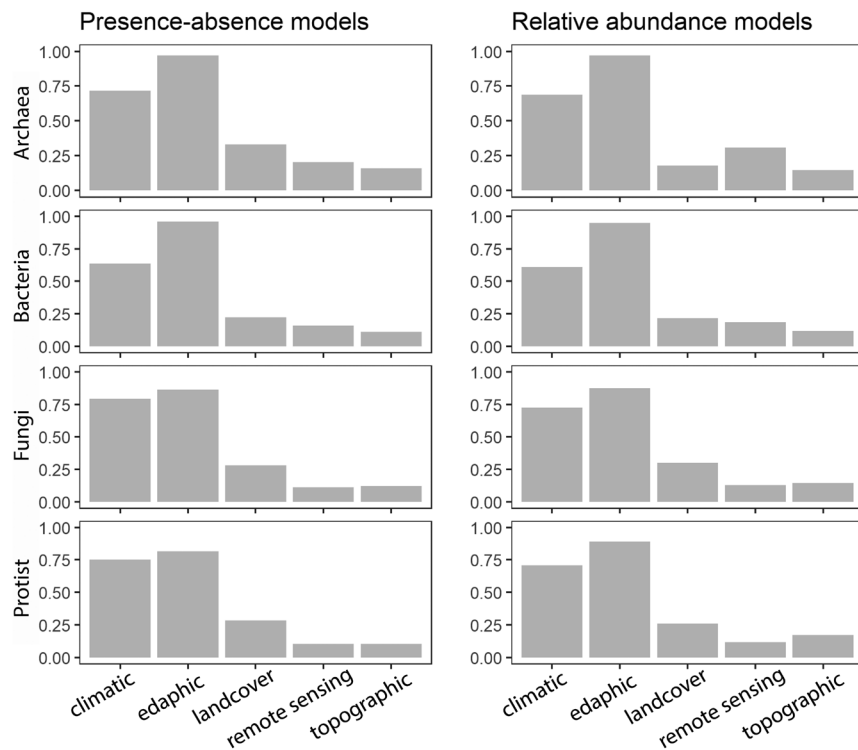
Figure 5. Proportion of generalised linear models selecting covariates within edaphic, climatic, land cover, remote-sensing, and topographic groups of covariates for bacteria, archaea, fungi, and protist. See Supporting information for corresponding figures for other modelling algorithms.

represent statistically more appropriate models. However, our preliminary tests using centered-log ratio on a randomly sampled subset of ASVs did not show improvement (Supporting information). Additionally, zero-inflated models of semi-quantitative data (Guisan and Harrell 2000), as proposed by Guisan et al. (1998) for abundance-dominance measures and Irvine et al. (2016) for plant cover, could be adapted to model ordinal classes of microbial ASVs that were transformed to centered-log ratios. Moreover, given the very large number of ASVs and the diversity in the distribution of their observation values, using a single pipeline on all the ASVs would imply the need to automate the fine tuning of each model to the statistical characteristics of each ASV, likely resulting in high computing time. Despite these constraints, increased accuracy could potentially be obtained by allowing models to test different statistical distributions and hyperparameters.

A common application of species distribution models is predicting the distribution of modelled entities outside of sampling locations and time (i.e. 'projections'; Guisan and Thuiller 2005). This kind of application could benefit microbial biogeography by allowing spatial predictions and future projections of community composition based on environmental conditions. Our results showed that the presence–absence patterns of our soil-borne ASVs were highly dependent on the edaphic conditions in the soil. Consequently, projections in time and space of mountain soil-borne microorganisms would necessitate the development of edaphic maps and associated scenarios of change (Mod et al. 2021). Yet, mapping

soil properties is not an easy task, even under current conditions (Cianfrani et al. 2018). SoilGrid maps (Hengl et al. 2017) represent a possibility, but their resolution (250 m) is currently not precise enough for local study areas, especially in rugged mountain landscapes as in the western Swiss Alps (Buri et al. 2020). Moreover, deriving future predictions with models that include soil covariates will not be possible until scenarios of soil changes are also concurrently developed (as can be currently found for climate and land-use). Yet, soil evolution under global change is still rather uncertain (Mod et al. 2021, Rumpf et al. unpubl.). While some studies predict an acidification of mountain soils due to pollution (Hédl et al. 2011), others predict more mitigated responses of soil pH and carbon and nitrogen content (Davidson and Janssens 2006, Trumbore and Czimczik 2008, Rocci et al. 2021), with a lag between climatic changes and edaphic changes (Ladau et al. 2018, Mod et al. 2021, Rumpf et al. unpubl.). 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 predict the presence–absence of many microbial ASVs and the relative abundance for a far more limited number. Both presence–absence and relative abundance approaches explore different aspects of the microbial ASV distribution patterns and can be helpful in ecological research on soil function and management. However, care should be given to measures of uncertainty in predictions, before giving too much credit to the

actual predicted values obtained from models, particularly for relative abundance. These models, at least presence–absence or presence-only ones, pave the way for the development of maps to predict the spatial distribution of soil ASVs in future soil and landscape scenarios. The value of these maps will lie in their ability to inform the public about microbial biodiversity conservation and land management. In this context, fine-scale maps of soil edaphic covariates, as well as future scenarios, should be generated, because of their importance as the main drivers of soil microbial ASV distribution.

Author contributions

Valentin Verdon: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Project administration (lead); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review and editing (equal). **Lucie Malard**: Data curation (equal); Project administration (supporting); Supervision (supporting); Visualization (supporting); Writing – original draft (supporting); Writing – review and editing (equal). **Flavien Collart**: Methodology (supporting); Project administration (supporting); Supervision (supporting); Visualization (supporting); Writing – original draft (supporting); Writing – review and editing (equal). **Antoine Adde**: Data curation (equal); Formal analysis (supporting); Methodology (supporting); Software (supporting); Writing – original draft (supporting); Writing – review and editing (equal). **Erika Yashiro**: Data curation (supporting); Methodology (supporting); Writing – review and editing (equal). **Enrique Lara Pandi**: Data curation (supporting); Methodology (supporting); Writing – review and editing (equal). **Heidi Mod**: Data curation (supporting); Formal analysis (supporting); Methodology (supporting); Writing – review and editing (equal). **David Singer**: Data curation (supporting); Writing – review and editing (equal). **Hélène Niculita-Hirzel**: Data curation (supporting); Writing – review and editing (equal). **Nicolas Guex:** Data curation (supporting); Formal analysis (supporting); Methodology (supporting); Software (supporting); Writing – review and editing (equal). **Antoine Guisan:** Conceptualization (supporting); Funding acquisition (lead); Methodology (supporting); Project administration (lead); Resources (lead); Supervision (lead); Visualization (supporting); Writing – original draft (supporting); Writing – review and editing (equal).

Transparent peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ecog.07086.

Data availability statement

The raw sequence data are available under NCBI bioproject no. PRJNA810480 and no. PRJEB30010. All codes are available from the Zenodo Digital Repository: https://doi.org/10.5281/zenodo.11034811. Raw data and modelling results for each individual ASV are available on Figshare: https://doi.org/10.6084/m9.figshare.23674758 (Verdon et al. 2024).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Abarenkov, K., Zirk, A., Piirmann, T., Pöhönen, R., Ivanov, F., Nilsson, R. H. and Kõljalg, U. 2022. Full UNITE+INSD dataset for fungi ver. 16.10.2022. – 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. and Guisan, A. 2023. Too many candidates: embedded covariate selection procedure for species distribution modelling with the covsel R package. – Ecol. Inform. 75: 102080.

Aitchison, J. 1982. The statistical analysis of compositional data. – J. R. Stat. Soc. B 44: 139–160, https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

Allouche, O., Tsoar, A. and Kadmon, R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

Alzarhani, A. K., Clark, D. R., Underwood, G. J. C., Ford, H., Cotton, T. E. A. and Dumbrell, A. J. 2019. Are drivers of root-associated fungal community structure context specific? – ISME J. 13: 1330–1344, https://doi.org/10.1038/s41396-019-0350-y.

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. and Rahbek, C. 2019. Standards for distribution models in biodiversity assessments. – Sci. Adv. 5: aat4858.

Austin, M. P. 1971. Role of regression analysis in plant ecology. – Proc. Ecol. Soc. Aust. 6: 63–75.

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Modell. 157: 101–118, https://doi.org/10.1016/S0304-3800(02)00205-3.

Averill, C., Anthony, M. A., Baldrian, P., Finkbeiner, F., Van Den Hoogen, J., Kiers, T., Kohout, P., Hirt, E., Smith, G. R. and Crowther, T. W. 2022. Defending earth's terrestrial microbiome. – Nat. Microbiol. 7: 1717–1725, https://doi.org/10.1038/s41564-022-01228-3.

Bahram, M. et al. 2018. Structure and function of the global topsoil microbiome. – Nature 560: 233–237.

Ballantyne, A., Smith, W., Anderegg, W., Kauppi, P., Sarmiento, J., Tans, P., Shevliakova, E., Pan, Y., Poulter, B., Anav, A., Friedlingstein, P., Houghton, R. and Running, S. 2017. Accelerating net terrestrial carbon uptake during the warming hiatus due to reduced respiration. – Nat. Clim. Change 7: 148–152.

Baquero, F., Coque, T. M., Galán, J. C. and Martinez, J. L. 2021. The origin of niches and species in the bacterial world. – Front. Microbiol. 12: 657986.

Bardgett, R. D. and Van Der Putten, W. H. 2014. Belowground biodiversity and ecosystem functioning. – Nature 515: 505–511.

Bastiaanssen, T. F. S., Quinn, T. P. and Loughman, A. 2023. Bugs as features (part 1): concepts and foundations for the compositional data analysis of the microbiome–gut–brain axis. – Nat. Mental Health 1: 930–938, https://doi.org/10.1038/s44220-023-00148-3.

Bay, S. K., McGeoch, M. A., Gillor, O., Wieler, N., Palmer, D. J., Baker, D. J., Chown, S. L. and Greening, C. 2020. Soil bacterial communities exhibit strong biogeographic patterns at fine taxonomic resolution. – mSystems 5: e00540–20, https://doi.org/10.1128/mSystems.00540-20.

Benoit, D., Jackson, D. A. and Ridgway, M. S. 2018. Assessing the impacts of imperfect detection on estimates of diversity and community structure through multispecies occupancy modeling. – Ecol. Evol. 8: 4676–4684, https://doi.org/10.1002/ece3.4023.

Birkhofer, K. et al. 2012. General relationships between abiotic soil properties and soil biota across spatial scales and different land-use types. – PLoS One 7: e43292.

Botella, C., Joly, A., Bonnet, P., Munoz, F. and Monestiez, P. 2021. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. – Methods Ecol. Evol. 12: 933–945, https://doi.org/10.1111/2041-210X.13565.

Breiner, F. T., Guisan, A., Bergamini, A. and Nobis, M. P. 2015. Overcoming limitations of modelling rare species by using ensembles of small models. – Methods Ecol. Evol. 6: 1210–1218.

Breiner, F. T., Nobis, M. P., Bergamini, A. and Guisan, A. 2018. Optimizing ensembles of small models for predicting the distribution of species with few occurrences. – Methods Ecol. Evol. 9: 802–808.

Broennimann, O. and Guisan, A. 2024. CHclim25 – a spatially and temporally very high resolution climatic dataset for Switzerland. – Earth Syst. Sci. Data Discuss: https://doi.org/10.5194/essd-2024-79, preprint.

Bryant, D. A. 2019. Phototrophy and phototrophs. – In: Schmidt, T. M. (ed.), Encyclopedia of microbiology, 4th edn. Academic Press, pp. 527–537.

Buri, A., Grand, S., Yashiro, E., Adatte, T., Spangenberg, J. E., Pinto-Figueroa, E., Verrecchia, E. and 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. – J. Biogeogr. 47: 1143–1153.

Callahan, B. J., McMurdie, P. J. and Holmes, S. P. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. – ISME J. 11: 2639–2643, https://doi.org/10.1038/ismej.2017.119.

Carter, A., Kearney, M., Mitchell, N., Hartley, S., Porter, W. and Nelson, N. 2016. Modelling the soil microclimate: does the spatial or temporal resolution of input parameters matter? – Front. Biogeogr. 7:fb_27849.

Cavicchioli, R. et al. 2019. Scientists' warning to humanity: microorganisms and climate change. – Nat. Rev. Microbiol. 17: 569–586.

Chevalier, M., Zarzo-Arias, A., Guélat, J., Mateo, R. G. and Guisan, A. 2022. Accounting for niche truncation to improve spatial and temporal predictions of species distributions. – Front. Ecol. Evol. 10: 944116.

Cianfrani, C., Buri, A., Verrecchia, E. and Guisan, A. 2018. Generalizing soil properties in geographic space: approaches used and ways forward. – PLoS One 13: e0208823.

Collart, F. and Guisan, A. 2023. Small to train, small to test: dealing with low sample size in model evaluation. – Ecol. Inform. 75: 102106.

Collart, F., Broennimann, O., Guisan, A. and Vanderpoorten, A. 2023. Ecological and biological indicators of the accuracy of species distribution models: lessons from European bryophytes. – Ecography 2023: e06721.

Crowther, T. W. et al. 2016. Quantifying global soil carbon losses in response to warming. – Nature 540: 104–108.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J. 2007. Random forests for classification in ecology. – Ecology 88: 2783–2792.

Davidson, E. A. and Janssens, I. A. 2006. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. – Nature 440: 165–173.

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. and Bardgett, R. D. 2012. Abiotic drivers and plant traits explain landscape-scale patterns in soil microbial communities. – Ecol. Lett. 15: 1230–1239.

Degrune, F., Theodorakopoulos, N., Colinet, G., Hiel, M.-P., Bodson, B., Taminiau, B., Daube, G., Vandenbol, M. and Hartmann, M. 2017. Temporal dynamics of soil microbial communities below the seedbed under two contrasting tillage regimes. – Front. Microbiol. 8: 1127, https://doi.org/10.3389/fmicb.2017.01127.

Deiner, K., Walser, J.-C., Mächler, E. and Altermatt, F. 2015. Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. – Biol. Conserv. 183: 53–63.

Deng, H. and Runger, G. 2013. Gene selection with guided regularized random forest. – Pattern Recognit. 46: 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. and Zimmermann, N. E. 2020. Spatial modelling of ecological indicator values improves predictions of plant distributions in complex landscapes. – Ecography 43: 1448–1463.

Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. – Global Ecol. Biogeogr. 23: 1472–1484, https://doi.org/10.1111/geb.12216.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitáo, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D. and Lautenbach, S. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. – Ecography 36: 27–46.

Dormann, C. F., Bobrowski, M., Dehling, D. M., Harris, D. J., Hartig, F., Lischke, H., Moretti, M. D., Pagel, J., Pinkert, S., Schleuning, M., Schmidt, S. I., Sheppard, C. S., Steinbauer, M. J., Zeuss, D. and Kraan, C. 2018. Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. – Global Ecol. Biogeogr. 27: 1004–1016, https://doi.org/10.1111/geb.12759.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.-P. and Guisan, A. 2011. Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches: predicting plant species richness. – Divers. Distrib. 17: 1122–1131.

Dubuis, A., Giovanettina, S., Pellissier, L., Pottier, J., Vittoz, P. and Guisan, A. 2013. Improving the prediction of plant species

distribution and community composition by adding edaphic to topo-climatic variables. – J. Veg. Sci. 24: 593–606.

Edgar, R. C. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. – Bioinformatics 34: 2371–2375. https://doi.org/10.1093/bioinformatics/bty113

Efron, B. 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. – J. Am. Stat. Assoc. 78: 316–331.

Efron, B. and Tibshirani, R. 1997. Improvements on cross-validation: the .632+ bootstrap method. – J. Am. Stat. Assoc. 92: 548–560.

Elith, J., Leathwick, J. R. and Hastie, T. 2008. A working guide to boosted regression trees. – J. Anim. Ecol. 77: 802–813.

Ferrier, S. and Guisan, A. 2006. Spatial modelling of biodiversity at the community level. – J. Appl. Ecol. 43: 393–404

Fierer, N. and Jackson, R. B. 2006. The diversity and biogeography of soil bacterial communities. – Proc. Natl Acad. Sci. USA 103: 626–631.

Franklin, J. 2010. Mapping species distributions: spatial inference and prediction, 1st edn. – Cambridge Univ. Press.

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. and Penders, J. 2020. How to count our microbes? The effect of different quantitative microbiome profiling approaches. – Front. Cell. Infect. Microbiol. 10: 403.

Gardner, W., Mulvey, E. P. and Shaw, E. C. 1995. Regression analysis of counts and rates: poission, overdispersed. – Psychol. Bull. 118: 392–404.

Giner, C. R., Forn, I., Romac, S., Logares, R., De Vargas, C. and Massana, R. 2016. Environmental sequencing provides reasonable estimates of the relative abundance of specific picoeukaryotes. – Appl. Environ. Microbiol. 82: 4757–4766.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. and Egozcue, J. J. 2017. Microbiome datasets are compositional: and this is not optional. – Front. Microbiol. 8: 2224.

Greenacre, M. 2021. Compositional data analysis. – Annu. Rev. Stat. Appl. 8: 271–299. https://doi.org/10.1146/annurev-statistics-042720-124436

Greenwell, B., Boehmke, B., Cunningham, J. and GBM Developers. 2022. gbm: generalized boosted regression models. – R package ver. 2.1.8.1, https://CRAN.R-project.org/package=gbm.

Griffiths, R. I., Thomson, B. C., Plassart, P., Gweon, H. S., Stone, D., Creamer, R. E., Lemanceau, P. and Bailey, M. J. 2016. Mapping and validating predictions of soil bacterial biodiversity using European and national scale datasets. – Appl. Soil Ecol. 97: 61–68.

Guillou, L. et al. 2012. The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. – Nucleic Acids Res. 41: D597–D604.

Guisan, A. and Harrell, F. E. 2000. Ordinal response regression models in ecology. – J.Veg. Sci. 11: 617–626. https://doi.org/10.2307/3236568

Guisan, A. and Hofer, U. 2003. Predicting reptile distributions at the mesoscale: relation to climate and topography: predicting reptile distributions at the mesoscale. – J. Biogeogr. 30: 1233–1243.

Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.

Guisan, A. and Rahbek, C. 2011. SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. – J. Biogeogr. 38: 1433–1444.

Guisan, A., Theurillat, J.-P. and Kienast, F. 1998. Predicting the potential distribution of plant species in an alpine environment. – J. Veg. Sci. 9: 65–74.

Guisan, A., Edwards, T. C. and Hastie, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. – Ecol. Modell. 157: 89–100.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J. M. C., Aspinall, R. and Hastie, T. 2006. Making better biogeographical predictions of species' distributions. – J. Appl. Ecol. 43: 386–392. https://doi.org/10.1111/j.1365-2664.2006.01164.x

Guisan, A., Graham, C. H., Elith, J., Huettmann, F. and the NCEAS Species Distribution Modelling Group. 2007a. Sensitivity of predictive species distribution models to change in grain size. – Divers. Distrib. 13: 332–340.

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S. and Peterson, A. T. 2007b. What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? – Ecol. Monogr. 77: 615–630.

Guisan, A., Thuiller, W. and Zimmermann, N. E. 2017. Habitat suitability and distribution models: with applications in R. – Cambridge Univ. Press.

Guo, C., Lek, S., Ye, S., Li, W., Liu, J. and Li, Z. 2015. Uncertainty in ensemble modelling of large-scale species distribution: effects from species characteristics and model techniques. – Ecol. Model. 306: 67–75. https://doi.org/10.1016/j.ecolmodel.2014.08.002

Guo, F., Lenoir, J. and Bonebrake, T. C. 2018. Land-use change interacts with climate to determine elevational species redistribution. – Nat. Commun. 9: 1–7.

Hadly, E. A., Spaeth, P. A. and Li, C. 2009. Niche conservatism above the species level. – Proc. Natl Acad. Sci. USA 106: 19707–19714.

Hallman, T. A. and 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. – Divers. Distrib. 26: 315–328.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. The elements of statistical learning. – Springer.

Hédl, R., Petřík, P. and Boublík, K. 2011. Long-term patterns in soil acidification due to pollution in forests of the eastern Sudetes Mountains. – Environ. Pollut. 159: 2586–2593.

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. and Kempen, B. 2017. SoilGrids250m: global gridded soil information based on machine learning. – PLoS One 12: e0169748.

Hernandez, P. A., Graham, C. H., Master, L. L. and Albert, D. L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773–785.

Herrmann, M., Wegner, C. E., Taubert, M., Geesink, P., Lehmann, K., Yan, L., Lehmann, R., Totsche, K. U. and Küsel, K. 2019. Predominance of Cand. Patescibacteria in groundwater is caused by their preferential mobilization from soils and flourishing under oligotrophic conditions. – Front. Microbiol. 10: 1407.

Horrigue, W., Dequiedt, S., Prévost-bouré, N. C., Jolivet, C., Saby, A., Arrouays, D., Bispo, A., Maron, P. and Ranjard, L. 2016. Predictive model of soil molecular microbial biomass. – Ecol. Indic. 64: 203–211.

Hugerth, L. W. and Andersson, A. F. 2017. Analysing microbial community composition through Amplicon sequencing:fFrom sampling to hypothesis Testing. – Front. Microbiol. 8: 1561. https://doi.org/10.3389/fmicb.2017.01561

Hutchinson, G. E. 1957. Concluding remarks. – Cold Spring Harb. Symp. Quant. Biol. 22: 415–427.

Irvine, K. M., Rodhouse, T. J. and Keren, I. N. 2016. Extending ordinal regression with a latent zero-augmented beta distribution. – J. Agric. Biol. Environ. Stat. 21: 619–640.

Jiao, S., Peng, Z., Qi, J., Gao, J. and Wei, G. 2021. Linking bacterial-fungal relationships to microbial diversity and soil nutrient cycling. – mSystems 6: e01052-20.

Jones, R. T., Robeson, M. S., Lauber, C. L., Hamady, M., Knight, R. and Fierer, N. 2009. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. – ISME J. 3: 442–453.

Kaboré, O. D., Godreuil, S. and 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. – Front. Cell. Infect. Microbiol. 10: 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. and Hartley, I. P. 2014. Temperature sensitivity of soil respiration rates enhanced by microbial community response. – Nature 513: 81–84.

King, A. J., Freeman, K. R., McCormick, K. F., Lynch, R. C., Lozupone, C., Knight, R. and Schmidt, S. K. 2010. Biogeography and habitat modelling of high-alpine bacteria. – Nat. Commun. 1: 53.

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. 2024. SWECO25: a cross-thematic raster database for ecological research in Switzerland. – Sci. Data 11: 21.

Kunin, W. E. 1998. Biodiversity at the edge: a test of the importance of spatial "mass effects" in the Rothamsted Park Grass experiments. – Proc. Natl Acad. Sci. USA 95: 207–212. https://doi.org/10.1073/pnas.95.1.207

Ladau, J., Shi, Y., Jing, X., He, J. S., Chen, L., Lin, X., Fierer, N., Gilbert, J. A., Pollard, K. S. and Chu, H. 2018. Existing climate change will lead to pronounced shifts in the diversity of soil prokaryotes. – mSystems 3: e00167-18

Landis, J. R. and Koch, G. G. 1977. The measurement of observer agreement for categorical data. – Biometrics 33: 159–174.

Lauber, C. L., Hamady, M., Knight, R. and Fierer, N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. – Appl. Environ. Microbiol. 75: 5111–5120.

Lauber, C. L., Ramirez, K. S., Aanderud, Z., Lennon, J. and Fierer, N. 2013. Temporal variability in soil microbial communities across land-use types. – ISME J. 7: 1641–1650. https://doi.org/10.1038/ismej.2013.50

Lavrinienko, A., Jernfors, T., Koskimäki, J. J., Pirttilä, A. M. and Watts, P. C. 2021. Does intraspecific variation in rDNA copy number affect analysis of microbial communities? – Trends Microbiol. 29: 19–27.

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Østerås, M., Schrenzel, J. and François, P. 2009. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. – J. Microbiol. Methods 79: 266–271.

Lee-Yaw, J., McCune, L., Pironon, S. and Sheth, S. N. 2022. Species distribution models rarely predict the biology of real populations. – Ecography 2022: e05877.

Lembrechts, J. J., Nijs, I. and Lenoir, J. 2019. Incorporating microclimate into species distribution models. – Ecography 42: 1267–1279.

Lembrechts, J. J., Broeders, L., de Gruyter, J., Radujković, D., Ramirez-Rojas, I., Lenoir, J. and Verbruggen, E. 2020. A framework to bridge scales in distribution modeling of soil microbiota. – FEMS Microbiol. Ecol. 96: fiaa051.

Liang, Q., Mod, H. K., Luo, S., Ma, B., Yang, K., Chen, B., Qi, W., Zhao, Z., Du, G., Guisan, A., Ma, X. and Le Roux, X. 2023. Taxonomic and functional biogeographies of soil bacterial communities across the Tibet plateau are better explained by abiotic conditions than distance and plant community composition. – Mol. Ecol. 32: 3747–3762. https://doi.org/10.1111/mec.16952

Lipson, D. A., Schmidt, S. K. and Monson, R. K. 1999. Links between microbial population dynamics and nitrogen availability in an alpine ecosystem. – Ecology 80: 1623–1631. https://doi.org/10.1890/0012-9658(1999)080[1623:LBMPDA]2.0.CO;2

Lubbe, S., Filzmoser, P. and Templ, M. 2021. Comparison of zero replacement strategies for compositional data with large numbers of zeros. – Chemom.Intell. Lab. Syst. 210: 104248. https://doi.org/10.1016/j.chemolab.2021.104248

Malard, L. A., Mod, H. K., Guex, N., Broennimann, O., Yashiro, E., Lara, E., Mitchell, E. A. D., Niculita-Hirzel, H. and 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 Biol. Biochem. 169: 108674.

Marshall, L., Carvalheiro, L. G., Aguirre-Gutiérrez, J., Bos, M., de Groot, G. A., Kleijn, D., Potts, S. G., Reemer, M., Roberts, S., Scheper, J. and Biesmeijer, J. C. 2015. Testing projected wild bee distributions in agricultural habitats: predictive power depends on species traits and habitat type. – Ecol. Evol. 5: 4426–4436.

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. and Guisan, A. 2021. Soil protist function varies with elevation in the Swiss Alps. – Environ. Microbiol. 24: 1689–1702

McCune, J. L., Rosner-Katz, H., Bennett, J. R., Schuster, R. and Kharouba, H. M. 2020. Do traits of plant species predict the efficacy of species distribution models for finding new occurrences? – Ecol. Evol. 10: 5001–5014.

Merges, D., Bálint, M., Schmitt, I., Böhning-Gaese, K. and Neuschulz, E. L. 2018. Spatial patterns of pathogenic and mutualistic fungi across the elevational range of a host plant. – J. Ecol. 106: 1545–1557. https://doi.org/10.1111/1365-2745.12942

Mod, H. K. et al. 2020. Greater topoclimatic control of above- versus below-ground communities. – Global Change Biol. 26: 6715–6728.

Mod, H. K., Buri, A., Yashiro, E., Guex, N., Malard, L., Pinto-Figueroa, E., Pagni, M., Niculita-Hirzel, H., van der Meer, J. R. and Guisan, A. 2021. Predicting spatial patterns of soil bacteria under current and future environmental conditions. – ISME J. 15: 2547–2560.

Mod, H. K., Scherrer, D., Luoto, M. and Guisan, A. 2016. What we use is not what we know: environmental predictors in plant

distribution models. – J. Veg. Sci. 27: 1308–1322. https://doi.org/10.1111/jvs.12444

Murali, A., Bhargava, A. and Wright, E. S. 2018. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. – Microbiome 6: 140.

Navarrete, A. A., Kuramae, E. E., De Hollander, M., Pijl, A. S., Van Veen, J. A. and Tsai, S. M. 2013. Acidobacterial community responses to agricultural management of soybean in Amazon forest soils. – FEMS Microbiol. Ecol. 83: 607–621.

Nottingham, A. T., Whitaker, J., Turner, B. L., Salinas, N., Zimmermann, M., Malhi, Y. and Meir, P. 2015. Climate warming and soil carbon in tropical forests: insights from an elevation gradient in the Peruvian Andes. – BioScience 65: 906–921.

Nottingham, A. T., Bååth, E., Reischke, S., Salinas, N. and Meir, P. 2019. Adaptation of soil microbial growth to temperature: using a tropical elevation gradient to predict future changes. – Global Change Biol. 25: 827–838.

Nunan, N., Wu, K., Young, I. M., Crawford, J. W. and Ritz, K. 2003. Spatial distribution of bacterial communities and their relationships with the micro-architecture of soil. – FEMS Microbiol. Ecol. 44: 203–215.

Pearce, J. and Ferrier, S. 2001. The practical value of modelling relative abundance of species for regional conservation planning: a case study. – Biol. Conserv. 98: 33–43.

Pellissier, L., Niculita-Hirzel, H., Dubuis, A., Pagni, M., Guex, N., Ndiribe, C., Salamin, N., Xenarios, I., Goudet, J., Sanders, I. R. and Guisan, A. 2014. Soil fungal communities of grasslands are environmentally structured at a regional scale in the Alps. – Mol. Ecol. 23: 4274–4290. https://doi.org/10.1111/mec.12854

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M. and Araújo, M. B. 2011. Ecological niches and geographic distributions. – Princeton Univ. Press.

Philippot, L., Spor, A., Hénault, C., Bru, D., Bizouard, F., Jones, C. M., Sarr, A. and Maron, P. A. 2013. Loss in microbial diversity affects nitrogen cycling in soil. – ISME J. 7: 1609–1619.

Pinto-Figueroa, E. A., Seddon, E., Yashiro, E., Buri, A., Niculita-hirzel, H., van der Meer, J. R. V. D. and 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. – Front Microbiol. 10: 656.

Pradervand, J.-N., Dubuis, A., Pellissier, L., Guisan, A. and Randin, C. 2014. Very high resolution environmental predictors in species distribution models: moving beyond topography? – Prog. Phys. Geogr. 38: 79–96.

Qiao, H., Peterson, A. T., Ji, L. and Hu, J. 2017. Using data from related species to overcome spatial sampling bias and associated limitations in ecological niche modelling. – Methods Ecol. Evol. 8: 1804–1812.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. – Nucleic Acids Res. 41: D590–D596.

Redford, K. H. 2023. Extending conservation to include Earth's microbiome. – Conserv. Biol. 37: e14088. https://doi.org/10.1111/cobi.14088

Regos, A., Gagne, L., Alcaraz-Segura, D., Honrado, J. P. and Domínguez, J. 2019. Effects of species traits and environmental predictors on performance and transferability of ecological niche models. – Sci. Rep. 9: 4221.

Ren, B., Hu, Y., Chen, B., Zhang, Y., Thiele, J., Shi, R., Liu, M. and 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. – Sci. Rep. 8: 5619.

Rocci, K. S., Lavallee, J. M., Stewart, C. E. and 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. – Sci. Total Environ. 793: 148569.

Scherrer, D. and Guisan, A. 2019. Ecological indicator values reveal missing predictors of species distributions. – Sci. Rep. 9: 3061. https://doi.org/10.1038/s41598-019-39133-1

Schmidt, P.-A., Bálint, M., Greshake, B., Bandow, C., Römbke, J. and Schmitt, I. 2013. Illumina metabarcoding of a soil fungal community. – Soil Biol. Biochem. 65: 128–132.

Schröder, B. 2008. Challenges of species distribution modeling belowground. – J. Plant Nutr. Soil Sci. 171: 325–337.

Seppey, C. V. W., Broennimann, O., Buri, A., Singer, D., Mitchell, E. A. D., Hirzel, H. N. and Guisan, A. 2020. Soil protist diversity in the Swiss Western Alps is better predicted by topoclimatic than by edaphic variables. – J. Biogeogr. 2019: 866–878.

Seppey, C. V. W., Lara, E., Broennimann, O., Guisan, A., Malard, L., Singer, D., Yashiro, E. and Fournier, B. 2023. Landscape structure is a key driver of soil protist diversity in meadows in the Swiss Alps. – Landscape Ecol. 38: 949–965.

Serna-Chavez, H. M., Fierer, N. and Van Bodegom, P. M. 2013. Global drivers and patterns of microbial abundance in soil. – Global Ecol. Biogeogr. 22: 1162–1172.

Smith, A. B., Godsoe, W., Rodríguez-Sánchez, F., Wang, H. H. and Warren, D. 2019. Niche estimation above and below the species level. – Trends Ecol. Evol. 34: 260–273.

Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. K. and 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 Res. 43: D593–D598.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H. W. and Richards, T. A. 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. – Mol. Ecol. 19: 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. – Science 240: 1285–1293.

Tay, J. K., Narasimhan, B. and Hastie, T. 2023. Elastic net regularization paths for all generalized linear models. – J. Stat. Softw. 106: 1.

Terrat, S., Horrigue, W., Dequietd, S., Saby, N. P. A., Lelièvre, M., Nowak, V., Tripied, J., Régnier, T., Jolivet, C., Arrouays, D., Wincker, P., Cruaud, C., Karimi, B., Bispo, A., Maron, P. A., Prévost-Bouré, N. C. and Ranjard, L. 2017. Mapping and predictive variations of soil bacterial richness across France. – PLoS One 12: 5–8.

Tessarolo, G., Lobo, J. M., Rangel, T. F. and Hortal, J. 2021. High uncertainty in the effects of data characteristics on the performance of species distribution models. – Ecol. Indic. 121: 107147.

Thuiller, W., Brotons, L., Araújo, M. B. and Lavorel, S. 2004. Effects of restricting environmental range of data to project current and future species distributions. – Ecography 27: 165–172.

Tian, L., Cai, T., Goetghebeur, E. and Wei, L. J. 2007. Model evaluation based on the sampling distribution of estimated absolute prediction error. – Biometrika 94: 297–311.

Tôrres, N. M., De Marco, P., Santos, T., Silveira, L., De Almeida Jácomo, A. T. and 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. – Divers. Distrib. 18: 615–627.

Trumbore, S. E. and Czimczik, C. I. 2008. An uncertain future for soil carbon. – Science 321: 1455–1456.

Van Couwenberghe, R., Collet, C., Pierrat, J.-C., Verheyen, K. and Gégout, J.-C. 2013. Can species distribution models be used to describe plant abundance patterns? – Ecography 36: 665–674.

Vaulot, D., Geisen, S., Mahé, F. and Bass, D. 2022. pr2-primers: an 18S rRNA primer database for protists. – Mol. Ecol. Resour. 22: 168–179.

Verdon, V., Malard, L., Collart, F., Adde, A., Yashiro, E., Pandi, E. L., Mod, H., Singer, D., Niculita-Hirzel, H., Guex, N. and Guisan, A. 2024. Data from: Can we accurately predict the distribution of soil microorganism presence and relative abundance? – Figshare Repository, https://doi.org/10.6084/m9.figshare.23674758.

Von Däniken, I., Guisan, A. and Lane, S. 2014. RechAlp.vd: une nouvelle plateforme UNIL de support pour la recherche trans-disciplinaire dans les Alpes vaudoises. – Bull. Soc. Vaudoise Sci. Nat. 94: 175–178

Waldock, C., Stuart-Smith, R. D., Albouy, C., Cheung, W. W. L., Edgar, G. J., Mouillot, D., Tjiputra, J. and Pellissier, L. 2022. A quantitative review of abundance-based species distribution models. – Ecography 2022: e05694.

Williams, J. W. and Jackson, S. T. 2007. Novel climates, no-analog communities, and ecological surprises. – Front. Ecol. Environ. 5: 475–482. https://doi.org/10.1890/070037

Wisz, M. S. et al. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. – Biol. Rev. 88: 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Wood, S. N. 2017. Generalized additive models: an introduction with R, 2nd edn. – CRC Press/Taylor and Francis.

Yashiro, E., Pinto-Figueroa, E., Buri, A., Spangenberg, J. E. and Adatte, T. 2016. Local environmental factors drive divergent grassland soil bacterial communities in the Western Swiss Alps. – Appl. Environ. Microbiol. 82: 6303–6316.

Yashiro, E., Pinto-Figueroa, E., Buri, A., Spangenberg, J. E., Adatte, T., Niculita-hirzel, H., Guisan, A. and 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. – Sci. Rep. 8: 5758.