Supplement for Microbial Ecological Forecasting Manuscript

Methods Summary

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 use all microbial groups present in at least 50% of samples within our global calibration data sets at each level of taxonomy (phylum to genus) (*sensu* Delgado-Baquerizo *et al.* 2018). Bacterial datasets had many more taxa than could be reasonably analyzed using this criterion, so instead we 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 modeled the abundances of ectomycorrhizal, saprotrophic, wood saprotrophic, plant pathogenic and animal pathogenic fungi. These fungi play key roles in plant nutrient acquisition, decomposition, plant and animal health (Nguyen *et al.* 2016). We intended to model arbuscular mycorrhizal fungi as well, however due to known biases against arbuscular mycorrhizal fungi in the ITS primers we used for assessing fungi (Lekberg *et al.* 2018), these fungi were not sufficiently represented within our data sets. 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), and 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 global scale (Tedersoo *et al.* 2014; Bahram *et al.* 2018). Soil covariates included pH, which was associated with every soil sample, and percent 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 harbor radically different soil fungal communities, and potentially bacterial communities, than trees that do not (Tedersoo *et al.* 2014). 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 harbor different soil microbial communities than non-forests, and coniferous forests are known to harbor their own

suite of root associated fungi (Tedersoo *et al.* 2014). 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 (Bahram *et al.* 2018; Delgado-Baquerizo *et al.* 2018). Ideally, we would have incorporated more covariates, including but not limited to micronutrient concentrations, fine root biomass, soil porosity, and more. All of these covariates likely influence soil microbial communities at both small and large spatial scales (Tedersoo *et al.* 2014). However, we are limited by the covariates that have been observed within both our calibration and validation data sets.

Once models were trained on calibration data set, we validated models using data collected across the National Ecological Observatory Network (NEON). NEON hierarchically samples soil microbial communities. Three soil cores are collected and analyzed within 10 plots across 11 observatory sites for which there was sufficient data at the time of this analysis. This allowed us to validate forecasts at core, plot and site scales. Importantly, all model validation was performed without the model ever "seeing" the validation data set. The validation data set is only used to quantify model accuracy, and never used in the model calibration process.

There are important methodological differences within and between calibration and validation data sets. How soils were collected (aggregated vs. separated soil horizons), how communities were amplified (differences in fungal primers) and differences in sequencing technology (Roche 454 vs. Illumina Hiseq vs. Illumina Miseq platforms) may drive substantial mismatch between calibration and validation data sets. We describe 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 with very different methods (Ramirez *et al.* 2018). Furthermore, discrepancies between calibration and validation data sets are also a feature of our analysis. We aim 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. Therefore, we chose not to divide a single data set 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 ever looking at validation data.

Calibration data - global soil fungal observations: We calibrated soil fungal forecast models using data from a global sampling of soil microbial communities (Tedersoo *et al.* 2014). We focused on 131 observations within Northern Temperate latitudes in an effort to increase the similarity between our calibration and validation (i.e. NEON) data sets (Supplementary Figure 4). In this calibration data set, forty 5-cm diameter soil cores were taken to 5cm depth within a ~2500m^2 circular plot at each sampling site. All soil cores were then homogenized, air-dried and stored on silica before grinding and DNA extraction. ~2.0g of ground soil were extracted using the PowerMax Soil DNA Isolation kit (MoBio, Carlsbad, CA USA). 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 the GS-FLX+ technology and Titanium chemistry as implemented by Beckman Coulter. Soil C and N concentrations were quantified using an elemental analyzer. Soil pH was measured in a 1N HCl solution. Authors reported the relative abundance of Ectomycorrhizal plants at each site. Site mean annual temperature (MAT) and mean annual precipitation (MAP) were taken from the Wordclim2 global data set (Fick & Hijmans 2017). NPP was taken from the MODIS global data set (Running *et al.* 2011). Sequence data were obtained from the short read archive (SRA) database and information necessary to link sequence data to environmental covariates were provided in supplementary data files form original publications or by contacting study authors directly. Extensive field sampling and chemical analysis details can be found in the original publication (Tedersoo *et al.* 2014). Raw fungal sequence data were processed using the dada2 bioinformatic pipeline and de-replicated into exact sequence variants (ESVs, Callahan *et al.* 2016). ESVs were then assigned to taxonomic and functional groups. Taxonomy was assigned using the RDP classifier (Wang *et al.* 2007), paired with the UNITE database for fungi (Kõljalg *et al.* 2013). Fungi were assigned to ectomycorrhizal, saprotrophic, wood saprotrophic, plant pathogenic or animal pathogenic functional groups using the FUNGuild database, which links taxonomy to function (Nguyen *et al.* 2016). Fungal observations were rarefied to 1,000 reads per sample, and samples with fewer than 1,000 reads were removed from the analysis.

Calibration data - global soil bacterial observations: We calibrated bacterial forecast models using a dataset compiled from a global sampling study (Delgado-Baquerizo *et al.* 2018) as well as a collection of 30 studies synthesized by Ramirez *et al.* (2018). We subsetted data to northern temperate latitudes in an effort to better match the sampling extent of the NEON sampling, our validation data set (Supplementary Figure 4). Samples were collected between 2003 and 2015, at a variety of soil depths (median depth: 10cm). Location and pH measurements were available for all samples. Site MAT and MAP were taken from the Wordclim2 global data set (Fick & Hijmans 2017). NPP was taken from the MODIS global data set (Running *et al.* 2011). The relative basal area of ectomycorrhizal trees was derived from the spatial product presented in Steidinger *et al.* (2019). Samples from murine stool, desert, or arctic environments were excluded, as well as samples sequenced using Roche 454 technology (which was noted in Ramirez et al. 2018 to present strong biases against common phyla). Our resulting calibration dataset included 1629 samples from 22 studies. Global sampling data from Delgado et al. (2018) was processed using the dada2 bioinformatic pipeline and de-replicated into exact sequence variants (ESVs, Callahan et al. 2016). ESVs were rarefied to 10,000 reads (as in Delgado et al. 2018), and samples with fewer than 10,000 reads were removed from the analysis. Taxonomy was assigned using greengenes database (DeSantis et al. 2006). The synthesis dataset retrieved from Ramirez et al. (2018) 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 re-processing raw sequences, so we used the former, which had a higher sample size. Taxonomic assignments were then used to assign functional groups using the following sources: Presence of complete genomic N-cycling pathways (nitrification, dissimilatory nitrate reduction, denitrification, dissimilatory nitrite reduction, assimilatory nitrate reduction, assimilatory nitrite reduction, and nitrogen fixation) was reported by Albright et al (2019); genera were assigned to an N-cycling functional group if any species within the genera had complete pathways for any step of these processes (i.e. the first or second step of denitrification). Cellulolytic taxa were similarly assigned at the genus level using a dataset from Berlemont and Martiny (2013); 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 the literature

review from Ho et al. (2017), with finer-scale taxonomic classifications superceding 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.

Validation data - National Ecological Observatory Network (NEON) observations: To validate our forecasts, we collected soil microbiome observations and environmental covariates from NEON (National Ecological Observatory Network 2018). For this analysis, we only used the most currently available NEON data from 13 sites 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 3 soil cores are sampled per plot, from 10 plots nested within a site. Soils are sampled to 30cm depth, 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 (≥2cm diameter) to a depth such that the total soil depth sampled (organic plus mineral) equals 30 ±1cm. Soils for molecular analysis are frozen at -80° C. ~2.0g DNA was extracted per soil subsample for microbial community characterization using a MoBio PowerSoil Kit (MoBio, Carlsbad, CA). 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 2x250 base-pair paired end chemistry. Soil C and N concentrations were measured on an elemental analyzer. 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 (Averill *et al.* 2018). 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 (National Ecological Observatory Network 2018). Sequence data were processed using the dada2 bioinformatic pipeline and de-replicated into exact sequence variants (ESVs, Callahan *et al.* 2016). ESVs were then assigned to phylogenetic and functional groups. Phylogeny was assigned using the RDP classifier (Wang *et al.* 2007), paired with the UNITE database for fungi (Kõljalg *et al.* 2013), or the greengenes database for bacteria (DeSantis *et al.* 2006). Functional groups were assigned using taxonomy as done for calibration data sets.

Statistical modeling in-sample: We modeled either taxonomic or functional groups of bacteria or fungi using a Dirichlet multivariate regression model (Pawlowsky-Glahn *et al.* 2015). The Dirichlet distribution is the multivariate generalization of the beta distribution, and allowed 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 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 (Smithson & Verkuilen 2006; Cribari-Neto & Zeileis 2010). 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 data set (Johnson *et al.* 1997). However, when these models were fit, they performed poorly compared to Dirichlet-only fits to transformed data. Rarefying samples to a common sequence depth improved multinomial-Dirichlet model fits, however parameters for low abundance groups frequently failed to converge.

Species abundances were modeled as a linear combination of predictors and parameters, mapped to the Dirichlet distribution using a log-link function.

$$\log(\alpha) = X\beta$$

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-N 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 (Ramirez *et al.* 2017). This was not necessary for fungal models, as all data came from a single study. In interpreting these Dirichlet parameters, the vector of mean predicted relative abundances for the i_{th} observation is given by

$$\mu_i = \frac{\alpha_{i,\cdot}}{\sum \alpha_{i,\cdot}}$$

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

$$y_i \sim Dir(\alpha_{i,\cdot})$$

Where, y_i is the vector of observed taxonomic or functional group relative abundances for the i_{th} 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 (Plummer 2003). JAGS models were fit through the runjags package for R statistical software (Denwood 2016).

Bayesian statistical forecasting out-of-sample: NEON soil microbial observations are made at the individual core scale. Because 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 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, and the fundamental challenge of orchestrating a continental scale observation network, there are missing covariate observations in our data set. In an effort to account for missing data and retain as many microbial observations as possible, our statistical forecast included a missing data model (Gelman & Hill 2007). When data were missing, they were estimated based on 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 an observation was absent for an entire site, it was assigned a mean and uncertainty based on 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 based on hierarchical models.

Forecasts at the core, plot and site scale are based on 10,000 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 there respective distributions. In the event we did not have an uncertainty for a given covariate (i.e. soil chemical data at the core scale) we assigned a very low uncertainty (standard deviation = 0.1% of median observation) to that observation to facilitate Monte Carlo sampling.

Forecast validation: 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 (Pawlowsky-Glahn *et al.* 2015). For each microbiome prediction, we also plot 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 under-confident (Dietze 2017).

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, as described in equation 2, 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 (Dietze 2017). We plot 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 Figure 3).

Visualizing predictor importance: We modeled hundreds of fungal and bacterial taxa. To facilitate visualization of which predictors were important for predicting which phylogenetic and functional

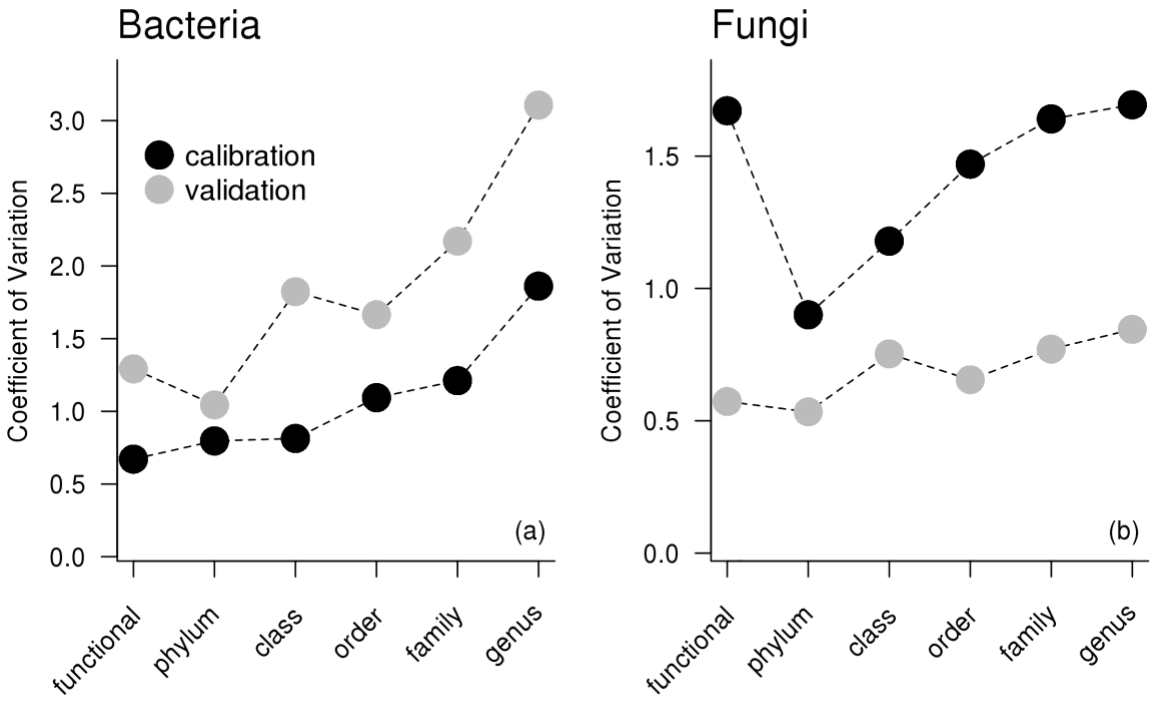
groups, we performed a principle components analysis (PCA) on fitted model parameters. Parameter values for all functional or phylogenetic groups were collapsed into a single matrix. PCA was performed on this matrix using the prcomp function for R statistical software (R Core Team 2017). Parameter values were zero centered and scaled proportional to their variance in order 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 visualize 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 (Figure 3, 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 distributions using Moran's I, a statistic that estimates the degree of spatial autocorrelation in a response variable (Moran 1950), for all functional and phylogenetic groups modeled. Moran values were calculated using distance matrices of group relative abundances and physical distances in meters using the Moran.I function within the ape package for R statistical software (Paradis *et al.* 2004). We then aggregated observed Moran's I for each grouping of microbes (genus to phylum as well as functional groups) to understand phylogenetic 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 vs. observed models of MAT and MAP using linear regression, fit in a Bayesian framework using JAGS software (Plummer 2003; Denwood 2016). We observed that variation in MAP observations increased with elevation, so we fit a model where MAP observation uncertainty scaled with elevation. This allowed 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 artifact 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 Figure 2). This allowed us to understand if predictability patterns across spatial scales based on models fit to site-level data were driven by the spatial scale of calibration data.



Supplementary Figure 1. 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. We note the coefficients were lower for fungal validation models, both because absolute root mean square error was lower in validation fits, and because the abundance of organisms was, on average, lower in calibration fits, which increases the calculated coefficient of variation.

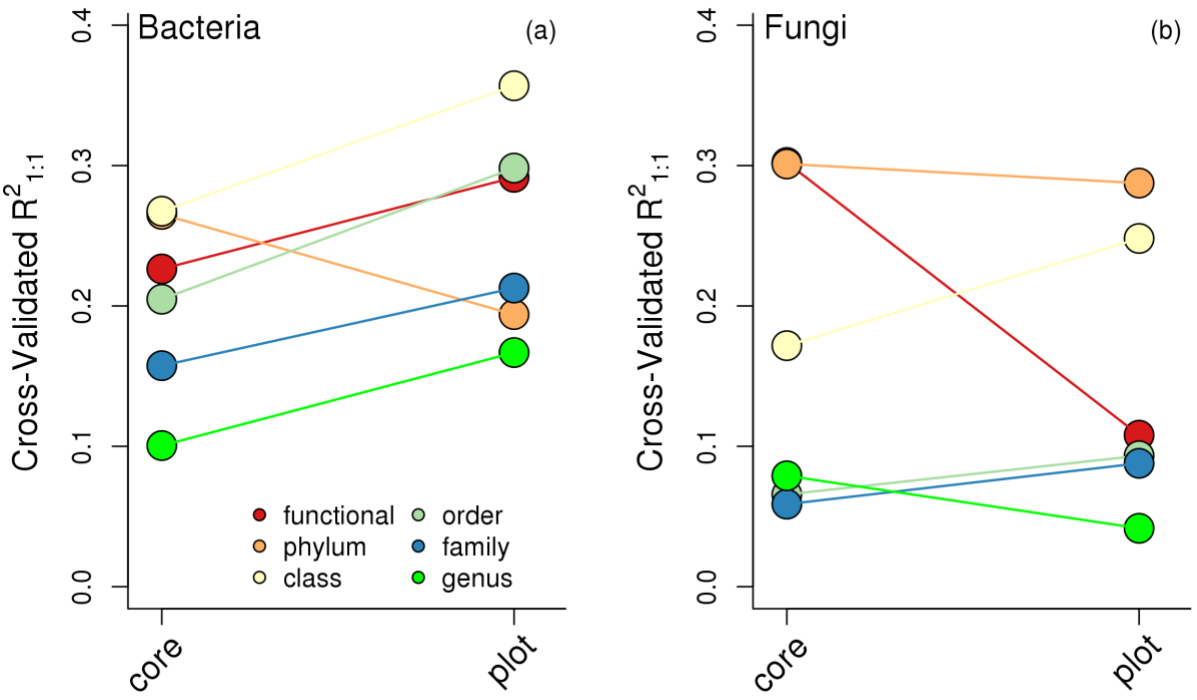
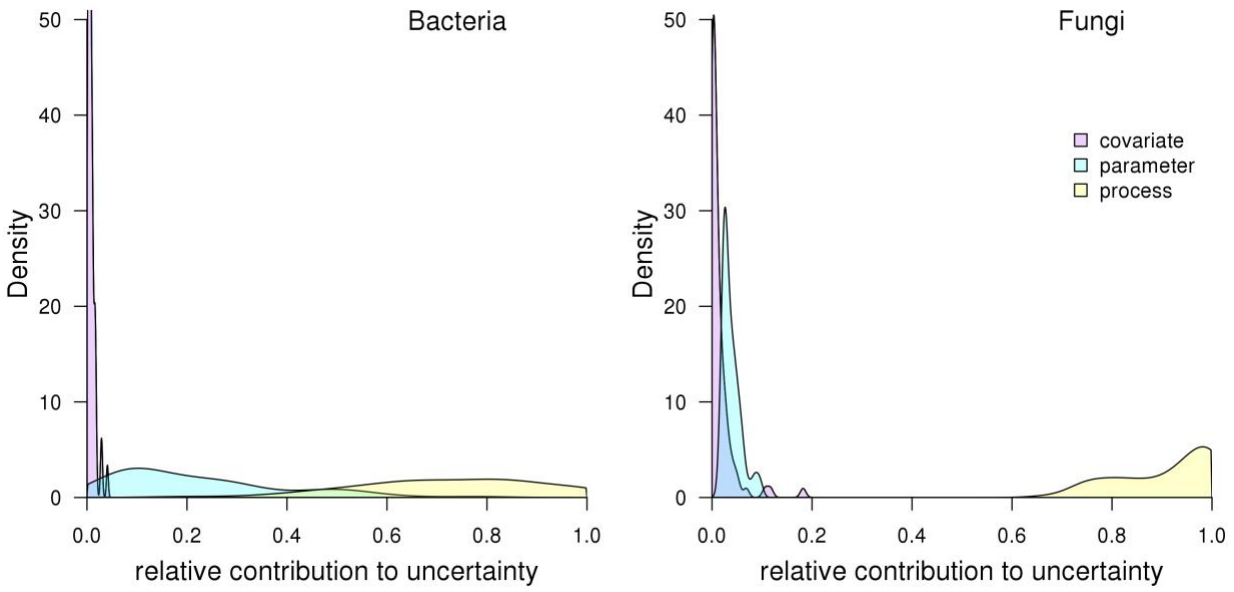
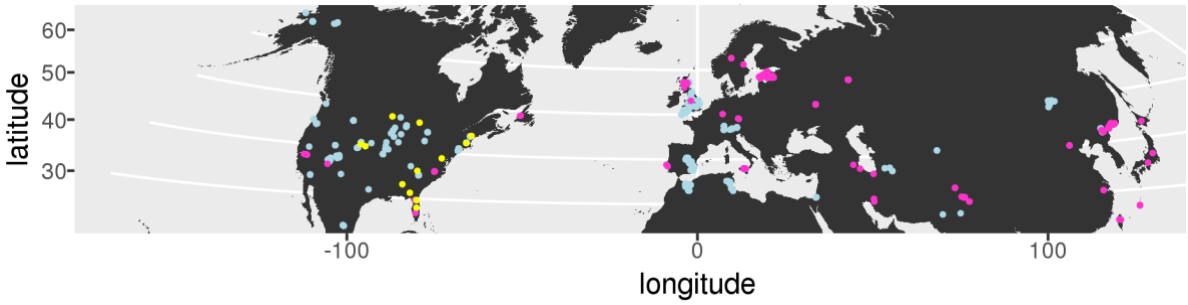


Figure 2. Cross-validation within the NEON dataset. Mean cross-validated R_2 relative to the 1:1 prediction across functional and taxonomic groups for (a) bacteria and (b) fungi.



Supplementary Figure 3. Density plot of variance decomposition for all (a) bacterial and (b) fungal groups modeled at the site level.



Supplementary Figure 4. 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.