nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-021-01445-9

Soil microbiome predictability increases with spatial and taxonomic scale

Colin Averill[1,2,3]✉, Zoey R. Werbin[1,2], Kathryn F. Atherton[1,4], Jennifer M. Bhatnagar[1,5] and Michael C. Dietze[2,5]

Soil microorganisms shape ecosystem function, yet it remains an open question whether we can predict the composition of the soil microbiome in places before observing it. Furthermore, it is unclear whether the predictability of microbial life exhibits taxonomic- and spatial-scale dependence, as it does for macrobiological communities. Here, we leverage multiple large-scale soil microbiome surveys to develop predictive models of bacterial and fungal community composition in soil, then test these models against independent soil microbial community surveys from across the continental United States. We find remarkable scale dependence in community predictability. The predictability of bacterial and fungal communities increases with the spatial scale of observation, and fungal predictability increases with taxonomic scale. These patterns suggest that there is an increasing importance of deterministic versus stochastic processes with scale, consistent with findings in plant and animal communities, suggesting a general scaling relationship across biology. Biogeochemical functional groups and high-level taxonomic groups of microorganisms were equally predictable, indicating that traits and taxonomy are both powerful lenses for understanding soil communities. By focusing on out-of-sample prediction, these findings suggest an emerging generality in our understanding of the soil microbiome, and that this understanding is fundamentally scale dependent.

Soil microorganisms control critical ecosystem processes, from agricultural productivity and animal disease transmission to greenhouse gas emissions[1]. It is increasingly clear that the identity of microorganisms within the soil environment determines the type and rate of these environmental processes[2–4]. Understanding microbial biogeography—the spatial distribution of microbial taxa across the planet—is therefore critical to understanding ecosystems and the processes that they regulate, yet tremendous fine-scale spatial heterogeneity in soil microbial communities[5] has led to scepticism about our ability to predict the presence or abundance of key types of microorganisms in soil[6]. Rapid advancement in DNA-sequencing technology has revolutionized our understanding of how the soil microbiome is shaped by environmental factors such as resource availability, host preference and climate[6–8]. Furthermore, the microbiology of some of the most economically and societally influential soil microorganisms (for example, human pathogens and moulds) has been studied for nearly 150 years[9]. Nevertheless, it remains unclear whether or not this information can help us to confidently predict the composition of different microbial groups in locations that have never been observed. To incorporate microbial biodiversity into regional and global scale analyses of ecological community and ecosystem properties, we need to be able to predict the presence and abundance of soil microbial community members and quantify our confidence that these predictions are accurate.

It is probable that our ability to predict the soil microbiome is scale dependent, yet it is unclear whether this scale dependence is similar to, or fundamentally different from, scale dependence observed in macrobiological communities. For example, the relative importance of deterministic versus stochastic ecological processes and, therefore, predictability exhibits remarkable spatial and taxonomic scale dependence in plant and animal communities[10–13]. Signatures of deterministic, environmental filtering processes in macrobiological communities become more apparent at larger spatial and higher taxonomic scales. For example, it can be difficult to predict the identity of any particular tree in a forest, analogous to trying to predict the outcome of a single coin flip. However, predicting the relative abundance of a tree species among thousands of trees across the landscape, analogous to trying to predict the outcome of thousands of coin flips, is possible. Whether these relationships function as general rules of life that extend across biology remains an open question[14], as multiple features of microbial biology may generate fundamentally different ecological scaling relationships. Microbial habitat preferences can evolve and change frequently[15], which may rapidly erode taxonomic signal, leading to greater predictability at lower, rather than higher, taxonomic scales (Fig. 1b compared with 1a). Furthermore, there is tremendous diversity even within soil cores—our smallest scale of soil microbial observation[7,8]. This spatial scale is still enormous for most microorganisms, such that comparing even larger scales may be analogous to aggregating distinct biogeographical regions of plant and animal communities, which could erode environmental signal with spatial scale (Fig. 2).

Furthermore, functional-trait-based frameworks in ecology have repeatedly been put forward as more predictive than strictly taxonomic ones[16,17]. Microbial functional groups (such as nitrogen fixers and mycorrhizal fungi) capture convergent patterns in habitat preferences and ecological function across disparate lineages[16,18,19], such that functional trait groupings may better describe the variation in community composition linked to environmental conditions compared with taxonomy alone[20,21] (Fig. 1c). However, microbial diversity within functional groups is vast, and microbial functional groups that have been defined to date may be overly broad or too simplistic[22]. Functional trait frameworks require a priori knowledge of which functional traits are most important for determining

[1]Department of Biology, Boston University, Boston, MA, USA. [2]Department of Earth & Environment, Boston University, Boston, MA, USA. [3]Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland. [4]Graduate Program in Bioinformatics, Boston University, Boston, MA, USA. [5]These authors jointly supervised this work: Jennifer M. Bhatnagar, Michael C. Dietze. ✉e-mail: colin.averill@usys.ethz.ch

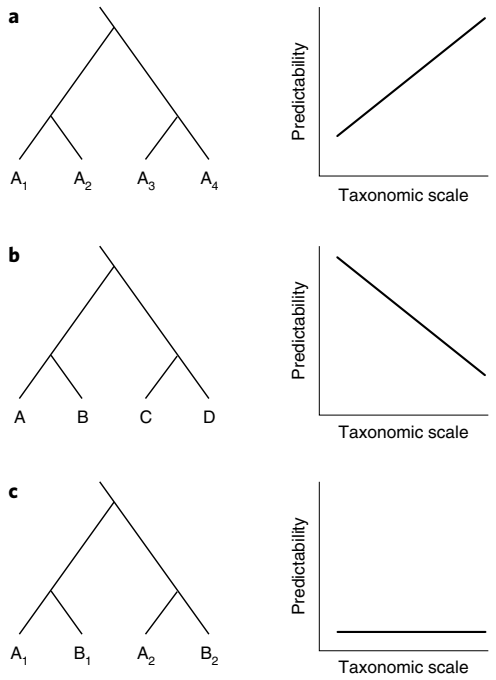


Fig. 1 | How can predictability vary as a function of taxonomic scale? **a**, Species A_1–A_4 are all habitat A specialists. Ecological drift (stochastic processes) prevents the prediction of any particular species as a function of the environment—the presence of habitat A can result in any combination of these four species. However, predictability increases as a function of taxonomic scale, as habitat A specialists are monophyletic. **b**, Species A–D are all habitat specialists in habitats A–D, respectively. In this case, predictability is highest at the lowest taxonomic scale as no other species shares these habitat preferences, and considering higher taxonomic scales results in combining taxa with different habitat preferences. **c**, Species A_1 and A_2 are both habitat A specialists, whereas species B_1 and B_2 are both habitat B specialists. In both cases, ecological drift prevents prediction at the species level. Furthermore, as taxonomic scale increases, these taxa remain unpredictable, as considering higher taxonomic scales results in combining taxa with different habitat preferences. Predicting these taxa requires a priori knowledge about which taxa are habitat A versus B specialists (that is, assigning these taxa to functional groups).

microbial fitness and its environmental sensitivity[17]. It is therefore unclear whether our current functional groupings of microorganisms are more or less predictable than taxonomic ones.

If environmental controls of microbial relative abundances and the associated scale dependence are truly general, then we should be able to predict the abundance of microbial taxa in locations before observing them. Prediction has been challenging in soil microbiome science, as only very recently have multiple independent large-scale community surveys become available that enable the validation of predictive community models. In this Article, we combine recently generated independent, large-scale datasets on soil microbial community composition with an ecological forecasting framework[23] to generate out-of-sample predictive models of soil microbial communities. We calibrate Bayesian statistical models to global surveys of 134 soil fungal and bacterial taxonomic and functional groups, then validate spatial predictions in a separate, independent continental-scale microbial community composition survey (that is, an out-of-sample prediction) from the US National Ecological Observatory Network (NEON). We built forecasts for all of the fungal and bacterial groups present in at least 50% of calibration samples. Models included commonly measured climate, soil and ecological covariates that have been linked to microbial diversity and composition and are available at large spatial scales (mean annual temperature and mean annual precipitation, remotely sensed net primary productivity, the presence or absence of forest vegetation, soil pH, soil percentage of carbon, soil ratio of carbon to nitrogen and the relative abundance of ectomycorrhizal trees)[7,8]. Separate models were constructed for different functional and taxonomic scales (from the phylum to genus levels) and validated at multiple spatial scales (core, plot and site level). Importantly, model development was conducted without ever 'seeing' the validation data and validation data were examined only once predictions were made. In contrast to hold-out and cross-validation approaches, for which sampling and measurement biases are shared between the model training and validation data, the NEON validation data provide a test dataset that is truly independent from the development of our predictive statistical model. With this approach, we tested our ability to predict, rather than describe, soil microbial communities and their associated scale dependence.

Results

Here we show our ability to predict soil microbial abundances was greatest at the largest spatial scales, consistent with patterns found within plant and animal communities[10,12]. Predictability consistently increased with spatial scale across the nested NEON sampling design for the majority of microbial groups modelled (Fig. 3), although we emphasize that not all microbial groups were predictable. When examining the relative abundance of individual microbial groups, observations were overdispersed relative to predictions (fewer than 95% of observations fell within the 95% predictive interval) at core and plot scales (Fig. 4) and the prediction error decreased as the spatial scale increased. We considered the possibility that the relationship between spatial scale and microbiome predictability may be an artefact of the models being trained on site-scale soil microbiome observations (that is, our global calibration datasets). However, when we recalibrated core and plot-scale models to 70% of the NEON data and then used these models to predict the remaining 30% of observations at the identical spatial scale, we found that predictability consistently increased and prediction error decreased with spatial scale within this dataset as well (Supplementary Fig. 1, Extended Data Fig. 1). Although this hold-out validation approach using the NEON data lacks the strengths of our independent validation set, it confirms that the spatial scale dependence observed within our larger analysis is not an artefact of sampling design. We suspect that considering many soil cores in aggregate—whether in the field or computationally—overwhelms the very high spatial heterogeneity in soil at small spatial scales[5]. Ecological theory predicts that stochastic processes are more likely to dominate community assembly at finer spatial scales[10]. However, our findings imply that extreme fine-scale heterogeneity in soil microbial communities at centimetre scales[5] does not prevent prediction at larger spatial scales and that inclusion of microbiome information into ecosystem and Earth-system models may be within reach. Furthermore, the increasing predictability of ecological communities with spatial scale across kingdoms of life seems to represent a general scaling relationship that may hold across biology.

We also found that our ability to predict soil fungi increases with taxonomic scale, where high-level fungal taxonomic groups (that is, classes and phyla) were, on average, more predictable and had lower prediction error than low-level taxonomic groups (that is, families and genera; Fig. 5a and Supplementary Fig. 2, Extended Data Fig. 2), a finding that held across spatial scales (Fig. 5c). This result parallels relationships found among macro-organisms, despite the extraordinary taxonomic diversity harboured by fungal phyla, suggesting that there is a general taxonomic scaling pattern in biology that crosses multiple kingdoms of life. To further examine this finding, we estimated Moran's *I*, which is a metric of

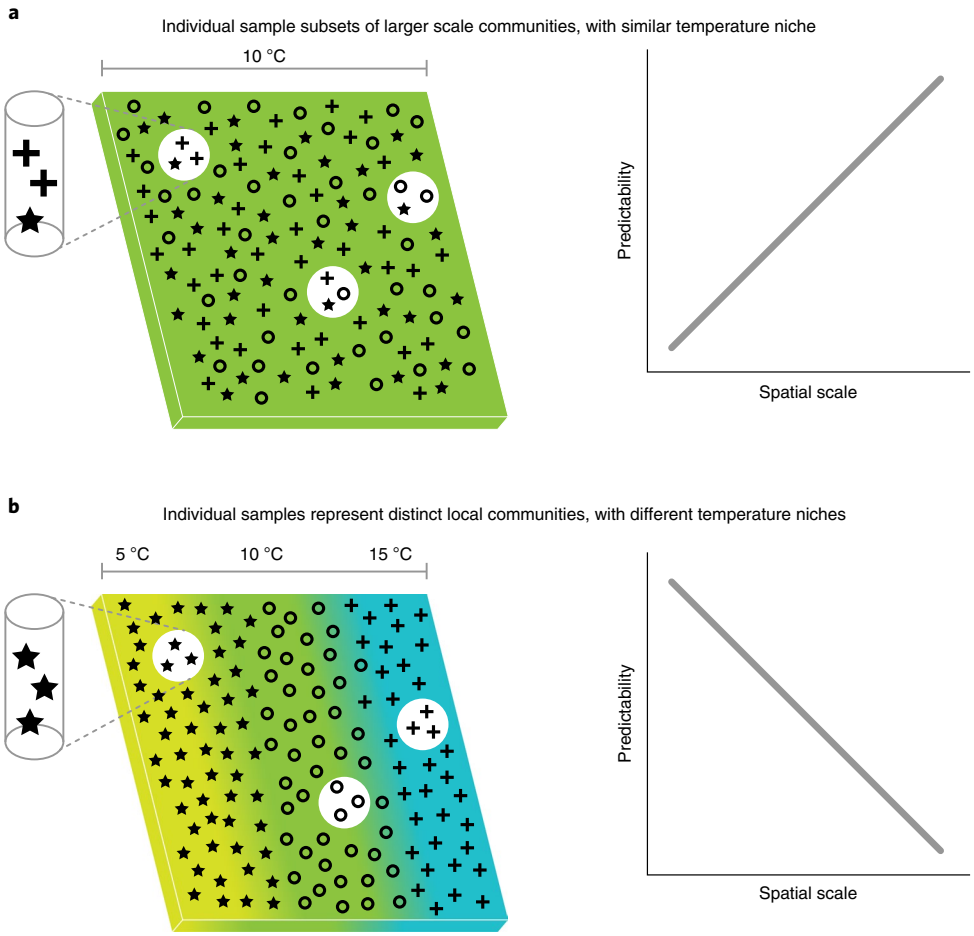


Fig. 2 | How can predictability vary as a function of spatial scale? a, If soil cores are subsamples of a larger scale community, then aggregating multiple soil cores to the plot scale will increase our ability to predict soil microbial communities as a function of the environment, as estimates of community composition and environment improve. **b**, If individual soil cores are capturing distinct microbial communities with different habitat preferences, then aggregating multiple cores to the plot scale will decrease our ability to predict soil microbial communities as a function of the environment, as mixing cores blurs community–environment relationships.

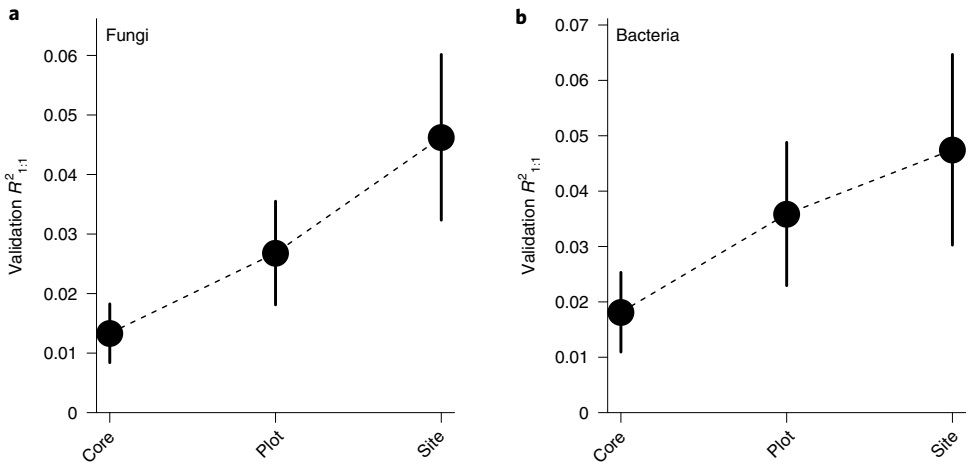


Fig. 3 | Predictability and spatial scale. a,**b**, Mean validation R^2 value relative to the 1:1 line at core, plot and site scales across all predicted fungal (**a**) and bacterial (**b**) groups. Data are mean ± 1 s.e.m. We emphasize that the mean values are low because they consider many groups that did not validate out of sample; however, many groups could be predicted substantially better (Fig. 2).

spatial autocorrelation, for the relative abundances of all microbial groups. Fungal functional groups and high-level fungal taxonomic groups had a higher Moran's *I* (greater spatial autocorrelation and lower patchiness) across the continent compared with lower-level taxonomic groups, consistent with patterns found in plant and animal communities (Fig. 5b). These patterns may emerge if trait

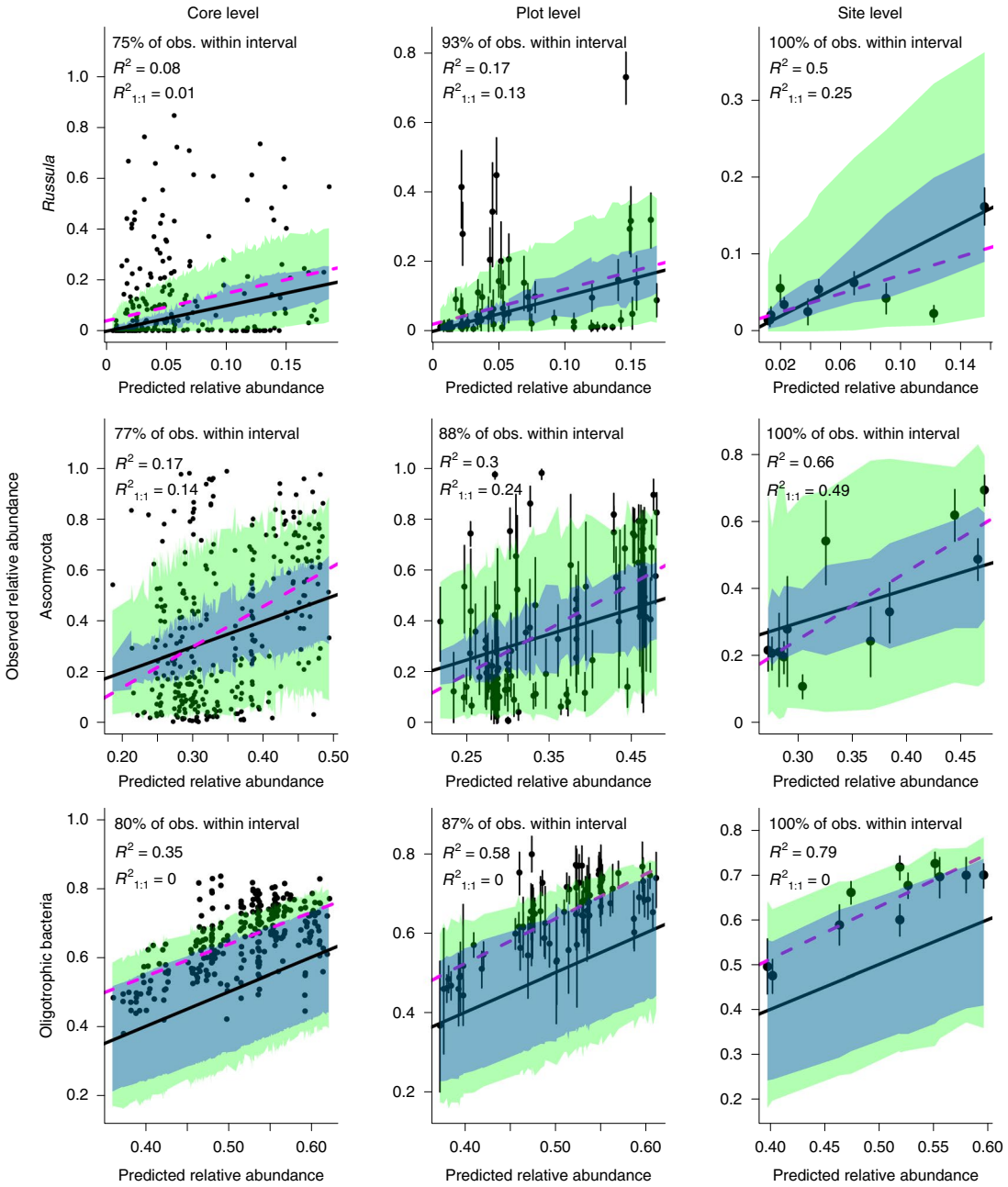


Fig. 4 | Predictability and taxonomic scale. Predicted versus observed relative abundance in the NEON out-of-sample validation data across core, plot and site scales for three representative groups—the fungal genus *Russula*, the fungal phylum Ascomycota and oligotrophic bacteria, a bacterial functional group. The 1:1 relationship is shown in black and the best fit is shown in purple. The shaded blue region represents the 95% credible interval of the mean estimate, and the green region represents the 95% predictive interval, where 95% of observations (obs.) are expected to fall. Points represent the observed relative abundance at each respective scale, and the errors bars show the 95% credible intervals of the mean. We report the percentage of observations that fall within the 95% prediction interval, the R^2 value of the best fit line between predicted and observed, as well as the R^2 value of observations relative to the 1:1 line.

conservatism is coupled with environmental filtering[12,13]; for example, if habitat preferences evolve early in the evolutionary history of an organism and biotic interactions among individuals are less important to the establishment of taxa within a local community than historical environmental preferences[24]. However, traits linked to fundamental microbial habitat preferences can also evolve frequently[21] and this may account for the high predictability of some low-level taxonomic groups (Fig. 4).

The evidence for taxonomic scaling among bacterial groups was mixed and inconclusive. Although in-sample fits did exhibit taxonomic scale dependence, this failed to hold once we attempted to validate models out of sample using the NEON dataset (Fig. 5d). This occurred despite the fact that bacterial spatial autocorrelation exhibited similar taxonomic scale dependence to fungi within the NEON dataset (Fig. 5e). However, although out-of-sample fits for bacteria were generally correlated with predictions, they were substantially biased, resulting in low predictive accuracy (Fig. 4, bottom row). We suspect that future efforts will overcome these biases by better accounting for differences in molecular methods computationally and by using molecular standards in routine microbiome data collection.

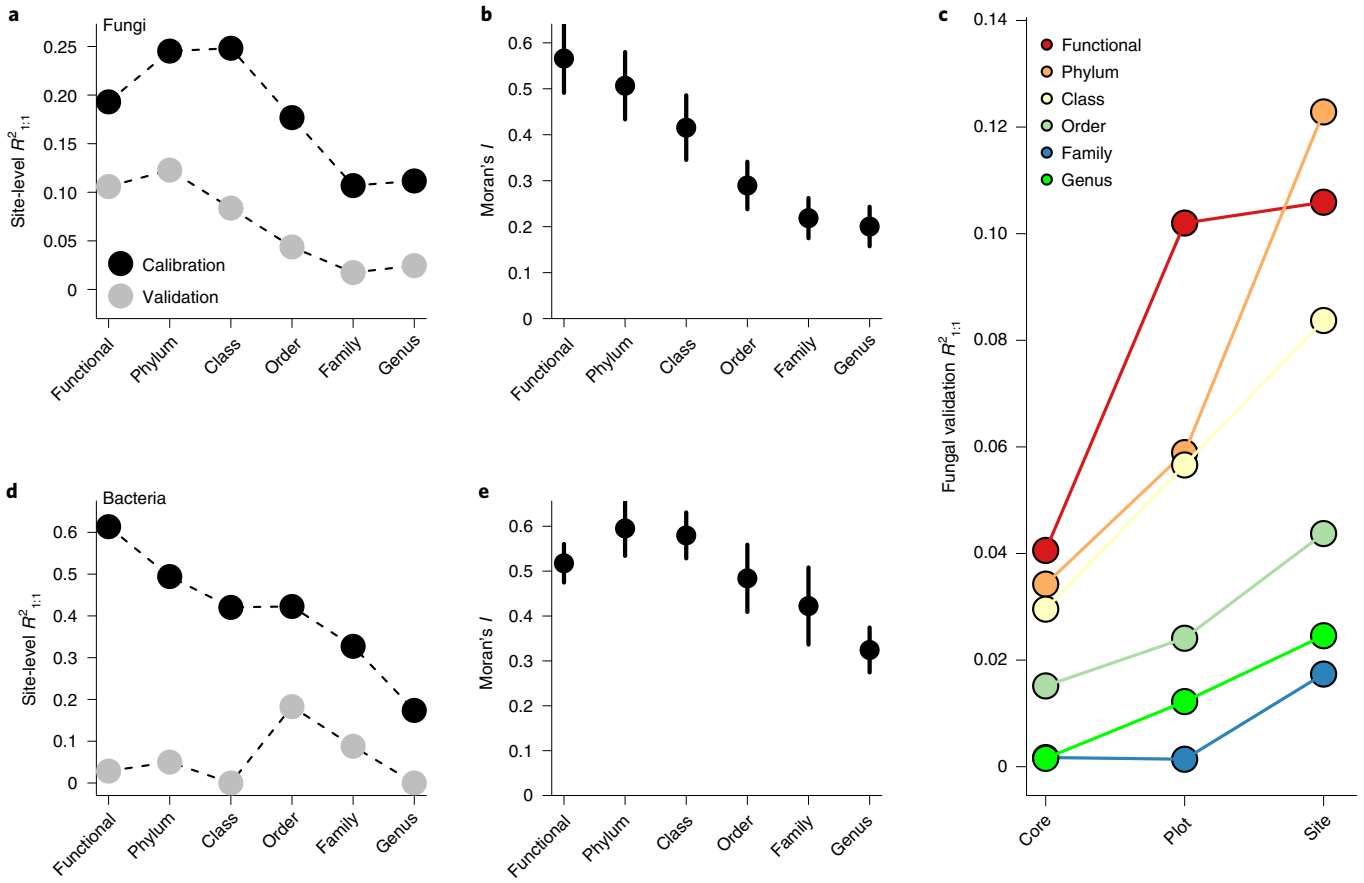


Fig. 5 | Average predictability of soil fungi and bacteria in the continental United States. a,**d**, Confidence in bacterial (**d**) and fungal (**a**) models (measured as R^2 values relative to the 1:1 line) on the basis of in-sample calibration models (black points) and out-of-sample validation of these same models to NEON observations aggregated to the site level (grey points). **b**,**e**, Moran's *I*, a metric of spatial autocorrelation of fungal (**b**) and bacterial (**e**) groups across the NEON network. The error bars represent 95% bootstrap confidence intervals. Greater Moran's *I* values are linked to reduced 'patchiness' and potentially greater spatial predictability. **c**, Fungal out-of-sample validation R^2 relative to the 1:1 line on the basis of validation using core-, plot- or site-scale observations across the NEON network.

We found that fungal functional groups were as predictable as high-level taxonomic groups (that is, fungal phyla) and that this pattern held across spatial scales (Fig. 5a). This pattern suggests that the functional groups of fungi assessed here may be of use to those seeking to incorporate features of soil fungal communities into models of ecosystem functions at large spatial scales. Bacterial functional groups, although generally more predictable than taxonomic groups in our calibration dataset, were not as predictable as fungal functional groups using the independent NEON survey validation data (Fig. 5d). However, raw correlations between predicted bacterial functional group relative abundances and out-of-sample observations, although biased, were higher for bacterial compared with fungal functional groups at the site level (Figs. 4 (bottom row) and 5a,d and Supplementary Data 1). This finding implies that, if study-to-study-level methodological variation can be overcome, prediction of bacterial groups in completely independent datasets might be possible.

We discovered large heterogeneity in predictability across microbial groups, yet reasonable accuracy ($R^2_{1:1}$ values of 50% and 55%) for the most predictable fungal and bacterial groups, respectively. We emphasize predictive accuracy is measured relative to the 1:1 observed–predicted line (rather than a best-fit line), so accuracy assessments are both qualitative and quantitative. For example, oligotrophic bacteria, although well correlated with model predictions, are underpredicted by our models and therefore have a low $R^2_{1:1}$ value (Fig. 4, bottom row). The most highly predictable microbial groups included the bacterial phylum Proteobacteria, the fungal order Pleosporales and the fungal phylum Ascomycota (Supplementary Data 1). We found some evidence that the most predictable fungal groups were the most sensitive to environmental parameters included in the prediction models. Looking at the variability in model parameters across taxa, there was wide variation in which features of the environment particular microorganisms were sensitive to, indicating that different environmental factors probably select for different soil microbial community members (Supplementary Fig. 3a,d, Extended Data Fig. 3). However, the more a fungal group was sensitive to soil pH, the relative abundance of ectomycorrhizal trees and, to a lesser extent, climate, the better we could model a particular fungal group within the calibration dataset (Supplementary Fig. 5b,c). This is consistent with previous studies in which these factors had strong correlations with overall fungal community structure[7]. Among bacteria, there was no single environmental sensitivity that was strongly linked to predictability in-sample and, if anything, larger environmental sensitivities were associated with lower accuracy (Supplementary Fig. 3e,f; Extended Data Fig. 3).

Many microbial groups that correlated well with environmental covariates within the calibration dataset (that is, in-sample validation using our global dataset) could not be predicted out-of-sample in the NEON dataset. Whereas 61% of fungal groups and 81% of

bacterial groups had at least 10% of variation explained in calibration datasets, only 18% of fungal groups and 4% of bacterial groups retained this accuracy once compared with independent validation data, relative to the 1:1 line. When we considered raw correlations between predictions and validation data, we found that 50% of fungal groups and 67% of bacterial groups became predictable at this threshold, indicating that, for many microbial groups, forecasts were qualitatively accurate (predictions and observations were correlated) but quantitatively biased (forecasts consistently under- or overpredicted relative abundances). For example, we observed a negative relationship between Acidobacteria relative abundance and pH in both calibration and validation datasets, yet, for the same pH, Acidobacteria are more abundant within the NEON dataset (Supplementary Fig. 4a, Extended Data Fig. 4). Thus, our model calibrated without the NEON data consistently underpredicts the relative abundance of this group, even though our model does capture the negative relationship between pH and Acidobacteria abundance. Most individual studies also observed qualitatively similar, but quantitatively different, relationships between Acidobacteria relative abundance and soil pH (Supplementary Fig. 4b). Thus, although qualitative predictions of many microbial taxa can be made (that is, there should be more or less of a particular lineage in a given location), quantitative predictions remain elusive. More-accurate qualitative versus quantitative predictions were more common among bacterial taxa and functional groups compared with among fungal taxa (Supplementary Data 1). This finding is consistent with previous research showing that the composition of bacterial communities can vary strongly by study[25], suggesting that bacteria have a higher sensitivity to differences in sample preparation or sequencing, an issue that could be overcome through the addition of positive controls and standards as part of routine microbiome analysis[26]. Poor out-of-sample predictability for many microorganisms raises concerns about the large-scale mapping efforts of individual taxa, operational taxonomic units or sequence variants based solely on in-sample statistical calibrations. Important methodological differences between our calibration and validation datasets, such as differences in sampling methodology and sequencing platforms, may explain some of the discrepancies that we observed between calibration and validation model fits. However, many microbial groups were predictable in the NEON dataset, implying that poor predictions for some groups are probably due to more than methodological differences.

To understand what may be driving prediction uncertainty, we decomposed calibration model uncertainty into process, sampling and observation sources, as these same calibration models were used to make predictions at all spatial and taxonomic scales. Uncertainty in predictions of all microbial groups was dominated by uncertainty in the ecological processes governing community structure; other sources of uncertainty (observation precision and sampling effort) had a small role (Supplementary Fig. 5, Extended Data Fig. 5). However, we emphasize that sufficient technical information was available only for climate covariates, and so uncertainty in observing other covariates, and uncertainty in sequencing microbial communities themselves, will be lumped into residual process uncertainty. Furthermore, within process uncertainty, we cannot yet distinguish model inadequacy from fundamentally stochastic ecological and evolutionary processes, which are more likely to have a role at finer spatial and taxonomic scales[27,28]. Yet, the increasing predictability of microbial groups with taxonomic and spatial scale implies a decreasing importance of stochastic processes at larger scales. Future studies focused on directly quantifying the relative importance of deterministic versus stochastic ecological processes for microbial community composition would be extremely valuable, as they could help to place an upper limit on the deterministic predictability of soil microbial community members, as well as how this limit varies across scales. Inclusion of additional covariates that capture features of ecological networks and interactions, as well as more dimensions of environmental filtering and species interactions, will probably improve prediction accuracy.

Discussion

Prediction in ecology is still rare, but increasingly urgent if ecological information is to inform future decision making[29]. This is especially true for soils, which are essential to global food and timber production, regulate the world's biogeochemical cycles and are perhaps the most biologically diverse environments on Earth[30], yet represent one of the greatest uncertainties in predicting Earth function[31,32]. Validation of our models with independent data (that is, out-of-sample validation) was critical for assessing the predictability of individual microbial groups. Although our current forecasts show that environmental relationships are predictive of spatial variation in many microbial groups, it remains to be determined whether spatial relationships can be used to forecast how the soil microbiome may change in time. Ongoing temporal monitoring of the microbiome across the NEON network, coupled with a predictive temporal forecasting framework, will enable us to test the validity of space-for-time substitutions in soil microbiome science. More broadly, forecasts developed to make predictions in time enable environmental monitoring data to play a role in hypothesis testing and hold the potential to further advance microbiome science beyond a basic natural history understanding of microbial biogeography. By demonstrating how the predictability of the microbiome varies across scales, this research will help to transition microbial ecology from a descriptive science to a predictive science that can be leveraged to better understand and predict Earth system processes.

Methods

Summary of the methods. To test the predictability of the soil microbiome, we focused our analysis on a subset of cosmopolitan fungal and bacterial taxonomic and functional groups. For fungal taxa, we used all of the microbial groups present in at least 50% of the samples within our global calibration datasets at each level of taxonomy (phylum to genus, sensu Delgado-Baquerizo et al.[6]). Bacterial datasets had many more taxa than could be reasonably analysed using this criterion, so we instead used the ten most frequently observed bacterial groups at each taxonomic level (phylum to genus). Furthermore, we binned fungal and bacterial taxa into functional groups of particular interest to the soil microbial ecology community. For fungi, we modelled the abundances of ectomycorrhizal, saprotrophic, wood saprotrophic, plant pathogenic and animal pathogenic fungi. These fungi have key roles in plant nutrient acquisition, decomposition, and plant and animal health[19]. We intended to model arbuscular mycorrhizal fungi as well; however, due to known biases against arbuscular mycorrhizal fungi in the internal transcribed spacer (ITS) primers we used for assessing fungi[33], these fungi were not sufficiently represented within our datasets. For bacteria, we binned taxa into the following functional groups: N-cyclers (nitrification, dissimilatory nitrate reduction, denitrification, dissimilatory nitrite reduction, assimilatory nitrate reduction, assimilatory nitrite reduction and nitrogen fixation), C-cyclers (cellulolytic, ligninolytic, chitinolytic and methanotroph), copiotrophs and oligotrophs. Each bacterial taxon could belong to multiple functional groups, except for copiotrophs and oligotrophs, which were mutually exclusive.

Models were fit as a function of environmental covariates that are commonly measured at large spatial scales, and which have been shown to be associated with the composition of soil fungal and bacterial communities at the global scale[7,8]. Soil covariates included pH, which was associated with every soil sample, and percentage of carbon (%C) and the carbon to nitrogen ratio (C:N), which were available only for fungal soil samples. At the plot and site scale, we focused on the relative abundance of ectomycorrhizal-associated trees, as trees that form ectomycorrhizal symbioses harbour radically different soil fungal communities and potentially bacterial communities compared with trees that do not[7]. In fungal models, additional vegetation characteristics included whether or not a site was a forest and whether or not conifers were present at a site as binary predictors, as forests generally harbour different soil microbial communities compared with non-forests, and coniferous forests are known to harbour their own suite of root associated fungi[7]. Finally, bacterial and fungal models included observations of mean annual temperature (MAT), mean annual precipitation (MAP) and net primary productivity (NPP), as these predictors have been shown to be important in previous analyses of global-scale soil microbial community composition[6,8]. All variables were used in the final analyses and, given their inclusion was based on a priori hypotheses, no variable selection was performed. Ideally, we would have incorporated more covariates, including but not limited to micronutrient

concentrations, fine root biomass and soil porosity. All of these covariates probably influence soil microbial communities at both small and large spatial scales[7]. However, we are limited by the covariates that have been observed within both our calibration and validation datasets. Furthermore, we considered testing for interactions among variables; however, while our calibration set includes hundreds of observations, our models already include a high number of covariates compared with observations. As global microbiome datasets grow and become more accessible, we will be able to test for more-complex relationships between the environment and soil microbiome composition.

Once models were trained on the calibration dataset, we validated models using data collected across the NEON. NEON hierarchically samples soil microbial communities. Three soil cores are collected and analysed within 10 plots across 12 observatory sites for which there was sufficient data at the time of this analysis. This enabled us to validate forecasts at core, plot and site scales. Importantly, all model validation was performed without the model ever 'seeing' the validation dataset. The validation dataset was used only to quantify model accuracy, and never used in the model calibration process. The spatial distribution of observations used for calibration and validation is presented in Supplementary Fig. 6 and Extended Data Fig. 6.

There are important methodological differences within and between calibration and validation datasets. How soils were collected (aggregated versus separated soil horizons), how communities were amplified (differences in fungal primers) and differences in sequencing technology (Roche 454 versus Illumina HiSeq versus Illumina MiSeq platforms) may drive substantial mismatch between calibration and validation datasets. We describe the study methodology in detail below. Ideally, all soils would be sampled in the same way, and measurements made using the same analytical methods. This is almost never the case in soil microbiome science or ecology in general. However, other scientists have successfully merged independent studies collected using very different methods[25]. Furthermore, discrepancies between calibration and validation datasets are also a feature of our analysis. We aimed to evaluate the general predictability of the soil microbiome in a way that has the potential to extend to future observations at NEON, as well as other completely independent studies. Thus, we chose not to divide a single dataset collected for a single study into calibration and validation subsets. Validating our models with completely independent data, collected by a different team with a different set of objectives, is a strong test of model performance, as well as our basic understanding of microbiome science. Models were calibrated without ever seeing validation data, and we did not validate forecasts until all model calibration was complete. Predictions were made before looking at the validation data.

Calibration data. *Global soil fungal observations.* We calibrated soil fungal forecast models using data from a global sampling of soil microbial communities[7]. We focused on observations within Northern Temperate latitudes in an effort to increase the similarity between our calibration and validation (that is, NEON) datasets (Supplementary Fig. 4, Extended Data Fig. 4). In this calibration dataset, 40 soil cores (diameter, 5 cm) were taken to 5 cm depth within a ~2,500 m^2 circular plot at each sampling site. All soil cores were then homogenized, air-dried and stored on silica before grinding and DNA extraction. Around 2.0 g of ground soil was extracted using the PowerMax Soil DNA Isolation kit (MoBio). Soil fungi were PCR-amplified using forward and degenerate reverse primers targeting the ITS2 region were designed to match >99.5% of all fungi. Fungal amplicons were sequenced on the 454-pyrosequencing platform using GS-FLX+ technology and Titanium chemistry as implemented by Beckman Coulter. Soil C and N concentrations were quantified using an elemental analyser. Soil pH was measured in a 1 N HCl solution. The authors reported the relative abundance of Ectomycorrhizal plants at each site. Site MAT and MAP were taken from the Wordclim2 global dataset[34]. NPP was taken from the MODIS global dataset[35]. Sequencing data were obtained from the Sequence Read Archive (SRA) database and information necessary to link sequence data to environmental covariates was provided in the supplementary data files of the original publications or by contacting the study authors directly. Extensive field sampling and chemical analysis details can be found in the original publication[7]. Raw fungal sequencing data were processed using the dada2 bioinformatic pipeline and dereplicated into exact sequence variants (ESVs)[36]. ESVs were then assigned to taxonomic and functional groups. Taxonomy was assigned using the RDP classifier[37], paired with the UNITE database for fungi[38]. Fungi were assigned to ectomycorrhizal, saprotrophic, wood saprotrophic, plant pathogenic or animal pathogenic functional groups using the FUNGuild database, which links taxonomy to function[19]. Fungal observations were rarefied to 1,000 reads per sample, and samples with fewer than 1,000 reads were removed from the analysis. The final fungal calibration dataset included 128 unique observations.

Global soil bacterial observations. We calibrated bacterial forecast models using a dataset compiled from a global sampling study[6] as well as a previous synthesis of 30 studies[25]. We subsetted data to northern temperate latitudes in an effort to better match the sampling extent of the NEON sampling, our validation dataset (Supplementary Fig. 6, Extended Data Fig. 6). Samples were collected between 2003 and 2015 at a variety of soil depths (median depth = 10 cm). Location and pH measurements were available for all of the samples. Site MAT and MAP were taken from the Wordclim2 global dataset[34]. NPP was taken from the MODIS global dataset[35]. The relative basal area of ectomycorrhizal trees was derived from a spatial product[39]. Samples from murine stool, desert or arctic environments were excluded, as well as samples sequenced using Roche 454 technology (which is known to present strong biases against common bacterial phyla[25]). Our resulting calibration dataset included 1,638 samples from 22 studies. Global sampling data from Delgado et al.[6] were processed using the dada2 bioinformatics pipeline and dereplicated into ESVs[36]. ESVs were rarefied to 10,000 reads as described previously[6] and samples with fewer than 10,000 reads were removed from the analysis. Taxonomy was assigned using the Greengenes database[40]. The synthesis dataset retrieved from Ramirez et al.[25] included merged and standardized taxonomy files from all studies. The authors reported that their 'name-matched' relative-abundance dataset performed similarly to a dataset created by reprocessing raw sequences, so we used the former, which had a larger sample size. Taxonomic assignments were then used to assign functional groups using the following sources: the presence of complete genomic N-cycling pathways (nitrification, dissimilatory nitrate reduction, denitrification, dissimilatory nitrite reduction, assimilatory nitrate reduction, assimilatory nitrite reduction and nitrogen fixation)[41]; genera were assigned to an N-cycling functional group if any species within the genera had complete pathways for any step of these processes (that is, the first or second step of denitrification). Cellulolytic taxa were similarly assigned at the genus level using a dataset[42]; the presence of any glycoside hydrolases genes for cellulose deconstruction was used to assign a genus to the 'cellulolytic' functional group. Other C-cycling groups (ligninolytic, chitinolytic and methanotroph) were assigned using a literature review. Copiotroph and oligotroph functional groups were assigned using a literature review[43], with finer-scale taxonomic classifications superseding broader-scale classifications; only assignments for copiotrophs and oligotrophs were mutually exclusive, but taxa could be assigned to any number of N-cycling and C-cycling functional groups. Finally, we emphasize that there are more bacterial functional groups that are important for understanding ecological and biogeochemical processes than considered here, yet we are constrained by a lack of information in the literature on the taxa that constitute functional groups other than those that we considered here. Many important functional traits, such as glycogenesis (which can be a proxy for stress tolerance[44]), may vary considerably across taxa and define additional functional groups that critically regulate ecosystem processes. Moreover, our approach requires taxonomic groups to map to discrete functional groups, which omits potential quantitative variation in microbial genomic investment in these strategies. We see quantitative functional traits as a next frontier in applying forecasting approaches to soil microbial communities, particularly through the integration of shotgun metagenomic datasets.

Validation data: NEON observations. To validate our forecasts, we collected soil microbiome observations and environmental covariates from NEON[45]. For this analysis, we used only the most currently available NEON data at the time of data acquisition, sampled in 2014 or 2016, during the peak greenness sampling (rather than during seasonal transition periods). The NEON sampling design is hierarchical. During each sampling, three soil cores are sampled per plot from ten plots nested within a site. Soils are sampled to a depth of 30 cm where possible. If a soil organic horizon is present, it is sampled using a square frame. Mineral soils are sampled using a circular soil corer (diameter, ≥2 cm) to a depth such that the total soil depth sampled (organic plus mineral) equals 30 ± 1 cm. Soils for molecular analysis are frozen at −80 °C. Around 2.0 g DNA was extracted per soil subsample for microbial community characterization using a MoBio PowerSoil Kit (MoBio). Soil fungi were characterized by PCR-amplifying the ITS1 region using the ITS1f–ITS2 primer pair. Soil bacteria were characterized by PCR-amplifying the 16S region using the 515FB–806R primer pair. Fungal and bacterial amplicons were sequenced using an Illumina MiSeq sequencer and v2 2 × 250 base-pair paired-end chemistry. Soil C and N concentrations were measured using an elemental analyser. Soil pH was measured in water as a 1:2 or 1:4 weight:weight ratio for mineral and organic horizon soils, respectively. We determined the relative basal area of ectomycorrhizal trees at each site (if present) using basal area measurements, species identities and a key that links tree species identities to mycorrhizal associations[46]. Full NEON soil sampling methods are described in NEON documents—'NEON TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling' as well as 'NEON TOS Science Design for Terrestrial Microbial Diversity'[45]. Sequencing data were processed using the dada2 bioinformatics pipeline and dereplicated into ESVs[36]. ESVs were then assigned to taxonomic and functional groups. Taxonomy was assigned using the RDP classifier[37], paired with the UNITE database for fungi[38] or the Greengenes database for bacteria[40]. Functional groups were assigned using taxonomy as performed for calibration datasets.

After all filtering criteria, the final fungal validation set included 317 unique observations at the core level, nested within 100 unique study plots, in turn nested within 12 NEON sites. For bacteria, the final validation set included 288 unique soil cores, nested within 83 plots, in turn nested within 11 NEON sites.

Statistical modelling in-sample. We modelled either taxonomic or functional groups of bacteria or fungi using a Dirichlet multivariate regression model[47]. The

Dirichlet distribution is the multivariate generalization of the beta distribution, and enabled us to model multiple functional or taxonomic groups simultaneously while accounting for covariance among group abundances due to the 'sum to 1' constraint of compositional data (all relative abundances of taxa within a sample must sum to 1). The Dirichlet model cannot handle relative abundance values of zero, so we transformed values to be on the open interval (0,1) and then rescaled values such that the sum of taxa relative abundances within a sample summed to one[48,49]. We attempted to avoid rarefaction and transformation by fitting a multinomial Dirichlet distribution, which accounts for variation in sequence depth across samples and allowed zeros to be present in our dataset[50]. However, when these models were fit, they performed poorly compared with Dirichlet-only fits to transformed data. Rarefying samples to a common sequence depth improved multinomial-Dirichlet model fits; however, parameters for low-abundance groups still failed to converge.

Taxonomic or functional group abundances were modelled as a linear combination of predictors and parameters, mapped to the Dirichlet distribution using a log-link function.

$$\log[\alpha] = X\beta \quad (1)$$

where α is a N-by-k matrix of Dirichlet parameters for k taxonomic or functional groups, N is the total number of observations, X is a N-by-j matrix of predictor values and β is a j-by-k matrix of parameters. For bacterial models, we also included a random effect of study to capture technical biases introduced by sequencing platform, primer choice and amplicon region[25]. This was not necessary for fungal models, as all data came from a single study. Thus, a given taxonomic or functional group would be modelled as:

$$\log[\alpha_k] = \beta_{k1} \times \%C + \beta_{k2} \times C{:}N + \beta_{k3} \times pH + \beta_{k4} \times relEM + \beta_{k5} \times forest + \beta_{k6} \times conifer + \beta_{k7} \times MAT + \beta_{k8} \times MAP + \beta_{k9} \times NPP \quad (2)$$

where β_{k1-9} are parameters that are linked to taxonomic or function group k. %C is the percentage of soil carbon by mass, C:N is the mass ratio of soil carbon to nitrogen, relEM is the relative abundance of ectomycorrhizal associated trees within a plot, forest is a binary predictor of whether or not the plot is classified as a forest, conifer is a binary predictor of whether or not coniferous vegetation is present, MAT is mean annual temperature, MAP is mean annual precipitation and NPP is the MODIS-derived net primary production. We re-emphasize that the bacterial models did not include terms for soil %C or C:N.

In interpreting these Dirichlet α values, the vector of mean predicted relative abundances for the Nth observation is given by

$$\mu_N = \frac{\alpha_{N,\cdot}}{\sum \alpha_{N,\cdot}} \quad (3)$$

and the predictive variance decreases as $\sum \alpha_{N,\cdot}$ increases. The final Dirichlet models were then specified as

$$y_N = \mathrm{Dir}(\alpha_{N,\cdot}) \quad (4)$$

Where, y_i is the vector of observed taxonomic or functional group relative abundances for the ith observation.

When possible, we included estimates of covariate uncertainty and sampled from covariate distributions when fitting models to account for covariate observation uncertainty. In practice, this resulted in our models incorporating only MAT and MAP uncertainty, as NPP, soil chemical observation uncertainties and tree basal area observation uncertainties were not reported. Statistical models were implemented in a Bayesian framework using JAGS, a Bayesian programming language[51]. JAGS models were fit using the runjags package for R statistical software[52]. Fungal models were fit using 3 Markov chains, each with 200 adaptive iterations, 2,000 burn-in iterations and 1,000 sample iterations. Bacterial models were also fit using 3 Markov chains, each with 60,001 adaptive iterations, 15,002 burn-in iterations and 5,003 sample iterations. We confirmed appropriate chain convergence by checking that all Gelman–Rubin diagnostic statistics were below 1.1. Bacterial models used substantially more MCMC iterations compared with fungal models because many more groups were modelled simultaneously at each level of taxonomy and function.

Bayesian statistical forecasting out-of-sample. NEON soil microbial observations are made at the individual core scale. As the calibration datasets are based on many pooled soil cores at the site scale, and because we were interested in how scale in and of itself affected the predictability of the soil microbiome, we made and validated NEON forecasts at the core, plot and site scales. Given the early stage of NEON sampling, as well as the fundamental challenge of orchestrating a continental scale observation network, there are missing covariate observations in our dataset. In an effort to account for missing data and retain as many microbial observations as possible, our statistical forecast included a missing data model[53]. When data were missing, they were estimated on the basis of a hierarchical model of each predictor. Therefore, if a core-level observation was missing, but it had been observed at the plot and site scale, this information was used to constrain the distribution of the missing observation. In the event that an observation was absent for an entire site, it was assigned a mean and uncertainty value on the basis of all observations across all sites. Plot and site-scale forecasts required hierarchically aggregating covariates observed at the core and plot scale, respectively. These aggregated covariates were also assigned uncertainties on the basis of on hierarchical models. Spatial scales vary from site to site but, in general, soil cores are 5.08 ± 1.27 cm diameter and 30 cm deep (or until bedrock). Microbial sampling plots are $20\,m \times 20\,m$ and include 3 soil cores sampled per plot. NEON sites vary from 5 km^2 to 50 km^2 and each include 10 sampling plots.

Forecasts at the core, plot and site scale are based on 1,000 ensemble draws of parameter and covariate distributions. Parameter draws were made by sampling the rows of the MCMC output of our model to account for parameter covariance. Covariates were drawn from their respective distributions. In the event that we did not have an uncertainty for a given covariate (that is, soil chemical data at the core scale), we assigned a very low uncertainty (s.d. = 0.1% of median observation) to that observation to facilitate Monte Carlo sampling.

Forecast validation. Models of all taxonomic and functional groups were validated at the core, plot and site scales. To validate forecasts at the plot and site scales, we hierarchically aggregated microbiome observations of taxonomic or functional relative abundances to the plot and site scales using a simple hierarchical Dirichlet model that estimated mean abundances at each level[47]. For each microbiome prediction, we also plotted a 95% credible interval and 95% predictive interval. The 95% credible interval represents our uncertainty of the mean microbial relative abundance at a given core, plot or site location. The 95% predictive interval represents where we expect 95% of all observed values to fall within. By comparing how many forecasted observations fall within the 95% predictive interval, we can assess whether our estimated forecast uncertainty is over- or underconfident[23].

Variance decomposition. To understand the dominant sources of uncertainty in our forecasts, we repeated forecasts, sequentially turning off process, covariate and parameter uncertainty. Parameter and covariate uncertainty represent uncertainty introduced by drawing from parameter and covariate distributions, respectively. Process uncertainty represents the uncertainty introduced into our forecast by passing our matrix of a_i estimates through the Dirichlet distribution, and reflects residual error that cannot be attributed to parameter or covariate uncertainty. We re-ran forecasts at a mean set of covariates, and estimated variances sequentially turning off each source of uncertainty[23]. We plotted this variance decomposition by normalizing the variance estimated in each case, by an estimate of the total variance with all sources of variation (process, parameter, covariate) turned 'on' (Supplementary Fig. 3, Extended Data Fig. 3).

Visualizing predictor importance. We modelled hundreds of fungal and bacterial taxa. To facilitate visualization of which predictors were important for predicting which phylogenetic and functional groups, we performed a principal components analysis on fitted model parameters. Parameter values for all functional or phylogenetic groups were collapsed into a single matrix. Principal components analysis was performed on this matrix using the prcomp function for R statistical software[54]. Parameter values were zero-centred and scaled proportional to their variance to facilitate comparison among variables. For the calibration datasets, we regressed the absolute magnitude of each prediction against each microbial group's R^2. We visualized the single predictor most tightly linked to in-sample predictability and also report the ability of each predictor in the model to predict calibration R^2 (Fig. 3; see the main text).

Diagnosing spatial signal across functional and taxonomic scales. Once calibration models had been fit and validated out of sample, we estimated spatial signal in fungal and bacterial observations across the NEON network using Moran's I, a statistic that estimates the degree of spatial autocorrelation in a response variable[55], for all functional and phylogenetic groups modelled. Moran values were calculated using distance matrices of group-relative abundances and physical distances in metres using the Moran.I function within the ape package for R statistical software[56]. We then aggregated observed Moran's I for each grouping of microorganisms (genus to phylum as well as functional groups) to understand the taxonomic and functional patterns in spatial autocorrelation.

Climate uncertainty estimates for the WorldClim2 dataset. We developed an uncertainty product for the WorldClim2 dataset, using the raw data provided by the original authors. To do so, we extracted observed MAT and MAP for each site used to develop the WorldClim2 tool. We then fit predicted versus observed models of MAT and MAP using linear regression, fit in a Bayesian framework using JAGS software[51,52]. We observed that variation in MAP observations increased with elevation, so we fit a model where MAP observation uncertainty scaled with elevation. This enabled us to quantify climate observation uncertainties and propagate these uncertainties through our analysis.

Cross-validating spatial patterns using NEON data. Models used to forecast to the NEON network are calibrated to observations made at the site scale. Therefore,

a failure to predict NEON microbial abundances at core and plot scales may be an artefact of the dataset our prior models were calibrated to. To assess this, we performed a cross-validation using only NEON network data. We refit models to either 50% of the core-level NEON observations or 70% of the plot-level NEON observations, and used these models to predict the remaining observations at the core or plot scale (Supplementary Fig. 1, Extended Data Fig. 1). This enabled us to understand whether predictability patterns across spatial scales based on models fit to site-level data were driven by the spatial scale of calibration data.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data used to train statistical models are either publicly available in associated studies or were provided on request to original study authors. All data used to validate models are publicly available through the National Ecological Observatory Network data portal (https://data.neonscience.org/). We will provide raw and processed data on request for purposes of replicating the findings of this study.

Code availability

All code needed to process raw data and to replicate these analyses is available at GitHub (https://www.github.com/colinaverill/NEFI_microbe).

Received: 6 October 2020; Accepted: 17 March 2021;
Published online: 22 April 2021

References

1. Schlesinger, W. H. & Bernhardt, E. S. *Biogeochemistry: an Analysis of Global Change* (Elsevier/Academic Press, 2012).
2. Fernandez, C. W., Langley, J. A., Chapman, S., McCormack, M. L. & Koide, R. T. The decomposition of ectomycorrhizal fungal necromass. *Soil Biol. Biochem.* **93**, 38–49 (2016).
3. Glassman, S. I. et al. Decomposition responses to climate depend on microbial community composition. *Proc. Natl Acad. Sci. USA* **115**, 11994–11999 (2018).
4. Mushinski, R. M. et al. Microbial mechanisms and ecosystem flux estimation for aerobic NO_y emissions from deciduous forest soils. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1814632116 (2019).
5. Prosser, J. I. Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology. *Nat. Rev. Microbiol.* **13**, 439–446 (2015).
6. Delgado-Baquerizo, M. et al. A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).
7. Tedersoo, L. et al. Global diversity and geography of soil fungi. *Science* **346**, 1256688 (2014).
8. Bahram, M. et al. Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).
9. Drews, G. The roots of microbiology and the influence of Ferdinand Cohn on microbiology of the 19th century. *FEMS Microbiol. Rev.* **24**, 225–249 (2000).
10. Chase, J. M. Spatial scale resolves the niche versus neutral theory debate. *J. Veg. Sci.* **25**, 319–322 (2014).
11. Ricklefs, R. E. & Renner, S. S. Global correlations in tropical tree species richness and abundance reject neutrality. *Science* **335**, 464–467 (2012).
12. Cavender-Bares, J., Keen, A. & Miles, B. Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. *Ecology* **87**, S109–S122 (2006).
13. Cavender-Bares, J., Kozak, K. H., Fine, P. V. A. & Kembel, S. W. The merging of community ecology and phylogenetic biology. *Ecol. Lett.* **12**, 693–715 (2009).
14. Ladau, J. & Eloe-Fadrosh, E. A. Spatial, temporal, and phylogenetic scales of microbial ecology. *Trends Microbiol.* **27**, 662–669 (2019).
15. Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4**, 457–469 (2003).
16. Diaz, S. & Cabido, M. Plant functional types and ecosystem function in relation to global change. *J. Veg. Sci.* **8**, 463–474 (1997).
17. Violle, C. et al. Let the concept of trait be functional! *Oikos* **116**, 882–892 (2007).
18. Fierer, N., Bradford, M. A. & Jackson, R. B. Toward an ecological classification of soil bacteria. *Ecology* **88**, 1354–1364 (2007).
19. Nguyen, N. H. et al. FUNGuild: an open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecol.* **20**, 241–248 (2016).
20. Whittaker, R. H. *Communities and Ecosystems* (Macmillan, 1975).
21. Gibbons, S. M. Microbial community ecology: function over phylogeny. *Nat. Ecol. Evol.* **1**, 0032 (2017).
22. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl Acad. Sci. USA* **113**, 5970–5975 (2016).
23. Dietze, M. C. *Ecological Forecasting* (Princeton Univ. Press, 2017).
24. Losos, J. B. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecol. Lett.* **11**, 995–1003 (2008).
25. Ramirez, K. S. et al. Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nat. Microbiol.* **3**, 189–196 (2018).
26. Smets, W. et al. A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biol. Biochem.* **96**, 145–151 (2016).
27. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ. Press, 2001).
28. Leibold, M. A., Urban, M. C., De Meester, L., Klausmeier, C. A. & Vanoverbeke, J. Regional neutrality evolves through local adaptive niche evolution. *Proc. Natl Acad. Sci. USA* **116**, 2612–2617 (2019).
29. Dietze, M. & Lynch, H. Forecasting a bright future for ecology. *Front. Ecol. Environ.* **17**, 3 (2019).
30. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
31. Todd-Brown, K. E. O. et al. Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences* **10**, 1717–1736 (2013).
32. Todd-Brown, K. E. O. et al. Changes in soil organic carbon storage predicted by Earth system models during the 21st century. *Biogeosciences* **10**, 18969–19004 (2013).
33. Lekberg, Y. et al. More bang for the buck? Can arbuscular mycorrhizal fungal communities be characterized adequately alongside other fungi using general fungal primers? *New Phytol.* **220**, 971–976 (2018).
34. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
35. Running, S., Mu, Q. & Zhao, M. *MOD17A3 MODIS/Terra Net Primary Production Yearly L4 Global 1km SIN Grid V055. NASA EOSDIS Land Processes DAAC* (NASA, 2011); https://cmr.earthdata.nasa.gov/search/concepts/C198653829-LPDAAC_ECS.html
36. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
37. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
38. Kõljalg, U. et al. Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* **22**, 5271–5277 (2013).
39. Steidinger, B. S. et al. Climatic controls of decomposition drive the global biogeography of forest-tree symbioses. *Nature* **569**, 404–408 (2019).
40. DeSantis, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
41. Albright, M. B. N., Chase, A. B. & Martiny, J. B. H. Experimental evidence that stochasticity contributes to bacterial composition and functioning in a decomposer community. *mBio* **10**, e00568-19 (2019).
42. Berlemont, R. & Martiny, A. C. Phylogenetic distribution of potential cellulases in bacteria. *Appl. Environ. Microbiol.* **79**, 1545–1554 (2013).
43. Ho, A., Lonardo, D. P. D. & Bodelier, P. L. E. Revisiting life strategy concepts in environmental microbial ecology. *Microbiol. Ecol.* https://doi.org/10.1093/femsec/fix006 (2017).
44. Wang, L. & Wise, M. J. Glycogen with short average chain length enhances bacterial durability. *Naturwissenschaften* **98**, 719–729 (2011).
45. *Soil Microbe Community Composition (DP1.10081.001)* (National Ecological Observatory Network (NEON)); https://data.neonscience.org
46. Averill, C., Dietze, M. C. & Bhatnagar, J. M. Continental-scale nitrogen pollution is shifting forest mycorrhizal associations and soil carbon stocks. *Glob. Change Biol.* **24**, 4544–4553 (2018).
47. Pawlowsky-Glahn, V., Egozcue, J. J. & Tolosana-Delgado, R. *Modelling and Analysis of Compositional Data* (John Wiley & Sons, 2015).
48. Smithson, M. & Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods* **11**, 54–71 (2006).
49. Cribari-Neto, F. & Zeileis, A. Beta regression in *R. J. Stat. Softw.* **34**, 1–22 (2010).
50. Johnson, N. L., Kotz, S. & Balakrishnan, N. *Discrete Multivariate Distributions* (Wiley, 1997).
51. Plummer, M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proc. 3rd International Workshop on Distributed Statistical Computing* 1–8 (2003); http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf
52. Denwood, M. J. runjags: an R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *J. Stat. Softw.* **71**, 1–25 (2016).
53. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge Univ. Press, 2007).
54. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
55. Moran, P. A. P. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).

56. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

Acknowledgements

The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle Memorial Institute. C.A., Z.R.W., M.C.D. and J.M.B. were supported by NSF Macrosystems Biology (no. 1638577). C.A. was supported by an Ambizione Grant (no. PZ00P3_179900) from the Swiss National Science Foundation. K.F.A. was supported by the Boston University BRITE Bioinformatics REU program. D. Maynard gave feedback on an earlier version of this manuscript. L. Stanish helped to access and interpret microbial data from the NEON Network. J. Luecke designed and illustrated Figs. 1 and 2.

Author contributions

C.A., J.M.B. and M.C.D. conceived the study. C.A., Z.R.W. and K.F.A. performed all analysis and computation. All of the authors wrote the manuscript collaboratively.

Competing interests

The authors declare no competing interests.

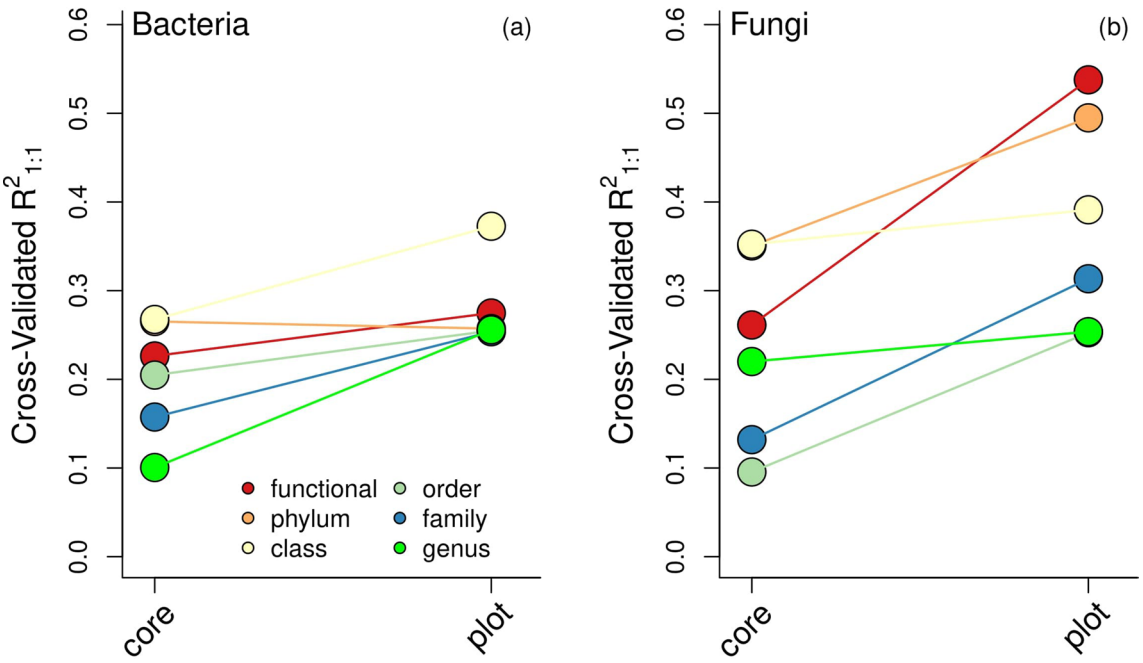
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-021-01445-9.

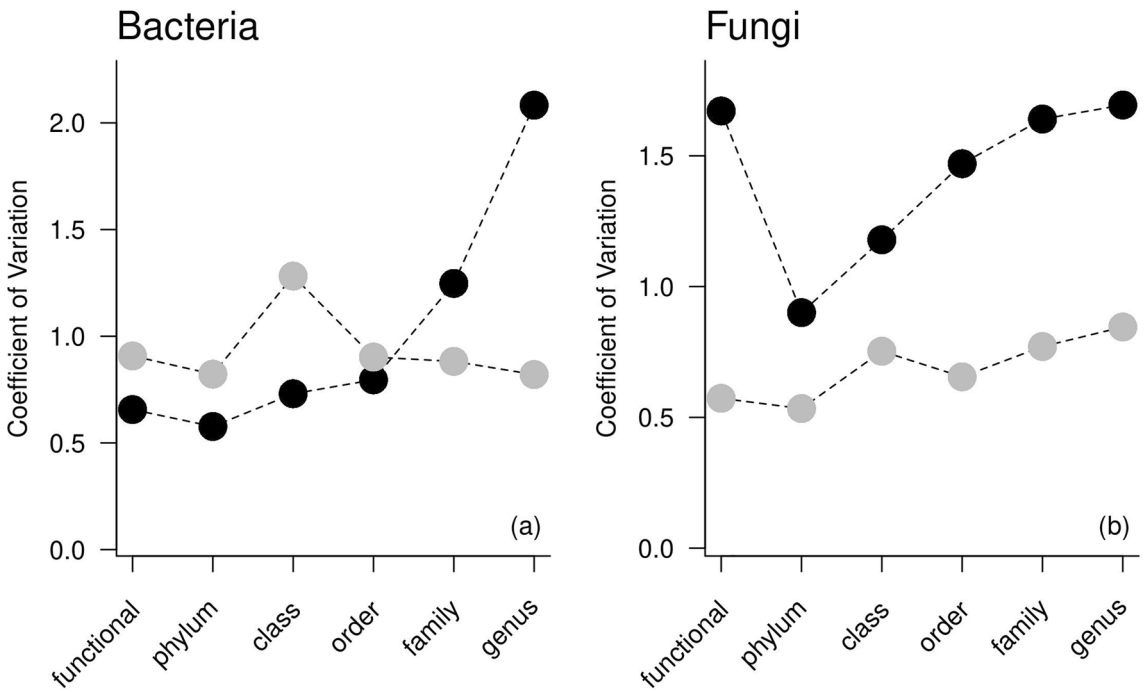
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-021-01445-9.

Correspondence and requests for materials should be addressed to C.A.

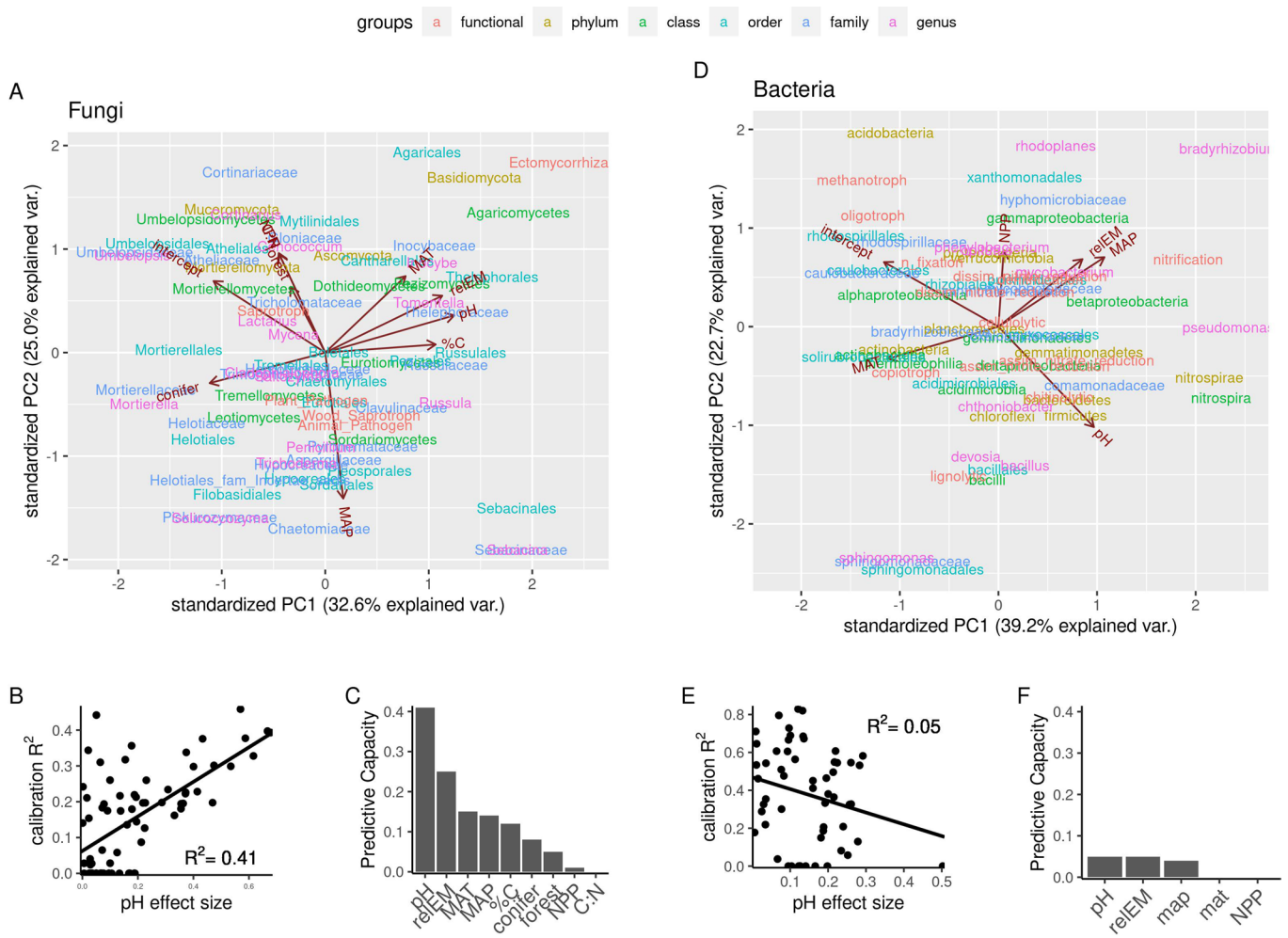
Peer review information *Nature Ecology & Evolution* thanks Xiaofeng Xu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



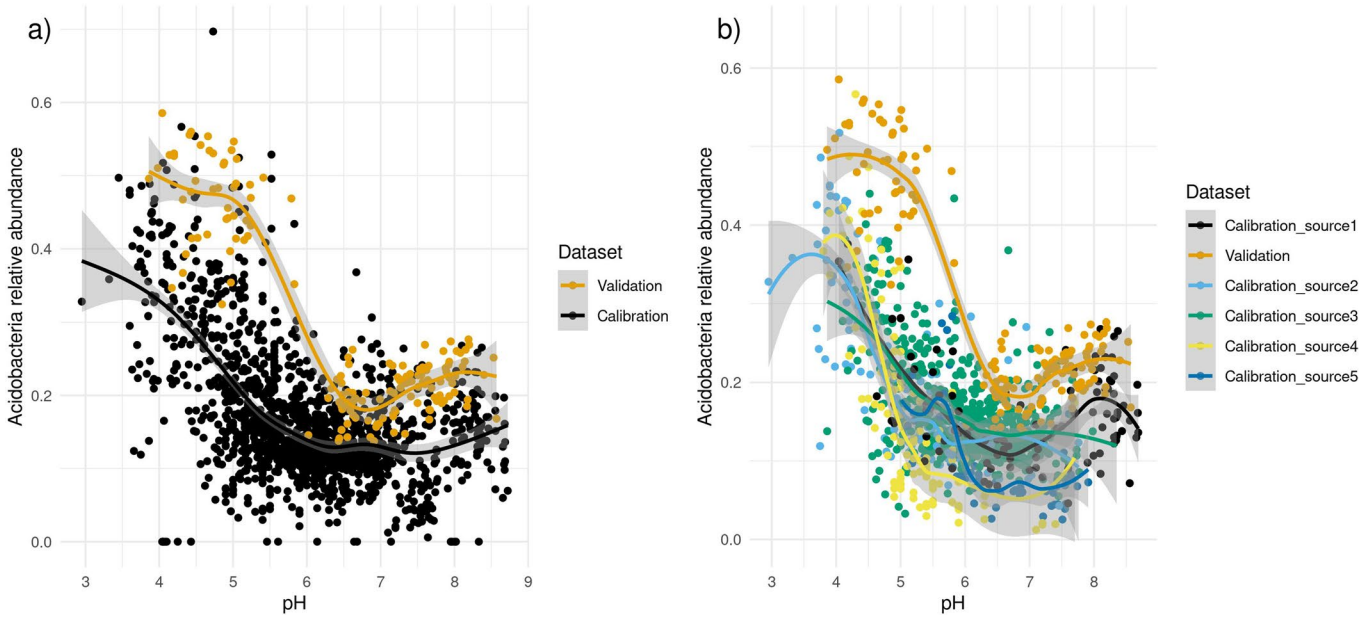
Extended Data Fig. 1 | Cross-validation within the NEON dataset. Mean cross-validated R^2 relative to the 1:1 prediction across functional and taxonomic groups for (**a**) bacteria and (**b**) fungi. All models were trained on 70% of NEON core or plot level data, and the validated using the remaining 30% of the data.



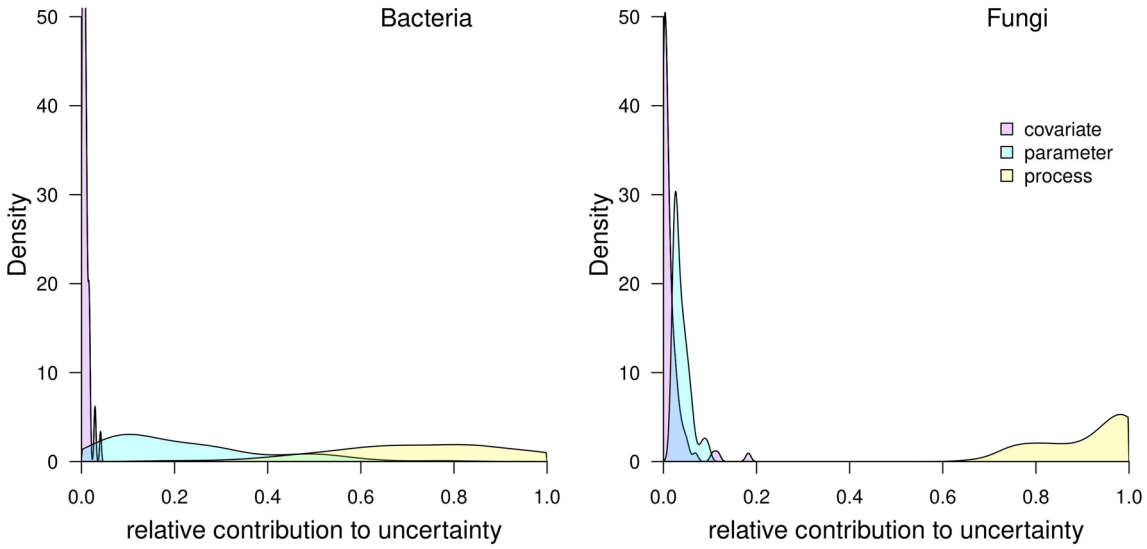
Extended Data Fig. 2 | Coefficient of variation across taxonomic and functional groupings. Coefficient of variation of model predictions vs. observations across functional and taxonomic groups, both in and out of sample for (**a**) bacteria and (**b**) fungi.



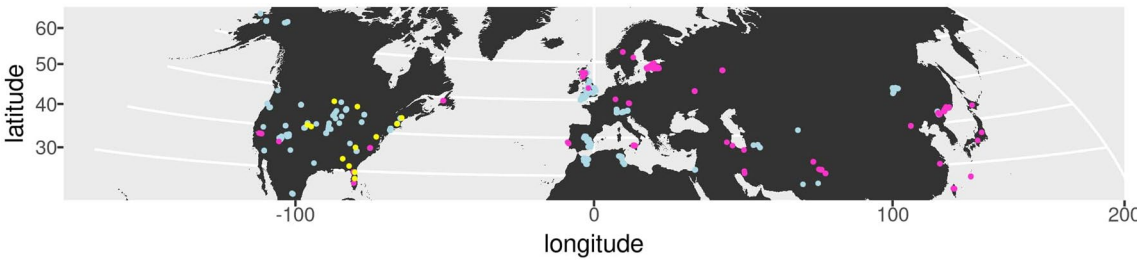
Extended Data Fig. 3 | Principal component analysis of microbial environmental sensitivities. Principal component analysis of phylogenetic and functional group parameter values in the global calibration dataset for (**a**) fungi and (**d**) bacteria. Factor importance in principal component space is indicated by the direction and length of factor vectors. We visualize the strongest correlation between an individual factor effect size and predictability and the calibration dataset (**b**,**e**), as well as the correlations for all factors (**c**,**f**). Factors include net primary productivity (NPP), whether or not conifers are present (conifer), whether or not a site is a forest (forest), mean annual temperature (MAT), mean annual precipitation (MAP), soil pH (pH), soil percent carbon (%C), soil carbon to nitrogen ratio (C:N), and the relative abundance of ectomycorrhizal trees (relEM).



Extended Data Fig. 4 | Qualitatively similar but quantitatively different relationships between Acidobacteria and soil pH. Relative abundance of bacterial phylum Acidobacteriaplotted as function of soil pH, highlighting differences in trends between independent sources. **a**, Values from combined calibration dataset and validation dataset, with points and loess curves colored by dataset. The relationship between Acidobacteria and pH within the validation data, sourced from the National Ecological Observatory Network, appears to have strong a systematic bias; however, due to the compositional nature of amplicon sequencing data, it is difficult to determine the source of biases for any given taxon. **b**, Values from a subset of 5 independent datasets used in calibration, with points and loess curves colored by dataset.



Extended Data Fig. 5 | Variance decomposition. Density plot of variance decomposition for all (**a**) bacterial and (**b**) fungal groups modeled at the site level.



Extended Data Fig. 6 | Distribution of samples used in this analysis. Distribution of sampling sites used in this analysis. Sites used for fungal model calibration are in pink, sites used for bacterial model calibration are in blue, and NEON sites used for validation are in yellow.

nature research

Corresponding author(s): Colin Averill

Last updated by author(s): Mar 3, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All code used to work up raw data to analysis ready products, analyze data and generate figures can be found at https://github.com/colinaverill/altSS_forest_mycorrhizas.
Data analysis	All code used to work up raw data to analysis ready products, analyze data and generate figures can be found at https://github.com/colinaverill/altSS_forest_mycorrhizas.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All forest data used in this analysis are available from the FIA data mart All code used to work up raw data to analysis ready products, analyze data and generate figures can be found at https://github.com/colinaverill/altSS_forest_mycorrhizas. All forest data used in this analysis are available from the FIA data mart (https://apps.fs.usda.gov/fia/datamart/). All code used to work up raw data to analysis ready products, analyze data and generate figures can be found at https://github.com/colinaverill/altSS_forest_mycorrhizas. All environmental covariate data sources are publicly available, and detailed in Supplementary Data File 1.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study uses observational data of US forest composition to understand how arbuscular and ectomycorrhizal symbioses influence growth, recruitment and survival of trees. We test for positive con-mycorrhizal feedbacks in these processes.
Research sample	All data used in this set of analyses comes from the US Forest Service's Forest Inventory and Analysis database, version 728. We further subsetted to sites within the Eastern US where re-measurement intervals are standardized, which facilitated fitting of demographic models and demographic simulations, as done in previous analyses (Figure 1). We only considered sites where all subplots were in a forested condition and excluded plantations as well as any sites with evidence of active management or harvesting. Trees were assigned AM or EM association as in previous analyses. Once mycorrhizal associations were assigned, we further subsetted the data set to sites where >90% of basal area was associated with either AM or EM associated trees. Therefore, a plot with 50% EM basal area and 50% AM basal area would be included in this analysis, however any plot with >10% mixed mycorrhizal, arbutoid, ericoid or non-mycorrhizal basal area would be excluded. This filter excluded 15% of all forested sites. Further investigation showed 2/3 of excluded sites (i.e. 10 of the 15%) were due to a high abundance of Populus, a well-known dual AM-EM tree genus. The remainder of exclusions were primarily driven by trees where mycorrhizal strategy was unknown. Our final dataset included 6,965 unique forest sites, re-censused at a ~5-year interval, with complete environmental covariates, comprised of 200,363 trees.
Sampling strategy	The FIA surey is designed as a gridded random sample of US forests.
Data collection	Raw forest inventory data were collected by foresters employed by the U.S. Forest Service. Environmental covariate data were extracted based on latitude/longitude and the geographic data layers described in Supplementary Data File 1.
Timing and spatial scale	Forest data were taken from the Eastern US. Plot remeasurement period was standardized to be 4.9-5.1 years apart.
Data exclusions	We further subsetted to sites within the Eastern US where re-measurement intervals are standardized, which facilitated fitting of demographic models and demographic simulations, as done in previous analyses (Figure 1). We only considered sites where all subplots were in a forested condition and excluded plantations as well as any sites with evidence of active management or harvesting. Trees were assigned AM or EM association as in previous analyses. Once mycorrhizal associations were assigned, we further subsetted the data set to sites where >90% of basal area was associated with either AM or EM associated trees. Therefore, a plot with 50% EM basal area and 50% AM basal area would be included in this analysis, however any plot with >10% mixed mycorrhizal, arbutoid, ericoid or non-mycorrhizal basal area would be excluded. This filter excluded 15% of all forested sites. Further investigation showed 2/3 of excluded sites (i.e. 10 of the 15%) were due to a high abundance of Populus, a well-known dual AM-EM tree genus. The remainder of exclusions were primarily driven by trees where mycorrhizal strategy was unknown. Our final dataset included 6,965 unique forest sites, re-censused at a ~5-year interval, with complete environmental covariates, comprised of 200,363 trees. Exclusion criteria were established prior to fitting any statistical models.
Reproducibility	This is an observational study, and therefore no experiments were performed or reproduced.
Randomization	Because this is an observational study, no randomization was performed.
Blinding	Because this is an observational study using data collected by others, no blinding was performed.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging