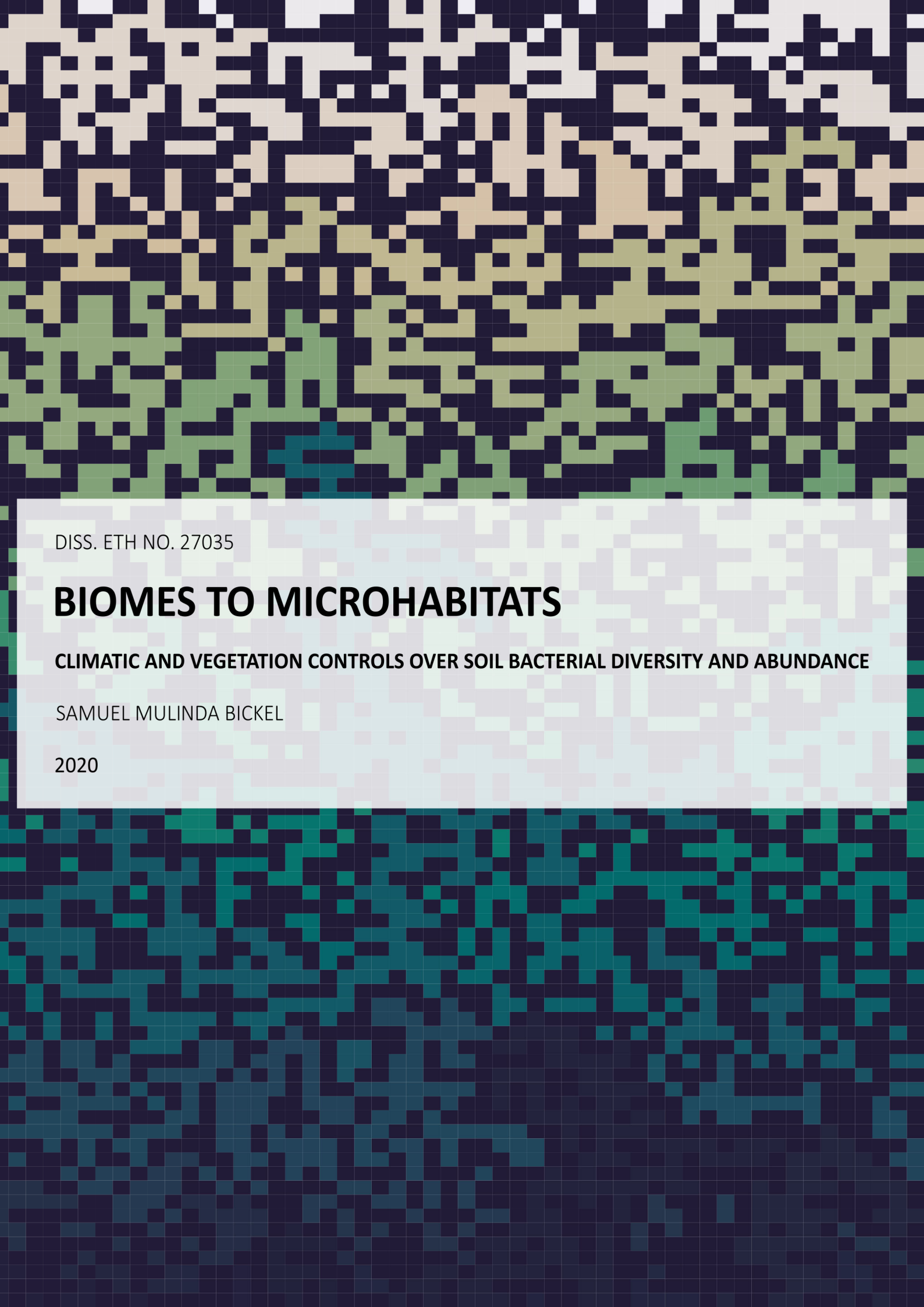


DISS. ETH NO. 27035

BIOMES TO MICROHABITATS

CLIMATIC AND VEGETATION CONTROLS OVER SOIL BACTERIAL DIVERSITY AND ABUNDANCE

SAMUEL MULINDA BICKEL

2020

DISS. ETH NO. 27035

BIOMES TO MICROHABITATS – CLIMATIC AND VEGETATION CONTROLS OVER SOIL BACTERIAL DIVERSITY AND ABUNDANCE

a thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

SAMUEL MULINDA BICKEL

MSc in Environmental Sciences (ETH Zurich)

born on *23.11.1987*

citizen of

Switzerland, Bubikon ZH

accepted on the recommendation of

Prof. Dr. Dani Or

Prof. Dr. Richard Bardgett

Prof. Dr. Noah Fierer

Dr. Naoise Nunan

2020

Content

Abstract 5

Zusammenfassung 6

Acknowledgement 7

Motivation 8

Introduction 9

1 Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes 13

1.1 Introduction 14

1.2 Results 16

1.2.1 Estimation of soil bacterial carrying capacity 16

1.2.2 Modeling bacterial diversity considering climate and soil 17

1.2.3 Species abundance distribution varies with hydration status 19

1.2.4 Global patterns of soil bacterial habitat diversity 20

1.2.5 Disentangling soil bacterial abundance and diversity 21

1.3 Discussion 22

1.4 Materials and Methods 25

1.4.1 Soil bacterial carrying capacity derived from NPP 25

1.4.2 Soil bacterial abundance dataset 25

1.4.3 Soil bacterial diversity datasets 26

1.4.4 Estimating soil specific 'climatic' water content 26

1.4.5 Estimation of aqueous habitat size distribution 27

1.4.6 Calculation of bacterial species diversity 29

1.4.7 Spatially-explicit individual-based model (SIM) 30

2 A hierarchy of environmental covariates control the global biogeography of soil bacterial richness 32

2.1 Introduction 33

2.2 Results and Discussion 34

2.2.1 Univariate analysis of bacterial richness 35

2.2.2 Multivariate general additive model (GAM) of bacterial richness 37

2.2.3 Varying proportions of low abundance species 38

2.2.4 Global patterns of soil bacterial richness 40

2.3 Conclusions 41

2.4 Materials and Methods 42

2.4.1 Data collection and processing 42

2.4.2 Metadata-based filtering 42

2.4.3 Primer-based filtering 42

2.4.4 Denoising 43

2.4.5 Taxonomy assignment for filtering of archaea 43

2.4.6 Rarefaction and estimation of diversity 44

2.4.7 Covariates 44

2.4.8 Correlation and clustering 45

2.4.9 Generalized additive models 45

2.4.10 Causal additive models 46

2.4.11 Prediction of global maps using tree-based algorithms 46
3 The chosen few – variations in common and rare soil bacteria across biomes 47
3.1 Introduction 48
3.2 Results 50
3.2.1 Relative abundance and prevalence of common and rare soil bacteria. 50
3.2.2 Rarity of soil bacterial species driven by climatic water contents. 51
3.2.3 How is bacterial species dominance reduced in dry soil? 53
3.2.4 Spatial patterns of bacterial rarity and functional consequences. 55
3.3 Discussion 56
3.4 Materials and Methods 58
3.4.1 Soil bacterial community data. 58
3.4.2 Classification of common and rare bacteria. 58
3.4.3 Climatic data of sampling locations. 59
3.4.4 Spatially-explicit individual-based model (SIM). 59
4 How soil bacterial microgeography affects community interactions and soil functions 61
4.1 Introduction 62
4.2 Results 65
4.2.1 Average cell density and community sizes linked to rainfall patterns and vegetation 65
4.2.2 Community size distribution based on spatial clustering of bacterial cells 66
4.2.3 Physical distances between bacterial communities limit trophic interactions 68
4.2.4 Variations in community sizes shape the proportion of anoxic bacterial communities across biomes 69
4.3 Discussion 71
4.4 Materials and Methods 74
4.4.1 Average cell density based on diffusion and distance to POM 74
4.4.2 Conversion of cell densities using soil particle surface area 75
4.4.3 Soil microcosm experiment 75
4.4.4 Image analysis for determination of cell locations 76
4.4.5 Spatially-explicit individual-based model of bacterial growth on soil particle surfaces 76
4.4.6 Clustering of proximal cells for estimation of community size distributions 77
4.4.7 Spatial cell aggregation model – community size distribution 77
Summary and Outlook 79
References 81
Curriculum Vitae 89
Appendix 91

Abstract

The diversity and abundance of soil bacteria are highly dynamic and vary considerably across scales and biomes with significant effects on soil ecological functioning. Soil bacterial communities are composed of a few abundant species, with most of their richness associated to rare species with largely unknown ecological roles. The thesis incorporates key environmental ingredients that affect soil bacterial abundance and diversity into a mechanistic modeling framework that links soil, climate and carbon inputs. The fragmentation of the soil aqueous-phase is directly linked to bacterial diversity found under different soils and climates. Soil bacterial diversity peaks at intermediate water contents with numerous aqueous habitats that remain well supplied by plant derived carbon. We employ statistical modeling of recent global soil bacterial datasets to test the dependency of bacterial richness on key soil and climatic attributes. Results confirm the well-established association of bacterial richness with soil pH and reveal a hierarchy among covariates. Climatic soil water content has been proposed to create links between aqueous micro-habitats and climatic conditions. Surprisingly, rare bacterial species that are present at low relative abundances exhibit high sensitivity to environmental conditions. A novel classification of common and rare soil bacteria suggests consistent changes of rarity as found in observations and predicted by the mechanistic model. Results show an increase in rare bacterial species proportions in drier soils with lower carbon inputs. A shift in bacterial species composition results from suppressed activity of common species leading to more even distributions of species abundances in arid soils. The novel modeling framework predicts general tendencies of soil bacterial abundance and diversity by considering microscale processes based on only few environmental variables. The results here pave the way for systematic incorporation of microscale processes and their effects on bacterial life across scales; from soil grain surfaces to terrestrial biomes.

Zusammenfassung

Die Diversität und Abundanz von Bodenbakterien sind sehr dynamisch und variieren beträchtlich über verschiedene Skalen und Biome hinweg mit signifikantem Einfluss auf die ökologische Funktion des Bodens. Bodenbakteriengemeinschaften setzen sich aus einigen wenigen, reichlich vorhandenen, Arten zusammen. Der grösste Teil ihrer Vielfalt besteht aus seltenen Arten mit weitgehend unbekannter ökologischer Rolle. In dieser Dissertation werden die wichtigsten Umweltfaktoren, welche die Diversität und Abundanz von Bodenbakterien beeinflussen, in ein mechanistisches Modellierungssystem integriert. Dazu werden Bodeneigenschaften, Klima und Kohlenstoffeinträge berücksichtigt. Die Fragmentierung der wässrigen Phase im Boden steht in direktem Zusammenhang mit der bakteriellen Artenvielfalt. Diese erreicht ihren Höhepunkt bei mittleren Bodenwassergehalten wo zahlreiche, isolierte, aquatische Habitate gut mit pflanzlichem Kohlenstoff versorgt bleiben. Statistische Modellierung von globalen Bodenbakteriendatensätzen testet die Abhängigkeit des Bakterienreichtums von wichtigen Boden- und Klimaattributen. Die Ergebnisse bestätigen die gut etablierte Assoziation der bakteriellen Artenvielfalt mit dem Boden pH-Wert und zeigen zudem eine Hierarchie zwischen den Umweltfaktoren auf. Der klimatische Bodenwassergehalt wurde vorgeschlagen um aquatische Mikrohabitate mit klimatischen Bedingungen zu verknüpfen. Überraschenderweise zeigen seltene Bakterienarten, die in geringen relativen Häufigkeiten vorkommen, eine hohe Empfindlichkeit gegenüber Umweltbedingungen. Eine neue Klassifizierung von häufigen und seltenen Bodenbakterien lässt auf konsistente Veränderungen der Seltenheit schließen, in Übereinstimmung mit empirischen Beobachtungen und wie sie durch das mechanistische Modell vorhergesagt wurden. Die Ergebnisse zeigen eine Zunahme des Anteils seltener Bakterienarten in trockeneren Böden mit geringerem Kohlenstoffeintrag. Eine Verschiebung in der Zusammensetzung der Bakterienarten ergibt sich aus der unterdrückten Aktivität der häufigen Arten, was zu einer gleichmäßigeren Verteilung der Artenhäufigkeit in trockenen Böden führt. Der neuartige Modellierungsansatz sagt allgemeine Tendenzen der Häufigkeit und Vielfalt von Bodenbakterien voraus, indem mikroskalige Prozesse auf der Grundlage nur weniger Umweltvariablen berücksichtigt werden. Die Ergebnisse ebnen den Weg für die systematische Einbeziehung mikroskaliger Prozesse und ihrer Auswirkungen auf das bakterielle Leben über verschiedene Skalen hinweg; von der Oberfläche von Bodenkörnern bis hin zu terrestrischen Biomen.

Acknowledgement

I want to express my greatest gratitude to Prof. Dr. Dani Or who provided me with the opportunity to graduate under his supervision. His critical thinking and mentorship challenged me to continuously improve my work and I have learned a lot; scientifically and personally. I am very thankful for the many interesting discussions we had and I truly appreciate the direct communication and swift feedback throughout my PhD. I would also like to thank Prof. Dr. Richard Bardgett, Prof. Dr. Noah Fierer and Dr. Naoise Nunan for taking the time to serve on the expert committee that enabled the nice discussion during the defense of my thesis. Further, I would like to thank Dr. Peter Lehmann, Dr. Andreas Papritz, Dr. Stan Schymanski and Dr. Robin Tecon for valuable inputs and scientific discussions. In addition, I am very grateful for the collaborations with Dr. Siul Ruiz and master students Xi Chen and Jingyu Wang. My special thanks are extended to Dani Breitenstein and Hans Wunderli for the excellent technical assistance in the laboratory and the pleasant conservations during numerous lunch and coffee breaks. I also want to thank the entire team at STEP for the time we could spend together. My deepest gratitude and appreciation go to my partner and wife Dr. Minsu Kim. Her unconditional support and understanding are most important to me. I am looking forward to our future family life with our daughter Lia and our son Kai. I am very thankful to my parents Irene and Christoph and my sister Flurina for their support. Lastly, I want to thank my friends for the many nice moments shared.

Motivation

Soil bacteria constitute a large proportion of the global biomass and their abundance and diversity are associated with soil ecological functioning and ecosystem services. How soil bacterial habitats at the microscale are shaped by soil type, climate and land-use remains uncertain. Insights into the fundamental processes that render soils functional and enable terrestrial life are of vital importance for understanding and sustaining ecosystems. Soils are the interface to a large fraction of the earth's biomass that, in turn, affect global water, carbon and energy fluxes. Interactions of biological agents within this complex, multi-phase environment affect the spatial variability of biomass across scales. Access to resources and dispersal distances are both related to average soil transport properties, which are controlled by climate and soil-type, shaping bacterial growth as well as the potential for interspecific interactions. A comprehensive study on bacterial life in soil requires microscale information for disentangling abundance and diversity. A single soil sample offers a vast living space for numerous bacterial communities with highly localized interactions within micro-habitats that preclude direct inference of soil microbiome functioning from macroecological patterns. The mismatch in spatial scales between climatic drivers and bacterial cells as living organisms on soil grain surfaces motivates the development of a novel modeling framework that preserves information of smaller scales and allows for prediction of soil bacterial abundance and diversity across terrestrial biomes.

Introduction

Soil carbon decomposition is largely controlled by carbon molecular composition, microbial biomass and the physical environment[1]. Diverse soil microorganisms carry a seemingly unlimited number of functional genes[2] that translate to yet uncertain soil metabolic potential[3]. Bacteria are a major component of belowground biomass[4] and dominate soil functional capacity[2] compared to other soil microorganisms. The soil microbial diversity has been directly linked to its functional capacity via measuring the sizes of the taxonomic and genetic pools[5] in which many specific functions are being carried by soil bacteria[2] often present at low abundances[6]. The abundance of soil microbial biomass carbon is related to mean annual precipitation and temperature[7] under constrained microbial cellular stoichiometry[8]. Similarly, soil bacterial biomass follows the carbon input by primary productivity that varies with soil depth[8,9]. An upper bound on soil bacterial cell density ('carrying capacity') can be obtained from consideration of carbon input fluxes by primary productivity that are also related to the distribution of rainfall and soil hydration conditions[9]. The soil hydration state affects bacterial activity by controlling cell transport and diffusion of carbon sources[10,11]; resulting in heterogeneous distributions of cell densities[12] at sub-millimeter scales.

Bacterial cells live in soil pores and on soil grain surfaces. Despite often very large numbers (10^{10} cells per gram of soil), they are rarely directly observed at the scale of their habitat. Water held in soil pores and in thin water films adsorbed to soil grains (here the 'soil aqueous-phase') is frequently fragmented and constitutes the heterogeneous living-spaces of soil bacteria that harbor many species and communities within these 'aqueous habitats'. The definition of ecological communities (a group of potentially interacting species) requires information on the nature of soil bacteria interaction potential, which depends on cell density and spatial configuration. Ecological interactions in competing bacterial communities alter bacterial diversity and are mediated by soil water content[5,9]. At the scale of bacterial habitats, the spatially heterogeneous soil-aqueous phase governs the distribution of local carrying capacity and causes non-linearities in observed carbon fluxes[13].

The macroscopically measured soil water content determines the sizes and numbers of spatially isolated aqueous micro-habitats that enable many co-existing bacterial communities or populations in a cubic centimeter of soil[9]. In addition, bacterial growth, motility and dispersal ranges are defined by the soil aqueous-phase connectivity that controls the interaction potential of bacterial species[14]. Constraints to dispersal dictate that in most soils (often unsaturated) we should expect clustered distributions of cells limited by growth that are concentrated around carbon sources ('hot-spots')[15]. The ubiquitous spatial isolation between such communities reduces opportunities for interactions among species and is mediated by the microscale distribution of aqueous habitats that itself is controlled by soil type and climatic rainfall patterns. This causes a discrepancy in spatial scales

between the soil bacterial habitats and measurements of bacterial abundance, diversity and ecosystem functioning. The mismatch in spatial scales precludes direct inference of interactions in the soil microbiome based only on macroscopic observation of diversity and abundance[16]. How climatic drivers and soil properties shape bacterial life in spatially distributed communities at small scales is the central question of interest.

In this work we seek to quantify the effects of soil aqueous phase connectivity on bacterial diversity and abundance across biomes. Information on biome characteristics, such as rainfall patterns and vegetation should link to physical properties of bacterial aqueous habitats; from climatic scales to microhabitats. This is important for placing empirical observations of bacterial abundance and diversity in context of the heterogeneous soil environment where they affect soil ecosystem functioning. The main objective of this thesis lies on the development of a modeling framework that preserves the information on bacterial microscale distributions to quantitatively predict macroscopic observations of soil bacterial diversity and abundance. Soil bacterial habitats are expected to be larger and better connected in wet soils that are also well supplied with carbon by vegetation along the soil profile (Fig. 1). This enables high soil bacterial abundance and leads to increased opportunities for interactions that shape bacterial diversity and should vary across biomes.

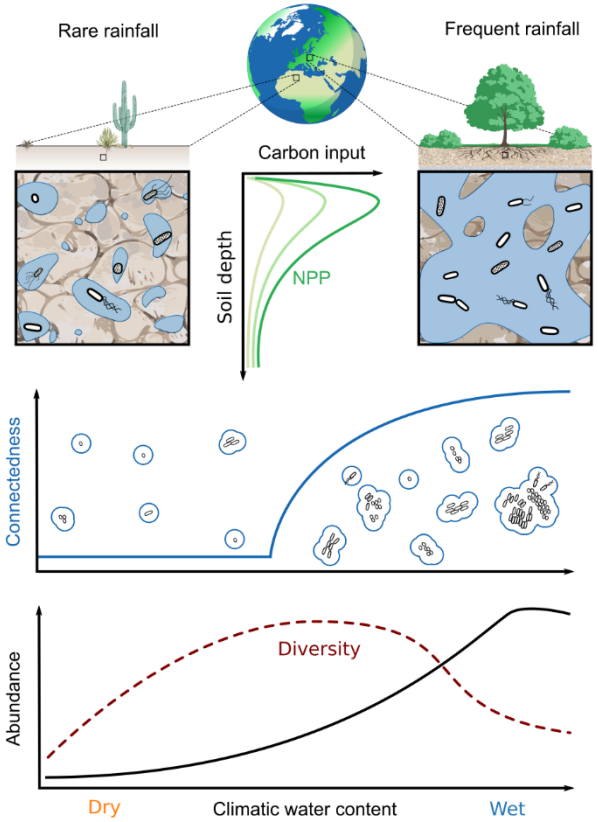


Fig. 1 Soil bacterial abundance and diversity are expected to vary across biomes. Rainfall patterns and carbon input by vegetation are characteristic to terrestrial biomes. Net primary productivity (NPP) that decays rapidly with soil depth maintains belowground bacterial abundance. Dry soils support small communities of bacteria that are spatially isolated. Increased aqueous phase connectivity leads to increased abundance that provides opportunities for interactions and is expected to affect bacterial diversity.

We capitalize on the generality and predominant role of the soil-aqueous phase in shaping bacterial habitats and on the availability of globally distributed data sets for spatial mapping and model evaluation. To determine a climatically representative soil water content ('climatic water content') we suggested a simple model based on the average frequency of rainfall and soil specific water storage that is subjected to evaporative losses[9,17]. This climatic water content changes gradually across biomes and serves as a proxy for the average soil aqueous phase connectedness. It also co-varies, to some extent, with primary productivity. The global distributions of the main variables and their relations are illustrated in Figure 2.

With estimates of soil bacterial carrying capacity, the climatic water content was used to predict the global distribution of soil bacterial diversity[9] (Chapter 1, *published*). A statistical analysis of soil bacterial biogeography supports a predominant role of climatic water contents and revealed a hierarchy of environmental covariates that can explain variations in bacterial richness[17] (Chapter 2, *published*). Detailed modeling of soil bacterial species abundance distributions and a novel classification of common and rare bacteria, enabled the discovery of systematic shifts in bacterial rarity with climatic water contents (Chapter 3, *submitted*). Furthermore, we could model and observe spatially clustered bacterial cell distributions on soil surfaces and developed relations that link macroscopic quantities (carrying capacity, water content) to the distribution of bacterial community sizes and interaction potential (Chapter 4, *submitted*). Finally, a summary and short outlook are presented.

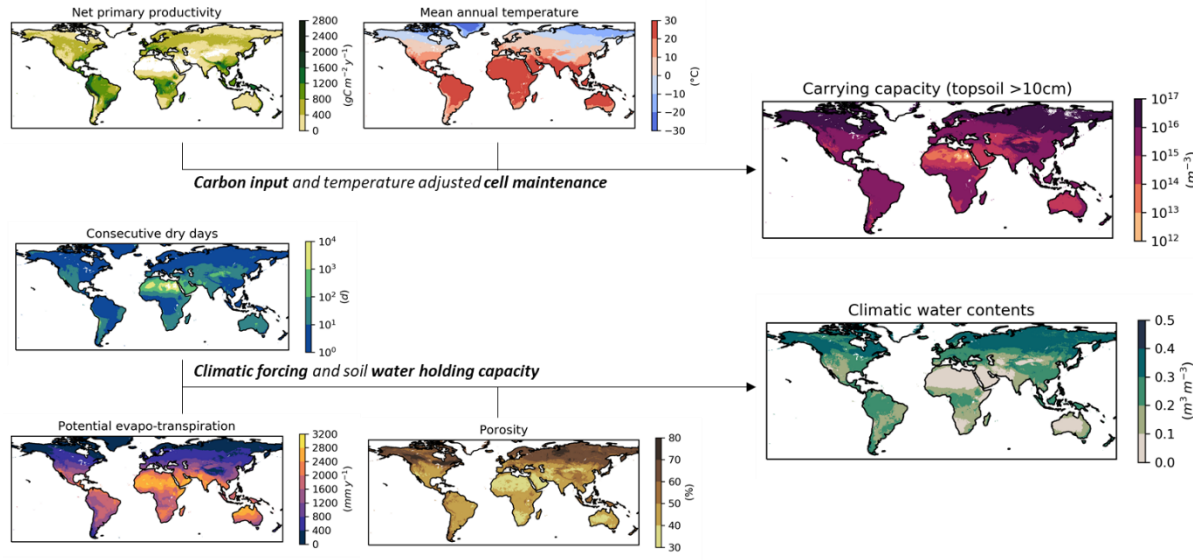


Fig. 2 Key variables used for modeling soil bacterial abundance and diversity. Carbon input by net primary productivity and temperature adjusted cell maintenance are used to estimate upper bounds on bacterial cell density ('carrying capacity'). Climatic forcing (number of consecutive dry days and potential evapotranspiration) together with soil porosity are used to estimate climatic water contents. Global data from several resources were compiled at 0.1° x 0.1° resolution (SoilGrids[18], WorldClim[19] and MSWEP[20]).

1 Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes

Samuel Bickel and Dani Or

Nature Communications 11, no. 1 (January 8, 2020): 1–9. https://doi.org/10.1038/s41467-019-13966-w.

Abstract

Soil bacterial diversity varies across biomes with potential impacts on soil ecological functioning. Here, we incorporate key factors that affect soil bacterial abundance and diversity across spatial scales into a mechanistic modeling framework considering soil type, carbon inputs and climate towards predicting soil bacterial diversity. The soil aqueous-phase content and connectivity exert strong influence on bacterial diversity for each soil type and rainfall pattern. Biome-specific carbon inputs deduced from net primary productivity provide constraints on soil bacterial abundance independent from diversity. The proposed heuristic model captures observed global trends of bacterial diversity in good agreement with predictions by an individual-based mechanistic model. Bacterial diversity is highest at intermediate water contents where the aqueous phase forms numerous disconnected habitats and soil carrying capacity determines level of occupancy. The framework delineates global soil bacterial diversity hotspots; located mainly in climatic transition zones that are sensitive to potential climate and land use changes.

1.1 Introduction

Soil hosts unparalleled bacterial diversity, ranking highest among all other compartments of the biosphere[5,18,19]. The number of bacterial phylotypes ranges between 10^2 to 10^6 per gram of soil[5,18,20], with high values similar to the diversity in all of earths environments[19]. This immense richness is often attributed to soil's intrinsically heterogeneous physical and chemical micro-environments[21–25]. The complex structure of soil pores offers numerous refugia for hosting diverse bacterial species[25]. This wide range of microhabitats is particularly important for maintaining the rare components of the soil microbiome. Low abundance bacterial species play important roles in key biogeochemical processes[6,26] and serve as a 'seed bank' for species richness[27]. Microbial diversity is manifested both at the scale of soil grains[24] and at very large scales across climatic regions and terrestrial biomes[5,28,29]. These observations often include variations in microbial biomass that responds to resource availability and affects bacterial diversity at all scales[30–32]. For example, well-established observations of microbial abundance variations with soil depth[8] could confound inferences of bacterial richness by promoting the detection of low abundant species in resource rich environments.

Quantifying the roles of soil factors, such as soil texture, porosity and hydration conditions in relation to climate and vegetation cover is an important step towards disentangling bacterial diversity and abundance as suggested by recent empirical evidence[32]. Soil chemical properties such as pH[5,29,32,33] and organic carbon content[30–32] together with climatic attributes, such as aridity index[30], precipitation[5,32] and temperature[28], have been identified as important explanatory variables. Yet, the rapid expansion of soil bacterial diversity datasets has not been met with similar development of predictive models for interpretation of the observed spatial patterns[34]. Improved predictability of soil bacterial diversity could be essential for understanding soil bacterial functioning; from contributions to soil respiration[26,35] to the resistance of bacterial communities to invasion by pathogens[36].

Such endeavors invariably require development of mechanistic frameworks for systematic incorporation of the various factors that affect soil bacterial diversity. In this study, we capitalize on recent empirical[5,24,28,30,32,37] and theoretical developments[23,38,39] to generalize the role of soil aqueous microhabitat fragmentation and its nearly universal role in mediating bacterial diversity across soil types and climatic conditions. To characterize the average conditions in soils and facilitate long-term predictions, we define a soil climatic water content that combines rainfall patterns and volumetric soil water holding capacity into a well-defined attribute. This measure considers the average duration between soil wetting events important for diversity maintenance (see Methods). Under a wide range of climatic conditions, soils remain unsaturated with the bacterial aqueous habitats fragmented to varying degrees based on soil type and rainfall dynamics (amount and frequency). A critical hypothesis is that the microscale arrangement of water retained in soil pores defines the size distribution and

connectedness of aqueous bacterial habitats that, in turn, affect diffusion rates of substrates, the rates and spatial extents of cell motility[39,40] and opportunities for cell-to-cell interactions[41]. The objective of this study was to formalize the influence of these abiotic factors in a heuristic framework that enables quantitative representation of soil bacterial abundance and diversity at scales ranging from grains to watersheds and beyond.

The core of the model is the quantification of numbers and sizes of aqueous bacterial habitats considering climatic water contents and soil types. We use concepts of percolation theory to describe the size distribution of aqueous patches[38] that could support bacterial cells. Soil organic carbon input flux, derived from the net primary productivity (NPP), and mean annual temperature (MAT) are used to estimate a soil carrying capacity that defines limits for the abundance of bacterial cells (Fig. 1.1). For simplicity, we first assume that each isolated aqueous patch is inhabited by a single bacterial phylotype (hereafter referred to as 'species'). This heuristically enables estimation of bacterial diversity based on the species abundance distribution (SAD) deduced from the size and number distribution of microscale aqueous habitats. The framework expresses soil bacterial diversity at two interlinked spatial scales: at the single aqueous habitat scale and at the soil sample scale that can contain many isolated aqueous habitats.

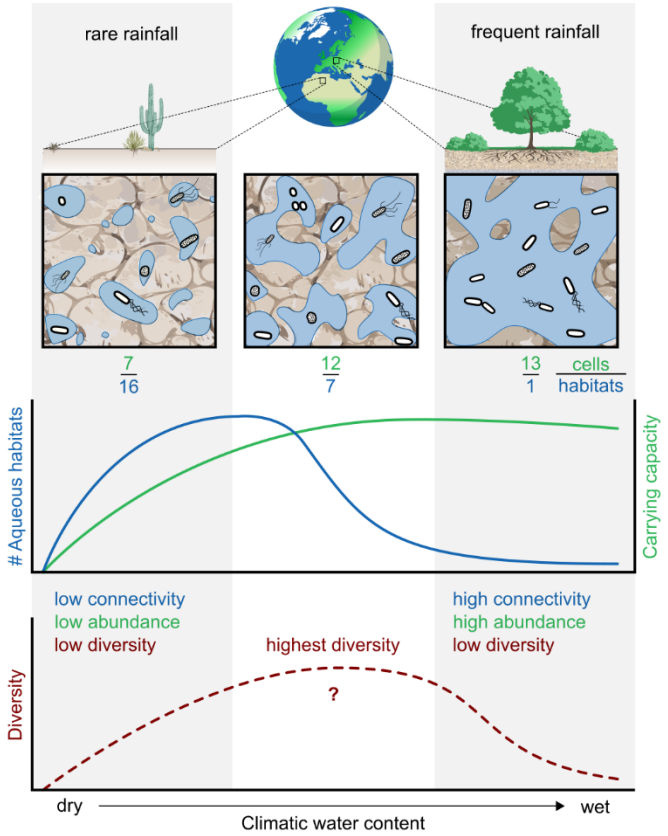


Fig. 1.1 Illustration of aqueous habitat fragmentation and carrying capacity in relation to climatic water contents. In regions where rainfall is frequent, the soil aqueous phase is largely connected and provides a common habitat for cells of different bacterial species. In soils of drier regions, the aqueous phase is increasingly fragmented and offers a large number of distinct habitats. When the soil becomes sufficiently dry almost all aqueous habitats are physically isolated and might contain only a few species. Additionally, the total number of cells that can be maintained (potential carrying capacity) is reduced and smaller patches become uninhabited. The specific carrying capacity in a biome is based on carbon input flux and temperature that establish an upper bound on bacterial cell density (rarely realized in any particular location due to other limiting factors). The numbers below each panel indicate the number of cells per number of habitats. Diversity is expected to drop in dry regions with low cell abundance and in wet regions with high habitat connectivity.

Modeled trends of soil bacterial carrying capacity and diversity are compared to empirical observations[8,18,20] across terrestrial biomes and suggest a peak in bacterial diversity at intermediate climatic water contents. To evaluate predictions by this aqueous-phase fragmentation-based heuristic model (HM), we employ a detailed, spatially-explicit individual-based model (SIM) that mechanistically simulates bacterial communities growing on hydrated soil surfaces[23,39]. The SIM enables systematic variations of hydration conditions and tracks the growth and life history of multiple species interacting on soil grain surfaces (see Materials and Methods).

The simple HM does not differentiate between the roles of legacy and environmental conditions in shaping soil bacterial diversity. As evidenced from the choice of climatic averaging and the implicit representation of species with no taxonomic attribution, the focus lies on the role of aqueous habitats and their average connectivity. Other factors at play such as soil chemistry and the presence of larger organisms are not modeled. We refer to 'microbes' for aspects that apply to all microbial life in soil (bacteria, fungi, protists and viruses), and specifically to bacteria for modeling and quantification of diversity and abundance. Summarizing, we propose a hydration-centered modeling framework that considers the interplay of climatic water content; carbon input flux and temperature in shaping soil microhabitats and thus bacterial diversity.

1.2 Results

1.2.1 Estimation of soil bacterial carrying capacity

We evaluated theoretical estimates of soil bacterial carrying capacity using previously published measurements of soil microbial carbon[8]. The heuristic model (HM) assumes that a certain proportion of the annual NPP-derived organic carbon input is allocated to bacteria (24% of NPP for bacterial respiration[42,43]). We found that varying the range of expected values (14-30% of NPP[42]) had little impact on estimates of carrying capacity. A constant value of this respiratory fraction was therefore considered based on mechanistic model simulations[42]. We employ a basic estimate of bacterial cell maintenance rate of 1.5 gC gC_{cell}^{-1} y^{-1} (≈10^{-4} gC gC_{cell}^{-1} h^{-1}) and adjust it according to the local mean annual temperature (MAT)[44] to account for different climatic regions. Combining local annual NPP and adjusted cell maintenance rate, we derive estimates of soil bacterial carrying capacity as upper bounds for soil bacterial cell density (Fig. 1.2 a). Despite the many simplifying assumptions, we obtain reasonable estimates of potential soil bacterial carrying capacity that are comparable with observations of realized bacterial cell density across a range of environmental conditions. Model estimates of soil carrying capacity for three values of MAT are depicted in Figure 1.2 a (representing the median of three groups: ≤ 0 °C, 0 - 15 °C, > 15 °C with -2, 9 and 19 °C, respectively). Observed cell densities tend to be higher for colder regions as considered by the HM. We note that soil bacterial cell density is expected to vary with soil depth due to the distribution of organic carbon flux from the soil

surface and distribution by plant roots[8]. Soil bacterial carrying capacity decreases steeply with depth and was represented parametrically by a lognormal distribution (μ = 0.18, σ = 1.00) (Fig. 1.2 b). The lognormal distribution provided a better global representation of the average topsoil carrying capacity (upper 10 cm, A1 Supplementary Figure 1) over the previously reported exponential model[8]. It is important to keep in mind that the estimated soil carrying capacity was defined independently from bacterial diversity and values were calculated globally based on NPP, MAT and soil depth.

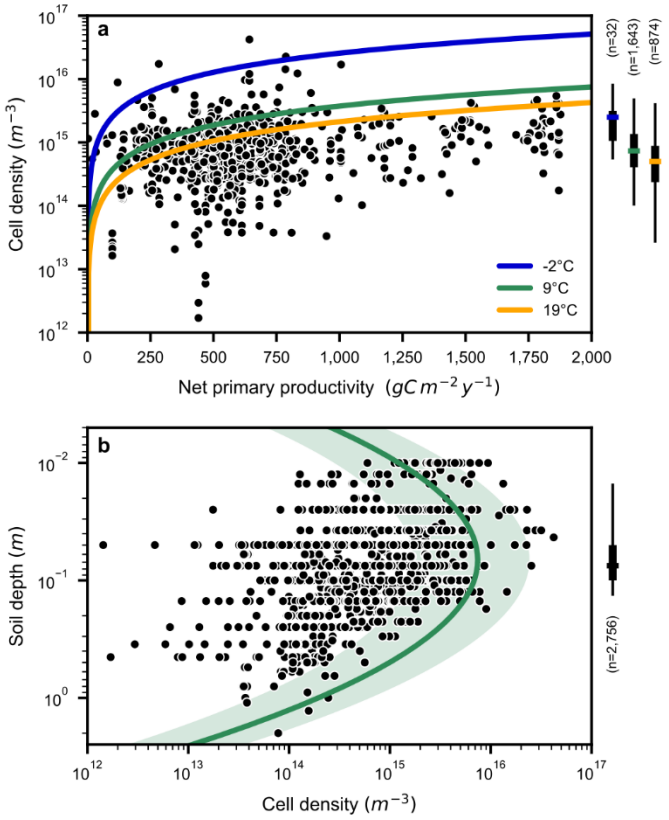


Fig. 1.2 Soil bacterial abundance varies in relation to carbon input, temperature and soil depth. a, Bacterial cell density at soil carrying capacity as a function of net primary productivity (NPP) with model estimates sensitive to mean annual temperature (MAT) (solid lines). Estimates are compared with measured data of microbial biomass[8] converted to bacterial cell density and are grouped by temperature (MAT ≤ 0 °C, 0 °C > MAT ≤ 15 °C, MAT > 15 °C). Each group's median is reported in the figure legend in blue, green and orange, respectively. The distributions of cell densities are indicated for each temperature group as the central 50 and 95% range. **b**, Variations of bacterial cell density with soil depth. The log-normal fit provides bounds on cell density (carrying capacity) for intermediate MAT (solid line) and for the central 95% of NPP (shaded area). Observed estimates of cell density are reported for their average sampling depth. Most samples were taken above 10 cm as shown in the boxplot. Source data are provided as a Source Data file.

1.2.2 Modeling bacterial diversity considering climate and soil

The simple heuristic model (HM) was developed in two conceptual steps. We first assumed only a single species per aqueous habitat. This approach, although useful as a heuristic, exhibited some limitations for large aqueous habitats under wet conditions (see comparison of species abundance distributions below).We thus adapted the model to allow multiple species in large habitats by assigning the number of species N_{sp} proportional to the length scale of a habitat of size s ($N_{sp} \sim s^{1/d}$, d = 2 or 3 = dimensionality). Hence, the HM links species richness to the soil aqueous-phase fragmentation via percolation theory and accommodates the possibility of multiple species per habitat. For most unsaturated conditions the refined formulation does not alter the prediction since small habitats are likely to host only a single species. In the following we refer to the multispecies HM if not stated otherwise. We have used median values of global soil carrying capacity to describe trends

in soil bacterial diversity across soil types and climatic regions. Comparisons of model estimates with empirical observations of bacterial richness obtained from the studies of Thompson *et al.* (EMP)[18] and Delgado-Baquerizo *et al.* (DEL)[20] are depicted in Figure 1.3 along with the mechanistic predictions by the SIM. We have expressed mean soil hydration status via the climatic water content that is a proxy for average soil wetness and habitat connectivity. Soil and climatic variables were compiled from different sources (A1 Supplementary Table 1) with matched geographical coordinates and soil depths for the samples. We present soil bacterial richness (total number of types) and note that taxonomic assignment was absent for the phylotypes detected in EMP. Bacterial richness was binned by water contents because some hydration conditions were overrepresented (bin width: 0.05). Since richness in the EMP data was measured at different soil depths, they were also grouped to top and sub-soil (<25 cm and ≥25 cm). Exact number of samples per group are reported in A1 Supplementary Table 2. The EMP data displays a tendency of lower values of richness in the sub-soil (Fig. 1.3 a). In the DEL dataset, measurements were taken at the same soil depth, and soil pH is reported instead (Fig. 1.3 b). We observe a strong tendency of lower soil pH in climatically wetter soils. The results depict an average decrease in bacterial richness where the soil becomes saturated as also predicted by the HM for median soil carrying capacity (Fig. 1.3 a and b). The modeled sensitivity to soil carrying capacity is shown for a scenario of reduced cell densities (e.g. less carbon input to deeper soil layers; Fig. 1.3 a – dashed line). We emphasize that parameters were not fitted to observed diversity data, but rather are based on mean values for soil properties (porosity θ_s = 0.49 and 0.47; sample length L = 5 and 6 mm; textural length δ = 0.07 and 0.1 mm; for EMP and DEL, respectively). Additionally, we used a fixed value for the critical water content ($\theta_c \approx 0.15$) and a threshold for the number of cells N_{cell} needed to model occupancy of potential habitats ($N_{cell} > 4000$). Lastly, we compared the aqueous-phase fragmentation-based HM to numerical simulations of the SIM. We simulated the spatially-explicit growth and movement of individual cells in a diverse bacterial community on heterogeneous soil pore surfaces. Qualitatively, both HM and SIM predict similar trends of variations in bacterial richness with soil hydration conditions as estimated from the EMP and DEL datasets (Fig. 1.3 a and b). In addition to removing single cells (singletons) from the simulated communities, the modeled species counts were rarefied to 5000 and 1000 for comparison with EMP and DEL, respectively. To compare with the DEL dataset, simulated bacterial richness is reported only for the 512 most abundant species and describes the observed invariance of richness towards low climatic water contents (Fig. 1.3 b). The discrepancy in water contents where richness peaks (between HM and SIM) is attributed to the dimensionality of the models (three for HM, two for SIM) and is well captured by the percolation-based HM in two dimensions (A1 Supplementary Figure 2).

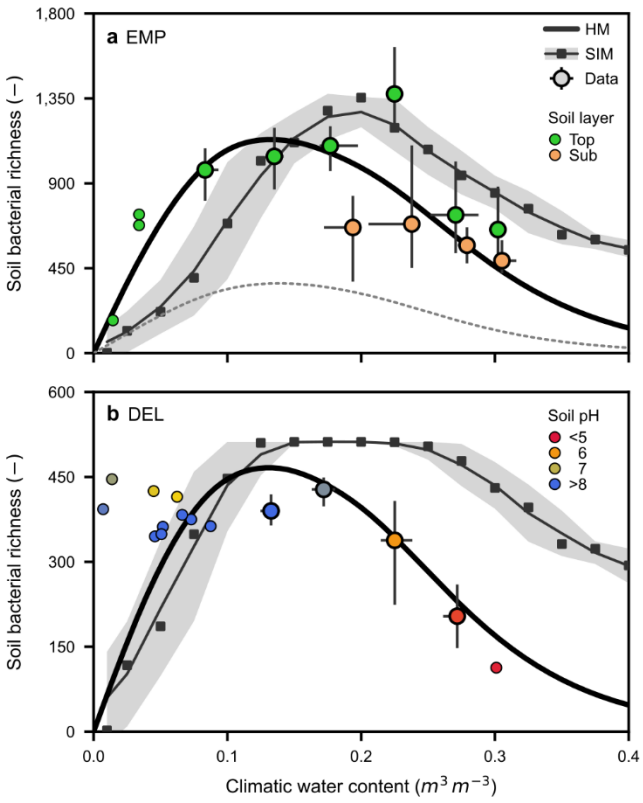


Fig. 1.3 Observed and predicted variations in soil bacterial diversity with climatic water content. **a** and **b**, Estimates of bacterial richness from two studies are binned by climatic water contents (bin width: 0.05) and the median and interquartile range are reported (circles and bars, respectively). The exact number of samples per group are listed in Supplementary Table 2. Individual data points are shown for bins containing less than ten samples (small circles). The solid black lines correspond to predictions by the fragmentation-based heuristic model (HM) for median carrying capacity specific to each dataset. The square symbols, thin solid line and shading (mean, rolling mean ± SD, n = 12) depict simulated bacterial richness using the spatially-explicit individual-based model (SIM) for different water contents. **a**, Bacterial richness from the Earth Microbiome Project (Thompson *et al*. - EMP)[21] was reported for different soil depths and thus grouped accordingly (<25 and ≥25 cm, top- and subsoil, respectively). The dashed line represents a model scenario with reduced carrying capacity by considering only the subsoil. **b**, Soil bacterial richness from a recent study (Delgado-Baquerizo *et al*. - DEL)[23]. Colors indicate reported soil pH, which has been shown to be affected by climate[52]. For comparison with the DEL dataset, only the top 512 species were considered in the SIM. Source data are provided as a Source Data file.

1.2.3 Species abundance distribution varies with hydration status

We quantified variations in bacterial species abundance distribution (SAD) with soil attributes and climatic water contents in comparison with empirical estimates from the EMP and DEL datasets (A1 Supplementary Figure 3). Here we used soil properties and carrying capacity specific for each geographical location and soil depth. The results show good alignment of the single species model predictions with observed relative SADs and resulted in Pearson correlation values of 0.84 (n = 230) and 0.76 (n = 218) for the EMP and DEL datasets, respectively (A1 Supplementary Figure 3 a and b). Nevertheless, the single species HM erroneously predicts a higher proportion of the most abundant species than observed. We attribute this systematic overestimation to the stringent assumption of one single species per aqueous (micro-) habitat. This discrepancy suggests that the single species per aqueous habitat assumption may not hold for very large aqueous habitats in wet soil that could host multiple species. To rectify this limitation, we considered a scenario where the number of species N_{sp} is assumed proportional to the size s of an aqueous habitat ($N_{sp} \sim s^{1/3}$). This relaxed occupancy assumption improved Pearson correlations to values of 0.88 (n = 230) and 0.84 (n = 218) for the EMP and DEL datasets, respectively (A1 Supplementary Figure 3 c and d). Predictions by the HM for ranked SADs compare qualitatively with observations that were grouped by average hydration conditions (A1

Supplementary Figure 4). An increase in dominance of the most abundant bacterial species is visible in the ranked SADs of both datasets under sufficiently wet conditions (A1 Supplementary Figure 4 b and c).

1.2.4 Global patterns of soil bacterial habitat diversity

Motivated by the general agreement with observations of bacterial richness and the SADs produced by the HM, we used highly resolved global datasets for soil properties, NPP and precipitation as inputs to estimate global patterns of soil bacterial habitat richness (Fig. 1.4 a). Recall that a central element of the model is the link between the number of distinct aqueous habitats per soil volume and soil bacterial richness. Additionally, we considered the sizes of aqueous habitats to yield spatially resolved distributions of the Shannon index of bacterial diversity patterns (Fig. 1.4 b). We note that the modeled soil bacterial diversity follows constraints imposed by local soil carrying capacity where high bacterial cell numbers are associated with locally high NPP and low cell maintenance requirements. Both diversity indices exhibit spatial patterns with distinct regions of increased diversity associated with climatic transition zones (e.g., the Sahel). This pattern is more pronounced when considering the Shannon index and suggests that soil bacterial community evenness, indicative of how equally habitats are partitioned, is sensitive to soil wetness. Such an association is also observed empirically where evenness decreases with increasing climatic water contents (Pearson r = -0.17 and -0.43 for EMP and DEL, respectively; A1 Supplementary Figure 5a).

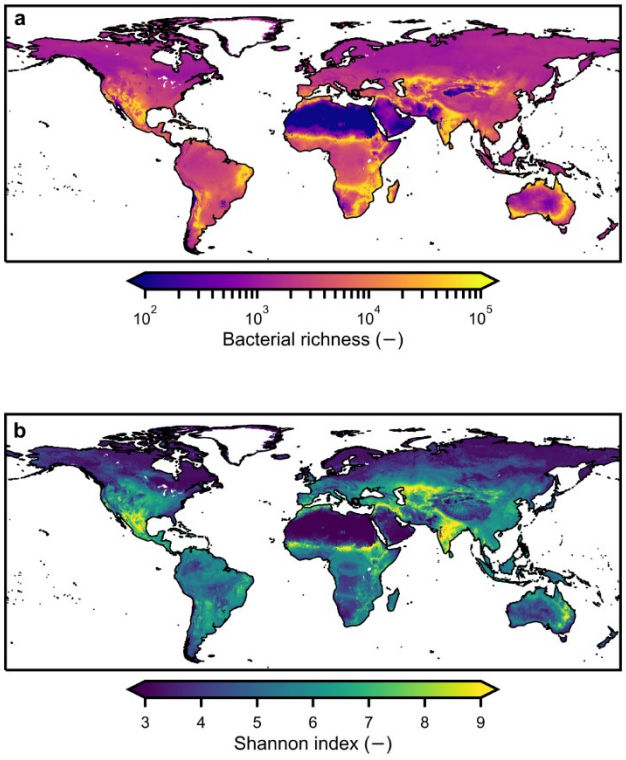


Fig. 1.4 Modeled global biogeography of soil bacterial diversity. Global patterns are modeled based on aqueous microhabitats in the top 25 cm considering climate, NPP and soil type. **a**, Global map of predicted soil bacterial richness. High values correspond to more heterogeneous soil environments, potentially harboring a larger number of habitats. **b**, Global distribution of Shannon index for estimated bacterial diversity. In addition to richness, the Shannon diversity index considers the relative abundance of unique habitats. Higher values of the Shannon index could translate to more even bacterial communities.

1.2.5 Disentangling soil bacterial abundance and diversity

To address the challenge of disentangling bacterial abundance and diversity, we compared bacterial community evenness with climatic water content and carrying capacity (Fig. 1.5). Evenness decreases gradually with climatic water content and with increasing soil carrying capacity (Fig. 1.5, A1 Supplementary Figure 1.5 b). The results are consistent with the tendency of wetter conditions being associated with an increase in cell densities and was confirmed (with no prior assumptions) using detailed mechanistic modeling (SIM) for small spatial and short temporal scales (A1 Supplementary Figure 6). In the aqueous-phase fragmentation-based HM, predicted bacterial cell densities are independent of climatic water contents. This could result in unrealistic values relative to empirical observations. We therefore used pairs of values for carrying capacity and climatic water contents to constrain the HM for evenness prediction (Fig. 1.5). Considering the relation between climatic water content and soil carrying capacity highlights the sensitivity of HM predictions to bacterial cell density as also observed in the mechanistic simulation results of the SIM. The dependency of cell density on climatic water content in the SIM results in a persistent decrease of evenness with increasing water content (A1 Supplementary Figure 7). When considering paired values of water content and cell densities obtained from the SIM, the simpler HM captures the simulated trends reasonably well (A1 Supplementary Figure 7). Although beyond the scope of this study, we observed that pre-processing measurements of relative species abundance may affect diversity metrics such as richness and evenness, which alters the apparent tendencies (A1 Supplementary Figure 8).

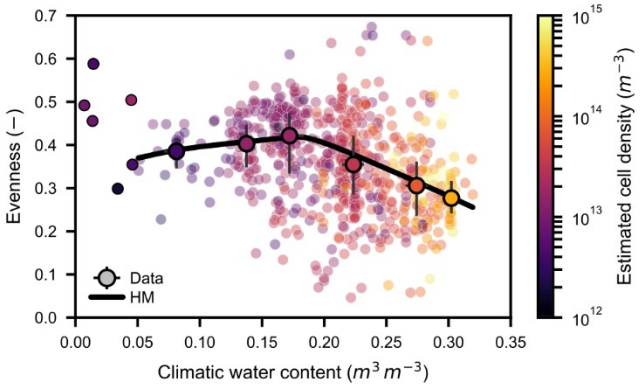


Fig. 1.5 Bacterial community evenness decreases with carrying capacity and climatic water contents. Evenness from two independent studies is shown together with estimated cell density (carrying capacity). Samples were aggregated by latitude, longitude and soil depth (EMP[21], n = 484 and DEL[23], n = 218). The median and interquartile ranges (colored symbols and bars) are displayed for groups of water contents (bin width: 0.05, number of samples see Supplementary Table 2). Individual data points are shown for bins containing less than ten samples (small circles) and samples with cell density lower than 10^{12} m^{-3} were removed. Evenness predicted by the heuristic model (HM) is calculated using paired values of climatic water content and carrying capacity (evaluated for every sample). Using the joint data of water content and cell density as model input, the HM reproduces the observed tendency of evenness. A locally weighted scatterplot smooth (LOWESS) of modeled evenness is shown for the HM predictions (solid line). Source data are provided as a Source Data file.

1.3 Discussion

The heuristic nature of the aqueous-phase fragmentation-based model (HM) precludes comparison of bacterial richness and abundance on a per sample basis, as climatic assumptions and associated large-scale variables are not likely to apply at a particular sampling location and time. Nonetheless, the proposed HM captures the salient features of global trends in bacterial richness related to climate, biome and soil type. Our estimate of soil bacterial cell density represents an upper bound on soil bacterial abundance (carrying capacity) and shows general agreement with measurements of soil bacterial biomass carbon[8]. It tracks the temperature dependency of reaction rates[44] and provides an independent measure of maximal cell density that is sensitive to climate and organic carbon input by vegetation. Bacterial diversity increases towards lower values of climatic water contents (i.e., with increased aridity[30]), as long as soil bacterial life is not limited by low organic carbon input. Assuming a constant soil bacterial carrying capacity, we can attribute much of the variations in bacterial richness to the microscale behavior of soil hydration conditions (Fig. 1.3). Surprisingly, the trends of bacterial richness for both surveys EMP[18] and DEL[20] were very similar despite their different objectives and processing protocols of the genetic information; namely the use of amplicon sequence variants in EMP and operational taxonomic units in DEL (Fig. 3 a and b). We note that the values of bacterial richness in the DEL dataset saturate towards lower values of climatic soil hydration (Fig. 1.3 b). This is likely due to the truncation of species richness used in that study which focuses on the most abundant soil bacteria[20]. These, highly abundant species, might be the last to disappear under reduced carrying capacity and therefore do not show a decline towards dry conditions. The data available at low climatic water contents are sparse and do not provide support for the predicted steep decline of bacterial diversity as soil becomes dry that was previously reported with increased aridity at large scales[30]. However, a significant decrease in bacterial richness was also observed in a recent statistical meta-analysis for climatic scales[17] and could be confirmed using the SIM (Fig. 1.3 b). Additionally, it has been reported that bacterial diversity declines sharply with moisture in dry soils of Antarctica[37] and decreases with soil relative humidity along transects of the Atacama desert[45]. Microcosm experiments revealed an increase in richness with moisture that peaks at intermediate water contents that promote rare bacterial species[46]. Similarly, bacterial richness was highest at intermediate climatic water contents where isolated aqueous habitats are numerous and sufficiently well supplied by diffusion to realize the soil carrying capacity (Fig. 1.3). This observation is supported by the mechanistic simulation results of the SIM, which explicitly considers the dynamics and spatial structure of the bacterial community (Fig. 1.3). The generality of the aqueous-phase fragmentation-based approach permits comparison of systems with different dimensionality and can account for the

shift of maximal richness towards higher water contents when comparing the HM with the two-dimensional simulation of bacterial life on hydrated surfaces by the SIM (A1 Supplementary Figure 2). Increasing the organic carbon input and thus soil bacterial abundance seems to support higher diversity of soil microorganisms[30]. This is in line with the observation of decreasing bacterial richness with soil depth (Fig. 1.3 a) that is often attributed to diminishing carbon inputs with depth (Fig. 1.2 b). However, considering the various interacting factors at play, the general picture might be more complicated. An increase in soil carrying capacity may not necessarily translate to increased bacterial diversity as evidenced by declining community evenness (Fig. 1.5, A1 Supplementary Figure 5). This could be due to dominance of a few species that may cluster near nutrient hot-spots[47], or loss of oligotrophic species that would be outcompeted in well-connected and dense communities[25]. We observe sensitivity of bacterial evenness to climatic water contents (Fig. 1.5), also in relation to soil carrying capacity (A1 Supplementary Figure 5). However, care should be taken regarding the interpretation of bacterial richness and evenness, since biases introduced by data processing and sampling could depend on the shape of the underlying SAD (A1 Supplementary Figure 8). Mechanistic models, such as HM and SIM, are valuable tools to investigate such dependencies as illustrated by considering only the most abundant species (Fig. 1.3 b) or increasing sampling effort and removing species present at low abundance (A1 Supplementary Figure 8 a and b, respectively). Nonetheless, an inherent tradeoff between availability of nutrients and protection by spatial isolation appears to play an important role in the establishment and maintenance of high soil bacterial diversity[17,32,47]. In other words, the relation between bacterial abundance and diversity is only positive when the aqueous phase is fragmented and spatial isolation suppresses the dominance of few species. As aqueous microhabitats become connected following soil rewetting by rainfall or irrigation, competition and other trophic interactions between bacterial cells are likely to reduce soil bacterial diversity (Fig. 1.3 a and b) by reducing the communities evenness (Fig. 1.5). Many other factors such as pH[5,18,29,33], nutrient composition[21], carbon sources distribution[22,47], stoichiometric constraints[8,29] and metabolic dependencies[48] shape soil bacterial abundance and diversity and could contribute to the discrepancy between our HM and empirical observations. Our study suggests that some of those factors might be associated with climatic hydration conditions. Interestingly, we find that soil samples exhibiting high bacterial diversity at intermediate climatic water contents coincide with near neutral pH values. In contrast, samples at low and high climatic water contents show high (basic) and low (acidic) pH tendencies, respectively (Fig. 1.3 b). This is supported by studies that relate soil pH with differences in soil water balance at climatological timescales[49]. We consequently expect soil pH to result from differences between precipitation and evapotranspiration as described by climatic water contents (A1

Supplementary Figure 9). Teasing apart such confounding associations requires detailed statistical analysis and experimental validation, which are best conducted in dedicated studies.

Using a single parameter set, largely based on standard percolation theory combined with data on soil properties, our HM predicts SADs that closely resemble empirical observations (A1 Supplementary Figures 3 and 4). Nevertheless, the increased aqueous-phase connectedness in climatically wet soils may also promote interactions that are suppressed under spatial isolation of dry conditions[37]. Processes that support bacterial species coexistence across small distances are not captured by the present model and would result in persistent underestimation of bacterial diversity (unless provisions are introduced as done for very large aqueous habitats – see A1 Supplementary Figures 2 and 3). Another inherent limitation of the analyses presented here is the focus on soil bacteria ignoring the interplay with other soil microorganisms that comprise Earth's microbiome[34]. For example, fungi could play an important role in modifying soil bacterial habitats[5] and are currently only considered in the partitioning of microbial carbon.

The framework presented in this study captures the salient spatial trends in soil bacterial diversity at climatic timescales and provides insights into effects of habitat fragmentation on the prevalence of bacterial interactions in natural soil. This is particularly important for the interpretation of species co-occurrence and interspecific interactions[48]. Such interactions between different species become possible only for conditions supported by the soil aqueous-phase connectedness[37]. This promotes diversity by enabling macroscopic co-existence[21,23,38] in soil bacterial communities competing for space and a common resource.

A unique aspect of the HM is the ability to bridge scales from soil pores to biomes where information at both scales is preserved. Further investigations are required to test some of the model implications at different scales. For example, elucidating the dependency of cell microscale distribution on soil type and hydration conditions could provide insights into the processes shaping bacterial interactions in soil. Additionally, taking into account factors affecting the partitioning of carbon at the ecosystem scale could refine model estimates of bacterial abundance beyond potential carrying capacity. Nonetheless, modeling climate and soil specific bacterial diversity offers a useful reference for comparing effects of climatic shifts (e.g. in temperature, precipitation) or land use change (e.g. in intensity of agricultural management or restoration to natural ecosystems) on soil bacterial communities that could guide future exploration of the soil bacterial micro- and macro geography.

1.4 Materials and Methods

In the following, we provide a detailed overview of the methods used in the study and list key assumptions. Although the heuristic model (HM) uses a yearly timescale for climatic averaging, the framework could be applied to finer and more resolved datasets. The global predictions of soil bacterial diversity were based on a 0.1°x0.1° grid to harmonize raster layers. For a description of data sources see A1 Supplementary Table 1. Variables added to the datasets of point measurements are taken at the native, highest spatial resolution of the respective property. Where necessary and not explicitly stated, missing values were imputed using the mean value of the corresponding variable.

1.4.1 Soil bacterial carrying capacity derived from NPP

The flux of carbon into the soil is taken from the MODIS NPP dataset[50]. We have used mean annual values (2000-2015). Missing values (e.g. desert) were imputed with values obtained from the Miami model[51] using parameters fitted to the non-missing values of MODIS NPP. Only an average fraction (ϵ = 0.24) of the total NPP entering the soil column is available for bacterial respiration[42,43]. The vertical distribution of microbial carbon in the soil column follows the distribution of plant roots[8]. This allowed us to impose the depth z at which most of the carbon is released by integrating over the sampled interval dz and calculating the fraction of NPP available for bacteria at a particular depth ($\mathrm{NPP}_{\mathrm{b},z} = \epsilon \frac{\mathrm{NPP}}{d_{\mathrm{soil}}} F_z = \epsilon \frac{\mathrm{NPP}}{d_{\mathrm{soil}}} \int f(z)\, dz$). The factor F_z denotes the fraction of carbon available at a particular depth and is described by $f(z)$ for the entire depth of the soil profile considered (d_{soil} = 1 m). Assuming no net growth of the bacterial community so that only energy requirements for maintenance metabolism are satisfied, permits computation of maximal bacterial cell density ρ_{cell} (m^{-3}). This soil carrying capacity supported by the input flux of carbon is calculated using equation (1).

$$\rho_{\mathrm{cell}}(z,T) = \frac{\mathrm{NPP}_{\mathrm{b},z}}{f_T m M_{\mathrm{c}}} \quad (1)$$

Using a constant mass of carbon per cell M_{c} and by fitting maintenance rate m, we calculated the bacterial cell density ρ_{cell}. Temperature dependency was implemented as a factor f_T based on the Schoolfield model[44] using mean annual temperature (MAT) from the WorldClim dataset[52].

1.4.2 Soil bacterial abundance dataset

Xu *et al.* (XU)[8] compiled a dataset for the abundance of soil carbon associated with microbial biomass. This was used here as a reference for bacterial abundance for a range of geographical locations. We considered the relation between the soils carbon to nitrogen (C:N) ratio and the proportion of bacterial biomass to total microbial biomass[53]. Total microbial biomass carbon contains mainly fungal and bacterial carbon ($C_{\mathrm{mic}} \approx C_{\mathrm{F}} + C_{\mathrm{B}}$). A piece wise linear function was used to describe the ratio of fungal to bacterial carbon ($R_{\mathrm{FB}} = C_{\mathrm{F}} / C_{\mathrm{B}}$) with varying C:N ratio of the soil organic matter. This ratio was taken as a constant below C:N = 18.4 (R_{FB} = 5, see[42]) and increases with a slope of 0.5 above said

value[53]. From R_{FB} the relative proportion of bacterial biomass f_B was calculated ($f_B = 1 / (R_{FB} + 1)$). A carbon content per cell[54] of $M_c = 8.6x10^{-14}$ g C was used in all conversions of soil bacterial biomass and for the estimation of soil carrying capacity. To determine the decay of carbon input in the soil profile (f_z) we first averaged the bacterial biomass per soil depth. Averaging was necessary to avoid putting more weight on more frequently sampled depths. Values were integrated from the soil surface to the maximum depth of two meters. This cumulated bacterial biomass was normalized by its total sum to obtain the cumulative fraction of biomass with soil depth. For parameter estimation, we fit the cumulative lognormal distribution to the cumulative fraction of bacterial biomass yielding $\mu = 0.18$ and $\sigma = 1.00$ for parametrization of F_z. We chose a lognormal distribution as it gave a better fit to the vertical distribution of measured bacterial biomass than the previously used exponential model (A1 Supplementary Figure 1). The global maintenance rate was subsequently estimated by fitting equation (1) for the soil carrying capacity to measurements of soil bacterial biomass carbon[8] using inputs of local $NPP_{b,z}$ and MAT. The optimization yielded a maintenance rate of $m = 1.5$ gC gC_{cell}^{-1} y^{-1}.

1.4.3 Soil bacterial diversity datasets

Two datasets of bacterial species/phylotype abundances based on 16S rRNA sequencing were employed in this study. Data from the Earth Microbiome Project as published by Thompson *et al.* (EMP)[18] and data collected by Delgado-Baquerizo *et al.* (DEL)[20] were used to estimate bacterial diversity. Diversity was calculated on the data 'as provided' using the procedures outlined below. Except some samples in the EMP dataset had to be removed due to misclassification or unsuitable conditions. The following procedure was applied to filter the EMP data based on metadata: Samples labeled as 'Soil(non-saline)' where selected if the environmental material was either 'soil' or 'bulk soil'. We then removed samples containing the features 'oil contaminated soil' or 'extreme high temperature habitat'. Tables of sampled abundances of phylotypes were then used as published (90 bp qc filtered and rarified to 5000 for EMP (n = 2871) and the top 511 phylotypes after taxonomic assignment for DEL (n = 237)). Variables relevant to soil and climate were added according to reported geographical coordinates and soil depth resulting in 484 and 218 sites for EMP and DEL, respectively. The mass of soil is taken from the extraction protocol used in the studies. For DEL 0.25 g of soil and for EMP an average of 0.175 g were chosen.

1.4.4 Estimating soil specific 'climatic' water content

A metric for the average hydration conditions relies on estimation of a representative value of water content based on rainfall patterns. We use a simplified approach where the periods in which soil drains or dries following a rain event are calculated. We apply a threshold to the precipitation time series to remove small wetting events that immediately evaporate and estimate the time in between rain events. The average duration between events is the characteristic dry down for given geographical

locations. During this time, water mass is lost at a constant rate determined by (mean daily) potential evapotranspiration (PET) resulting in an exponential reduction of average water content within the considered soil profile (d_{soil} = 1 m). We assume for simplicity that a daily temporal resolution is compatible with the cessation of internal drainage of most soils. Hence, climatic soil water content does not exceed field capacity (a stable water content after internal drainage becomes negligible). For simplicity, we define the volumetric field capacity θ_{FC} (V_{water}/V_{soil} in $m^3\ m^{-3}$) as half of the porosity θ_s (V_{void}/V_{soil} in $m^3\ m^{-3}$). The latter is obtained using an empirical (pedo-transfer) function[55] that relates commonly measured soil properties (sand-, silt-, clay- contents and bulk density[56]) to soil porosity. The MSWEP[57] precipitation records of 37 years (1979–2016) are used to derive average rainfall quantities per wetting-drying cycle. The spatial resolution of the precipitation data is roughly 11 km at the equator and the temporal resolution is given at a sub-daily (3 hourly) timescale. The data is down sampled to daily resolution as the dynamics of soil wetting and drying relevant for the bacterial habitat are expected to be within this timescale. Further, the precipitation time series is subjected to a threshold taken from estimates of PET[58] based on temperature and radiation[52] to identify wetting events. The run lengths between wetting events are measured and averaged across wetting cycles. The key result of the analysis is the mean time interval between rainfall events τ (an ensemble average) for every location. This quantity combined with daily PET ($m\ d^{-1}$) were used to deduce the climatic water contents θ_τ (V_{water}/V_{soil} in $m^3\ m^{-3}$) according to equation (2).

$$\theta_\tau = \theta_{\mathrm{FC}} \mathrm{e}^{-\alpha<\tau>} \text{ with } \alpha = \frac{\mathrm{PET}}{d_{\mathrm{soil}}\,\theta_{\mathrm{FC}}} \quad (2)$$

The significance of θ_τ is that it combines rainfall patterns, PET, and soil properties over climatic time scales and provides a measure of the average hydration conditions experienced by soil bacteria in a particular geographical location (A1 Supplementary Figure 9).

1.4.5 Estimation of aqueous habitat size distribution

We estimated the size distribution of distinct aqueous habitats based on soil properties and hydration conditions (e.g., climatic water content). Soil water content was treated as the aqueous-phase occupancy probability *p* (the probability of finding a water filled pore or roughness element) that, in turn, enabled the application of standard percolation theory to represent the characteristics of aqueous bacterial habitats (sizes and numbers). We considered the soil as a three-dimensional lattice (two-dimensional (2D) for comparison with the SIM) with a critical occupancy probability and universal exponents that determine the number of (aqueous) patches and their sizes[59]. The critical percolation threshold p_c was multiplied by the soil void fraction (or saturated water content θ_s) to account for soil porosity[60]. The critical water content is thus defined by equation (3) and could be expressed as critical saturation S_c (4) to remove the dependency on θ_s.

$$\theta_c = \theta_s p_c \quad (3)$$

$$S_c = \frac{\theta_c}{\theta_s} = p_c \quad (4)$$

The size distribution of aqueous patches $n_s(p)$ was assumed to follow general proportionalities of percolation theory (5-7)[59]:

$$n_s(p) \sim s^{-\tau} e^{-\frac{s}{s_\xi}} \quad (5)$$

$$s_\xi \sim |p_c - p|^{-\frac{1}{\sigma}} \quad (6)$$

$$P^\infty \sim (p - p_c)^\beta \quad (7)$$

With the patch size s (number of sites/pores) for $s >> 1$, Fisher exponent $\tau \approx 2.18$ (2D: $\tau = 187/91$), cutoff exponent $\sigma \approx 0.45$ (2D: $\sigma = 36/91$) and cutoff size s_ξ[59]. P^∞ is the fraction of the domain occupied by a spanning (algebraically infinite) patch with exponent $\beta \approx 0.41$ (2D: $\beta = 5/36$). The patch sizes follow a power law distribution at $p = p_c$. Away from this critical point when the cutoff size s_ξ is exceeded, patches shrink with decreasing water content ($p < p_c$) or merge and grow when approaching saturation ($p > p_c$) as patches of size $s > s_\xi$ become exponentially scarce. Although, the prediction is strictly valid only for p close to p_c, we assume such relations to hold for the range of conditions considered. The occupancy probability p was thus substituted with climatic water content θ_t and p_c with a critical water content $\theta_c \approx 0.15$ that correspond to a simple cubic lattice with porosity $\theta_s \approx 0.5$ (triangular lattice in 2D; $\theta_c \approx 0.25$).

To account for different soil types, a characteristic length scale δ is estimated based on the geometric mean diameter of soil particles[61]. This length scale is used for normalization of the aqueous patch size distribution in the range of water contents and patch sizes relevant for bacterial life. The soil type length scale δ and the system size L were considered (soil domain or sample size); here we used the mass of soil sampled m_{soil} and bulk-density ρ_{soil} specific to soil type (8). The total number of candidate sites N_0 in the sampled soil was then determined from simple geometry considering the dimensionality d = 2 or 3 (9).

$$L = \left(\frac{m_{soil}}{\rho_{soil}}\right)^{\frac{1}{d}} \quad (8)$$

$$N_0 = \frac{L^d}{\delta^d} \quad (9)$$

We approximated the behavior of the percolation transition using a bounded logistic curve that provides a smooth function $\hat{P}^\infty$

$$\hat{P}^\infty = \frac{\theta}{1+e^{-k(\theta-\theta_c)}} \quad (10)$$

where k describes the 'sharpness' of the transition (k = 16 for all calculations). The total size of aqueous clusters or potential habitats N_s was normalized as follows:

$$N_s^0(\theta, N_0) = \frac{\theta - \hat{P}^\infty}{\sum_1^{N_0} s\, n_s(\theta)} \quad (11)$$

$$N_s(\theta, s) = N_s^0(\theta, N_0)\, s\, n_s(\theta) \quad (12)$$

Thus requiring, by pre-factor N_s^0, that the total volume of aqueous patches conserve the volume of soil water at a given state of hydration. For practical reasons, subsequent calculations of aqueous patches proceed by removing the largest patch after normalization (this large patch biases the counting of habitats in a sample).

1.4.6 Calculation of bacterial species diversity

The distribution of aqueous patches derived from percolation theory and their properties defined the degree of spatial isolation and restricted the number of potential habitats. Both aspects were expected to alter the bacterial diversity patterns observed in natural soils. The estimated aqueous patch sizes and their prevalence defined the distribution of bacterial habitats. Together with carrying capacity we can estimate the number of cells within a single (habitat) size class *s* (13).

$$N_{\mathrm{cell},s} = \rho_{\mathrm{cell}}\, s\, \delta^d \quad (13)$$

Aqueous patches with cell count below a prescribed threshold (or limit of detection, $N_{\mathrm{cell}} < 4000$ for comparisons with empirical data) were removed from the total number of potential habitats N_s. Conceptually this can be interpreted as the discrete nature of bacterial cells that limits counts to integers greater than one. Empirically, there exists a lower limit of detection and a minimal number of cells from a single species (>>1) is needed to contribute to the measurement of bacterial richness. Initially, we assumed that only a single species occupies a patch by outcompeting possible co-inhabitants. Herby, the modeled species abundance distribution (SAD) follows the distribution of aqueous habitats with abundances bounded by carrying capacity within a defined volume of soil. Subsequently we introduced the possibility of multiple species occupying large aqueous patches (in proportion to their size and dimension; $N_{\mathrm{sp}} \sim s^{1/d}$, $d = 2$ or 3) to correct for model bias of over predicting the dominant species. The exponent (*1/d*) suggests that the number of species per habitat grows with the average distance between any two points selected randomly within a single habitat of size *s.* The limit of detection was not used for the comparison of SADs as the total number of habitats was truncated to the number of observed species.

Bacterial diversity was calculated in the general form[62] for all SADs (modeled and data):

$${}^qD = \left(\sum_{i=1}^{\mathrm{SR}} p_i^q\right)^{1/(1-q)} \quad (14)$$

With relative species abundance p_i and species richness SR. For $q = 0$ the equation corresponds to the weighted harmonic mean and equals the actual number of types (SR). The equation is not defined for $q = 1$ where the limiting form is described by the well-known Shannon index *H* (15) and evenness $E_{1,0}$ is calculated as defined by equation (16)[62].

$$\lim_{q\to 1} {}^qD = {}^1D = \exp(H) = \exp\left(-\sum_{i=1}^{\mathrm{SR}} p_i \ln(p_i)\right) \quad (15)$$

$$E_{1,0} = \frac{^{1}D}{^{0}D} \quad (16)$$

1.4.7 Spatially-explicit individual-based model (SIM)

An individual based approach was previously developed to model growth of diverse bacterial species on heterogeneous soil surfaces[23,39] and was adopted for the current study. The spatial domain was represented by a hexagonal grid with periodic boundary conditions (length L = 1 mm; area of a grid cell A_{hex} = 100 µm^2; and porosity θ_s = 0.49). Grid cells consisted of water holding elements with volumes drawn from a random uniform distribution (unif) that have a maximal size equal to the spacing of the grid (dx = 1.1x10^{-5} m). Thereby the modeled domain represents a slab of the soil pore space with a defined volume ($V_{soil} = L^2\ dx$). The bulk water content is prescribed to the domain as a control parameter and spatially distributed relative to the sizes of grid elements while conserving the total volume of water ($V_{water} = \sum V_{water,x,y}$). Based on the local volume of water, an average water film thickness h was calculated ($h_{water,x,y} = V_{water,x,y}/A_{hex}$). The heterogeneity of the water film thickness modified the mass transfer between grid cells by changing the cross-sectional area that contributed to the diffusive flux. Diffusion was solved using the implicit finite differences method with bacterial consumption represented as a sink term. Diffusivity is taken for a small molecule that is readily available for bacterial consumption (e.g. glucose) and does not vary spatially (D = 6.7x10^{-10} m^2 s^{-1}). The simulation period corresponded to eight days at a one-minute time step. Initial concentration of nutrients was constant in space and randomly replenished to initial concentration over time to mimic a fluctuating environment. The arrival of nutrient pulses was modeled as a Poisson process with an average rate of one arrival every four hours. The initial nutrient concentration was set to provide enough carbon to sustain a fixed cell density (10^{17} m^{-3}, corresponding to high carrying capacity) and was distributed evenly among nutrient pulses. The mass of nutrients locally available for bacterial consumption depended on the volume of water in a grid cell. All simulated bacteria were represented as elongating cylindrical capsules that consume a common carbon source dissolved in the aqueous phase. The diversity and multiple species i were prescribed in the model by varying Monod parameters (growth rate $\mu_{max,i}$, half saturation constant K_i - additionally maintenance rate $m_i := 0.01\ \mu_{max,i}$). Species specific parameters were randomly selected from uniform distributions of the Monod parameters (μ_{max} ~ unif(10^{-4} h^{-1}, 1.14 h^{-1}), K ~ unif(6.8 g m^{-3}, 680 g m^{-3})). All other parameters were held constant (mass of the cell m_{cell} = 9.5x10^{-13} g, mass at division $m_{div} = 2\ x\ m_{cell}$, yield Y = 0.5, cell radius r_{cell} = 0.5 µm). A single cell of each species was inoculated randomly on the domain at the beginning of the simulation (species richness SR at t = 0, SR_{t0} = 4096). Individual cells grew and divided along their axis with a slight asymmetry in mass to avoid complete synchrony (f_m ~ unif(0, 0.05), $m_{cell,1} = f_m\ m_{div}$ and $m_{cell,2} = (1\text{-}f_m)\ m_{div}$). All bacterial cells were subject to active and passive motion and could move continuously in the domain. Growth induced shoving represents the passive motion and was

implemented by displacing cells relative to their nearest neighbors (only considering the capsule geometry as n-spheres; no forces, e.g. capillary, friction, elastic, electrostatic, etc.). Shoving was not resolved to full relaxation due to the size of the domain, number of cells and the scale of interest (compromise between reduced computational demand and precision of the resulting spatial distributions). However, we implemented a simple rule to prevent local crowding: if the projected area of bacterial cells in a grid cell exceeded the area of the grid cell (A_{hex}), bacterial cells were randomly picked and moved to form a second layer (piling cells at the *z*- direction) from which they could 'drop' down again once space became available. Bacterial swimming motility was permitted where the aqueous phase was connected and the water film thickness exceeded cell diameter[40]. Cells aligned their motility trajectories along gradients of the nutrient field, whereas their velocity was modified by the water film thickness[40] and nutrient concentration[63]. Additionally, each velocity component (v_x, v_y) is independently multiplied with a random factor to allow for individual trajectories ($f_v \sim \text{unif}(0, 2)$). Integrating along the projected trajectory of each cell enabled consideration of varying water film thickness and prevented cells with high instantaneous velocity from 'jumping' across grid cells. At the end of the simulation the total number of cells and the number of cells per species were measured. To enable comparison of richness estimates from varying sample sizes (e.g. with observed species richness or simulations with different cell densities) total cell numbers were rarified to 5000 and 1000 counts, to compare with EMP and DEL, respectively. For comparison with the DEL dataset only the top 512 most abundant species were considered. Singletons, i.e. cells that were sampled only once when rarefying, were removed from the counts. The rarefication procedure was averaged across 15 trials to increase robustness of the diversity estimates. Only community evenness was also estimated without rarefication and removal of singletons as it affected the apparent community structure (A1 Supplementary Figure 8).

2 A hierarchy of environmental covariates control the global biogeography of soil bacterial richness

Samuel Bickel, Xi Chen, Andreas Papritz and Dani Or

Scientific Reports 9, no. 1 (August 20, 2019): 1–10. https://doi.org/10.1038/s41598-019-48571-w.

Abstract

Soil bacterial communities are central to ecosystem functioning and services, yet spatial variations in their composition and diversity across biomes and climatic regions remain largely unknown. We employ multivariate general additive modeling of recent global soil bacterial datasets to elucidate dependencies of bacterial richness on key soil and climatic attributes. Although results support the well-known association between bacterial richness and soil pH, a hierarchy of novel covariates offers surprising new insights. Defining climatic soil water content explains both, the extent and connectivity of aqueous micro-habitats for bacterial diversity and soil pH, thus providing a better causal attribution. Results show that globally rare and abundant soil bacterial phylotypes exhibit different levels of dependency on environmental attributes. Surprisingly, the strong sensitivity of rare bacteria to certain environmental conditions improves their predictability relative to more abundant phylotypes that are often indifferent to variations in environmental drivers.

2.1 Introduction

Delineating biogeographical patterns of soil bacterial richness could offer insights into potential links between natural bacterial community traits and belowground ecological functioning[64]. Various external drivers, land use and biome characteristics shape the soil bacterial community composition and structure. Spatial mapping of soil bacterial richness remains a challenge due to the high number of bacterial phylotypes and the sparse global coverage of available samples[65–67] that originate from only few biomes. The vast number of possibilities for community assembly across environments with high intrinsic heterogeneity limit inference of globally representative biogeographical patterns from small-scale measurements[66,68]. The establishment of reliable global maps of bacterial biogeography hinge on inclusion of ample sampling locations and tackle the hurdles of uneven sample sizes and primer biases in meta-analyses[69]. To overcome these limitations towards development of unbiased estimates of global bacterial richness patterns, require comprehensive and well-harmonized data sets. Additionally, the primary drivers for soil bacterial richness are often obscured by large uncertainty in measurements and by sensitivity of species richness to methodology and sampling protocol[65]. Identifying drivers of bacterial richness is particularly error-prone due to the metrics sensitivity to the detection of rare and low abundant species; thereby challenges data analysis and interpretation. One of the most common predictor (covariate) of soil bacterial diversity is the soil pH[29,33,70,71]. For near neutral soil pH, bacterial diversity peaks and then drops for acidic and basic soils[33]. Some have argued that such a pattern reflects increased abundance of specialist species in such environments or, alternatively, that pH is merely a proxy for other environmental factors[33]. Along with soil pH, many other environmental characteristics, such as mean annual precipitation and mean annual temperature are expected to affect soil microbial life, yet their effects are difficult to assess independently as they are often interlinked and only partly exhibited at scales relevant to soil bacterial habitats[10,14]. Soil hydration status has emerged as a primary factor affecting soil bacterial habitats[23,38], as supported by empirical observation[5,47,72,73]. The wetness of a soil affects the connectivity of the aqueous bacterial habitats[11], thereby modifying interactions and the motility of bacterial cells that in turn affect community composition and diversity. Yet few attempts have been made to statistically test the dependency of bacterial diversity on climatic soil moisture conditions at the global scale.

Three recently published datasets of soil bacterial community composition[5,18,28] combined with a consistent set of covariates (A2 Supplementary Table S1) permit the (i) systematic consideration of composite soil and climate variables that could reflect salient conditions of soil bacterial habitats, and (ii) enable a process-based understanding of the hierarchy in environmental factors that control soil bacterial richness. In this study, we (iii) analyze biogeographic trends to statistically test the explanatory power of composite variables, specifically climatic water content, with respect to soil

bacterial richness and (iv) predict global biogeographic trends using general additive models (GAM) and tree-based methods.

2.2 Results and Discussion

Merging the geo-referenced 16S rRNA sequence data resulted in 844 valid soil samples, of which, 320 representative sampling sites were obtained after sample aggregation (Fig. 2.1 a). Only bacterial diversity was analyzed, as the use of 16S rRNA sequences precludes the investigation of fungal diversity in the current study. Despite covering all 14 classified biomes of the world[74], sampling was not even, and some biomes and continents were under- or overrepresented (e.g., deserts contribute to about 18.9% of the terrestrial surface, yet only 6.3% of samples originated from these environments). From a total of 256,620 amplicon sequence variants (ASV) detected, we removed Archaea and unassigned sequences (at kingdom level, 1.55%) leaving 98.45% of bacterial ASVs. For ease of communication, we refer to the designated bacterial ASVs as "species" throughout the text. The widest range of species richness was observed in deserts (Fig. 2.1 b) and could be attributed to the wide span of variations in environmental conditions in such biomes[75]. The relatively low richness in montane grassland and tundra could be indicative of a non-monotonic relation between moisture availability and soil bacterial richness. Boreal forests (n=11) exhibited lower richness compared to tropical (n=23) and temperate forests (n=122; p=0.0311 and p=0.0063, respectively, Wilcoxon rank sum test). This latitudinal shift in species richness[5,28] suggests that temperature plays an important role in regulating bacterial richness. However, consideration of temperature alone provides no distinction between the richness observed in tropical and temperate forests (p=0.6575, Wilcoxon rank sum test), suggesting more complex interactions and mechanisms.

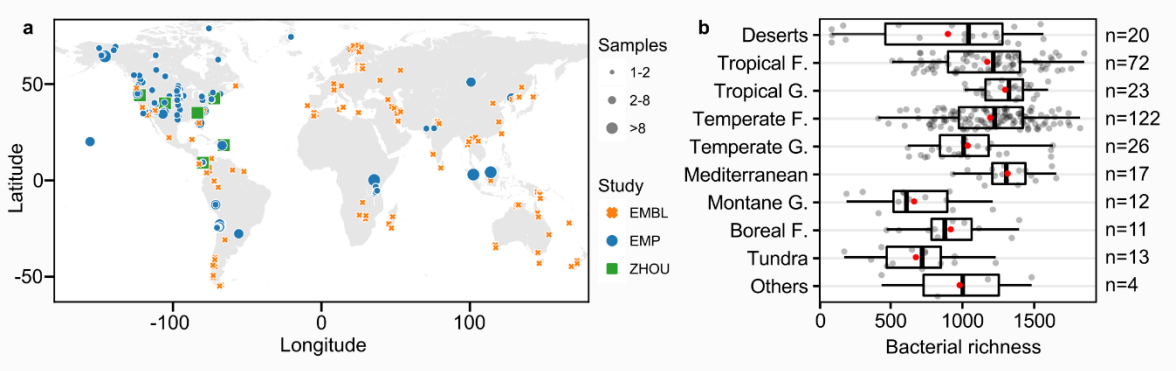


Fig. 2.1 Distribution of sites and representative samples obtained from three recent studies (EMBL [5], EMP [21], ZHOU [31]) used in this meta-analysis. a, Geographical locations of sites (n=320, by continent: AF=27, AS = 42, AU = 30, EU= 55, NA = 104, SA = 62). Size of the points represents the number of samples that were aggregated within 0.1° x 0.1° cells. Colors orange, blue and green represent the three studies EMBL, EMP and ZHOU respectively. **b,** Bacterial richness grouped by biomes (F. forest, G. grassland). Site values are shown in grey, while the red points represent mean values. Boxes show the inter quartile range (median as solid line) with bars indicating central 95%-range of values.

2.2.1 Univariate analysis of bacterial richness

We first evaluate trends of species richness considering climate and soil properties within univariate general additive modeling. Selected covariates were used that represent different aspects of the soil environment (A2 Supplementary Table S1). Climatic water content (CWC) represents the soil water storage capacity and climatic water balance based on the number of consecutive dry days (DRY) and potential evapotranspiration (PET) (A2 Supplementary Methods). It is a proxy for the soil's wetness, its dynamics and aqueous phase connectivity. Both shape the number of distinct aqueous habitats and their connectedness in a soil. We found an optimal CWC in the range of 0.15 to 0.20 where bacterial richness peaks (Fig. 2.2 a). A generally linear drop in richness seen towards low water availability is potentially due to nutrient limitations by the physically constrained diffusion processes and reduced carbon input. Soil pH exhibited a trend similar to the CWC with a peak near neutral values (pH 7, Fig. 2.2 b) as reported in previous studies[18,33]. We note, however, a strong linear association between pH and climatic water content (R^2 = 61%, n = 320, Fig. 2.2 c).

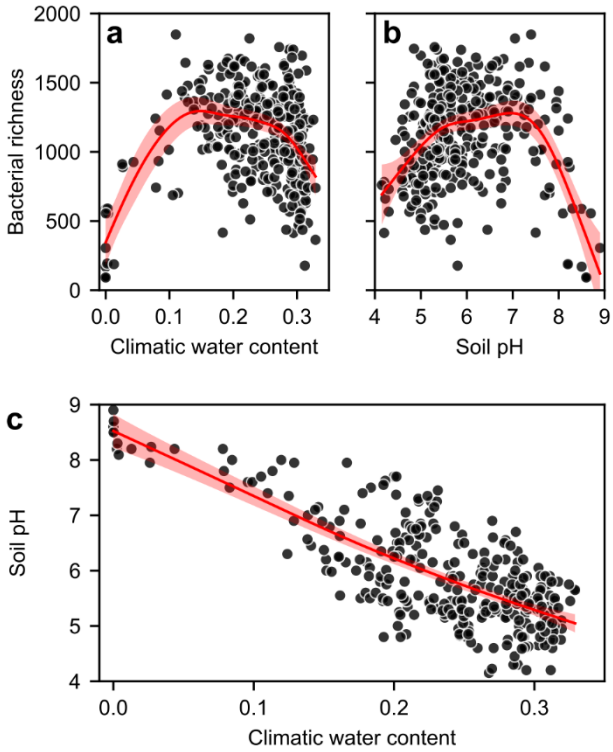


Fig. 2.2 Univariate general additive model (GAM) of soil bacterial richness. a, Relation between climatic water content and bacterial richness. Bacterial richness peaks in soils with intermediate climatic water contents (0.15–0.2) and drops in dry and wet soils (R^2 = 27.7%, RMSE = 298.1, AIC = 4557.5, EDF = 4.7). **b**, Commonly observed trend of bacterial richness peaking at near neutral conditions (pH 7) and showing distinct drops in acidic and basic soils (R^2 = 23.8%, RMSE = 306.0, AIC = 4574.0, EDF = 5.1). **c**, A strong linear association (adjusted R^2 = 60.8%, deviance explained 61.1%) is observed between climatic water contents and soil pH pointing to possible confounding effects of these covariates on bacterial richness. Shaded areas correspond to standard errors (n = 320).

Climatically humid regions tend to be acidic and dry regions basic. Such trends have been attributed to the difference between mean annual precipitation (MAP) and PET that determine the climatic soil water balance for the region[49]. A net accumulation of salt in soil (e.g. in arid regions) directly results from a negative water balance with more evaporation than precipitation. This increase of mineral concentrations enhances the soil pH buffering capacity and can result in high soil pH. With an increase in ionic strength we would also expect effects on bacterial physiology (e.g. increased osmotic pressures[10,14]) and possible, specialized adaptations to these environments. A recent study attempted to disentangle the effect of salts and soil pH on bacterial community composition and revealed a strong effect of salinity[76]. This may also suggest that previously reported dependencies of bacterial diversity on soil pH[5,29,33] could have been mediated by climatic soil water conditions via the accumulation of salts. Although pH is related to the suitability of bacterial habitats by increasing the tolerance (and competitive ability) of pH-adapted species[76], it might not be the underlying driver of bacterial diversity. This reasoning is based on the idea that competitive exclusion can only occur with some degree of habitat overlap and interactions between species. Under most conditions in natural soils the aqueous phase is largely fragmented and the (micro-) environments experienced by bacteria are not necessarily the same. This fragmentation permits coexistence and suppresses the elimination of inferior competitors and, hence, promotes bacterial diversity. The distinct optimality of bacterial richness related to soil wetness could be attributed to (i) resource limitation for extremely dry soils and (ii) the increased habitat connectivity that suppresses diversity by promoting competitive exclusion in wet soils. In this context, pH represents the chemical niche environment, a variable under control of primary (physical) factors, i.e. resulting from a soil's climatic water balance[49]. Temperature is another primary variable that might confound many processes. The mean annual temperature (MAT) is expected to alter species richness according to the metabolic theory of ecology[28,77]. This trend was manifested by a slight increase of richness with MAT peaking at 0–10 °C and 20–30 °C (A2 Supplementary Fig. S1), in agreement with a previous study[67]. One explanation for the lack of clear patterns could be that temperature not only modifies growth rates of bacterial cells, but also affects habitat connectedness via effects on precipitation and water balance. This may counteract the enhancing effect of temperature on richness in wet and warm regions (e.g. the Tropics) where bacterial habitats are frequently connected. Furthermore, despite the strong variation of MAT near the soil surface, the effective range at the sampled depth of 10 cm might be narrower due to the damping effects of soil and leads to a limited range of conditions experienced in bacterial habitats. Additionally, bacteria could be able to tolerate a wide range of temperatures. Bacterial richness was found to be driven by temperature near geothermal springs only beyond 70 °C[78]; conditions that are not frequently found in soil. Nonetheless, changes in light intensity (solar radiation, RAD) are strongly

correlated with temperature and latitude. A direct effect of light on bacterial richness would be expected by enabling growth of photoautotrophs and possible adaptation to high doses of UV light (or the lack thereof). Both effects could be masked by the presence of vegetation (e.g. NPP) that would intercept the solar radiation. We thus do not expect strong changes in the distribution of bacterial richness caused by light in vegetated environments and in sub surface soils (due to the strong attenuation of light). Nevertheless, the indirect effects of solar radiation should be well described by the used covariates (e.g. MAT and CWC) as light and water availability both shape the vegetation of an ecosystem. We used net primary productivity (NPP) to represent vegetation patterns at the ecosystem level and to characterize carbon input into subsurface bacterial habitats. NPP did not display a notable effect on species richness (slightly increasing richness up to 500 g C m^{-2} yr^{-1}, constant richness beyond, A2 Supplementary Fig. S1). Only in extreme environments, such as deserts and tundra, NPP seems to influence species richness.

2.2.2 Multivariate general additive model (GAM) of bacterial richness

The complexity of interactions among environmental factors, vegetation and soil microorganisms suggests that a single variable alone is not likely to explain the observed patterns of soil bacterial species richness. We therefore tested the robustness of the observed single-variable trends using a multivariate general additive model (GAM) with forward selection of covariates (Table 2.1, A2 Supplementary Fig. S2). The ranking of the most influential covariates remained consistent with the results of univariate GAM, with CWC slightly outperforming pH. Interestingly MAT occupied the third rank, suggesting that we were able to successfully capture combined effects on soil bacterial species richness. The goodness-of-fit statistics of the multivariate GAM using only the six selected covariates (R^2=35%, *RMSE*=283.7) were better than the statistics of any univariate GAM, suggesting that soil and climatic covariates provide additional information on species richness. Although we observed significant associations between bacterial diversity and environmental factors in uni- and multivariate modeling, these associations do not necessarily imply causation. To mitigate limitations of commonly used structural equation models (SEM) in discerning causal nonlinear effects, we have used a causal additive model (CAM)[79] to explore potential causes of soil bacterial diversity. We used this novel approach to generate a graph of inferred structural dependencies between covariates and bacterial richness (A2 Supplementary Fig. S3). By removing links between variables that are not considered significant ($p \leq 0.0005$), we can distinguish direct from indirect relations between covariates and bacterial richness; as variables that remain linked to richness directly and variables that are connected to richness via others. Compared to the results of the multivariate GAM, we obtained a similar set of covariates with direct effect on species richness, i.e. CWC and DRY. Surprisingly, pH and MAT were not selected as potential direct causes, implying that they may have weaker effects on species richness or

their associations with species richness were attributed to confounding effects. This approach enables further exploration of potential model structure. Nevertheless, care should be taken when interpreting inferred causal relationships as the method relies on the strong assumption of "no hidden variables" that are unknown in most natural environmental systems. Yet, it is noteworthy that no prior expectations or knowledge is imposed on the model structure, as is necessary with many SEM[5]. All direct and indirect links are deduced only from the observations with a given set of covariates. A drawback of this approach is that not all dependencies might be physically meaningful.

Table 2.1: Ranking of covariates determined by forward selection for the multivariate general additive model (GAM).

Step	*Selected*	*ΔAIC [a)]*	*p-value[b)]*
1	Climatic water content	-104.64	<0.0005
2	Soil pH	-19.43	<0.0005
3	Mean annual temperature	-16.82	<0.0005
4	Silt fraction	-7.18	0.0083
5	Consecutive dry days	-1.59	0.0385
6	Cation exchange capacity	-0.08	0.1497

***a**) Change of Akaike information criterion (ΔAIC) when the variable was added to a model that already contained the covariates listed above the current step; **b**) Likelihood ratio test of nested models*

2.2.3 Varying proportions of low abundance species

Thus far, we have focused on explaining bacterial species richness without considering environmental effects on species with different levels of abundance. We evaluated the performance of the univariate (CWC, pH) and multivariate GAM for metrics of diversity other than species richness and found a consistent increase in R^2 with increasing weight of species with low abundances (A2 Supplementary Fig. S4). The observation indicates that species with low abundance show greater sensitivity to environmental conditions than the species dominating within samples. To further evaluate effects of environmental variables on rare and common fractions of the soil bacterial populations, we split the species in to two groups by using a threshold (0.005%) of global relative abundance. For each sample, we computed the log-ratio of the number of rare and common species. A value of zero indicates that a sample contains the same number of rare and abundant species, and larger values indicate that the rare species are more numerous. We explored the dependence of the log-ratio on environmental covariates by univariate GAM (A2 Supplementary Table S2). Interestingly we find similar, but complementary trends for CWC and pH (Fig. 2.3 a and b). Most notably, a distinct drop in the number of rare species appears under elevated climatic soil water contents. This trend compares well with univariate and multivariate model results for species richness. The modeled dependencies of rare and common species diversity on climatic water content (Fig. 2.3 c) demonstrate a higher susceptibility of rare species to increased aqueous phase connectivity associated with high water contents. While the

common species remain abundant, the number of rare members of the soil bacterial community shows a steep decline towards wetter soil conditions. This discrepancy is weaker for soil pH where diversity of both rare and common species decreases at similar rates when approaching acidic conditions (Fig. 2.3 d). The gradual increase in the proportion of globally rare species under drier conditions (low CWC) is likely due to the more fragmented aqueous phase that may shelter bacterial species in small but numerous isolated aqueous habitats[23,38]. Alternatively, one might argue that the emergence of rare species under basic (high pH) — and possibly also very dry— conditions is attributed to the presence of specialist phylotypes capable of coping with such an environment[29,80]. However, if neutral pH would be favored by most bacterial species (i.e. leading to more diversity) we would expect less balanced soil bacterial communities with more of the rare species present around pH 7. Interestingly, the log-ratio does not increase again towards acidic conditions. Hence, acidic environments reduce diversity of rare and common species to a similar extent, and rare (specialist) species that benefit from weaker competition with common species seem to be missing. Although, information on many additional factors that could affect the presence of rare and common species (e.g. nutrient status of the soil) could not be included in the analysis, general tendencies could be identified using the variables considered. We thus conclude that aqueous habitat connectivity largely dominates the soil bacterial richness picture and should be taken into account together with additional factors when data is available.

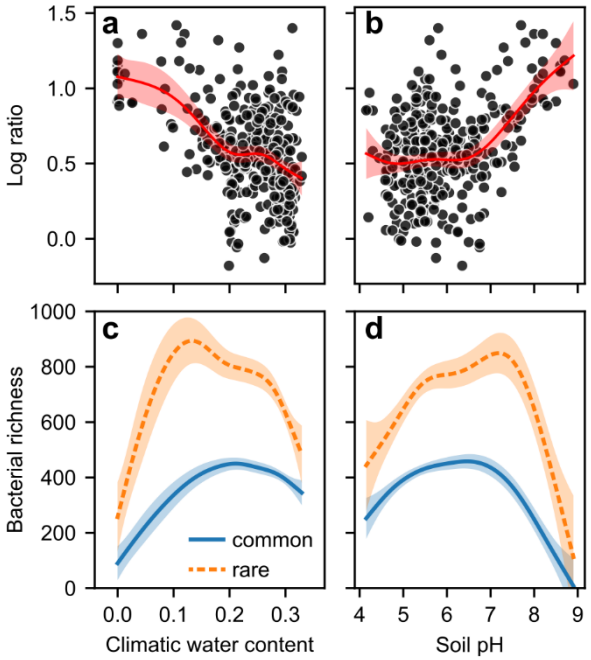


Fig. 2.3: Dependence of the log-ratio of number of rare and common species per sample on the two main predictors of bacterial richness. **a,** climatic water content (adjusted R^2 = 24.5%, deviance explained 25.5%, AIC = 70.5, EDF = 4.3) and **b,** soil pH (R^2 = 23.0%, deviance explained 23.9%, AIC = 77.0, EDF = 4.1). The log ratio is calculated by splitting species into two groups based on a threshold of global relative abundance (0.005%). A log ratio of zero indicates a balanced population, where the number of rare species per sample equals the number of common species. The modeled curves of both groups (rare and common) richness are shown for **c**, climatic water content and **d,** soil pH. Shaded areas correspond to standard errors (n = 320).

2.2.4 Global patterns of soil bacterial richness

The GAM used in this study accounts only for independent and additive effects of covariates on species richness. This may not be a realistic depiction of processes in natural ecosystems with numerous connections and interdependencies. Tree-based statistical models seem better suited to account for (higher-order) interactions between variables. For prediction of global maps of species richness, we therefore combined independently trained random forest and gradient boosting trees by simple averaging. The procedure was reinitialized and repeated ten times to stabilize the results and increase reproducibility. The tree-based model (R^2=40%, *RMSE*=261.5) performed better than the multivariate GAM (R^2=35%, *RMSE*=283.0) indicating that interactions between covariates are important for predicting species richness. Despite the considerably better performance, a large portion of variance remained unexplained. This is not unexpected, given the different sampling strategies and methodology of the studies. Additionally, covariates derived from remote sensing products and digital soil maps smooth the actual spatial variation of the respective characteristics and do not (yet) capture the full heterogeneity of natural soils. Another limitation of this study is the lack of fungal data. The data used does not permit analysis of fungal richness, and we can only speculate about potential, general trends. However, one study used in our dataset (EMBL)[5] investigated fungal diversity across biomes and report that fungal diversity does not peak in temperate regions (unlike bacterial diversity). The authors further suggest niche differentiation lead to contrasting responses of fungal diversity with precipitation and soil pH compared to bacterial diversity[5]. We thus would expect fungi to play a dominant role in vegetated soils with lower pH and high C:N ratios[5,53]. Such regions (biomes) are represented by high NPP and high climatic water contents. In these environments the aqueous phase connectedness could additionally enhance competition; potentially also between bacteria and fungi. The global map of predicted bacterial richness shows distinct regions of varying bacterial richness (Fig. 2.4). Tropical regions (e.g. the Amazon and the Congo Basin rainforests) exhibit remarkably lower bacterial richness highlighting the adverse effects of high levels of soil wetness on bacterial diversity. Lowest richness values were also found in regions where resources are most limiting, such as in the Sahara or the Atacama deserts. “Hotspots” of species richness lie in temperate regions and climatic transition zones where resource availability is not limiting and the aqueous phase remains fragmented, such as in the northern regions of India or in the Sahel. Tree-based methods provide a complementary approach to GAM as they efficiently handle higher order interactions between covariates and provide an efficient interface for spatial mapping. The implicit representation of covariate dependencies and model averaging, however, do not offer as much insight into the model structure as is possible with GAMs.

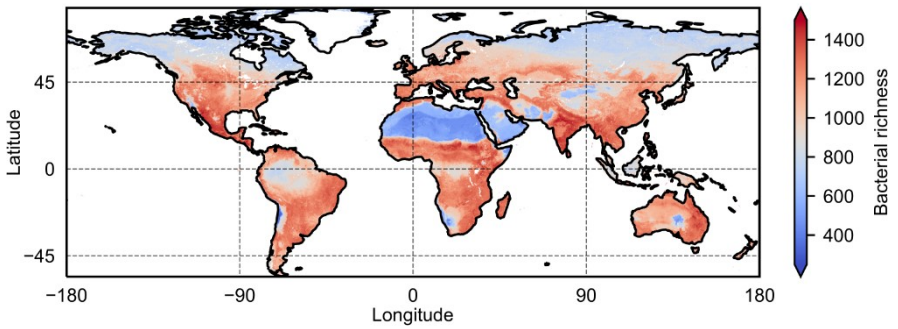


Fig. 2.4: **Global prediction of bacterial richness delineating spatial patterns of contrasting diversity** (R^2 = 40%, RMSE = 261.5). Tropics (e.g. Amazon, Congo) and northern higher latitudes (e.g. Siberia) show low bacterial richness. This is potentially linked to increased prevalence of frequently wet soils fostering connectedness of bacterial habitats. Low bacterial richness in deserts (e.g. Sahara, Atacama) is likely due to resource limitation. The highest bacterial richness is found in temperate regions and climatic transition zones (e.g. Sahel).

2.3 Conclusions

Incorporating the effects of soil and climate in the analysis of bacterial biogeography based on global datasets provides new insights into the key factors, namely climatic water content and pH that shape soil bacterial richness and community structure. The dominant role of climatic soil water content has not been fully recognized in previous studies. The inherent links between climatic soil water content and soil pH suggest that part of the soil bacterial diversity previously attributed to soil pH may reflect effects of climatic water content. We find that regions of intermediate climatic soil water content exhibit a peak in bacterial richness owing to the fragmentation of aqueous bacterial habitats that remain sufficiently supplied with resources, thus ensuring growth and protection from competitive exclusion. The results suggest that soil pH acts as a secondary driver of soil bacterial richness and represents a proxy of soil properties and climatic conditions. Placing local bacterial relative abundance in a global context provides fruitful insights into the biogeography of soil bacteria and the factors shaping spatial patterns of bacterial diversity. Especially the rare component of the soil bacterial community that contributes a large fraction of diversity is surprisingly predictable. This highlights the importance of environmental drivers, such as climatic water content, in shaping the genetic pool of potential functional capabilities by changing the size of the soil bacterial “seedbank”.

2.4 Materials and Methods

2.4.1 Data collection and processing

All 16S rRNA sequences of soil samples were obtained from three different studies. We hereafter use the terms EMBL (European Molecular Biology Laboratory)[5], EMP (Earth Microbiome Project)[18], and ZHOU[28] to refer to the sources of samples and metadata. Since sequences were different in terms of their representativity and amplification protocols, filters based on sample metadata, primer sequences as well as assigned taxonomy were applied to minimize methodological differences and maximize compatibility.

2.4.2 Metadata-based filtering

The metadata of soil samples (*n*=235, 7,974 and 126 for EMBL, EMP and ZHOU, respectively) were obtained from QIITA[81] and the European Nucleotide Archive ENA[82]. Although most soil samples were initially collected with the aim to study soil microbial communities, some of them could not be considered natural. The following procedures were applied to each study:

EMP: We selected representative samples carefully based on the metadata by removing potential artificial soils (e.g. sand filter in water purification system), managed soils (e.g. agricultural soil) and soils which cannot be considered as "natural" (e.g. soil samples taken from urban environments). Further, samples of Antarctic soils and from depth > 0.1 m were excluded due to limited information on local environments. The 16S rRNA sequences of all selected samples (n=587) were retrieved from ENA.

EMBL & ZHOU: No metadata-based filtering was done since all samples could be considered representative according to the criteria applied to EMP. 16S rRNA sequences for EMBL and ZHOU were obtained from ENA using study accession ID PRJEB19856 and PRJNA308872, respectively.

2.4.3 Primer-based filtering

EMBL, EMP and ZHOU used the marker gene sequencing method for amplification[83], yet their chosen primer sets and targeted regions of 16S rRNA differed substantially. To avoid primer biases, we only included samples which amplified the V4 region of 16S rRNA. Furthermore, two slightly different primer sets were used between studies, i.e. the original 515F-806R primer[84] and its modification[85]. The original primer (forward: GTGCCAGCMGCCGCGGTAA, reverse: GGACTACHVGGGTWTCTAAT) is known to be biased towards certain archaeal and bacterial groups, such as Crenarchaeota, Thaumarchaeota and SAR11[86,87]. The modified one adds one degeneracy in both the forward (GTGYCAGCMGCCGCGGTAA) and reverse (GGACTACNVGGGTWTCTAAT) primer to reduce those biases. However, most samples in EMP and ZHOU were published before the modified primer set came in use, whereas all samples in EMBL were amplified using the modified one. To make a valid

comparison, we either filtered particular sequences which could only be captured by the modified primer set (if the primers were retained in the raw sequences), or dropped the entire sample (if no information was available about the primers). We additionally removed sequences in which adapters could be identified (adapter contamination).

EMBL: All sequences in EMBL were raw and unjoined. We discarded pairs of sequences if GTGTCAGCMGCCGCGGTAA could be found in the forward reads or GGACTACGVGGGTWTCTAAT in the reverse reads (difference between the original and the modified primer). The forward and reverse reads were subsequently joined, trimmed and quality controlled (Phred threshold of three) using VSEARCH (QIIME2, 2018.8.0)[88], cutadapt[89] and split_libraries_fastq.py (QIIME1, 1.9.1)[84], respectively.

EMP and ZHOU: Unlike EMBL, sequences in EMP and ZHOU obtained from ENA were already preprocessed, i.e. de-multiplexed, and trimmed. Both of them were quality filtered with a Phred threshold of three using the script split_libraries_fastq.py (QIIME1, 1.9.1)[84,90].

2.4.4 Denoising

The Deblur (1.1.0) algorithm[91] was chosen to de-replicate sequences and remove potential sequencing errors. All sequences were trimmed to a length of 90 base pairs since most sequences in EMP had a length of 90 bases pairs, and the algorithm requires all sequences to have the same length. To strengthen the filtering rules, singletons per sample were removed before denoising by setting the min-size parameter to two. The algorithm corrected sequences based on a predefined error profile and returned amplicon sequence variants (ASV), which could be considered as putative error-free (representative) sequences for each sample. We adopted a method based on ASV instead of clustering sequences into operational taxonomic units (OTUs) because ASVs are (i) consistently labeled, thus facilitating meta-analysis of cross-study samples, and (ii) are not affected by the incompleteness of reference databases, hereby providing more accurate diversity estimates for bacterial communities [92–94]. A total of 256,620 unique ASV were identified with most of the sequences being relatively rare (14.94% observed only once and 70.79% less than ten times across all soil samples).

2.4.5 Taxonomy assignment for filtering of archaea

ASVs were assigned to taxonomic units using a multinomial Naive Bayes classifier (QIIME2, 2018.8.0), trained on the Greengenes 13_8[95], 99% OTUs (515F-806R region, 90 base pairs). Nevertheless, only 1.08% of the sequences could be assigned to a unique species designation. Sequences which were classified as archaea were removed, as they only contributed to a small proportion and may behave differently from bacteria[96]. Sequences that could not be classified confidently (<70%) at the lowest taxonomic levels (Kingdom) were discarded. Global singletons (observed only once across all samples) were dropped to remove potential errors and increase reliability.

2.4.6 Rarefaction and estimation of diversity

The optimal sequence rarefication depth (number of randomly drawn sequences without replacement from each sample) with respect to diversity was determined by a grid search over 2,500 to 15,000 (A2 Supplementary Fig. S5). After determining the rarefication depth, the procedure was repeated 100 times to increase reproducibility of ASVs abundance distributions. For each soil sample, diversity indices were calculated independently for each of the 100 rarefied ASVs tables and subsequently averaged[5]. The abundance of each ASVs was averaged over the 100 rarefied species abundance distributions, and thus may not be integer valued. We note that this procedure differs from common practices in ecological fields in which only one randomly generated rarefied ASVs (or OTUs) table is used for both diversity estimation and interpretation. From an ecological point of view, the randomness in the latter approach can be desired since in reality we would not have the ability to take multiple soil samples from the same site, amplify them independently and take the averaged diversity (corresponding to rarefying multiple times from an existing ASV or OTU table). However, from a statistical point of view, it lacks stability. In the foregoing analysis, we used the averaged (n=100) 7,500 ASVs as representative phylotypes for calculations of bacterial diversity in its general form (A2 Supplementary Methods).

2.4.7 Covariates

Soil properties were collected from 250 m SoilGrids[56] according to samples' geographical locations and soil depth. We did not use the on-site measured soil properties due to missing values and inconsistent methodologies of measurement across studies. Of additional concern was the comparison of variables measured at different scales. While it is common practice to compare remotely sensed covariates (e.g. temperature, primary productivity, precipitation) with sample scale measurements (e.g. pH, carbon-, nitrogen content) it is not desirable from a statistical point as the level of support varies. This can lead to misinterpretation of the relative variable importance with respect to their explanatory power and hereby would obscure our understanding of underlying processes. The mean annual net primary (NPP) productivity, obtained from MODIS 2000–2015[50] was used as a proxy for the net carbon influx and the distribution of land covers. Mean annual temperature (MAT) and solar radiation (RAD) were retrieved from WorldClim[52]. Mean annual precipitation (MAP) was estimated using MSWEP rainfall data[97]. Using mean monthly temperature and shortwave radiation as inputs, mean monthly potential evapotranspiration (PET) was calculated according to the empirical equation proposed by Jensen and Haise[58]. The empirical equation produced negative values at extremely low temperatures. These estimated negative PET are unrealistic and were replaced by zeros. The resulting mean monthly PET was averaged over one climatic year yielding the mean annual PET. The average number of consecutive dry days (DRY) was estimated from the MSWEP precipitation

time series. Briefly, daily precipitation was compared against the mean annual potential evapotranspiration (PET) to detect rainfall events that were expected to alter soil moisture conditions, i.e. exceeding the threshold set by PET. The values were reported as an absolute averaged spacing between rainfall events and could exceed one year. The available water capacity (AWC) in SoilGrids was derived based on a pedo-transfer function that depends on soil chemical conditions, e.g. soil pH (PH)[56]. Including soil chemistry in calculating AWC may potentially interfere with later interpretations. To avoid this, we alternatively estimated AWC by a function that only uses bulk density (BLD), organic carbon content (ORC), silt content (SLT) and clay content (CLY)[55]. Climatic water content (CWC) was introduced to describe the climatic state of soil wetness (A2 Supplementary Methods). It was calculated based on the assumption that the top one meter of soil can be fully replenished up to field capacity during rainfall events, and dry exponentially in consecutive days without rain (DRY). Summary of covariates is given in the A2 Supplementary Table S1.

2.4.8 Correlation and clustering

Spearman's rank correlation ρ_s was used to measure the pairwise correlation between covariates (A2 Supplementary Fig. S6). Covariates were then hierarchically clustered[98] according to their dissimilarity (distance), defined as $1 - |\rho_S|$. The inter-cluster distance was determined by the averaged dissimilarity of objects in different clusters (average linkage). The cluster size was selected by applying a dissimilarity threshold of 0.15. Within each cluster, only one covariate with the simple physical interpretation was retained (A2 Supplementary Fig. S6). Further, since sand (SND), silt (SLT) and clay content (CLY) are compositional, SND was discarded.

2.4.9 Generalized additive models

Generalized additive models (GAM) (R package mgcv, 1.8-24) were used to model the associations in both univariate and multivariate analysis[99]. Thin plate regression spline was chosen as basis function and the smoothing parameters were estimated by restricted maximum likelihood (REML). The dimension of the basis used for each smoothing term was not restricted (default parameter *k*). Forward selection in multivariate modeling was performed based on Akaike information criterion (*AIC*) and likelihood ratio tests (conditional on the estimated smoothing parameters). The double penalty approach of GAM was used for regularization. Covariates were considered as negligible in terms of contributions to model fits if their estimated degree of freedom were shrunk approximately to zero ($<10^{-3}$). The prediction performance was evaluated using leave-one-out cross-validated coefficient of determination (R^2) and root mean squared error (*RMSE*).

2.4.10 Causal additive models

Causal additive models (CAM) (R package CAM, 1.0) were used to infer the underlying data generating mechanism (causal structure) from observational data[79]. The model is a special case of the general structural equation model (SEM)[100], namely in that the structural equations are additive in variables and errors. The model further assumes no hidden variables, i.e. all variables involved in the data generating mechanism are observed, and absence of directed cycles in the causal graph. Since the dimension of the dataset was low (15 covariates, except SND), we did not use preliminary neighborhood selection (screening of covariates primarily aimed for reduction of computational time). Furthermore, in order to avoid using data twice (for both variable selection and inference after selection)[101,102], as well as the issue of "*p*-value lottery"[103–105], the last step (pruning of the directed graph) of CAM was combined with the multi-splits method[104]. Briefly, the method randomly splits data into training and testing sets; the training set is used for estimating the graph structure while the testing set is used for computing *p*-values of each covariate (repeated 100 times to avoid noisy selection of covariates and to stabilize the results).

2.4.11 Prediction of global maps using tree-based algorithms

Random forests (RF) (RandomForestRegressor in scikit-learn, 0.19.1) and gradient boosting trees (GB) (GradientBoostingRegressor in scikit-learn, 0.19.1) were used for prediction[106]. Hyperparameters (n_estimators, max_features, max_depth and min_samples_leaf) in both algorithms were optimized using cross validation (CV) with respect to R^2. Additionally, the learning rate in the boosting algorithm was set to a constant value of 0.05 since it can be compensated by the number of iterations. Independently trained random forest and gradient boosting trees were stacked by simple averaging. The generalization errors (R^2 and *RMSE*) were estimated using nested (ten by ten folds) CV, i.e. the inner CV selected the best-fit models (optimizing hyperparameters with respect to R^2) while the outer CV computed the test errors of the selected models. The entire procedure was repeated ten times using different random splits (or seeds) to increase stability. Using the estimated model, we predicted global bacterial richness at the full spatial coverage of covariates.

3 The chosen few – variations in common and rare soil bacteria across biomes

Samuel Bickel and Dani Or

Submitted

Abstract

Soil bacterial communities are dominated by a few abundant species while their richness is attributed to rare species with largely unknown ecological roles. Novel classification of common and rare soil bacteria reveals consistent changes of rarity across terrestrial biomes. Variations in rarity are driven by environmental conditions; prominently soil wetness. Observations and mechanistic model results show an increase in rare bacterial species proportions for drier climatic conditions and lower soil carbon inputs. Soil bacterial species compositional shift results from suppression of common species activity in dry soils with implications for carbon and nutrient turnover. Insights into soil and climatic drivers of the rare soil microbiome help unravel contributions to ecosystem functioning that vary across biomes.

3.1 Introduction

Bacterial communities are characterized by strongly skewed relative abundance distributions (RADs) with most phylotypes (or "species" for simplicity) present at low relative abundances[107] (RAs) providing important ecosystem functions[6]. Despite the vast richness of prokaryotic taxa[19], only a "chosen few" species are consistently prominent across soils from different environments[20,108]. The richness of soil bacterial communities is largely determined by rare species that constitute the long tail of the RAD, often associated with functional diversity[2,26,109] and specific ecosystem functions[110–114]. The functional potential of the soil microbiome is directly linked with its genetic composition[2] with strong "functional redundancy" such that the loss of few species would not alter ecosystem resilience[115]. Evidence suggests that rare bacterial species contribute to specific functional traits[2,109,116,117] and exhibit greater sensitivity to environmental factors relative to common species[17,118,119]. Notwithstanding the large body of information on the links between bacterial diversity and ecosystem functioning[26,110–113], only a few endeavored to distinguish between globally common and rare bacteria[108,116]. Rare bacterial species are often ignored[6] focusing on common species only in studies on RA patterns[20]. Operationally, the classification of common and rare species is based on their prevalence ("ubiquity") or their (relative) abundance. Prevalence measures the probability of detecting a species across soil samples, while RA measures the probability of encountering a species within a soil sample (~1 cm^3). Both aspects are important for assessing how likely it is to find a bacterial species (or group of species) in soils across biomes and link species prevalence and abundance with ecosystem functioning[64,108,120]. The processes and environmental factors that affect rare bacterial species remain largely unknown[118,121,122] or are overlooked[65]. The broad environmental range and fitness of common bacterial species[64,118] enable them to succeed across environmental conditions[119] thus these are often poor indicators for changes in community composition across biomes[65,123]. In contrast, rare soil bacterial species vary with climate and soil properties[17] and are found in only a few samples hence limiting data-driven inferences regarding details of their spatial distributions. Overall, the conditions under which rare species substantially contribute to ecosystem services remain understudied across the entire tree of life[124]. To better assess their role and properly attribute the contribution of the rare microbiome to ecological functioning we need a universal classification of common and rare bacterial species.

Here we seek to quantify what determines the proportions of common and rare soil bacterial species across biomes by: (i) developing a universal metric for global classification of rare and common soil bacteria based on their RA and prevalence; (ii) identifying patterns of richness and abundance in relation to environmental conditions, and (iii) employing a mechanistic model to quantify how climatic factors shape proportions of rare soil bacteria. The study is motivated by statistical models that

demonstrated increased explanatory power of environmental variables for bacterial diversity when additional weight was given to locally low abundant species[17]. In this study, we distinguish soil bacterial species as common and rare based on pooled RAD from all sampled soils across the globe. The classification quantifies the proportions of rare and common bacterial species in individual soil samples to demonstrate how both groups are affected by key environmental variables that define various biomes (Fig. 3.1); primarily the climatic water content[9] (CWC), the net primary productivity (NPP) and mean annual temperature (MAT). Evidence suggests that CWC plays a crucial ecological role in promoting bacterial diversity by the intrinsic fragmentation (or connectedness) of individual microscale soil aqueous habitats[9,17]. High values of CWC (wetter soils) support higher vegetation density and increase carbon inputs thereby enhancing soil carrying capacity and total bacterial biomass[7–9]. We use a spatially-explicit individual-based model (SIM) to simulate soil moisture dependent growth and motility of multispecies soil bacterial communities that rely on shared resources. Modeled species are represented by unique combinations of kinetic (Monod type) parameters that reflect their competitive ability under locally variable carbon source concentrations. Predictions of the SIM were evaluated for a range of soil moisture conditions that also affect diffusion from heterogeneous carbon sources via water films on hydrated soil surfaces. We hypothesize that drier soils with highly fragmented aqueous habitats and restricted diffusion of carbon suppress the activity of fast-growing common bacteria and lead to communities with more even RADs[38].

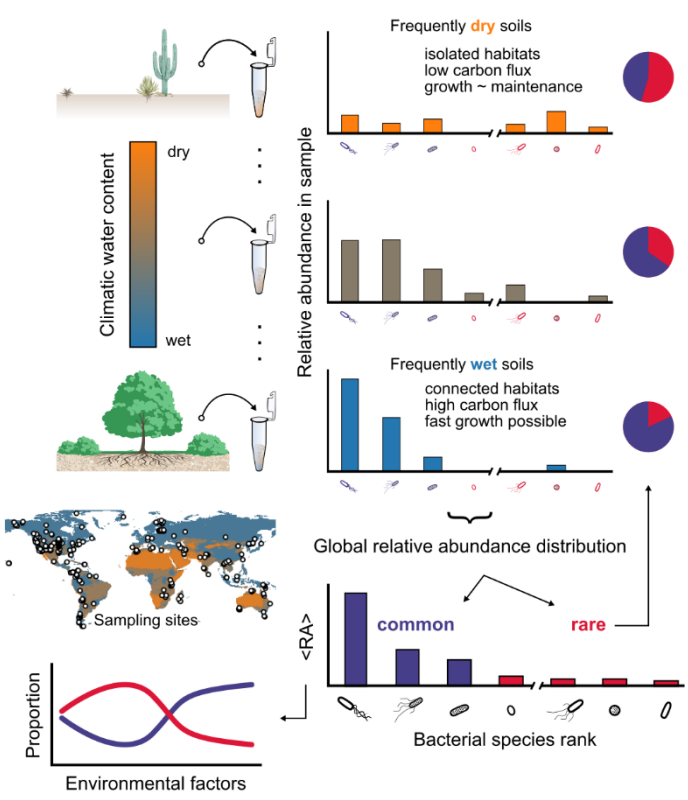


Fig. 3.1 Tracking globally common and rare soil bacteria across environmental conditions. We perform analysis of a global pool of bacterial relative abundance distributions (RADs) from soil samples across biomes with a wide range of climatic water contents. Bacterial species relative abundance (RA) from individual samples are used to obtain the global, average RAD and species ranking. Each species is classified as common or rare based on the global RAD. The proportions of common and rare bacteria are tracked across environmental factors. Particular focus lies on shifts in proportions with climatic water content that is a proxy for a soil's aqueous phase connectedness. Dry soils are expected to host communities with higher evenness that include many rare species by physically limiting growth to isolated aqueous habitats with low carbon fluxes. In frequently wet soils a "chosen few" common species are expected to grow rapidly and dominate the soil bacterial community.

3.2 Results

3.2.1 Relative abundance and prevalence of common and rare soil bacteria.

We have used published[17] genomic data (16S rRNA gene sequences) from soil samples[5,18,28] (n = 844) of 318 sites across major biomes and identified global patterns of common and rare bacteria. Common bacterial "species" (defined by 90 bp rRNA amplicon sequence variants) were distinguished using a global (across samples) threshold of RA based on minimizing cross-entropy[125], i.e. a threshold that minimizes the amount of information needed to reconstruct the RAD given the binary classification of common and rare species (Fig. 3.2 a). The resulting threshold to delineate RA of common species is remarkably consistent (0.019 ± 0.002%) and is comparable to previous, empirical or operationally defined thresholds based on RA[17,108,126]. Most bacterial species were classified as rare (99.6%) yet they make up only 42% of the global RA. The non-parametric threshold selection and resulting average proportion of rare species were robust even when using only ¼ of all samples available (A3 Supplementary Table S1). Soil bacterial community richness and cumulative RA of rare and common species varied among biomes indicating sensitivity to environmental conditions (Fig. 3.2 b and c). The differences in rare and common RAs were most pronounced for large changes in CWC values as also predicted by an aqueous-phase fragmentation-based heuristic model[9] evaluated for soils and climates of different biomes (Fig. 3.2 c). Generally, common species with high RA were more prevalent than rare species (Fig. 3.2 d). The average prevalence (median ± IQR) for common species (0.3 ± 0.2) was 300 times larger than for rare species (0.001 ± 0.003). The ratio of rare species richness to common species richness decreased significantly with more frequent rainfall (exponential R^2 = 0.19, Pearson r = -0.41, n = 318), indicating that community composition may vary with the climatic soil water content (Fig 3.2 e).

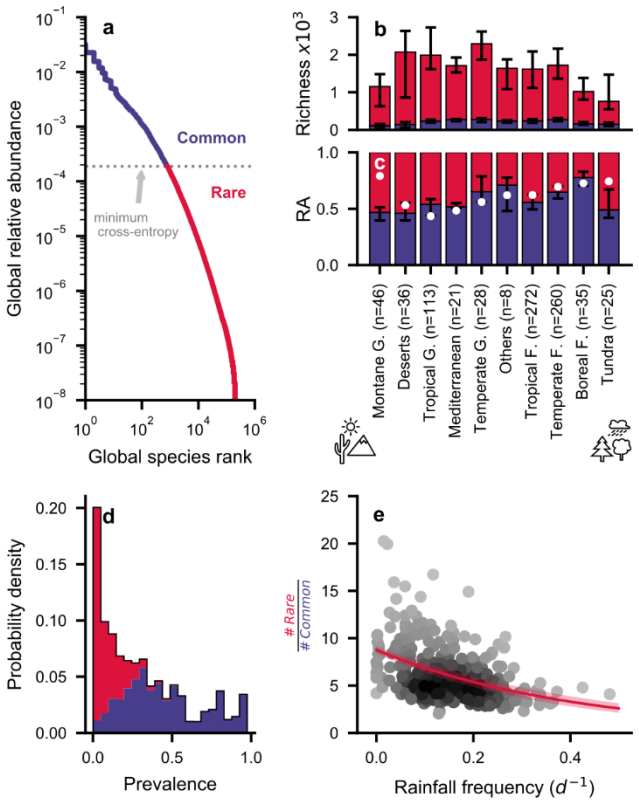


Fig. 3.2 Observed proportion and richness of common and rare bacteria across biomes. a, Global relative abundance distribution (RAD) of bacterial species (n = 844). The dashed line indicates the threshold based on minimum cross-entropy[125] that distinguishes common and rare species using only the global RAD (shown in purple and red, respectively). **b,** and **c,** Bacterial richness and relative abundance (RA) for common and rare bacteria vary across biomes. Stacked bars indicate the median ± IQR. The number of samples for each biome are reported (G = grassland, F = forest). **b,** Richness of rare bacteria varies more strongly among biomes compared to common bacteria while **c,** their RA seldom exceeds the RA of common species. This general tendency is also predicted by a recent heuristic model[9] (white circles). **d,** Prevalence, the fraction of samples in which a species occurs, is related to the global RAD. **e,** The ratio of rare to common species declines with increasing rainfall frequency for different sampling sites (exponential R^2 = 0.19; Pearson r = -0.41, n = 318).

3.2.2 Rarity of soil bacterial species driven by climatic water contents.

A few common species dominate bacterial communities in wet soils as also predicted by the mechanistic SIM that makes no assumptions regarding species composition or their relation to soil moisture conditions (Fig. 3.3 a). A gradual shift in RADs towards more even communities with larger proportions of rare species was observed in transition to dry soils (A3 SI Appendix, Fig. S1). We compare our computed CWC values based on rainfall frequency with yearly averaged soil moisture data obtained from climate model reanalysis (ERA5-land), the two estimates show good agreement for the overlapping range (A3 SI Appendix, Fig. S2). To detect changes in ranked community composition with CWC, each sample's RAD was compared to the global RAD. Spearman rank correlation was lower and Bray-Curtis community dissimilarity was higher for samples originating from drier environments (A3 SI Appendix, Fig. S3). This indicates changes in the subset and ranking of species. Additionally, the amount of information contained in the RAD of individual samples relative to the global RAD ("discrimination information") displayed a consistent decrease with increasing CWC as was previously postulated for reduced environmental heterogeneity[127] (A3 SI Appendix, Fig. S3). However, rare species proportions could be biased by the physiological state of the bacteria[65]. To test how the changes in rare proportions may be affected by the presence of inactive (dormant) bacteria, we removed from the modeling results cells that did not divide. The removal of these inactive cells (dormant or at maintenance rate state) resulted in a sharp decrease in the proportion of rare species

in very dry soils (Fig. 3.3 a). Bacterial cell density increases significantly under wet climatic conditions that also promote vegetation and carbon inputs as seen in model simulations and empirical estimates of maximal cell density (Fig. 3.3 b). This carrying capacity was estimated from carbon input by NPP and mean maintenance requirements of soil bacteria (adjusted for MAT) with no explicit dependency on CWC[9]. Considered independently, these two factors (NPP or MAT) did not exhibit clear tendencies for changes in proportions of common and rare bacterial species (A3 SI Appendix, Fig. S4). We examined effects of temperature using the SIM with temperature dependent bacterial growth[44,128]. Biome-specific CWC and MAT were used as boundary conditions for comparison with data and highlight the predominant influence of soil moisture on soil carrying capacity and the proportion of rare species (A3 SI Appendix, Fig. S5). We note, that CWC is affected by MAT via potential evapotranspiration that increases with higher temperatures. CWC therefore also co-varies with soil pH[9,17] due to the influence of climatic water balance on the value of soil pH[49] that is often reported as a key driver of bacterial diversity[18] and species abundance[20] (A3 SI Appendix, Fig. S4).

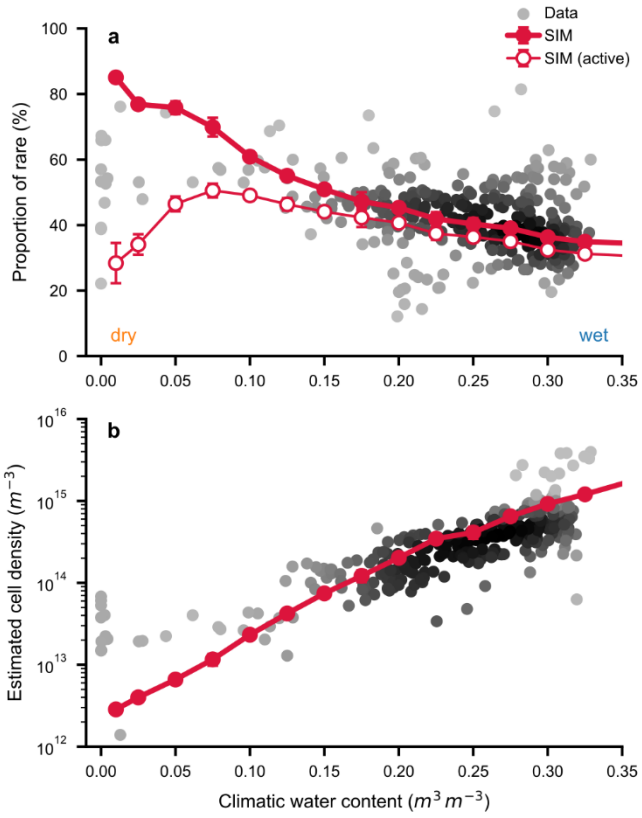


Fig. 3.3 Decline in observed soil bacterial rarity is mechanistically linked with hydration conditions and carrying capacity. a, The proportion of rare bacteria decreases with increasing climatic water contents (Spearman ρ =-0.36, n = 318; highest density of points shown in black). The decline in rare species proportion is predicted by a spatially-explicit individual-based model (SIM, mean ± SD, n = 5) and compares favorably with empirical observations. Considering only active cells in the SIM causes a drop in rare proportion under dry conditions (open symbols). **b,** Estimated cell density (potential carrying capacity) increases exponentially with climatic water contents. Cell density is calculated based on mean annual temperature, carbon input flux (net primary productivity) and bacterial maintenance requirements using independent data[8,9]. The prediction by the SIM makes no assumptions on the relation of cell density with soil hydration conditions.

3.2.3 How is bacterial species dominance reduced in dry soil?

A distinct shift in the soil bacterial RAD was observed for different (climatic) soil hydration conditions, with a smaller proportion of common species found in dry regions (Fig. 3.4 a). SIM results suggest that the common species become suppressed under dry conditions where their superior physiological traits cannot be expressed and thus their activity is equalized with less fit species[38] (Fig. 3.4 b). In other words, the simulated bacterial community composition under dry conditions becomes more even in terms of RA[38,129] and distribution of maximum growth rates. The total number of simulated individuals ranged from ~10^3 to 10^6 and closely followed the soil water content and average physiological parameters (maximal growth rate and carbon source affinity; A3 SI Appendix, Fig. S6). This implies that the soil bacterial community is capable of responding to changes in soil hydration conditions at timescales of days (as used in the SIM) from initial or background community composition associated with climatic time scales.

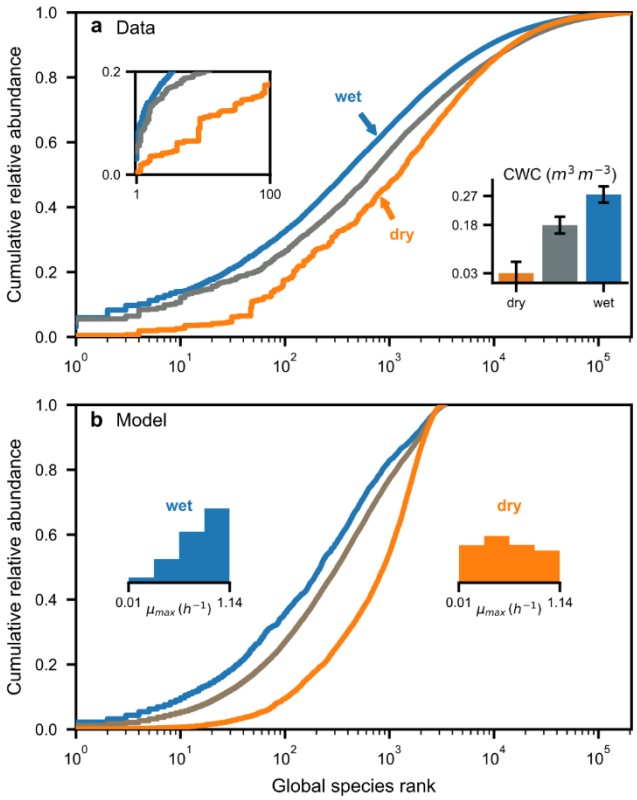


Fig. 3.4 Soil bacterial community shifts with water content – drier conditions suppress common species. a, The relative abundance distributions (RADs) of soil bacteria for three groups of climatic water contents (CWC; bars indicate mean ± SD) are shown as cumulative relative abundance[17]. Values are sorted by global species rank with one indicating the globally most abundant. The RAD displays a systematic shift towards high ranks under dry conditions. The inset figure on the left shows the 100 most abundant species on a linear scale. **b,** The spatially-explicit individual-based model (SIM) confirms the observed tendency. The distribution of modelled species maximum growth rates (μ_{max}) at the end of the simulation indicates that physiological differences are equalized under dry conditions while the increased relative abundance of common species under wet conditions corresponds to higher growth rates.

These findings are in qualitative agreement with previous observations of community activity under perturbation. We have used our classification scheme to track the proportions of rare species in a desert soil community[75] responding to a rainfall event (Fig. 3.5). The daily observations are comparable with simulations by the SIM and reflect proportions of bacterial species activity[75,130]. Following a winter rainfall event in the Negev desert, the activity of rare bacterial species dropped during soil wetting and recovered to initial values following soil drying (Fig. 3.5 a). The community displayed consistent shifts where the common species dominated under wet conditions but were suppressed when the soil was dry (Fig. 3.5 b). We note, however, that the proportion of rare species in this dataset appeared extremely small. This could be partially explained by the taxonomic assignment used which did not allow to resolve species with very low RAs. These "unassigned" species were removed from the analysis and caused the RA to not sum up to unity. More importantly, the RNA-based measurements exclude dormant species that could constitute a large proportion of the bacterial community in dry soils (Fig. 3.3 a).

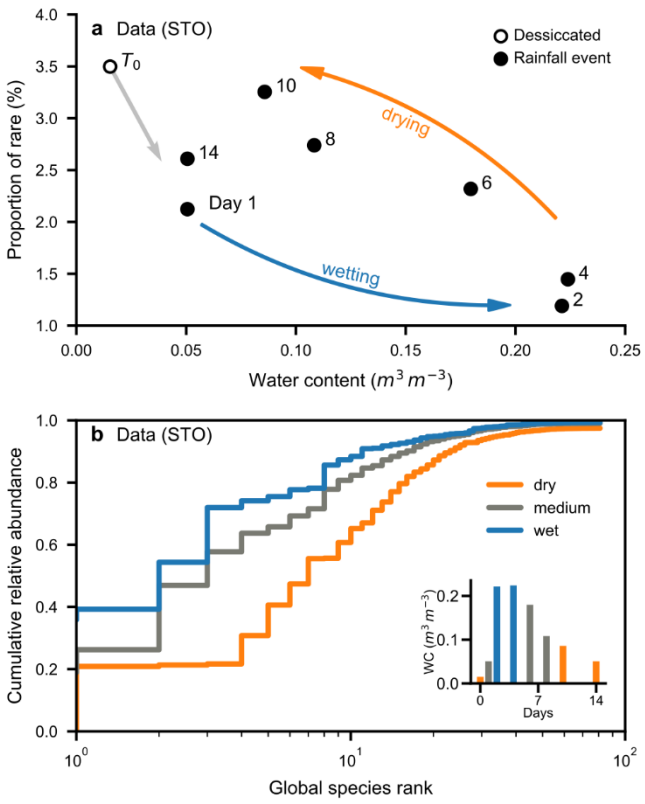


Fig. 3.5 Activity of common and rare soil bacteria shaped by rainfall. a, Short term shifts in proportion of rare species following a winter rain event in the Negev desert (day 1 and 2; T_0 marks full desiccation in summer) with subsequent return to initial community composition as previously reported (STO[75]). Each point represents averaged measurements of bacterial RNA abundance (n = 3). **b,** Samples were further aggregated to three hydration conditions[75] and are displayed as ranked, cumulative RA. Water contents (WC) at the time of sampling are shown in the inset figure.

3.2.4 Spatial patterns of bacterial rarity and functional consequences.

The proportion of soil samples of our dataset in which rare species jointly dominate community composition exhibit a steep transition with CWC (Fig 3.6 a). We attribute this transition to a critical (mean) water content (θ_c) above which the aqueous phase is frequently connected (estimated[60] as $\theta_c \approx \theta_s\, p_c$, with soil porosity[56] θ_s and $p_c \approx 0.31$ for site percolation on a simple cubic lattice[59]). Considering the universal role of water contents in structuring the soil bacterial microbiome we can map global regions where rare bacteria are, on average, likely to dominate (Fig 3.6 b). The climatic transition region is given by the central 95% of global θ_c values. Variations of ecosystem functions[113] and the association of CWC and bacterial diversity[9,17] coincide with this range. For example, the interplay in Glucose mineralization and β-Glucosidase activity reported[113] (A3 SI Appendix, Fig. S7). Data show that Glucose is readily mineralized under wet conditions whereas the activity of β-Glucosidase, used to decompose more complex carbon sources, has been reduced. Such specialized functions have been attributed to rare species present at low abundances[109,117]. Similar patterns could also be observed in a recent microcosm study with artificial diversity gradient where several ecosystem functions were associated with soil microbial richness[111] and (using our approach) with proportions of rare species (A3 SI Appendix, Fig. S8). Leaf litter decomposition was positively associated with the proportion of rare bacteria while other functions such as leaching of Nitrogen (N) or Phosphorus (P), and the uptake of N and P into Grasses, Legumes and Forbes displayed mixed yet systematic tendencies. These findings illustrate the multitude of potential associations of rare bacteria with specialized ecosystem functions and consider not only richness but the RA of rare members that may vary spatially.

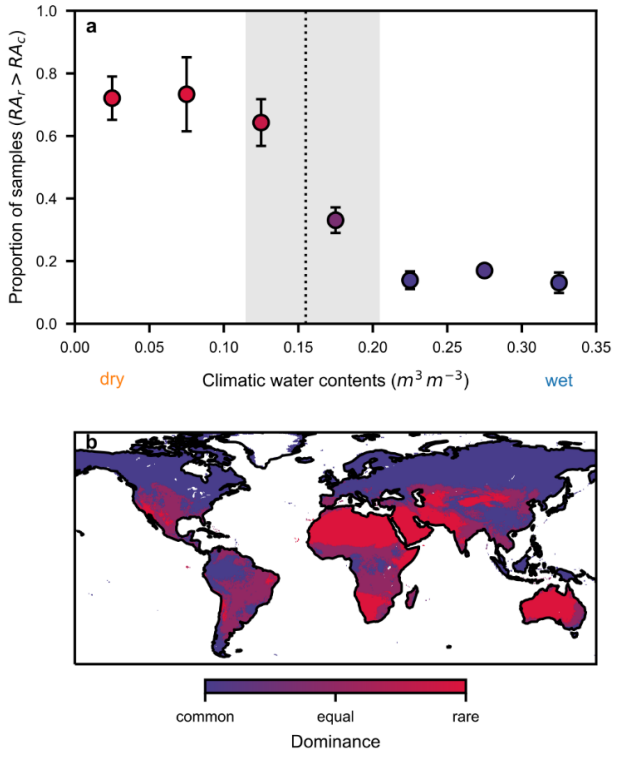


Fig. 3.6 Critical role of soil hydration and corresponding spatial patterns of rarity. a, The proportion of samples where the relative abundance (RA) of rare species (RA_r) exceeds the relative abundance of common species (RA_c) displayed a transition along a gradient of climatic water contents (CWC). Samples were binned by CWC (bin width: 0.05; mean ± SE). The dashed vertical line illustrates the median critical water content below which the aqueous phase is largely disconnected (shading represents the central 95% of global values). **b,** Regions based on CWC where rare bacteria are likely to dominate the community RA (red), where common and rare transition (purple) and where the common dominate (blue).

3.3 Discussion

The proposed non-parametric classification of rare and common soil bacterial species makes no assumptions regarding the shape of the RAD and permits relative comparison of data from different studies and model results. The classification is insensitive to "noise" among (rare) species with low abundance since it does not contain information that would affect the threshold selection[125]. This global classification consistently labels soil bacterial species across all soils and biomes. The large environmental range of common species with high RA is attributed to their intrinsic fitness[64] that supports their global prevalence (Fig. 3.2 d). Nevertheless, changes in species RA ranks occur frequently across environments (A3 SI Appendix, Fig. S3). Such variations could be linked to functional diversity of soil bacteria and to specific functional roles[109] and genetic potential[64,120]. Overall, we observe greater similarity in RADs with a decline in "discrimination information" towards wetter soils (A3 SI Appendix, Fig. S3). This suggests that bacterial community RADs resemble the global RAD in wet soils where communities follow more closely the global species ranking. Dry soils may harbor communities with even bacterial diversity in which species are sheltered in isolated aqueous micro-habitats[129] that impose "gene flow discontinuities"[131]. The sparse vegetation growing in arid soils further limit carbon inputs and soil carrying capacity thereby suppressing the fast growth of "chosen few" common bacterial species and leads to more even bacterial communities. A transition in processes governing community assembly has been postulated[127] for decreased environmental "randomness" as expected in wet soils with enhanced aqueous phase connectedness. A well connected soil aqueous phase (on average) also implies higher fluxes of nutrients and bacterial mobility (i.e. a "selection-dominated" regime[127]). In contrast, aqueous habitat fragmentation in dry soils imposes "randomness" (i.e. "drift-dominated" communities[127]). Reduced nutrient fluxes in dry soils constrain the physiological advantage of common species in agreement with recent experimental evidence[132] and as seen in the SIM results with more even distributions of abundances and maximal growth rates under dry conditions (Fig. 3.4 b). Growth limiting conditions in desert soils cause a large drop of rare species proportion when removing inactive cells from the simulation results (Fig. 3.3 a) consistent with observed low activity of rare species (Fig. 3.5). We should expect many bacterial cells in natural communities to be dormant and at low abundances[65,130] with particular functional implications for dry soils[114]. Sensitivity of rare species to environmental conditions may be explained by a hydration-centered modeling framework without assigning specific functional traits. Rare bacteria constitute a deep reservoir of physiological traits and we can expect their functional contributions to vary with CWC. Broader ecosystem functions, such as soil heterotrophic respiration, are widespread among bacterial species[2] and are likely to be associated with the activity of common species that make up most of the community biomass[124]. This is evidenced, for example, by rapid

saturation of CO_2 production with increasing bacterial richness in microcosm experiments[26,35]. The degradation of complex carbon sources, on the other hand, requires activity of specific enzymes that are thought to be contributed by metabolically versatile but rare bacteria[2,109,111,121,133] and could be associated with slower growth compared to mineralization of readily degradable sugars. Small differences in peak activity could amplify the RA of particular bacterial taxa to extents that make them common across samples. For example, few taxonomic groups (γ-Proteobacteria, Clostridia, Bacilli and Bacteroida) dominated bacterial activity in a desert community only while the soil was wet[75]. This short time span was enough to propel these taxa to prominence[114], causing them to be labeled as "common" across samples. In other words, the "chosen few" common species are "kings" when it is wet but remain hidden among the rare otherwise. Since rainfall events in deserts are very infrequent and offer very short windows of opportunity, we do not expect many globally common species to be detected in this biome and observe that rare species are on average seven times more numerous in regions characterized by low rainfall frequencies relative to common species (Fig 3.2 e). This indicates robustness of our procedure in delineating common species and successfully captured dynamics of soil bacterial community activity that could be manifested under climatic time scales. Soil moisture predictably alters the shape of the RAD and points to the variable's importance for disentangling effects of other environmental factors (e.g. carbon input, temperature) that are reconciled in the context of biome-specific hydration conditions and carrying capacity (A3 SI Appendix, Fig. S5). Carrying capacity increases for lower temperature by reducing maintenance and growth rates[134] but shifts caused by temperature are much smaller compared to those caused by soil hydration conditions in agreement with observations on the global drivers of soil microbial carbon[7]. High temperatures are further associated with drier soils that push many cells into dormancy due to reduced carbon fluxes. We distinguish environments in which bacterial species abundance is shaped primarily by physical constraints (fragmented aqueous habitats under dry conditions) with limited biomass production from environments where physiological traits could shape community composition (enhanced nutrient fluxes under wet conditions)[123,127]. Regions dominated by rare bacterial species (Fig 3.6 b) could harbor large functional potential that is readily expressed[114] under variations in climatic conditions and render the rare soil microbiome sensitive to environmental changes.

3.4 Materials and Methods

3.4.1 Soil bacterial community data.

A previously published dataset on soil bacterial community composition was used to delineate patterns of common and rare soil bacteria across biomes. The detailed methodology that was used to combine raw (16S rRNA V4) sequence data of soil samples from three studies[5,18,28] was previously described[17]. Briefly, the 16S rRNA sequences were de-replicated and de-noised after trimming to 90 bp length. Singletons were removed before de-noising for each individual sample (n = 844) resulting in a total of 256'620 unique amplicon sequence variants (ASV) of which 71% were observed less than ten times across samples. ASVs were assigned taxonomy using a multinomial Naive Bayes classifier trained on Greengenes 13_8, 99% OTUs (515F-806R region). Sequences that could not be classified confidently (<70%) as bacteria at the Kingdom level or sequences classified as archaea were discarded. Additionally, global singletons (observed only once across samples) have been removed. The resulting table of ASV abundance (referred to as "species" abundance) was then rarefied to a total count of 7'500 per sample. Independent rarefication was averaged for 115 realizations (differing slightly from previous[17] 100). Prevalence of each species was estimated as the number of non-zero rarefied counts *c* divided by the number of samples *n*. The sample relative abundance (proportion) *p* was obtained by dividing counts of species *i* for every sample *k* by the total counts (*N* = 7'500) as $p_{i,k} = \frac{c_{i,k}}{N}$. Species counts that were absent (below the limit of detection) where imputed with zero values with only negligible effects on other species proportions. Subsequently we obtained the global relative abundance *g* for each species by averaging across samples according to $g_i = \frac{\sum_{k=1}^{n} p_{i,k}}{n}$. We thus distinguish the local (e.g. sample) relative abundance distribution RAD from the global RAD that is subsequently used for classification of common and rare soil bacteria.

3.4.2 Classification of common and rare bacteria.

An algorithm for automatic threshold selection based on minimizing cross-entropy[125] was used to designate common and rare bacteria using only the global RAD. The algorithm was originally developed for image segmentation and was previously implemented (function "threshold_li" in scikit-image[135]). This approach makes no *a priori* assumption on the underlying distribution of values and provides the most unbiased estimate of the binary classification[125]. Here we use the obtained threshold value *t* to distinguish common and rare species based on each species global relative abundance g_i. The species *i* with $\{g_i | g_i \in [0,1], g_i \leq t\}$ were considered "rare" and species with $\{g_i | g_i \in [0,1], g_i > t\}$ are defined as "common". The relative abundance (RA) of rare (RA_r) and common (RA_c) species in a single sample are thus given by: $\mathrm{RA}_{\mathrm{r},k} = \sum_{g_i \leq t} p_{i,k}$ and $\mathrm{RA}_{\mathrm{c},k} = \sum_{g_i > t} p_{i,k}$ for proportions of rare and common species, respectively.

3.4.3 Climatic data of sampling locations.

Covariates for each topsoil (≤ 10cm) sample were added at their highest native resolution based on latitude and longitude using nearest neighbor interpolation as previously reported[17]. Net primary productivity (MODIS[50], averaged for 2000-2015) and mean annual temperature (WorldClim[52]) were used to estimate maximal cell density (potential carrying capacity) by dividing location specific soil carbon input flux by a temperature dependent[44], biomass carbon-specific maintenance rate (≈10^{-4} gC gC^{-1} h^{-1}) as previously described[9]. We have used climatic water contents (CWC) as a proxy for climatic soil hydration conditions and soil aqueous phase connectivity[9,17]. Values were based on global gridded precipitation time series (MSWEP[57], daily for 1979-2016 at 0.1° spatial resolution) that yield the average number of consecutive dry days (DRY) used for the calculation of CWC. Rainfall frequency was estimated by taking the inverse of DRY (f_{rain} = 1/DRY). The estimates of CWC were also compared to mean soil moisture obtained from recent climate reanalysis (ECMWF ERA5-land, 0-7 cm, monthly for 1981-2019 at 0.1° spatial resolution, https://doi.org/10.24381/cds.68d2bb30).

3.4.4 Spatially-explicit individual-based model (SIM).

An individual-based model was used to simulate growth of diverse bacterial species on heterogeneous soil surfaces[9,23,39]. Briefly, continuous growth and movement of individual cells of different species in the pore space of a defined soil volume (specified by area and thickness of a soil slab; 1 mm^2 and 11 µm) was simulated for 8 days at 1-min time steps with nutrients arriving on average every 4 h (replenishing carbon sources that potentially enable a maximum carrying capacity of around 10^{17} cells per m^3). A single cell of each species was initially placed randomly on the two-dimensional domain. At the end of the simulations, cells of each species were counted to obtain the RAD. We modified our previous implementation[9] (https://doi.org/10.5281/zenodo.3558542) to allow multiple nutrients ("carbon sources") to be consumed by different species. Each species *i* is represented by a set of Monod parameters (maximal growth rate μ_{max} and half saturation constant K) for all three carbon sources *j*. Resulting in three sets of parameters per species $(\mu_{max,i}, K_i)_j$. Considering four steps for the discretization of the parameter space this resulted in R = 3'360 species (possible permutations of parameter pairs). Cells could use carbon sources simultaneously with a partitioning of growth capacity loosely based on a previously developed model[136] generalized for arbitrary number of nutrients. The sub-additive growth rate of a species depending on carbon source concentration C_j is given by Eq. 1:

$$\mu_i = \frac{\sum_j \mu_{max,i,j}\chi_{i,j}}{(1+\sum_j \alpha_{i,j}\chi_{i,j})} \text{ with } \chi_{i,j} = \frac{C_j}{K_{i,j}+(1-\alpha_{i,j})C_j} \text{ and } \alpha_{i,j} = \frac{\mu_{max,i,j}}{\sum_j \mu_{max,i,j}} \quad (1)$$

The partitioning of cellular capacity (e.g. cell surface area for nutrient absorption) is described by normalizing Monod coefficients using Eq. 2:

$$\kappa_{i,j} = \frac{M_{i,j}}{\sum_j M_{i,j}} \text{ with } M_{i,j} = \frac{C_j}{K_{i,j}+C_j} \quad (2)$$

This leads to the definition of species-specific cellular maintenance rates $k_{m,i}$ that are a weighted fraction f_m of maximal, nutrient-specific growth rates described by Eq. 3:

$$k_{m,i} = f_m \sum_j \kappa_{i,j} \mu_{max,i,j} \quad (3)$$

The change in cell mass m_i of a species over time Δt is given by Eq. **4**:

$$\Delta m_i = \left(\mu_i - k_{m,i}\right) m_i \Delta t \quad (4)$$

Resource utilization $r_{i,j}$ is assigned a constant yield Y across species and nutrients (Eq. 5) that lead to changes in mass of nutrients m_j by converting local concentrations using the volume of water V_w (both vary spatially) as formulated in Eq. 6.

$$r_{i,j} = -\frac{\kappa_{i,j} \mu_i m_i}{Y} \quad (5)$$

$$\Delta m_j = \frac{\Delta C_j}{V_w} = \sum_{i=1}^{R} r_{i,j} \, \Delta t \quad (6)$$

For simulations with temperature dependency two additional factors based on the Schoolfield[44] model are used that reduce maximum growth and maintenance rates ($f_\mu(T)$,.$f_k(T)$, respectively). The factor affecting maintenance rates is assumed constant above the optimum temperature (unlike $f_\mu(T)$ that drops sharply above optimum temperature) but is otherwise identical. Parameters used for the temperature dependency are obtained from a recent study[128]. The Schoolfield[44] model was fitted to published data[128] of different species to obtain a single, average temperature response curve that was normalized to the maximal rate (rate at optimal temperature T_{opt}). Parameters used were: energy of activation (E_a = 0.4 ± 0.1 eV), energy of inactivation (E_h = 2.0 ± 0.1 eV) and temperature of inactivation (T_h = 31 ± 2 °C) that result in T_{opt} = 27 °C, which marks the temperature of the SIM without explicit temperature dependency. In simulations of biomes with subzero MAT a temperature of 0 °C was used instead of biome average MAT.

4 How soil bacterial microgeography affects community interactions and soil functions

Samuel Bickel and Dani Or

Submitted

Abstract

In contrast to rapid advances in resolving the global biogeography of soil bacteria, surprisingly little is known about how bacterial communities are distributed within the soil body at scales relevant to their interactions. The patchiness of nutrient sources and aqueous-phase configurations promote spatial clustering of bacterial cells in a few favorable locations leaving large soil volumes sparsely populated. We propose a heuristic framework for deducing microscale soil bacterial community sizes and spatial distributions from macroscopic soil traits, wetness and average bacterial cell densities. Results from an individual-based mechanistic model at high spatial resolution demonstrate sensitivity of bacterial community sizes to soil aqueous-phase connectivity and carbon fluxes. The proposed heuristic model links climatic and soil variables that shape soil bacterial microgeography across biomes. We have made direct observations of community sizes on natural soil surfaces under controlled hydration conditions in support of modelled spatial distributions. The sizes of bacterial communities and their spatial configurations at scales of diffusion-mediated trophic and other cell-cell interactions are critical for interpreting precursors for soil bacterial diversity and emergence of microbially-mediated soil functions under different climatic conditions.

4.1 Introduction

The distribution of organisms across spatial scales is of considerable ecological interest often studied with focus on large scale biogeographic patterns[33,137] and linking these to climate, terrain, vegetation and other spatial attributes. Applied to soil microorganisms, similar approaches have revealed drivers of microbial abundance[7,8] and diversity[5,33]. Soil bacterial abundance ranks high in the global biomass distribution[4] and is largely driven by precipitation[7], temperature and associated factors such as vegetation-derived primary productivity. Notwithstanding recent advances in identification of global microbial abundance patterns, the spatial distribution of bacterial cells and colonies at the submillimeter scale that matters most to interactions remains understudied[12,138–140]. In many applications, the analysis of soil bacterial communities is based on bulk soil samples of a few centimeters in scale that potentially mix otherwise spatially isolated and distinct populations[16,48]. Identifying the relative extent of spatial interactions at the scale of bacterial communities or cell clusters is important for interpreting measurements made at coarser scales, and for mechanistic understanding of soil microbiome functioning[1,16,141]. Given limitations to direct observation of microbial life in the opaque soil pores, mechanistic modeling at the resolution of individual cells and interacting colonies offers a means for bridging the present information gap and gaining new insights into yet-unobservable soil bacterial processes.

Soils are characterized by complex pore spaces with large specific surface area available for bacterial colonization[14]. Common bacterial cell densities (10^7-10^{10} bacterial cells per gram of soil[12]) may occupy less than 1% of the volume[142] of surface soil layers. The spatial distribution of bacterial cells is nonuniform[12] with soil bacteria exhibiting highly localized activity[13,143,144]. Similar patterns of spatial aggregation of bacterial cells have been observed in marine sediments[145] and were attributed to resource patchiness. Considering nutrient limitations in most soils[146], and the highly dynamic and restrictive aqueous diffusion pathways, suggest that the surfaces and volumes of most unsaturated soils are likely to be inhabited by small and resource-limited clusters of bacterial cells[12]. The highest bacterial cell densities have been associated with plant residues and particulate organic matter (POM) or adjacent to living plant roots of the rhizosphere[15]. At favorable locations, soil bacteria may attain cell densities similar to those found in biofilms[140,147] that form in water replete environments and host diverse consortia embedded in extracellular polymeric structures[147]. Evidence suggests, that a few dense bacterial colonies may contain a large proportion of soil bacterial biomass with the rest of the population distributed in numerous sparse settlements with only 10 to 100 cells each[12].

The harsh and heterogeneous soil environment promotes diverse bacterial ecology[146] as highlighted by the large number of biogeochemical processes[3] that occur within spatially distributed aqueous micro-habitats[141,143]. Certain processes and bacterial traits require close proximity among cells

(nanotube infrastructure[148], electron transport[149], gene transfer, cell-cell signaling[150]), whereas a host of general traits are likely to be constrained by resource availability and diffusion across distances that facilitate trophic interactions through the aqueous phase.

The physical distances that separate soil bacterial communities and the connectivity of the soil aqueous phase give rise to characteristic diffusion distances that are critical for the strength of trophic and other interactions (i.e., horizontal gene transfer). The consideration of diffusion distances ($L_D(\theta) = \sqrt{D_e(\theta)t}$) and diffusion times ($t_D(\theta) = L^2/D_e(\theta)$) for a given bacterial community separation length L provide quantitative measures that link soil bacterial micro-geography with specific ecological interaction potential. For example, in a fertile soil with high bacterial cell density of 10^{12} g^{-1}_{soil} and specific surface area of 100 m^2 g^{-1}_{soil} (loamy soil) we would expect up to 10^4 cells per mm^2 grain surface area. Considering the cells are distributed in colonies of 100 cells each, the average separation distance would be 100 µm. This physical distance may support trophic interactions among neighboring colonies under wet conditions where the characteristic diffusive length is of the order of 800 µm per day (with water content θ = 0.3, porosity θ_s = 0.5, bulk diffusivity D_0 = 10^{-10} m^2 s^{-1} and effective diffusivity $D_e(\theta) \approx 10^{-2}D_0$ using equation 8). In contrast, for drier soils or drier climate with lower cell density of 10^2 cells per mm^2 (10^8 g^{-1}_{soil} in sandy soils with 1 m^2 g^{-1}_{soil}), colonies would be separated by millimetric gaps for substrates to diffuse across with timescales of the order of years (with θ = 0.05, and effective diffusivity $D_e(\theta) \approx 10^{-4}D_0$).

The implications are that the distribution of POM in soil and the average aqueous-phase connectivity are key factors in ecological processes that involve exchanges across bacterial communities[1,141]. Additionally, certain soil biogeochemical fluxes become limited by the rates of gas transport, prominently oxygen. Such conditions may give rise to anoxic conditions in regions with high cell densities due to depletion of diffusion-limited oxygen[151]. The extent and number of such anoxic communities in a soil volume depends on the community sizes[151] and limitations to gas transport in relatively wet soil[152].

The lack of systematic data on soil bacterial distribution and limitations to direct observations at the microscale have motivated interest in advancing biophysical modeling for estimating community size and spatial distributions and their implications for diffusion mediated interactions in unsaturated soil. The specific objectives of the study were: (i) to quantify the spatial variation of soil bacterial cell density based on biome-specific soil carrying capacity and associated climatic water contents; (ii) to link bacterial community size distributions to average cell densities; and (iii) to provide direct observations of variations in bacterial density on soil surfaces under controlled experimental conditions.

We propose a soil bacterial interactions heuristic model (BIHM) that could link biome specific, vegetation and moisture dependent, bacterial cell densities with putative community sizes and spatial distributions under certain assumptions. We consider two main processes: the dependence of cell density on diffusive transport and soil bacterial carrying capacity[9], and spatial aggregation of bacterial cells due to growth[153] under constrained dispersal ranges[14] of unsaturated soils. The templates that governs soil bacterial communities and their interactions[132] are assumed to vary with the abundance of POM and climate-controlled hydration conditions that define the diffusive distances is soils across biomes (Fig. 4.1 a). Variations in macroscopic cell density are associated with microscopic bacterial community sizes; i.e., how many cells are clustered within 5 µm distance to their nearest neighbors depends directly on cell density (Fig. 4.1 b). A mechanistic and spatially-explicit individual-based model (SIM) was used to predict growth and dispersal of bacterial cells on two-dimensional hydrated soil surfaces that may result in spatially aggregated communities[9]. Numerical experiments were performed using cell densities comparable to estimates from microscale observations[12] and to global bacterial abundance data[8]. Simulations were used to obtain relations of bacterial community size distributions with water contents and cell densities for parametrization of the BIHM. In addition, we tested the feasibility for direct observation of bacterial spatial distribution on soil surfaces experimentally for varying nutrient and hydration conditions.

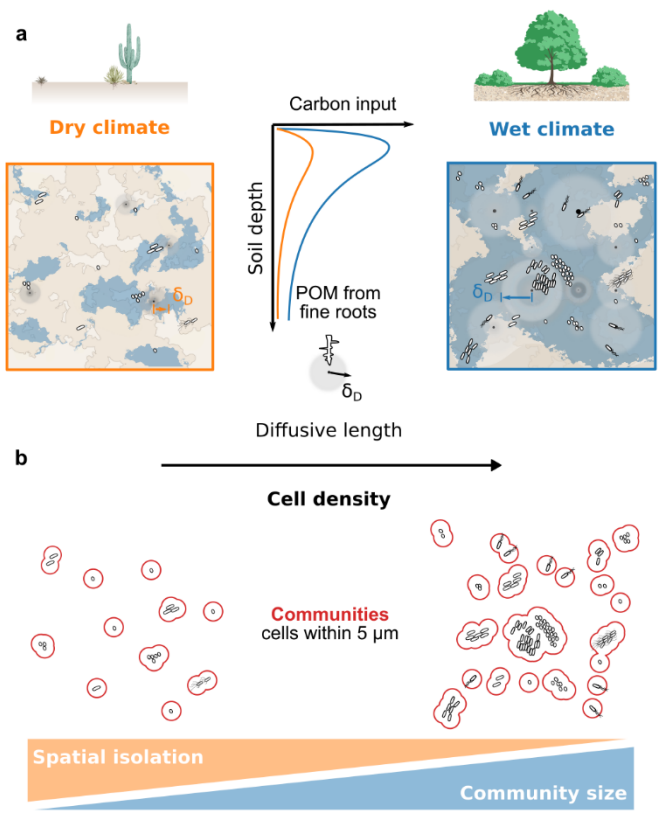


Fig. 4.1. Soil bacterial habitats and community sizes vary across biomes. a, Particulate organic matter (POM) derived from fine roots provides the carbon input to soil bacteria. Soils from dry environments with sparse distribution of POM host only few bacterial cells in fragmented aqueous habitats. Diffusive resource fluxes increase with soil aqueous phase connectivity and carbon input by vegetation towards wet environments. **b**, Localized bacterial growth and constrained motility cause strongly aggregated spatial distributions of bacterial communities at the microscale. Community sizes are expected to increase with decreased spatial isolation in soils with wet climate.

4.2 Results

4.2.1 Average cell density and community sizes linked to rainfall patterns and vegetation

Soil bacterial carrying capacity (potential cell density for given carbon input) decays with soil depth[139] following plant roots or soil POM distributions[8]. We assume that 35% of a biome's net primary productivity (NPP) is invested into new roots[154] and enters the soil via fine root fragments with yearly turnover[155]. Studies have shown that nearly 24% of this belowground NPP feeds soil bacterial biomass[42,43]. The average time between rainfall events in a given biome affects the distance across which bacterial cells could be supplied by nutrient diffusion considering transport limitations in unsaturated soils (e.g. the Millington-Quirk[152] effective diffusion model). Combining the information on total number of fine root fragments and diffusive distances, the number of cells maintained within diffusive spheres around discrete sources of POM (i.e., cell density, equation 12) was calculated for a range of hypothetical consecutive dry days (1-365 days). The results are depicted as a function of climatic water content for average soil and climatic conditions (Fig. 4.2 a). The model was evaluated for mean NPP and considers biomass to decay exponentially around every source of POM with a characteristic distance given by the diffusive length. Data on microbial biomass abundance[8] (topsoil, n = 429) was converted to bacterial cell density[9] and reported per bacteria accessible soil surface area (< 1% of soil surfaces[142]). Modeling results by the SIM support the increase in bacterial cell density with water contents and match the observed magnitudes (Fig. 4.2 a).

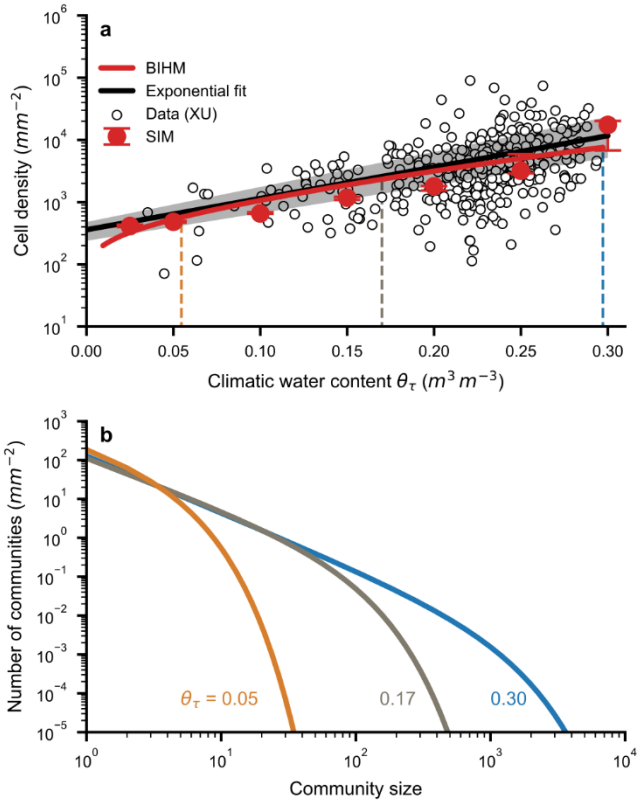


Fig. 4.2. Cell density and climatic water content shape soil bacterial communities. a, The number of cells per accessible soil surface area increases with climatic water content. Symbols indicate data (4) (XU) of soil microbial biomass carbon in the topsoil (< 100 mm) converted to estimates of bacterial cell density (black line - exponential fit, n = 429). The bacterial interaction heuristic model (BIHM, red line) considers soil properties (porosity, soil depth, specific surface area), climate (rainfall patterns, potential evapotranspiration, mean annual temperature) and vegetation (net primary productivity) to estimate average cell density that compares favorably with a spatially-explicit individual-based model (SIM, median ± IQR, n=9). **b**, Predicted soil bacterial community size distribution using the analytical model parametrized on results of the SIM. Soils with low climatic water contents contain few small communities. Towards wetter soils observations of communities with more than 100 cells become increasingly likely.

4.2.2 Community size distribution based on spatial clustering of bacterial cells

Simulation results by the mechanistic SIM that makes no assumptions regarding spatial cell distribution and preliminary observations in soil microcosms suggest that soil bacterial community sizes follow an exponentially truncated power law (Fig. 4.3). This distribution has been previously applied to describe animal group-sizes[156] and aggregation patterns of bacterial cells[153,157]. Soil bacterial community size distributions can be characterized by two parameters (equation 14); an exponent b and community cutoff size n_c that were estimated from simulation results and observations using maximum likelihood[158]. The parameters obtained from simulation results by the mechanistic SIM under a range of environmental conditions were used to link average soil bacterial cell density (a macroscopic quantity) with the distribution of bacterial community sizes (A4 Figure S1). The dependencies of n_c and b on cell density were used for general parametrization (equations 19 and 20, respectively) of the BIHM across biomes. Specifically, community size distributions were calculated for different climatic water contents and their corresponding bacterial cell densities (Fig. 4.2 b). Simulation results from the SIM were also compared with observations from our own microcosm experiments and with thin-section soil data from an independent study[12]. Overall, bacterial community size distributions were found to be highly skewed and display large variability for both empirical observations and simulation results (Fig 4.3 a and b). Results from our microcosm experiment with two nutrient conditions (sterilized tap water *W* and tryptic soy broth *TSB*) under two values of soil matric potential (-35 and -5 cm) show similar distributions across treatments with a slight increase in larger bacterial communities with addition of nutrients (Fig 4.3 a). Results from an independent study[12] (RAY), contained only few cells that resulted in a steep drop in the proportion of observed large bacterial communities (Fig 4.3 a). Simulation results from the SIM indicate strong variations in community sizes for a range of water contents and associated cell densities (Fig 4.3 b). Interestingly, the data points collapse by rescaling with the obtained parameters (b and n_c) suggesting that the proposed community size distribution (equation 14) describes both experimental data and SIM simulations (Fig 4.3 c and d). The black line represents the distribution using an average exponent (b = 1.65) and describes the central tendencies in community size data reasonably well (Fig. 4.3 c). A treatment from our experiment (-5 cm water, day 2; open blue circles) shows a deviation due to relatively low cell densities. Similarly, results from the independent study were not included here because the exponents could not be estimated reliably with the small number of communities analyzed (parameter estimation is often uncertain[156] for small n_c; A4 Figure S1 c). Nonetheless, the estimated parameters were considerably different from values expected from randomly distributed cells under varying cell densities (A4 Figure S1 a and c). Particularly, the larger n_c suggests larger community sizes at lower cell densities compared to randomly generated distributions.

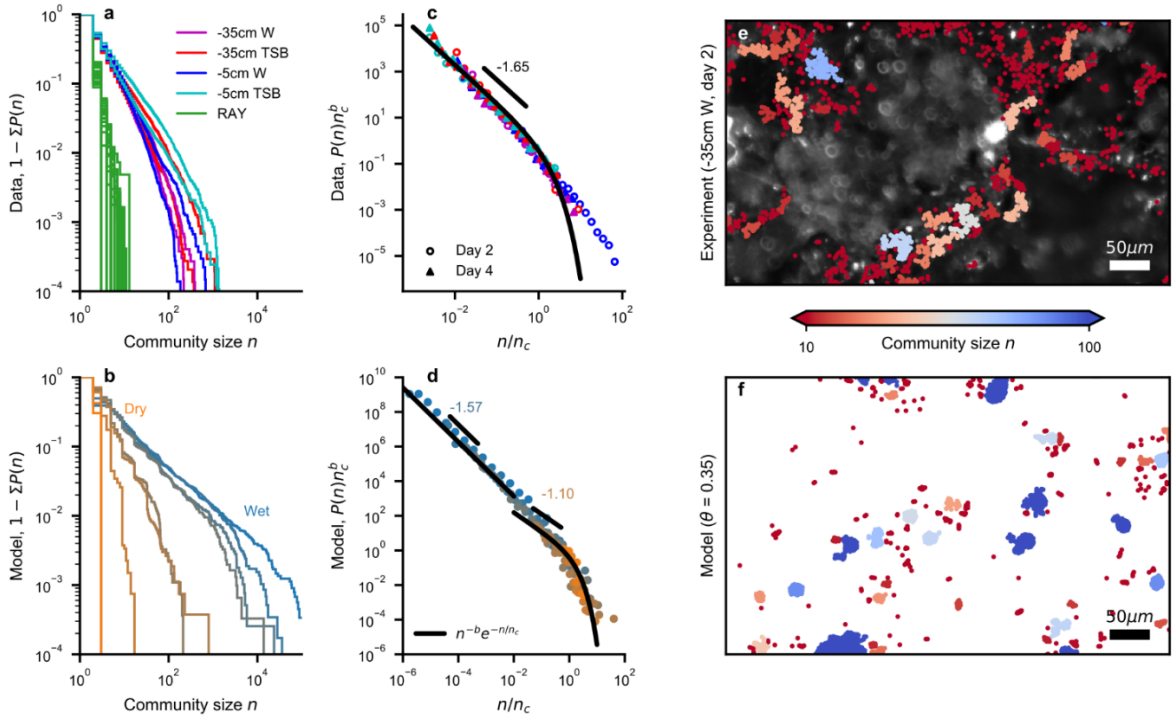


Fig. 4.3. Observed and modeled community size distribution. **a** and **b**, Complementary cumulative probability $P(n)$ of observing a community with *n* cells that are aggregated within 5 µm. **a**, Community size distributions from two days of microcosm measurements (purple, red, blue and cyan) under varying hydration (- 35 cm and -5 cm matric potential) and nutrient conditions (tap water W and tryptic soy broth TSB). Soil thin-section data from an independent study (7) (RAY) is shown in green. **b**, Distributions obtained for a range of water contents from a spatially-explicit individual-based model (SIM). **c** and **d**, $P(n)$ is rescaled using obtained parameters (exponent b and cutoff size n_c) and is shown with logarithmic bins for visual clarity. **c**, Microcosm data collapse after rescaling and are described using an average exponent ($b \approx 1.65$). Round symbols and triangles indicate measurements of day 2 and day 4, respectively. **d**, Distributions obtained from the SIM with different exponents above and below water content of 0.2 (b = 1.57 and 1.10 for wet and dry, respectively). **e**, Example of experimentally detected bacterial communities on soil surfaces. Only the regions in focus were analyzed (SYTO9 intensity in greyscale). Colors indicate community sizes (cell numbers, n). **f**, Spatial distribution of communities as obtained by the SIM.

Simulation results from the SIM indicate that the exponent b varies with soil water contents (Fig. 3 d), exhibiting lower values for water contents below ~0.2 (b = 1.10) compared with wetter conditions (b = 1.57). Small changes in community size distributions were observed in our microcosm experiments for different treatments, however, the largest community size increased consistently with cell density for both experiments and SIM simulations (A4 Figure S2 a). The proportion of cells associated with the largest bacterial community ($P_{largest}$) increased with increasing nutrients and hydration conditions (A4 Figure S2 b). Similarly, the proportion of isolated, single cells (P_{single}) decreased under wet conditions (A4 Figure S2 c). The unexpectedly high number of single cells observed in the wet treatment with no addition of nutrients could have been caused by the initial dilution of cells and increased dispersal opportunities in the soil (effectively reducing cell density and community detection). Examples of spatial cell distributions are shown for the microcosm experiment (Fig 4.3 e) and the SIM (Fig 4.3 f). In addition, images were taken at increased resolution to verify detection of cells in the microcosm experiment (A4 Figure S3). Overall, these findings suggest that it is feasible to

observe and quantify variations in soil bacterial community size distributions related to macroscopically-measured soil bacterial cell density. Increased soil bacterial cell density is linked to macroscopic quantities (e.g. due to higher water contents or increased carbon input) and was directly associated with changes in bacterial community size distributions. Soils receiving high precipitation with high NPP are expected to host a large proportion of biomass in only few communities thus affecting the nature of microbial interactions and framing the diversity picture[9].

4.2.3 Physical distances between bacterial communities limit trophic interactions

To quantify the conditions and strength of trophic interactions, we estimated average physical distances between bacterial communities attached to soil surfaces for different lower bounds on community sizes and climatic water contents (Fig. 4.4 a). Consistent with the assumption of spatially uniform POM distribution, the spatial distribution of bacterial communities in the bulk soil is also assumed uniform. This strong simplification facilitates the use of simple volume averaged macroscopic quantities such as effective nutrient diffusivities. Alternatively, biomass quantiles could be calculated as a function of distance to POM by setting appropriate bounds for integration of equation 11 to distinguish, for example, bulk soil communities from those inhabiting 'hot-spots'[15]. Here, we assumed bacterial communities are distributed uniformly on accessible soil surface area using macroscopic bacterial cell density values that vary with climatic water content (and associated NPP). The average distance between communities containing at least two cells was about 100 µm and did not vary much with soil wetness. However, the distance between larger communities (> 100 cells) increases rapidly as the soil becomes drier (on average) leading to an effective spatial isolation. To quantify the extent of temporal separation between such communities, we estimated the time required for a small molecule to diffuse across the average community separation distance (Fig. 4 b). This timescale increases from hours to months towards drier soils and could affect the distribution of shared resources and opportunities for cell-cell interactions. In particular, trophic interactions that rely on the exchange of diffusible compounds (e.g. a 'food chain', $A \rightarrow B \rightarrow C$) are expected to be suppressed in dry soils. This behavior was visible in results of the SIM where the amount of end product C strongly depended on the average water content (A4 Figure S4 a). Although enhanced trophic interactions and metabolite exchanges under wet conditions enabled higher bacterial richness, the Shannon index decreased towards higher water contents indicating reduced evenness with increased availability of resources[132] (A4 Figure S4 b). The rise in bacterial richness was comparable to the total number of communities and the number of multispecies communities (A4 Figure S4 c).

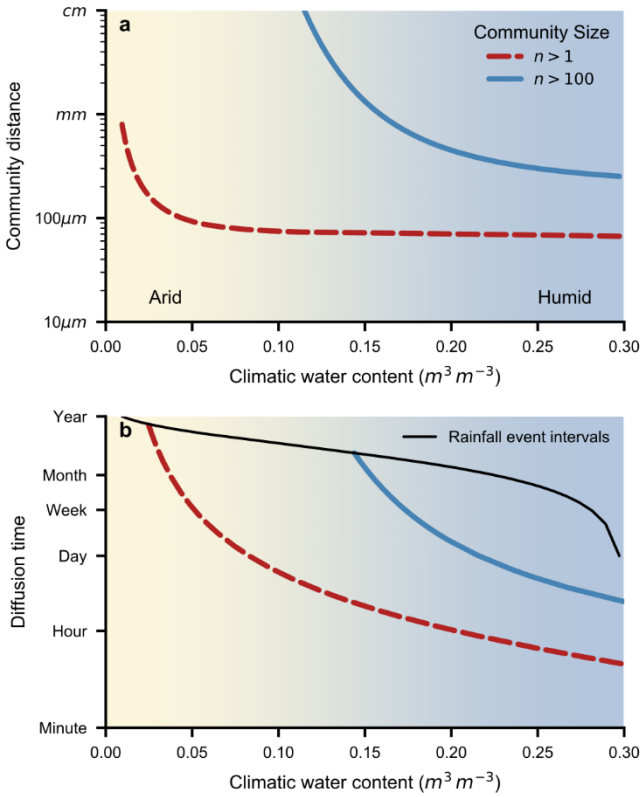


Fig. 4.4 Spatial and temporal scales of soil bacterial community separation. a, Modeled average distance between communities as a function of climatic water content for two lower bounds on community sizes n. The average distances between communities with two or more cells and between communities with more than 100 cells are shown in red and blue, respectively. **b**, Time needed for a small molecule to diffuse across the average distance between communities. The average time between soil wetting events is shown as an upper bound (black dashed line).

4.2.4 Variations in community sizes shape the proportion of anoxic bacterial communities across biomes

Simulation results by the SIM show how trophic interactions and bacterial diversity are strongly affected by hydration conditions[14] that also shape the distribution of biomass associated with different bacterial community sizes. We first demonstrate how community sizes vary across biomes before we apply the BIHM to quantify spontaneous occurrence of anoxic communities[141] to illustrate potential implications of community size variation on soil biogeochemical processes. We have used the BIHM to estimate the proportion of biomass associated with single cells and larger colonies (> 100 cells) for a range of climatic water contents with constant mean annual temperature (MAT) and constant mean NPP (Fig. 4.5 a). To illustrate how community size distribution varies with various environmental conditions, we evaluated the BIHM predictions for different temperatures (Fig. 4.5 b and c) and carbon inputs (Fig. 4.5 d and e). A decrease in MAT by 5 °C resulted in an increase in the proportion of large bacterial colonies at levels comparable to those for doubling NPP. This is consistent with enhanced soil carrying capacity[9] that enables large bacterial colonies in cold regions with high carbon inputs. A potentially important environmental factor was the proportion of dense bacterial colonies that could deplete oxygen in their core (i.e. where demand by respiration exceeds diffusive supply). For simplicity, we assume spherical bacterial colonies and estimate minimum colony radius needed to induce an anoxic core[151] assuming oxygen concentrations in the soil liquid phase reflect equilibrium with atmospheric levels (a conservative assumption for most soils). The BIHM could then estimate the

proportion of biomass associated with such anoxic cell clusters globally (Fig. 4.5 g) using spatially distributed soil properties[56,159], climate attributes[52,57] and vegetation carbon input[50] at 10 km resolution as previously described[9]. In addition, Copernicus global land cover data[160] (2019, dominant type) was used to compare the amount of biomass associated with anoxic communities for different biomes and land use classes (Fig. 4.5 f). Permafrost soils[161] were removed from the analysis because of limitations to diffusion that are currently not accounted for. The number of anoxic communities increased from 15 in bare soil to 5579 in closed forests, indicating communities with an anoxic core are expected to contain around 53000 cells on average. The predicted amount of bacterial biomass in anoxic communities was highest for closed forests followed by herbaceous wetlands and displayed high spatial variability for all classes. Reported values[162,163] of anaerobic cell counts were generally lower for a range of soils (around 10^6 ranging from 10^4 to 10^7 per gram of soil). One study also reported counts of anaerobic cells associated with plant residues in a rice paddy field[162] (between 10^8 and 10^{10} per gram of soil) that could mark an upper bound for very wet, organic matter rich soils.

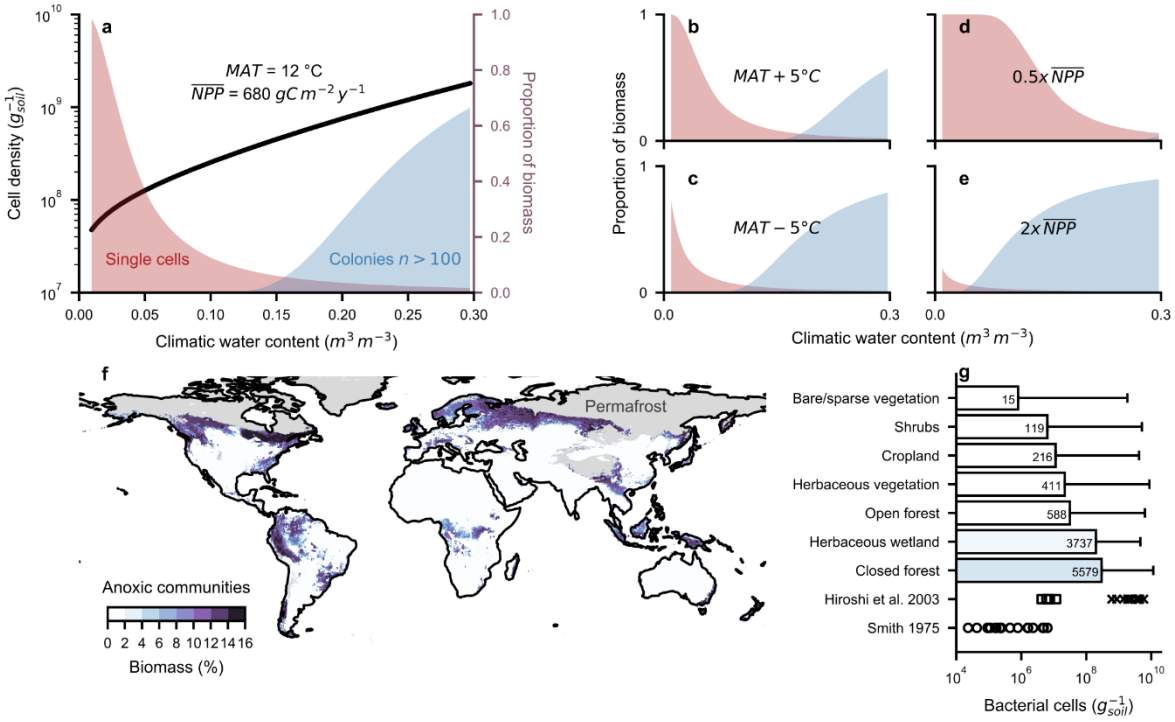


Fig. 4.5. Soil bacterial community size variations shape prevalence of anoxic communities. a, Modelled cell densities as a function of climatic water content for mean annual temperature (MAT) and mean annual net primary productivity (NPP). The second axis shows the proportion of bacterial biomass present as single cells and colonies with more than 100 cells in red and blue, respectively. **b**-**e**, Model sensitivity of biomass proportions. **b,** Increasing MAT by 5 °C and **c**, decreasing MAT by 5 °C. **d**, Half of mean annual NPP and **e**, twice the mean annual NPP. **e**, The predicted percentage of bacterial biomass associated with anoxic communities based on climatic, soil and single cell properties. Permafrost soils with mostly frozen water are indicated in grey and were excluded from further analysis. **g**, Area-weighted average number of bacterial cells predicted to be associated with anoxic communities for different land cover types. Whiskers indicate maximum values. Numbers on bars indicate the average number of anoxic communities. Data from two studies is shown for comparison. The number of anaerobic bacteria in various soils are reported by Smith (48) (n = 21). Hiroschi et al. (47) report cell densities of anaerobic bacteria for soil (square symbols, n = 5) and plant residues (crosses, n = 12) of a rice paddy field.

4.3 Discussion

Soil bacterial cell density varies across biomes governed by rainfall characteristics[7], vegetation derived carbon inputs[8] and temperature that define soil bacterial carrying capacity[9]. At the microscale, the connectivity of the heterogenous and dynamic aqueous phase often limits diffusive fluxes and cell dispersal ranges[14], thereby affecting bacterial abundance and traits[1,141]. Irrespective of the specific mechanisms, bacterial biomass is not uniformly distributed in soil[12,138]. The assumption of spatially uniform bacterial abundances at the sample scale often used for inferences of interactions and co-occurrence among taxa, requires careful consideration as it may bias[16] the picture of soil bacterial life under common soil conditions. Particularly in soils of drier climates where a few large communities may be supported by sparsely distributed POM, the fragmented aqueous phase limits potential for interactions (Fig. 4.4).

Evidence suggests that the size distribution of soil bacterial communities follow a power law[153] with a cutoff related to carrying capacity or overall cell density[153,156,157]. This representation is supported by simulation results from the SIM and direct observations in microcosm experiments (Fig. 4.3). The power law that characterizes community size distributions is compatible with a variety of processes ranging from collective motion[156,157] to growth models with preferential attachment (e.g. diffusion limited aggregation[164]) and is sensitive to spatial constraints[156,164]. Simulation results by the SIM show dependency of bacterial community size distributions on soil aqueous-phase connectivity. The dependency is manifested in predictions by the BIHM that is used to link average bacterial cell density to bacterial micro-geography on soil surfaces and enables generalization of soil bacterial community sizes across biomes (Fig. 4.5). Interactions with other soil microorganisms (viruses, protists, fungi) that may alter soil bacterial cell density are currently not considered in the BIHM. For example, competition with fungi for shared carbon sources could affect bacterial abundance and communities[5], particularly in ecosystems where fungi are increasingly prevalent[43] (e.g. forests).

Notwithstanding the many experimental and observational limitations, this study confirms that systematic measurement of spatial distributions of bacterial colonies at cell-level resolution are important for interpreting bacterially-mediated soil functions and are feasible on natural soil surfaces. The few observations at these small (but important) scales must be expanded to observe bacterial spatial configurations for a range of soil types and controlled conditions. As postulated in many studies and partially confirmed with our own observations, the sizes and spatial distribution of soil bacterial communities highlight the highly localized nature of interactions relevant to soil ecological functioning[1,139]. For example, interactions based on cell-cell contact[148–150] are likely limited by colony size[147] and restrict the extent of signaling[150] and metabolic activities[149] in bacterial communities. Most of the bulk soil is inhabited by small bacterial clusters[12] where the generally low metabolite exchanges

with neighboring communities and limited resources enhance overall diversity[132] by promoting co-existence of diverse bacterial species[9] in small soil volumes. The proportion of bacterial biomass associated with communities larger than one hundred cells increases disproportionally with increasing soil carrying capacity and more frequent rainfall events. Simulation results by the SIM predict enhancement of trophic interactions in wet soils with increased diffusive fluxes at the microscale (A4 Figure S4). Bacterial richness simultaneously increased with the total number of communities as a direct result of altered carbon fluxes that, in the long term, may shape the abundance of particular functional traits[141]. Interestingly, the proportion of small communities (< 100 cells) peaks at intermediate climatic water contents where soil bacterial diversity is highest[9]. The high number of small bacterial communities in many soils may also be reflected in sampled species abundance distributions and could be observed as increasing proportions of rare species in drier soils[17]. Even in marine sediments, we find that a large fraction of bacteria is attached to surfaces[140,165] with cell density 'hot-spots' that display distinct abundance distributions compared to those sampled from 'background' communities[145].

The implications are not only general for soil bacterial activity but also for specific physical and diffusive distances for activation of metabolic exchanges[3,48] and for the onset of anaerobic respiration[141,151]. Sufficiently large soil bacterial colonies may deplete oxygen in their cores (especially under restrictive gas diffusion in wet soils) and are frequent in soils with high carrying capacity such as the tropics and northern latitudes. These conditions delineate the expected extent of global terrestrial anoxic respiration that naturally emerges from these heuristic considerations (Fig. 4.5). Various refinements could be introduced to the simple calculations presented here including consideration of stoichiometric limitations (e.g. nitrogen) on carrying capacity and the use of distributed bacterial trait values (e.g. oxygen uptake rates) to estimate bacterially mediated processes. It is noteworthy that the bacterial cell and community densities presented here reflect average conditions in the soil body and require adaptations for certain extreme environments where carbon supply might differ and connectivity could be altered by extracellular polymeric substances, for example, in desert biocrusts[144]. Nonetheless, the large fluctuations in bacterial biomass across biomes[4] and associated metabolic capacities as indicated by the BIHM could partially explain the high uncertainty in soil greenhouse gas (GHG) emissions of different biomes[166], and affect the persistence of soil carbon[1,167]. A small increase in soil anoxic volume can greatly reduce carbon mineralization[167] and affect soil GHG emission at small scales[144]. The distribution of anoxic microsites is also related to land management[141] that might further affect soil carbon dynamics[1,13]. The heuristic model (BIHM) for linking bulk soil bacterial cell density with specific community size and spatial distributions could be extended to consider interactions of soil bacterial communities with other organisms relevant for soil

ecosystem functioning. For example, by considering bacterial communities as spatially distributed 'foraging grounds' for bacteriophages and soil fauna, respectively.

A unifying perspective is presented for linking key factors controlling soil bacterial community size distributions with spatial[148–150] and trophic[132,144,149] interactions. The spatial distribution of soil bacterial communities at the microscale are constrained by overall cell density and affected by aqueous phase connectivity and carbon flux[141]. The limited number of large communities suggests that considerable exchange of diffusible compounds is restricted to a few densely populated 'hot-spots' around sources of POM or in vicinity of plant roots. Such communities might even be large enough to spontaneously develop anoxic conditions that occur predominantly in resource-rich and frequently wet soils. In drier climate, ecosystem functioning can be largely attributed to isolated small settlements. Quantification of soil bacterial micro-geography provides tractable insight into the complexity[141] of bacterial habitats and provides spatial context to inferences of soil microbiome functioning[1,3,5].

4.4 Materials and Methods

4.4.1 Average cell density based on diffusion and distance to POM

The maximal number of cells maintained in a volume of soil ('soil carrying capacity') has been linked to carbon input by vegetation and cell specific maintenance rate that is modified by temperature[9]. It was estimated based on yearly averaged net primary productivity (NPP) that enters a section of the soil profile as new roots[154] (ξ = 0.35) and is available to soil bacteria[42,43] ($NPP_{b,z}$, with ϵ = 0.24) to a maximum soil depth d_{soil} = 1 m. The vertical distribution of carbon $f(z)$ is described using a log normal distribution with μ = 0.18 and σ = 1.00 as previously reported[9] (equation 1).

$$NPP_{b,z} = \xi\epsilon\frac{NPP}{d_{soil}}F_z = \xi\epsilon\frac{NPP}{d_{soil}}\int f(z)\,dz \quad (1)$$

Cell density at carrying capacity ρ_{CC} was estimated (equation 2) by assuming bacterial cells with mass M_c = 8.6x10^{-14} g C with maintenance rate m = 1.5 gC gC$_{cell}^{-1}$ y^{-1} and temperature sensitivity[44] f_T.

$$\rho_{CC}(z,T) = \frac{NPP_{b,z}}{f_T m M_c} \quad (2)$$

We assume that the main source of carbon for soil bacteria is POM derived from fine roots that have a turnover time of around one year[155]. The yearly total average volume of POM (V_{POM}) was estimated based on $NPP_{b,z}$ and the density of fine roots[155] ρ_{FR} = 0.5 g cm^{-3}. The yearly number of POM fragments N_{POM} was estimated based on a fine root diameter[155] d_{FR} = 0.5 mm. Assuming uniformly distributed root fragments we calculate an average distance to POM δ_{POM} (equations 3-5).

$$V_{POM} \cong \frac{NPP_{b,z}}{\rho_{FR}} \quad (3)$$

$$N_{POM} = \frac{V_{POM}}{d_{FR}^3} \quad (4)$$

$$\delta_{POM} = \left(\frac{V_{soil}}{N_{POM}}\right)^{\frac{1}{3}} \quad (5)$$

The (climatic) soil water content θ is defined as previously reported[9] (equation 6). The model assumes evaporation from a soil with saturated water content θ_s after drainage (at field capacity θ_{FC}) with a constant rate α (estimated by potential evapotranspiration PET) that is left for drying over some time t. A climatic average timescale τ over which the soil dries can be estimated as the number of consecutive dry days (based on precipitation[57] time series). The time between rainfall events during which the soil is wet was used to calculate the average number of wetting cycles per year N_{cyc} (equation 7).

$$\theta = \theta_{FC}e^{-\alpha t} \text{ with } \alpha = \frac{PET}{d_{soil}\,\theta_{FC}} \text{ and } \theta_{FC} \cong \frac{\theta_s}{2} \quad (6)$$

$$N_{cyc} = \frac{365}{\tau} \quad (7)$$

The total distance a small molecule could travel during a year when released from a point source of POM is related to soil effective diffusivity (Millington Quirk[152], equation 8). The area explored by a

particle with bulk diffusivity D_0 is obtained by integration over a drying cycle τ. The yearly diffusive distance is then obtained using the yearly number of wetting cycles N_{cyc} (equations 9-10).

$$D_e = D_0 \frac{\theta^{\frac{10}{3}}}{\theta_s^2} \quad (8)$$

$$A_{D\tau} = \int_0^{\tau} D_e(\theta, t)\, dt = 4\pi \frac{D_0 \theta_{FC}^{\frac{10}{3}}}{\theta_s^2} \frac{3}{10\alpha}\left(1 - e^{-\frac{10}{3}\alpha\tau}\right) \quad (9)$$

$$\delta_D = \sqrt{N_{cyc} A_{D\tau}} \quad (10)$$

The total number of cells that would be concentrated within the diffusive sphere around sources of POM was estimated considering N_{POM} and δ_D. The population of maintaining bacterial cells around a point source (POM) was assumed to decay radially with exponential rate δ_D^{-1}. Integration over radius r results in the expression for cell density ρ_c per number of POM sources N_{POM} and carrying capacity ρ_{CC} (equation 11 and 12).

$$\frac{\rho_c}{\rho_{CC}} = 4\pi N_{POM} \int_0^{\infty} e^{-\frac{r}{\delta_D}} r^2\, dr \quad (11)$$

$$\rho_c = \rho_{CC}\, 8\pi N_{POM} \delta_D^3 \quad (12)$$

4.4.2 Conversion of cell densities using soil particle surface area

Soil bacterial cell density is estimated using soil microbial biomass carbon[8] and was converted to bacterial cell density as previously described[9]. Bacterial cells are mostly attached to particle surfaces even in wet environments such as sediments[145,165]. In soil we expect most of the biomass to be attached to soil grain surfaces[140]. The specific soil-particle surface area SSA can be estimated using clay content f_{clay} and information on the dominant clay minerals. We consider proportions of Kaolinite, Illite and Smectite ($K, I, S,$) obtained from global maps[159] that dominate most of the soil clay fraction and are each associated with different surface areas; SA = 60, 200, 590 $m^2\ g^{-1}$, respectively. Only a fraction of the soil pore space is considered accessible to bacterial cells and we use 0.4% of the particle surface area[142] (η = 0.0038 ± 0.0005, n=6). The following equation 13 was used to estimate volumetric SSA (SSA_v) using soil bulk density[56] ρ_{soil} and includes a residual surface area of 1.1 $m^2\ g^{-1}$ of sand and silt particles:

$$SSA_v = \eta \rho_{soil}(f_{clay}(f_K SA_K + f_S SA_S + f_I SA_I + (1 - f_{dom})\overline{SA}) + 1.1) \text{ with } f_{dom} = f_K + f_S + f_I \quad (13)$$

4.4.3 Soil microcosm experiment

Natural soil (nutrient rich garden soil on ETH campus, 47° 22’ 43.8” N and 8° 32’ 53.6” E) was sampled between 5-10 cm depth in March 2018 and was subsequently sieved (< 2mm) after air drying for 3 h. The soil was incubated for four days at 28 °C on the porous surface model that allowed for controlled hydration and nutrient conditions. The experimental setup (porous surface model[168]) consisted of four ceramics with three holes (4 mm diameter and 3 mm depth) drilled in each that were filled with soil

(three replicate samples). Four treatments were applied by independently varying hydration conditions (-35 cm and -5 cm matric potential) and nutrient concentration (0% tryptic soy broth (TSB) and 100% TSB with autoclaved tap water). Soils were stained following the manufacturer's guidelines using SYTO9 to label DNA (Thermofisher Scientific; 3 µl were applied to each sample with a SYTO9 concentration of 10 µl incubated for 20 min). For image acquisition an epifluorescence microscope was used with a GFP filter cube (EVOS FL Auto, Life Technologies, Zug, Switzerland). From each ceramic, at least 9 images (3-4 for each hole, L ≈ 0.8 mm) were taken at 10x resolution (1 μm^2). Constant light settings were used throughout the experiment (light intensity = 10; exposure = 330 ms; gain was set to 0 dB). Images were taken to maximize the area in focus. Staining and imaging were done under suction (-50 cm matric potential) to remove excess water from the soil surface. Measurements were obtained after two and four days.

4.4.4 Image analysis for determination of cell locations

All images were analyzed with custom python scripts based on the SciPy[169] stack (including numpy, pandas and skimage[135]). Greyscale images were normalized to the range of pixel intensities (max-min). Images were denoised using (approximately) shift-invariant wavelet denoising ('cycle-spinning')[170] as implemented in the skimage function 'cycle_spin' with max_shifts = 9 and wavelet denoising[171] implemented in 'denoise_wavelet' using the Haar wavelet.

Images taken at 10x resolution where used to localize individual cells. First the area in focus was detected based on singular value decomposition[172] with a window size of 17 and retaining 7 most significant singular values. The resulting blur map was converted to a binary mask using cross-entropy thresholding ('threshold_li' in skimage[5]) that corresponds to the region in focus (effectively removing regions that contain no information). Holes and small objects were removed from the mask if they were smaller than 25 pixels. Positions of cells ('blobs') were detected using the Laplacian of Gaussian method as implemented in skimage[5]. A range of standard deviations was considered to detect local intensity peaks (σ = 0.4 to 7.8 in 40 steps). The smallest object could be represented by a standard deviation of 0.5 µm and the largest 10 µm. Coordinates of each cell were only used if they lie within the area in focus as determined by the blur mask. Total cell density was obtained by dividing the total number of cells by the area in focus. Cells were clustered as outlined below.

4.4.5 Spatially-explicit individual-based model of bacterial growth on soil particle surfaces

An individual-based model was used as previously implemented[9] with only minor modifications. Briefly, cells of multiple species (differing by kinetic parameters) grow by consuming from three carbon sources and move continuously (active swimming and passive shoving) on a heterogeneous, hydrated soil surface. Here, the number of species was reduced by coarsening the discretization of the cell physiological parameter space; resulting in an initial cell number and richness of 504 species.

Three carbon sources (A, B and C with associated yields Y = 0.25, 0.5 and 0.75) were considered to represent a trophic cascade ($A \rightarrow B \rightarrow C$). Difference in yields mimic species traits as observed for differing carbon use efficiencies depending on substrate uptake[173]. The carbon source with the lowest yield was provided initially and carbon mass was conserved by assigning the 'left over' carbon to the subordinate carbon type (e. g. $A \rightarrow B$ with efficiency $1 - Y_A$). A single carbon source was localized in the center of the domain (that represents a 1 by 1 mm soil surface) with constant concentration boundary condition (C_A = 25 g m^{-3}). The simulations were performed for a range of hydration conditions and a duration of 8 days at a 1 min time step. The total cell density, the coordinates of the cells were recorded and used for further analysis.

4.4.6 Clustering of proximal cells for estimation of community size distributions

Cells within a Euclidean distance of 5 µm were assigned to the same community. Agglomerative clustering (single linkage) was used where computationally feasible. For cell numbers exceeding 30'000, HDBSCAN was used instead (with similar parameters: min_cluster_size = 3, min_samples = 3, cluster_selection_epsilon = 5 µm, cluster_selection_method = 'eom'). The usage of HDBSCAN resulted in a worse detection of the smallest communities and was only necessary for cell locations obtained from the SIM (as cell numbers from experiments were lower). The spatial aggregation model (equation 13) was fitted to the distribution of community sizes using maximum likelihood to obtain estimates of b and n_c (as implemented in 'powerlaw'[158] with parameters discrete = True, discrete_approximation = 'xmax' and xmin = 1). Replicates were pooled to a single community size distribution to increase the counts of large communities (that are unlikely to be observed within small areas).

4.4.7 Spatial cell aggregation model – community size distribution

Spatial aggregation patterns of micro- and macro-organisms are often described by scaling relations[153,156,157]. An exponentially truncated power law is used to model community sizes assuming that individuals have the tendency to aggregate within a finite space[156]. The probability of having a group of n individuals is described as in equation 14 with exponent b and cutoff size n_c. A is a normalization constant given by constraint 15 with largest observed group size n_{max}.

$$P(n) = An^{-b}e^{-\frac{n}{n_c}} \quad (14)$$

$$1 = \sum_1^{n_{max}} P(n) \quad (15)$$

Similarly, the density fluctuations in growing bacterial colonies[157] and the distribution of bacterial community sizes on leaf surfaces[153] follow such cluster statistics where n_c is related to the total number of cells in the system[157]. The size of the largest community is therefore bounded and depends on the total number of individuals, for example, prescribed by carrying capacity or bulk cell density ρ_c. For infinite carrying capacity ($n_c \rightarrow \infty$) the relation only depends on b that is expected to change with spatial constraints[156]. In the case where b = 1 the distribution converges to the log-series[123]. In

soils we expect both parameters to be interdependent and vary with aqueous phase connectivity. The relation of b and n_c with water contents and carrying capacity are not known *a priori* and were determined using numerical simulations. In addition to the size distribution we can calculate the number of colonies $N(n)$ and cells $N_c(n)$ for every size class n using cell density ρ_C and equations (16-18).

$$p(n) = n^{1-b} e^{-\frac{n}{n_c}} \tag{16}$$

$$N(n) = \rho_C \frac{p(n)}{\sum p(n)} \tag{17}$$

$$N_c(n) = nN(n) \tag{18}$$

The dependency of n_c on cell density ρ_c was well described using equation 19 (with fitting parameters a, b) in agreement with empirical data (A4 Figure S1 a). The dependency of b on water contents in the SIM indicate a transition (A4 Figure S1 b). Parameters n_c and b are not independent (A4 Figure S1 c) and equation 20 was used for b (with fitting parameters α, β, γ).

$$n_c(\rho_c) = a\rho_c^b \tag{19}$$

$$b(\rho_c) = \begin{cases} 1, & b < 1 \\ \alpha \left(\frac{\rho_c}{n_c}\right)^{\beta} + \gamma, & b \geq 1 \end{cases} \tag{20}$$

Summary and Outlook

We have introduced a general model framework to link climate and soil type with the abundance distribution of soil bacteria and their diversity that compared favorably with observations and numerical simulations. Our model provides a mechanistic basis for observed biogeographical patterns of soil bacterial communities compatible with the explanatory power of soil pH, which, in the long term, results from a soil's climatic water balance. Soil bacterial diversity is highest in regions with intermediate climatic water contents where many isolated microhabitats are well supplied with carbon. Bacterial biomass increases with climatic water content across biomes, independent of bacterial diversity. Numerous isolated bacterial habitats shelter rare species from competition in relatively dry soils. Our model suggests that physiological differences among species can be equalized under the transport limiting conditions imposed by aqueous-phase connectivity and enable globally rare species to dominate arid regions at low absolute biomass. Non-linear shifts in bacterial species and biomass distributions are predicted to occur with changes in soil hydration and could have functional consequences for ecosystems that are sensitive to climate and land-use changes. The hydration centered model formulation allows to link the main components that control soil bacterial abundance and diversity to a few variables; offering a blueprint for adopting similar concepts to other organisms (e. g. earthworms; Appendix 5 *submitted*). Future work could extend beyond the monochromatic view of soil bacteria as individual agents and include a spectral perspective of spatial length scales that govern species across the tree of life. Nonetheless, our study provides a first step in the assignment of potential contributions to ecosystem functioning by spatially distributed soil bacterial communities with varying sizes and species compositions.

Short term dynamics corresponded well to long term climatic averages, yet it remains uncertain whether dynamic effects could be quantified and after what time convergence to climatic averages would occur. Incorporation of more refined models of soil hydration states (e.g. surface evaporation capacitor[174]) and partitioning of carbon inputs (e.g. primary productivity) would enable a dynamic model application that could be tested in the laboratory and at the field scale by measuring bacterial biomass and species abundance distribution; for example under changing irrigation patterns and distribution of POM.

Most importantly, the model implications for the spatial distribution of soil bacterial cells at small scales should be empirically verified. Genetic sequencing following image acquisition using modern microscopy techniques could be systematically applied to natural soils across environmental conditions to map the micro-geography of soil bacteria and gain direct access to observations of soil bacterial communities' spatial distributions at the sub-millimeter scale. Relating the evenness of the community size distribution to the evenness of the species abundance distribution provides insights

into the state of the soil microbiome. Beyond distances between colonies that shape interactions among bacteria, the size distribution of bacterial communities has implications for other organisms. For example, foraging nematodes or bacterial viruses should also be strongly affected by the spatial distribution of soil bacteria at small scales.

The general physical mechanism based on the distribution of resources in variably saturated environments (limiting growth and mobility) could be further applied to other micro- and macro-organisms. Extension of the model to habitats of soil fungi and other soil fauna would (only) require definition of characteristic length scales (e.g. pore sizes) and connectivity associated with physiological limitations to resource acquisition and movement that are specific to those organisms.

References

1. Lehmann, J. *et al.* Persistence of soil organic carbon caused by functional complexity. *Nature Geoscience* 1–6 (2020) doi:10.1038/s41561-020-0612-3.
2. Starke, R., Capek, P., Morais, D., Callister, S. J. & Jehmlich, N. The total microbiome functions in bacteria and fungi. *Journal of Proteomics* **213**, 103623 (2020).
3. Naylor, D. *et al.* Deconstructing the Soil Microbiome into Reduced-Complexity Functional Modules. *mBio* **11**, (2020).
4. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences* **115**, 6506–6511 (2018).
5. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).
6. Jousset, A. *et al.* Where less may be more: how the rare biosphere pulls ecosystems strings. *The ISME Journal* **11**, 853–862 (2017).
7. Serna-Chavez, H. M., Fierer, N. & van Bodegom, P. M. Global drivers and patterns of microbial abundance in soil: Global patterns of soil microbial biomass. *Global Ecology and Biogeography* **22**, 1162–1172 (2013).
8. Xu, X., Thornton, P. E. & Post, W. M. A global analysis of soil microbial biomass carbon, nitrogen and phosphorus in terrestrial ecosystems. *Global Ecology and Biogeography* **22**, 737–749 (2013).
9. Bickel, S. & Or, D. Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes. *Nat Commun* **11**, 1–9 (2020).
10. Or, D., Smets, B. F., Wraith, J. M., Dechesne, A. & Friedman, S. P. Physical constraints affecting bacterial habitats and activity in unsaturated porous media – a review. *Advances in Water Resources* **30**, 1505–1527 (2007).
11. Manzoni, S. & Katul, G. Invariant soil water potential at zero microbial respiration explained by hydrological discontinuity in dry soils. *Geophysical Research Letters* **41**, 7151–7158 (2014).
12. Raynaud, X. & Nunan, N. Spatial Ecology of Bacteria at the Microscale in Soil. *PLoS ONE* **9**, e87217 (2014).
13. Falconer, R. E. *et al.* Microscale heterogeneity explains experimental variability and non-linearity in soil organic matter mineralisation. *PloS one* **10**, e0123774 (2015).
14. Tecon, R. & Or, D. Biophysical processes supporting the diversity of microbial life in soil. *FEMS Microbiol Rev* **41**, 599–623 (2017).
15. Kuzyakov, Y. & Razavi, B. S. Rhizosphere size and shape: Temporal dynamics and spatial stationarity. *Soil Biology and Biochemistry* **135**, 343–360 (2019).
16. Armitage, D. W. & Jones, S. E. How sample heterogeneity can obscure the signal of microbial interactions. *The ISME Journal* **13**, 2639–2646 (2019).
17. Bickel, S., Chen, X., Papritz, A. & Or, D. A hierarchy of environmental covariates control the global biogeography of soil bacterial richness. *Sci Rep* **9**, 1–10 (2019).
18. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
19. Louca, S., Mazel, F., Doebeli, M. & Parfrey, L. W. A census-based estimate of Earth's bacterial and archaeal diversity. *PLOS Biology* **17**, e3000106 (2019).
20. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).
21. Zhou, J. *et al.* Spatial and Resource Factors Influencing High Microbial Diversity in Soil. *Appl. Environ. Microbiol.* **68**, 326–334 (2002).
22. Franklin, R. B. & Mills, A. L. Importance of spatially structured environmental heterogeneity in controlling microbial community composition at small spatial scales in an agricultural field. *Soil Biology and Biochemistry* **41**, 1833–1840 (2009).
23. Wang, G. & Or, D. Hydration dynamics promote bacterial coexistence on rough surfaces. *The ISME journal* **7**, 395–404 (2013).

24. Bach, E. M., Williams, R. J., Hargreaves, S. K., Yang, F. & Hofmockel, K. S. Greatest soil microbial diversity found in micro-habitats. *Soil Biology and Biochemistry* **118**, 217–226 (2018).
25. Vos, M., Wolf, A. B., Jennings, S. J. & Kowalchuk, G. A. Micro-scale determinants of bacterial diversity in soil. *FEMS Microbiology Reviews* **37**, 936–954 (2013).
26. Bell, T., Newman, J. A., Silverman, B. W., Turner, S. L. & Lilley, A. K. The contribution of species richness and composition to bacterial services. *Nature* **436**, 1157–1160 (2005).
27. Lennon, J. T. & Jones, S. E. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology* **9**, 119–130 (2011).
28. Zhou, J. *et al.* Temperature mediates continental-scale diversity of microbes in forest soils. *Nature Communications* **7**, 12083 (2016).
29. Kaiser, K. *et al.* Driving forces of soil bacterial community structure, diversity, and function in temperate grasslands and forests. *Scientific Reports* **6**, (2016).
30. Maestre, F. T. *et al.* Increasing aridity reduces soil microbial diversity and abundance in global drylands. *PNAS* **112**, 15684–15689 (2015).
31. Siciliano, S. D. *et al.* Soil fertility is associated with fungal and bacterial richness, whereas pH is associated with community composition in polar soil microbial communities. *Soil Biology and Biochemistry* **78**, 10–20 (2014).
32. George, P. B. L. *et al.* Divergent national-scale trends of microbial and animal biodiversity revealed across diverse temperate soil ecosystems. *Nature Communications* **10**, 1107 (2019).
33. Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *PNAS* **103**, 626–631 (2006).
34. Blaser, M. J. *et al.* Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. *mBio* **7**, (2016).
35. Yu, X., Polz, M. F. & Alm, E. J. Interactions in self-assembled microbial communities saturate with diversity. *The ISME Journal* 1 (2019) doi:10.1038/s41396-019-0356-5.
36. Elsas, J. D. van *et al.* Microbial diversity determines the invasion of soil by a bacterial pathogen. *PNAS* **109**, 1159–1164 (2012).
37. Lee, K. C. *et al.* Stochastic and Deterministic Effects of a Moisture Gradient on Soil Microbial Communities in the McMurdo Dry Valleys of Antarctica. *Front. Microbiol.* **9**, 2619 (2018).
38. Wang, G. & Or, D. A Hydration-Based Biophysical Index for the Onset of Soil Microbial Coexistence. *Scientific Reports* **2**, (2012).
39. Kim, M. & Or, D. Individual-Based Model of Microbial Life on Hydrated Rough Soil Surfaces. *PLOS ONE* **11**, e0147394 (2016).
40. Wang, G. & Or, D. Aqueous films limit bacterial cell motility and colony expansion on partially saturated rough surfaces: Aqueous films limit bacterial motion. *Environmental Microbiology* **12**, 1363–1373 (2010).
41. Tecon, R., Ebrahimi, A., Kleyer, H., Levi, S. E. & Or, D. Cell-to-cell bacterial interactions promoted by drier conditions on soil surfaces. *PNAS* **115**, 9791–9796 (2018).
42. Fatichi, S., Manzoni, S., Or, D. & Paschalis, A. A mechanistic model of microbially mediated soil biogeochemical processes - a reality check. *Global Biogeochemical Cycles* (2019) doi:10.1029/2018GB006077.
43. Fierer, N., Strickland, M. S., Liptzin, D., Bradford, M. A. & Cleveland, C. C. Global patterns in belowground communities. *Ecology Letters* **12**, 1238–1249 (2009).
44. Schoolfield, R. M., Sharpe, P. J. H. & Magnuson, C. E. Non-linear regression of biological temperature-dependent rate models based on absolute reaction-rate theory. *Journal of Theoretical Biology* **88**, 719–731 (1981).
45. Neilson, J. W. *et al.* Significant Impacts of Increasing Aridity on the Arid Soil Microbiome. *mSystems* **2**, (2017).
46. Banerjee, S. *et al.* Legacy effects of soil moisture on microbial community structure and N2O emissions. *Soil Biology and Biochemistry* **95**, 40–50 (2016).

47. Nunan, N., Leloup, J., Ruamps, L. S., Pouteau, V. & Chenu, C. Effects of habitat constraints on soil microbial community function. *Scientific Reports* **7**, (2017).
48. Zelezniak, A. *et al.* Metabolic dependencies drive species co-occurrence in diverse microbial communities. *PNAS* **112**, 6449–6454 (2015).
49. Slessarev, E. W. *et al.* Water balance creates a threshold in soil pH at the global scale. *Nature* **540**, 567–569 (2016).
50. Zhao, M., Heinsch, F. A., Nemani, R. R. & Running, S. W. Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote Sensing of Environment* **95**, 164–176 (2005).
51. Lieth, H. Modeling the Primary Productivity of the World. in *Primary Productivity of the Biosphere* (eds. Lieth, H. & Whittaker, R. H.) 237–263 (Springer Berlin Heidelberg, 1975).
52. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas: new climate surfaces for global land areas. *International Journal of Climatology* **37**, 4302–4315 (2017).
53. Waring, B. G., Averill, C. & Hawkes, C. V. Differences in fungal and bacterial physiology alter soil carbon and nitrogen cycling: insights from meta-analysis and theoretical models. *Ecology Letters* **16**, 887–894 (2013).
54. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences* **95**, 6578–6583 (1998).
55. Tóth, B. *et al.* New generation of hydraulic pedotransfer functions for Europe: New hydraulic pedotransfer functions for Europe. *European Journal of Soil Science* **66**, 226–238 (2015).
56. Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. *PloS one* **12**, e0169748 (2017).
57. Beck, H. E. *et al.* MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment. *Bull. Amer. Meteor. Soc.* **100**, 473–500 (2019).
58. Jensen, M. E. & Haise, H. R. Estimating Evapotranspiration from Solar Radiation. *Proceedings of the American Society of Civil Engineers, Journal of the Irrigation and Drainage Division* **89**, 15–41 (1963).
59. Stauffer, D. Scaling theory of percolation clusters. *Physics Reports* **54**, 1–74 (1979).
60. Scher, H. & Zallen, R. Critical Density in Percolation Processes. *The Journal of Chemical Physics* **53**, 3759–3761 (1970).
61. Shirazi, M. A. & Boersma, L. A Unifying Quantitative Analysis of Soil Texture1. *Soil Science Society of America Journal* **48**, 142–147 (1984).
62. Hill, M. O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **54**, 427–432 (1973).
63. Keller, E. & Segel, L. Model for chemotaxis. *Journal of Theoretical Biology* **30**, 225–234 (1971).
64. Barberán, A. *et al.* Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecology Letters* **17**, 794–802 (2014).
65. Meyer, K. M. *et al.* Why do microbes exhibit weak biogeographic patterns? *The ISME Journal* **12**, 1404–1413 (2018).
66. O'Brien, S. L. *et al.* Spatial scale drives patterns in soil bacterial diversity. *Environ Microbiol* **18**, 2039–2051 (2016).
67. Hendershot, J. N., Read, Q. D., Henning, J. A., Sanders, N. J. & Classen, A. T. Consistently inconsistent drivers of microbial diversity and abundance at macroecological scales. *Ecology* **98**, 1757–1763 (2017).
68. Griffiths, R. I. *et al.* The bacterial biogeography of British soils: Mapping soil bacteria. *Environmental Microbiology* **13**, 1642–1654 (2011).
69. Ramirez, K. S. *et al.* Detecting macroecological patterns in bacterial communities across independent studies of global soils. *Nature Microbiology* **3**, 189–196 (2018).

70. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Appl Environ Microbiol* **75**, 5111–5120 (2009).
71. Delgado-Baquerizo, M. & Eldridge, D. J. Cross-Biome Drivers of Soil Bacterial Alpha Diversity on a Worldwide Scale. *Ecosystems* (2019) doi:10.1007/s10021-018-0333-2.
72. Averill, C., Waring, B. G. & Hawkes, C. V. Historical precipitation predictably alters the shape and magnitude of microbial functional response to soil moisture. *Global Change Biology* **22**, 1957–1964 (2016).
73. Tecon, R. & Or, D. Bacterial flagellar motility on hydrated rough surfaces controlled by aqueous film thickness and connectedness. *Scientific Reports* **6**, 19409 (2016).
74. Olson, D. M. *et al.* Terrestrial Ecoregions of the World: A New Map of Life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience* **51**, 933–938 (2001).
75. Šťovíček, A., Kim, M., Or, D. & Gillor, O. Microbial community response to hydration-desiccation cycles in desert soil. *Scientific Reports* **7**, 45735 (2017).
76. Rath, K. M., Fierer, N., Murphy, D. V. & Rousk, J. Linking bacterial community composition to soil salinity along environmental gradients. *The ISME Journal* **13**, 836 (2019).
77. Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M. & West, G. B. Toward a Metabolic Theory of Ecology. *Ecology* **85**, 1771–1789 (2004).
78. Power, J. F. *et al.* Microbial biogeography of 925 geothermal springs in New Zealand. *Nature Communications* **9**, (2018).
79. Bühlmann, P., Peters, J. & Ernest, J. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics* **42**, 2526–2556 (2014).
80. Karimi, B. *et al.* Biogeography of soil bacteria and archaea across France. *Science Advances* **4**, eaat1808 (2018).
81. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods* **15**, 796 (2018).
82. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res* **39**, D28–D31 (2011).
83. Knight, R. *et al.* Best practices for analysing microbiomes. *Nature Reviews Microbiology* 1 (2018) doi:10.1038/s41579-018-0029-9.
84. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).
85. Walters, W. *et al.* Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems* **1**, (2016).
86. Apprill, A., McNally, S., Parsons, R. & Weber, L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology* **75**, 129–137 (2015).
87. Parada, A. E., Needham, D. M. & Fuhrman, J. A. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology* **18**, 1403–1414 (2016).
88. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
89. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
90. Bokulich, N. A. *et al.* Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* **10**, 57–59 (2013).
91. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, (2017).
92. Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* **12**, 118–123 (2010).

93. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581–583 (2016).
94. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11**, 2639–2643 (2017).
95. DeSantis, T. Z. *et al.* Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology* **72**, 5069–5072 (2006).
96. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* **87**, 4576 (1990).
97. Beck, H. E. *et al.* MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences Discussions* 1–38 (2016) doi:10.5194/hess-2016-236.
98. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967).
99. Wood, S. N. *Generalized Additive Models : An Introduction with R, Second Edition*. (Chapman and Hall/CRC, 2017). doi:10.1201/9781315370279.
100. Pearl, J. *Causality: Models, Reasoning and Inference*. (Cambridge University Press, 2009).
101. Dezeure, R., Bühlmann, P., Meier, L. & Meinshausen, N. High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi. *Statist. Sci.* **30**, 533–558 (2015).
102. Mandozzi, J. & Bühlmann, P. Hierarchical Testing in the High-Dimensional Setting With Correlated Variables. *Journal of the American Statistical Association* **111**, 331–343 (2016).
103. Meinshausen, N. & Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417–473 (2010).
104. Meinshausen, N., Meier, L. & Bühlmann, P. p-Values for High-Dimensional Regression. *Journal of the American Statistical Association* **104**, 1671–1681 (2009).
105. Wasserman, L. & Roeder, K. High-dimensional variable selection. *Ann. Statist.* **37**, 2178–2201 (2009).
106. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
107. Nemergut, D. R. *et al.* Patterns and Processes of Microbial Community Assembly. *Microbiology and Molecular Biology Reviews* **77**, 342–356 (2013).
108. Jiao, S., Chen, W. & Wei, G. Biogeography and ecological diversity patterns of rare and abundant bacteria in oil-contaminated soils. *Molecular Ecology* **26**, 5305–5317 (2017).
109. Rivett, D. W. & Bell, T. Abundance determines the functional role of bacterial phylotypes in complex communities. *Nat Microbiol* **3**, 767–772 (2018).
110. Saleem, M., Hu, J. & Jousset, A. More Than the Sum of Its Parts: Microbiome Biodiversity as a Driver of Plant Growth and Soil Health. *Annu. Rev. Ecol. Evol. Syst.* **50**, 145–168 (2019).
111. Wagg, C., Schlaeppi, K., Banerjee, S., Kuramae, E. E. & Heijden, M. G. A. van der. Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. *Nat Commun* **10**, 1–10 (2019).
112. Delgado-Baquerizo, M. *et al.* Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nat Commun* **7**, 10541 (2016).
113. Delgado-Baquerizo, M. *et al.* Multiple elements of soil biodiversity drive ecosystem functions across biomes. *Nat Ecol Evol* **4**, 210–220 (2020).
114. Aanderud, Z. T., Jones, S. E., Fierer, N. & Lennon, J. T. Resuscitation of the rare biosphere contributes to pulses of ecosystem activity. *Front Microbiol* **6**, (2015).
115. Allison, S. D. & Martiny, J. B. H. Resistance, resilience, and redundancy in microbial communities. *PNAS* **105**, 11512–11519 (2008).
116. Li, P. *et al.* Distinct Successions of Common and Rare Bacteria in Soil Under Humic Acid Amendment – A Microcosm Study. *Front. Microbiol.* **10**, (2019).
117. Puentes-Téllez, P. E. & Salles, J. F. Dynamics of Abundant and Rare Bacteria During Degradation of Lignocellulose from Sugarcane Biomass. *Microb Ecol* **79**, 312–325 (2020).

118. Nemergut, D. R. *et al.* Global patterns in the biogeography of bacterial taxa. *Environmental Microbiology* **13**, 135–144 (2011).
119. Clarke, R. T. & Murphy, J. F. Effects of locally rare taxa on the precision and sensitivity of RIVPACS bioassessment of freshwaters. *Freshwater Biology* **51**, 1924–1940 (2006).
120. Escalas, A. *et al.* Microbial functional diversity: From concepts to applications. *Ecology and Evolution* **9**, 12000–12016 (2019).
121. Kurm, V., Putten, W. H. van der, Boer, W. de, Naus-Wiezer, S. & Hol, W. H. G. Low abundant soil bacteria can be metabolically versatile and fast growing. *Ecology* **98**, 555–564 (2017).
122. Kurm, V. *et al.* Competition and predation as possible causes of bacterial rarity. *Environ Microbiol* **21**, 1356–1368 (2019).
123. Pueyo, S., He, F. & Zillio, T. The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecol Lett* **10**, 1017–1028 (2007).
124. Dee, L. E. *et al.* When Do Ecosystem Services Depend on Rare Species? *Trends in Ecology & Evolution* **0**, (2019).
125. Li, C. H. & Lee, C. K. Minimum cross entropy thresholding. *Pattern Recognition* **26**, 617–625 (1993).
126. Kurm, V., Putten, W. H. van der & Hol, W. H. G. Cultivation-success of rare soil bacteria is not influenced by incubation time and growth medium. *PLOS ONE* **14**, e0210073 (2019).
127. Fisher, C. K. & Mehta, P. The transition between the niche and neutral regimes in ecology. *Proceedings of the National Academy of Sciences* **111**, 13111–13116 (2014).
128. García, F. C., Bestion, E., Warfield, R. & Yvon-Durocher, G. Changes in temperature alter the relationship between biodiversity and ecosystem functioning. *PNAS* **115**, 10989–10994 (2018).
129. Treves, D. S., Xia, B., Zhou, J. & Tiedje, J. M. A Two-Species Test of the Hypothesis That Spatial Isolation Influences Microbial Diversity in Soil. *Microb Ecol* **45**, 20–28 (2003).
130. Campbell, B. J., Yu, L., Heidelberg, J. F. & Kirchman, D. L. Activity of abundant and rare bacteria in a coastal ocean. *Proceedings of the National Academy of Sciences* **108**, 12776–12781 (2011).
131. Chase, A. B. *et al.* Maintenance of Sympatric and Allopatric Populations in Free-Living Terrestrial Bacteria. *mBio* **10**, (2019).
132. Ratzke, C., Barrere, J. & Gore, J. Strength of species interactions determines biodiversity and stability in microbial communities. *Nat Ecol Evol* 1–8 (2020) doi:10.1038/s41559-020-1099-4.
133. Doud, D. F. R. *et al.* Function-driven single-cell genomics uncovers cellulose-degrading bacteria from the rare biosphere. *ISME J* **14**, 659–675 (2020).
134. Price, P. B. & Sowers, T. Temperature dependence of metabolic rates for microbial growth, maintenance, and survival. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4631–4636 (2004).
135. Walt, S. van der *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
136. Hermsen, R., Okano, H., You, C., Werner, N. & Hwa, T. A growth-rate composition formula for the growth of E.coli on co-utilized carbon substrates. *Mol. Syst. Biol.* **11**, 801 (2015).
137. Levin, S. A. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology* **73**, 1943–1967 (1992).
138. Dechesne, A. *et al.* A novel method for characterizing the microscale 3D spatial distribution of bacteria in soil. *Soil Biology and Biochemistry* **35**, 1537–1546 (2003).
139. *The spatial distribution of microbes in the environment*. (Springer, 2007).
140. Flemming, H.-C. & Wuertz, S. Bacteria and archaea on Earth and their abundance in biofilms. *Nature Reviews Microbiology* **17**, 247–260 (2019).
141. Neal, A. L. *et al.* Soil as an extended composite phenotype of the microbial metagenome. *Scientific Reports* **10**, 10649 (2020).
142. Hassink, J. Effects of soil texture and structure on carbon and nitrogen mineralization in grassland soils. *Biology and Fertility of Soils* **14**, 126–134 (1992).

143. Ruiz, S. *et al.* Image-based quantification of soil microbial dead zones induced by nitrogen fertilization. *Science of The Total Environment* 138197 (2020) doi:10.1016/j.scitotenv.2020.138197.
144. Kim, M. & Or, D. Microscale pH variations during drying of soils and desert biocrusts affect HONO and NH 3 emissions. *Nat Commun* **10**, 1–12 (2019).
145. Dann, L. M., Paterson, J. S., Newton, K., Oliver, R. & Mitchell, J. G. Distributions of Virus-Like Particles and Prokaryotes within Microenvironments. *PLOS ONE* **11**, e0146984 (2016).
146. van Elsas, J. D. *Modern Soil Microbiology*. (Taylor & Francis, 1997).
147. Paula, A. J., Hwang, G. & Koo, H. Dynamics of bacterial population growth in biofilms resemble spatial and structural aspects of urbanization. *Nature Communications* **11**, 1354 (2020).
148. Mamou, G., Malli Mohan, G. B., Rouvinski, A., Rosenberg, A. & Ben-Yehuda, S. Early Developmental Program Shapes Colony Morphology in Bacteria. *Cell Reports* **14**, 1850–1857 (2016).
149. McGlynn, S. E., Chadwick, G. L., Kempes, C. P. & Orphan, V. J. Single cell activity reveals direct electron transfer in methanotrophic consortia. *Nature* **526**, 531–535 (2015).
150. Larkin, J. W. *et al.* Signal Percolation within a Bacterial Community. *Cell Syst* **7**, 137-145.e3 (2018).
151. Wessel, A. K. *et al.* Oxygen Limitation within a Bacterial Aggregate. *mBio* **5**, (2014).
152. Millington, R. J. & Quirk, J. P. Permeability of porous solids. *Trans. Faraday Soc.* **57**, 1200–1207 (1961).
153. Monier, J.-M. & Lindow, S. E. Frequency, Size, and Localization of Bacterial Aggregates on Bean Leaf Surfaces. *Appl Environ Microbiol* **70**, 346–355 (2004).
154. Chapin, F. S., Matson, P. A. & Vitousek, P. M. Plant Carbon Budgets. in *Principles of Terrestrial Ecosystem Ecology* (eds. Chapin, F. S., Matson, P. A. & Vitousek, P. M.) 157–181 (Springer, 2011). doi:10.1007/978-1-4419-9504-9_6.
155. Sierra Cornejo, N., Hertel, D., Becker, J. N., Hemp, A. & Leuschner, C. Biomass, Morphology, and Dynamics of the Fine Root System Across a 3,000-M Elevation Gradient on Mt. Kilimanjaro. *Front. Plant Sci.* **11**, (2020).
156. Bonabeau, E., Dagorn, L. & Fréon, P. Scaling in animal group-size distributions. *PNAS* **96**, 4472–4477 (1999).
157. Zhang, H. P., Be'er, A., Florin, E.-L. & Swinney, H. L. Collective motion and density fluctuations in bacterial colonies. *PNAS* **107**, 13626–13630 (2010).
158. Alstott, J., Bullmore, E. & Plenz, D. powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions. *PLOS ONE* **9**, e85777 (2014).
159. Ito, A. & Wagai, R. Global distribution of clay-size minerals on land surface for biogeochemical and climatological studies. *Scientific Data* **4**, 170103 (2017).
160. Marcel Buchhorn *et al. Copernicus Global Land Service: Land Cover 100m: version 3 Globe 2015-2019: Product User Manual*. https://zenodo.org/record/3938963 (2020) doi:10.5281/zenodo.3938963.
161. Gruber, S. Derivation and analysis of a high-resolution estimate of global permafrost zonation. *The Cryosphere* **6**, 221–233 (2012).
162. Hiroshi, A., Izawa, T., Ueki, K. & Ueki, A. Phylogeny of numerically abundant culturable anaerobic bacteria associated with degradation of rice plant residue in Japanese paddy field soil. *FEMS Microbiol Ecol* **43**, 149–161 (2003).
163. Smith, L. Ds. Common Mesophilic Anaerobes, Including Clostridium botulinum and Clostridium tetani, in 21 Soil Specimens. *Appl Microbiol* **29**, 590–594 (1975).
164. Nicolás-Carlock, J. R. & Carrillo-Estrada, J. L. A universal dimensionality function for the fractal dimensions of Laplacian growth. *Scientific Reports* **9**, 1120 (2019).

165. Yamamoto, N. & Lopez, G. Bacterial abundance in relation to surface area and organic content of marine sediments. *Journal of Experimental Marine Biology and Ecology* **90**, 209–220 (1985).
166. Oertel, C., Matschullat, J., Zurba, K., Zimmermann, F. & Erasmi, S. Greenhouse gas emissions from soils—A review. *Geochemistry* **76**, 327–352 (2016).
167. Keiluweit, M., Wanzek, T., Kleber, M., Nico, P. & Fendorf, S. Anaerobic microsites have an unaccounted role in soil carbon stabilization. *Nature Communications* **8**, 1771 (2017).
168. Dechesne, A., Or, D., Gulez, G. & Smets, B. F. The Porous Surface Model, a Novel Experimental System for Online Quantitative Observation of Microbial Processes under Unsaturated Conditions. *Applied and Environmental Microbiology* **74**, 5195–5200 (2008).
169. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).
170. Coifman, R. R. & Donoho, D. L. Translation-Invariant De-Noising. in *Wavelets and Statistics* (eds. Antoniadis, A. & Oppenheim, G.) 125–150 (Springer, 1995). doi:10.1007/978-1-4612-2544-7_9.
171. Chang, S. G., Bin Yu & Vetterli, M. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing* **9**, 1532–1546 (2000).
172. Su, B., Lu, S. & Tan, C. L. Blurred Image Region Detection and Classification. in *Proceedings of the 19th ACM international conference on Multimedia* 1397–1400 (2011).
173. Saifuddin, M., Bhatnagar, J. M., Segrè, D. & Finzi, A. C. Microbial carbon use efficiency predicted from genome-scale metabolic models. *Nature Communications* **10**, 3568 (2019).
174. Or, D. & Lehmann, P. Surface Evaporative Capacitance: How Soil Type and Rainfall Characteristics Affect Global-Scale Surface Evaporation. *Water Resources Research* **55**, 519–539 (2019).
175. Stirling, G. & Wilsey, B. Empirical Relationships between Species Richness, Evenness, and Proportional Diversity. *Am. Nat.* **158**, 286–299 (2001).
176. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 (1948).
177. Simpson, E. H. Measurement of Diversity. *Nature* **163**, 688 (1949).
178. Whittaker, R. H. Dominance and Diversity in Land Plant Communities. *Science* **147**, 250 (1965).
179. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *PNAS* **113**, 5970–5975 (2016).

Curriculum Vitae

Personal information

First name, Surname:	Bickel, Samuel Mulinda		
Date of birth:	23.11.1987	Sex:	m
Nationality:	Swiss		
Researcher unique identifier:	https://orcid.org/0000-0002-9839-4591 https://publons.com/researcher/3540833/samuel-bickel/ https://scholar.google.com/citations?user=6JHewY8AAAAJ&hl=en&oi=ao		

Education

Year	Faculty/department - University/institution - Country
2020	Ph.D. candidate Environmental Sciences - ETH Zürich - Switzerland
2015	MSc Environmental Science - ETH Zürich - Switzerland
2011	BSc Environmental Science - ETH Zürich - Switzerland

Positions - current and previous

Year	Job title – Employer - Country
2015-present	Research assistant – ETH Zürich - Switzerland

Supervision of students

Master's students	Role - University/institution - Country
3	Co-supervisor - ETH Zürich - Switzerland

Other relevant professional experiences

Year	Description - Role
2019	Organization of the 12th PhD Congress for the Institute of biogeochemistry and pollutant dynamics (IBP) - organizing committee member
2017-2019	Teaching committee ETH - Deputy for the association of academic staff at ETH (AVETH)
2015-2018	Departmental teaching committee - Deputy for the association of academic staff of the Department Environmental Systems Science (VMUSYS)

Publication record

- Publications in peer-reviewed journals:
    - Lehmann, Peter, Samuel Bickel, Zhongwang Wei, and Dani Or. "Physical Constraints for Improved Soil Hydraulic Parameter Estimation by Pedotransfer Functions." *Water Resources Research* 56, no. 4 (2020): e2019WR025963. https://doi.org/10.1029/2019WR025963.
    - Bickel, Samuel, and Dani Or. "Soil Bacterial Diversity Mediated by Microscale Aqueous-Phase Processes across Biomes." *Nature Communications* 11, no. 1 (January 8, 2020): 1–9. https://doi.org/10.1038/s41467-019-13966-w.
    - Bickel, Samuel, Xi Chen, Andreas Papritz, and Dani Or. "A Hierarchy of Environmental Covariates Control the Global Biogeography of Soil Bacterial Richness." *Scientific Reports* 9, no. 1 (August 20, 2019): 1–10. https://doi.org/10.1038/s41598-019-48571-w.
- Conference items:
    - Ruiz, Siul, Samuel Bickel, Peter Lehmann, and Dani Or. Soil, "Climatic and Biophysical Constraints Determine Global Distribution and Activity Windows of Earthworms." *AGU Fall Meeting Abstracts*. American Geophysical Union, 2019.
    - Or, Dani, Samuel Bickel, and Peter Lehmann. "Physical Constraints Improve Soil Hydraulic Parameters Deduced from Pedotransfer Functions." *AGU Fall Meeting Abstracts*. American Geophysical Union, 2019.
    - Lehmann Grunder, Peter Ulrich, Samuel Bickel, and Dani Or. "Global Surface Evaporation – Insights and Opportunities." *EGU General Assembly Conference Abstracts*. Copernicus, 2019. https://doi.org/10.3929/ethz-b-000388731.
    - Bickel, Samuel, Jingyu Wang, and Dani Or. "Spatial Patterns of Soil Bacterial Abundance and Diversity - from Pores to Continents." *AGU Fall Meeting Abstracts*. American Geophysical Union, 2018.
    - Bickel, Samuel, and Dani Or. "Climate and Soil Properties Affect Diversity and Abundance of Soil Microbial Life across Scales – from Grains to Biomes." *EGU General Assembly Conference Abstracts.* Copernicus, 2018. https://doi.org/10.3929/ethz-b-000315679.
    - Ruette, Jonas von, Peter Lehmann, Linfeng Fan, Samuel Bickel, and Dani Or. "STEP TRAMM – A Modeling Interface for Simulating Localized Rainfall Induced Shallow Landslides and Debris Flow Runout Pathways." *EGU General Assembly Conference Abstracts.* Copernicus Publications, 2017. https://doi.org/10.3929/ethz-b-000227224.
    - Or, Dani, Samuel Bickel, and Peter Lehmann. "Linking Soil Type and Rainfall Characteristics towards Estimation of Surface Evaporative Capacitance." *AGU Fall Meeting Abstracts.* American Geophysical Union, 2017.
    - Bickel, Samuel, and Dani Or. "Growing under Pressure: Microbial Growth Rates under Physical Confinement." Edited by SystemsX.ch. Abstract Book, *3rd International SystemsX.Ch Conference on Systems Biology*. Zurich: SystemsX.ch – The Swiss Initiative in Systems Biology, 2017.
    - Bickel, Samuel, and Dani Or. "The Microgeography of Microbial Life in Soil across Climates and Biomes." *Abstract Book: 2016 International Symposium on Microbial Ecology (ISME 16)*. Montreal: International Symposium on Microbial Ecology, 2016.

Appendix

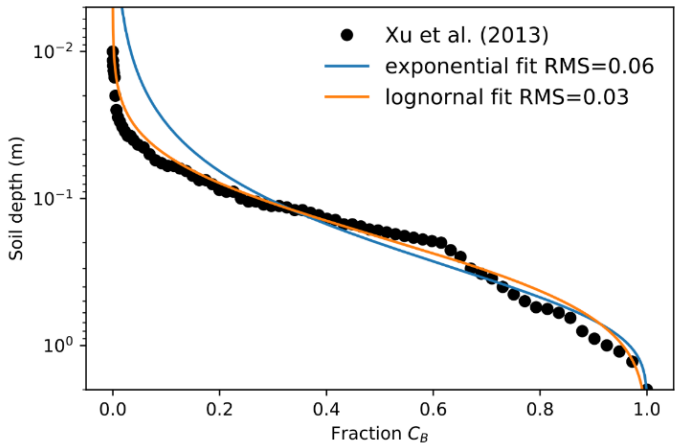
A1 Supplementary Information for: Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes

Supplementary Table 1 | Sources of global data and variables utilized in this study. For global maps, the data has been harmonized to a common grid of 0.1°x0.1° (≈ 11 km) determined by the MSWEP dataset.

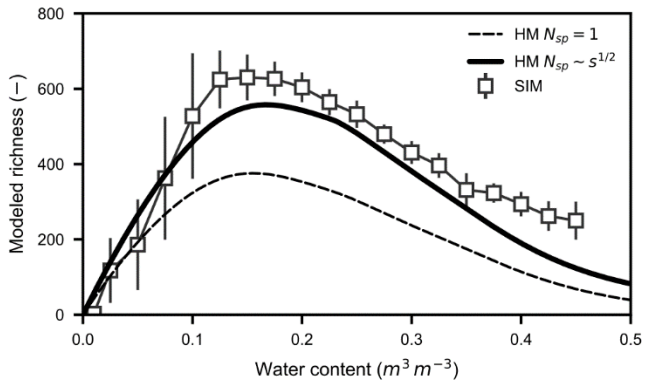
Name	Resolution	Record	Variables	References
MSWEP v2.1	0.1°, 3-hourly	1979-2016	mean annual precipitation, mean consecutive dry days	[57]
WorldClim v2	5 arcmin, monthly climatic	1970-2000	mean annual temperature, mean solar irradiance	[52]
SoilGrids	250m & 10km, 7 soil layers	NA	soil texture (sand, silt, clay), bulk density, pH	[56]
MODIS17	1km, annual	2000-2015	mean annual net primary production	[50]

Supplementary Table 2 | Number of samples for groups of climatic water contents (m^3 m^{-3}) in the diversity datasets. Datasets of bacterial diversity (Earth Microbiome Project - EMP[18], Delgado-Baquerizo *et al.* - DEL[20]) were grouped by climatic water contents (EMP additionally by soil depth; Top: <25cm, Sub: ≥25cm). The numbers of samples per group are reported.

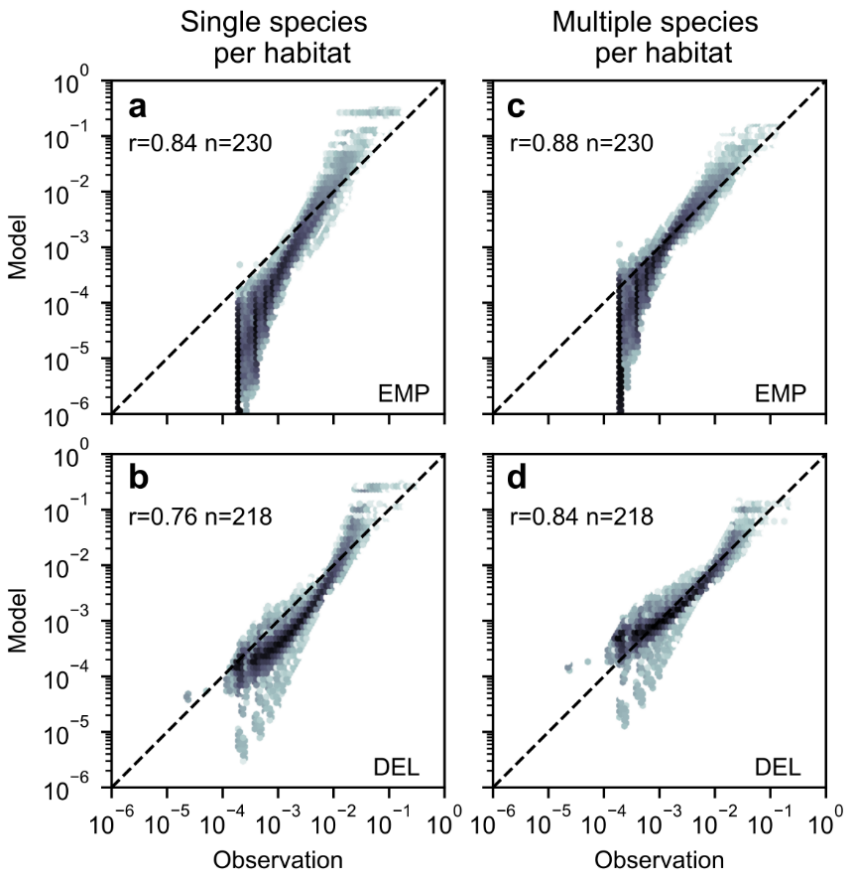
Study	**Climatic water content class**	**Soil layer**	**Number of samples**
DEL	(0.0, 0.05]	-	4
DEL	(0.05, 0.1]	-	6
DEL	(0.1, 0.15]	-	28
DEL	(0.15, 0.2]	-	104
DEL	(0.2, 0.25]	-	71
DEL	(0.25, 0.3]	-	23
DEL	(0.3, 0.35]	-	1
EMP	(0.0, 0.05]	Top	3
EMP	(0.05, 0.1]	Top	75
EMP	(0.1, 0.15]	Top	93
EMP	(0.15, 0.2]	Top	253
EMP		Sub	12
EMP	(0.2, 0.25]	Top	1594
EMP		Sub	44
EMP	(0.25, 0.3]	Top	562
EMP		Sub	93
EMP	(0.3, 0.35]	Top	28
EMP		Sub	11



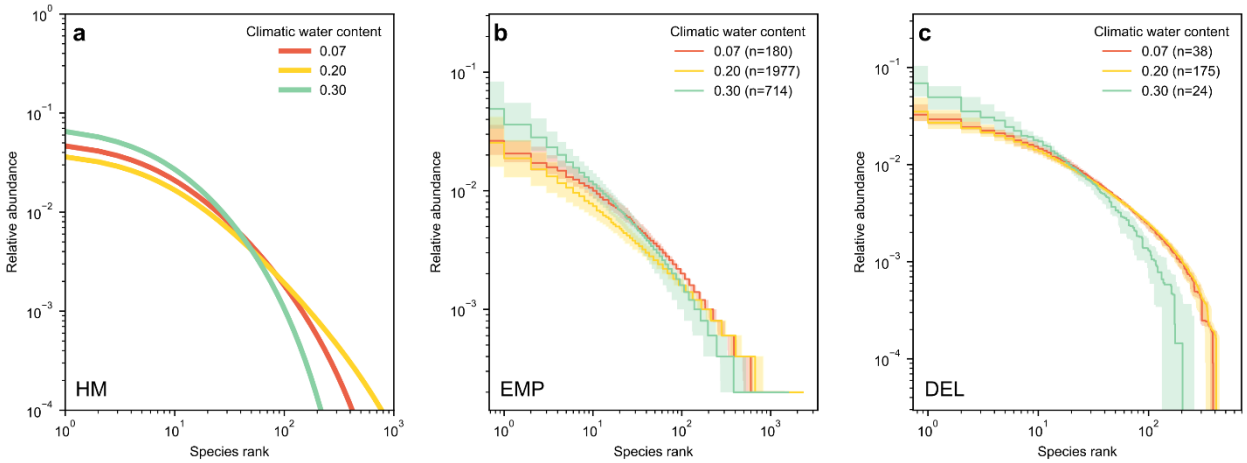
Supplementary Figure 1 | Decay of bacterial biomass carbon (C_B) with soil depth. The cumulative fraction of C_B taken from Xu *et al.* [8] is shown for a maximum depth of two meters (black symbols). The exponential model as reported by Xu *et al.* was fitted (blue line). A lognormal fit (orange line) shows an overall better alignment with the data, especially in the upper 10 cm. The root mean square errors (RMS) are reported in the legend.



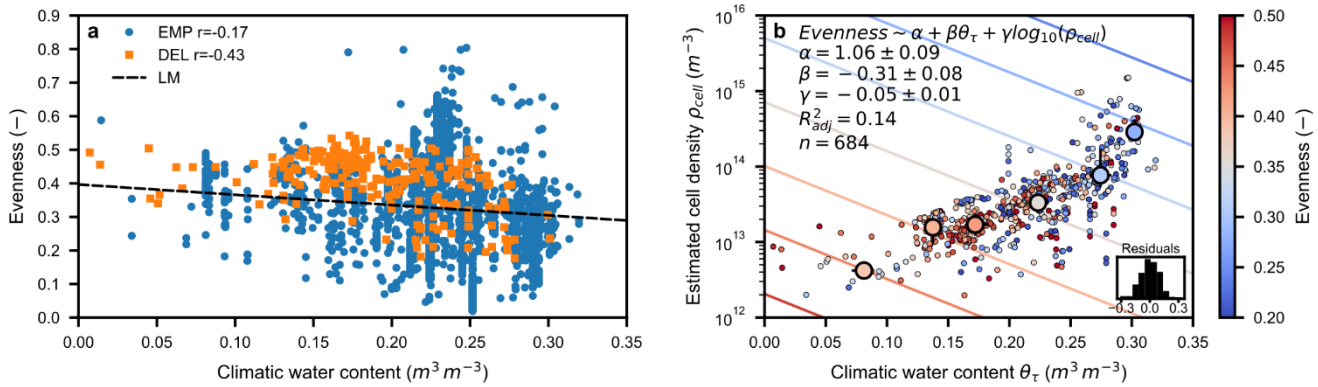
Supplementary Figure 2 | Comparison of the heuristic model (HM) with the spatially-explicit individual-based model (SIM) on surfaces (two dimensional domains). Square symbols and bars (mean ± SD, n = 12) depict richness predicted by the SIM rarified to 1000 counts. The aqueous-phase fragmentation-based HM (solid line) captures the trend in simulated richness with water content (m^3 m^{-3}). The proportionality of the number of species per habitat N_{sp} to the domain's dimensionality (surface or volume) and to the size s of the aqueous habitats ($N_{sp} \sim s^{1/2}$) may explain the difference between SIM and the single species HM (dashed line).



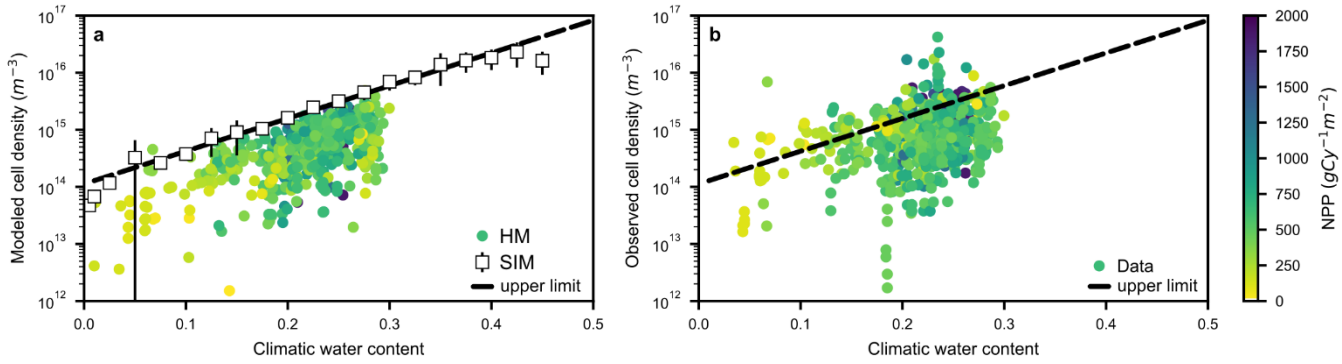
Supplementary Figure 3 | Modeled and observed soil bacterial species abundance distributions (SAD). Comparison of relative abundances from empirical observations (x-axis) with estimates of the aqueous-phase fragmentation-based heuristic model (HM; y-axis). Scenarios with single and multiple species per aqueous habitat are compared to observations. A 1:1 line and Pearson correlations are shown for both soil bacterial diversity datasets. **a**, Relative SADs from the Earth Microbiome Project (EMP) and **b,** from a recent study by Delgado *et al.* (DEL) considering a single species per habitat. The consideration of multiple species per habitat with the number of species N_{sp} proportional to the dimensionality and size s of the habitat ($N_{sp} \sim s^{1/3}$) improves the agreement with model predictions for both datasets; **c,** EMP and **d,** DEL, respectively.



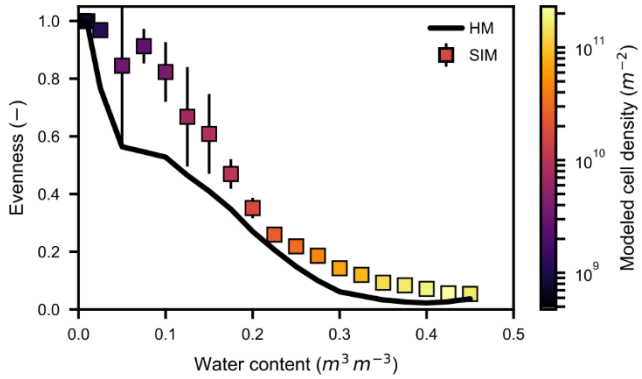
Supplementary Figure 4 | Qualitative comparison of theoretical and empirical species abundance distributions (SADs) for three climatic water contents (m^3 m^{-3}). **a,** SADs generated for median carrying capacity using the aqueous-phase fragmentation-based heuristic model (HM). Empirically observed SADs are grouped into equally spaced intervals of climatic water content (midpoint in legends) for **b,** the Earth Microbiome Project (EMP)[18] data and **c,** for a recent study by Delgado *et al.* (DEL)[20]. For each group the median (solid line) and the interquartile range (shading) as well as the number of samples are reported.



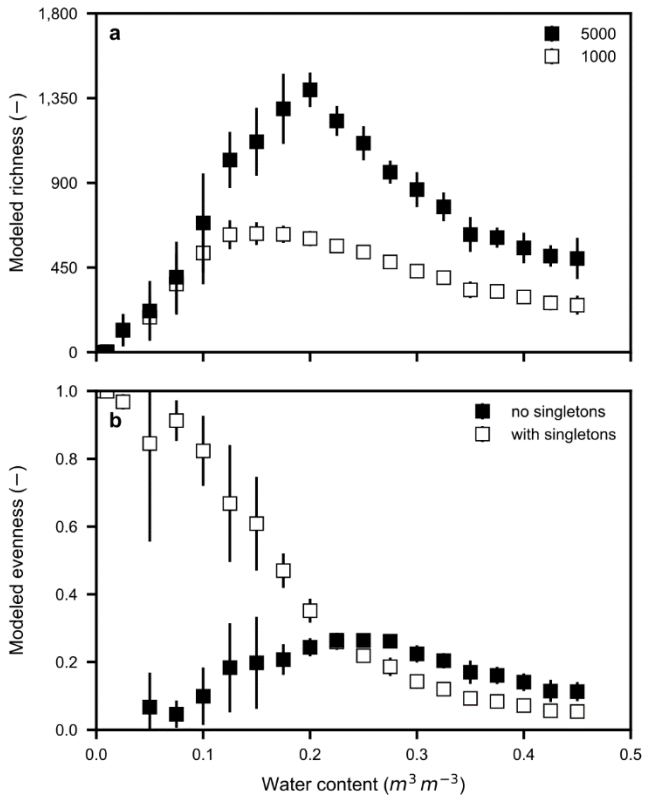
Supplementary Figure 5 | Empirically observed trends of soil bacterial evenness. **a**, Decrease of bacterial community evenness with climatic water content. Pearson correlation *r* for individual samples of both diversity datasets (EMP[18] n = 2871, DEL[20] n = 237) are indicated in the legend. The trend line shows a linear model (LM, see **b**) evaluated for median cell densities. **b**, The linear model was fitted to the empirical data for all sampled locations (n = 684) and the response surface of evenness is shown as a function of climatic water contents and cell densities (colored contours). Samples with cell densities lower than 10^{12} m^{-3} were removed prior to fitting the model as indicated in the figure. Negative slopes (β, γ) suggest that evenness is jointly reduced by increasing climatic water contents and cell density. Model residuals are not indicative of a persistent bias. Additionally, evenness is shown for bins of water contents (median ± IQR) to highlight the central tendency.



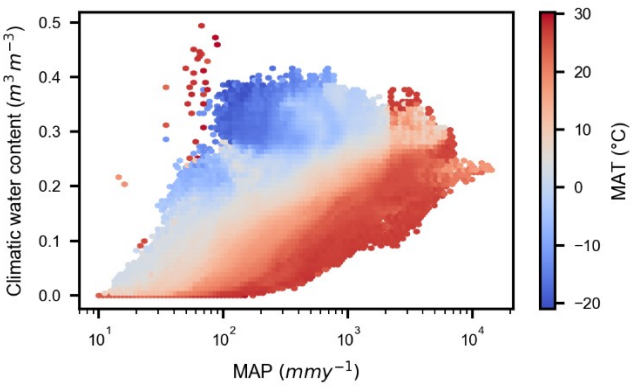
Supplementary Figure 6 | Relation of cell density with climatic water content (m^3 m^{-3}) based on net primary productivity (NPP). **a,** Modeled cell density as a function of NPP, mean annual temperature and soil depth using the heuristic model (HM; colored circles). The dashed regression line represents the mean tendency of the HM when integrated over the entire soil profile of 1 m and thus provides a theoretical upper bound. The dependency on climatic water contents was not explicitly modeled, it is rather indicative of the NPP's relation with hydration regime. The open black symbols are results of the spatially-explicit individual based model (SIM; mean ± SD, n = 12) where enough carbon to support a cell density of 10^{17} m^{-3} was prescribed. The dependency on water contents results from spatial variations in nutrient fluxes with water contents. **b,** Observed values of cell densities[8] are bounded by the theoretical upper limit.



Supplementary Figure 7 | Comparison of evenness estimated using the aqueous-phase fragmentation-based heuristic model (HM) and the spatially-explicit individual-based model (SIM) for different water contents and carrying capacity. The HM (solid line) is evaluated in two dimensions for every value pair of modeled cell density and water contents obtained from the SIM (square symbols and bars – mean ± SD, n = 12). Colors indicate modelled cell density from the SIM.



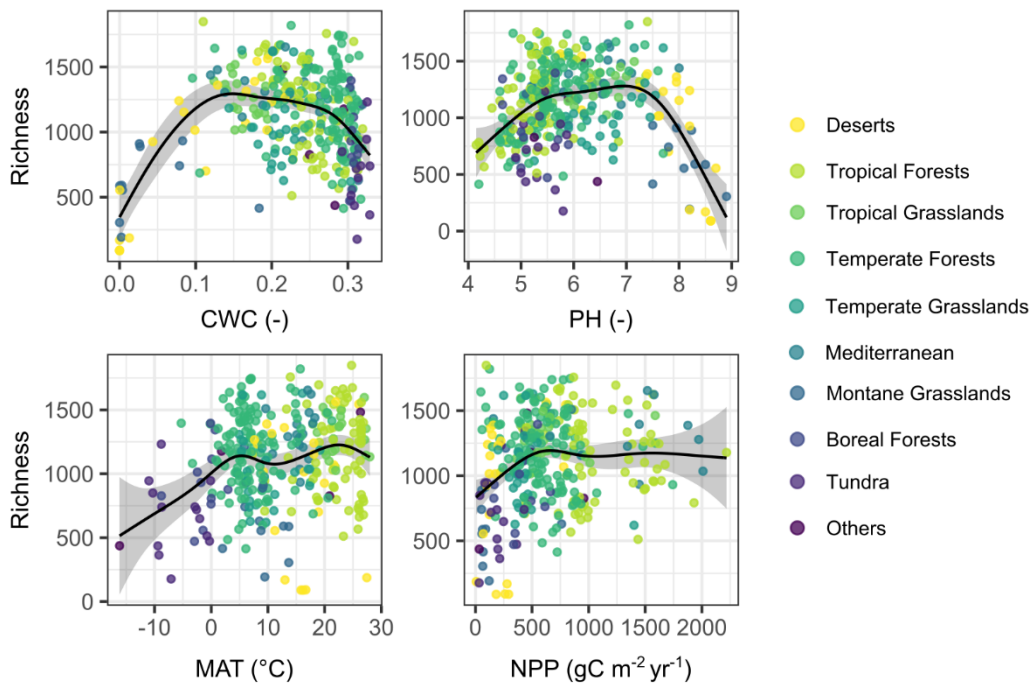
Supplementary Figure 8 | Pre-processing and sampling of simulated bacterial species abundance data may exert a strong effect on the deduced diversity metrics. **a** and **b**, Simulated soil bacterial diversity metrics using the spatially explicit individual-based model (SIM, mean ± SD, n = 12 different simulations). **a,** Rarefying to 5000 counts leads to higher magnitudes (compared to 1000 counts) of observed bacterial richness, yet the trends with water content remain consistent. **b**, Removing singletons (species sampled only once) exerts a strong influence on bacterial community evenness. This pre-processing step could distort the apparent relation of bacterial community evenness with climatic water content.



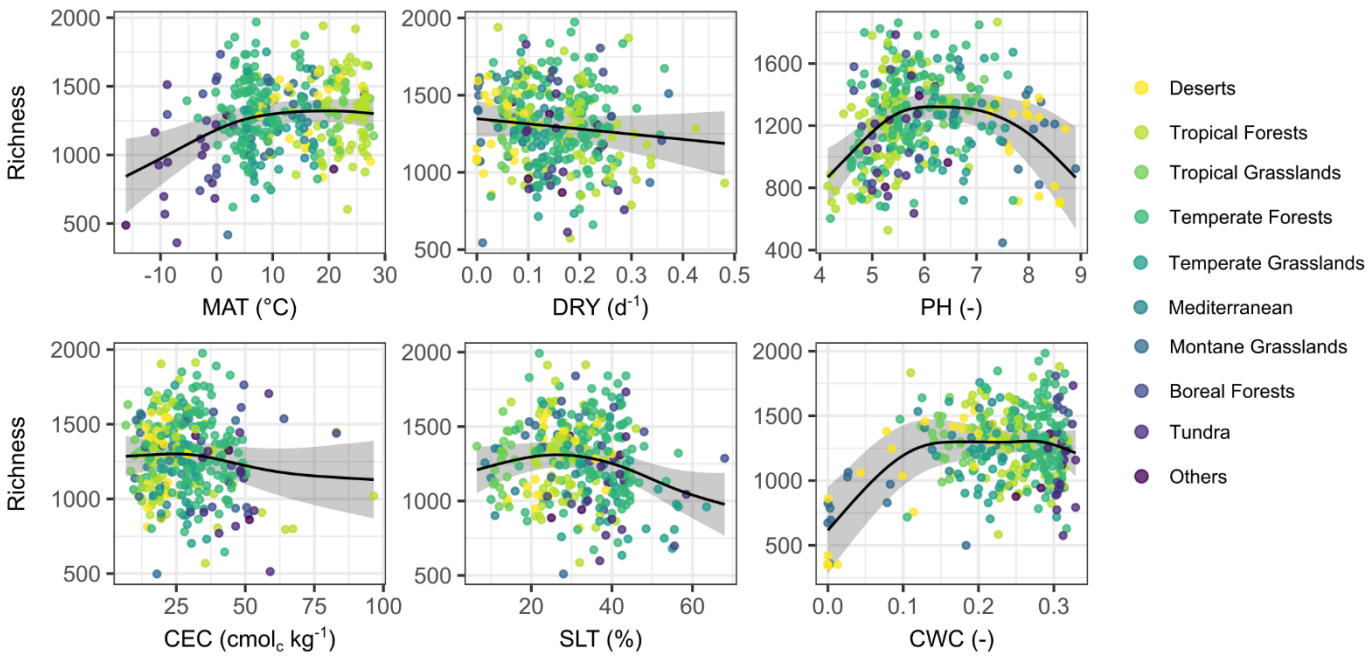
Supplementary Figure 9 | Global distribution of climatic soil water contents in relation with mean annual precipitation (MAP) and mean annual temperature (MAT). MAT spans climatic regions with different potential evapotranspiration (or aridity) and distinguishes locations (together with soil type) where climatic water contents may vary for the same MAP. For example, colder regions tend to require less MAP to attain relatively high climatic soil water contents.

A2 Supplementary Information for: A hierarchy of environmental covariates control the global biogeography of soil bacterial richness

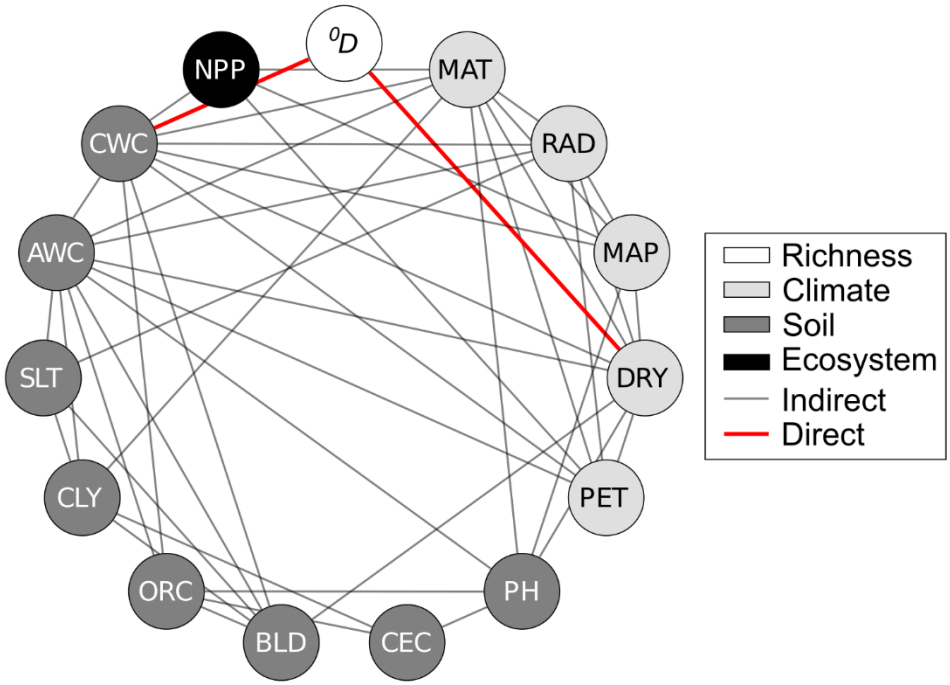
SI Figures



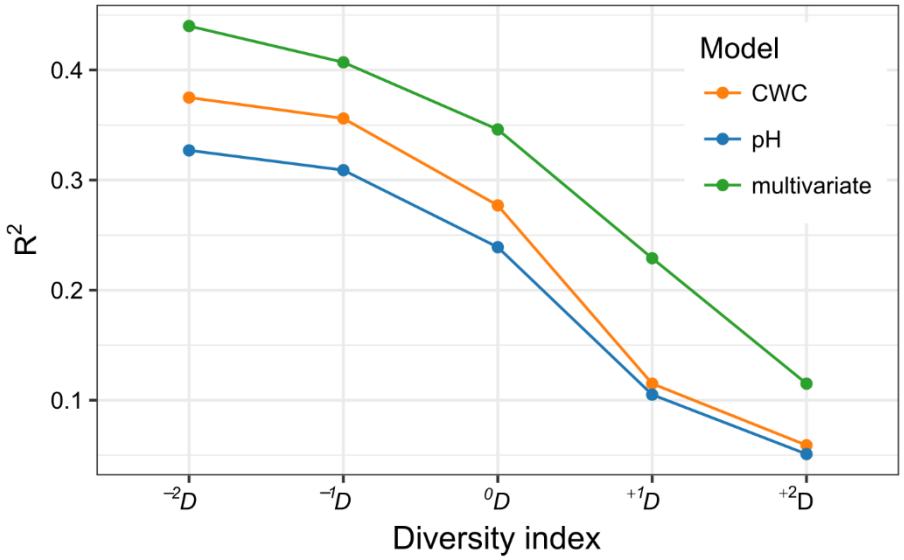
SI Fig. S1: Univariate GAM of selected variables. Colors indicate the sampled biomes. Bacterial richness as a function of climatic water content (CWC; R^2 = 27.7%, RMSE = 298.1, AIC = 4557.5, EDF = 4.7), soil pH (PH; R^2 = 23.8%, RMSE = 306.0, AIC = 4574.0, EDF = 5.1), mean annual temperature (MAT; R^2 = 5.9%, RMSE = 340.0, AIC = 4640.6, EDF = 4.9) and net primary productivity (NPP, R^2 = 5.7%, RMSE = 340.5, AIC = 4642.4, EDF = 3.7). Colors indicate the sampled biomes. Shaded areas correspond to standard errors (n = 320).



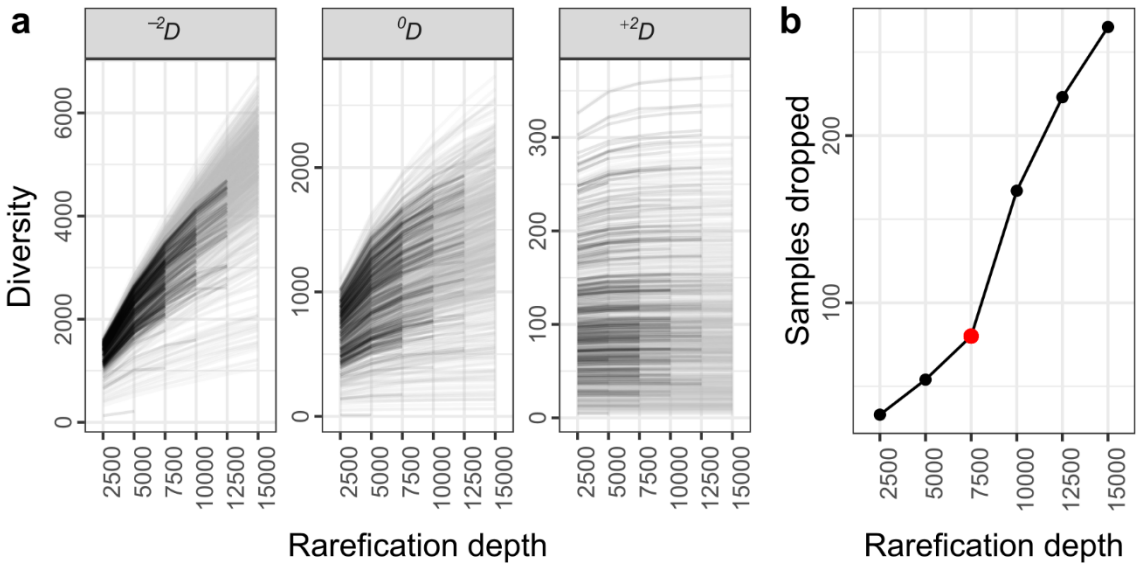
SI Fig. S2: Partial dependence plots of multivariate GAM for covariates temperature (MAT), frequency of dry days (DRY), soil pH (PH), cation exchange capacity (CEC), silt content (SLT), and climatic water content (CWC). Colors indicate the sampled biomes. Shaded areas correspond to standard errors (R^2 = 34.5%, RMSE = 283.6, AIC = 4517.8, n = 320).



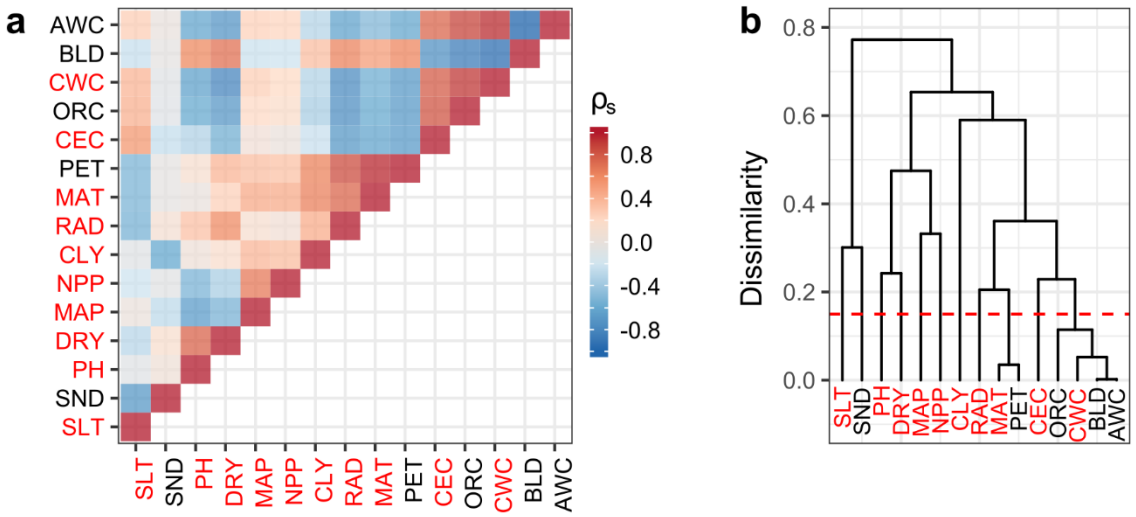
SI Fig. S3: Graph of dependencies estimated by the causal additive model (CAM) algorithm. Covariates are grouped by climate, soil and ecosystem properties. Bacterial richness (^0D) is the variable of interest and edges indicate inferred causal dependencies ($p \leq 0.0005$). The direct edges to soil bacterial richness are shown in red while all indirect edges are shown in grey.



SI Fig. S4: Change of goodness-of-fit (R^2) of univariate (climatic water content CWC, pH) and multivariate GAM for diversity indices ${}^{q}D$ that give dominant species more weight by increasing order q of the index.



SI Fig. S5: Species accumulation curves for varying rarefication (sampling) depths. (**a**) Different diversity metrics show varying response to sampling depth. More weights on abundant species (^{+2}D) leads to saturation of the metric with smaller rarefication depth. (**b**) Choice of sampling depth (red point) as a trade-off between the numbers of dropped samples and maximized rarefication depth.



SI Fig. S6: Spearman correlation among covariates. (**a**) Matrix of pairwise correlation coefficients. (**b**) Hierarchical clustering of covariates based on their dissimilarity. From insufficiently dissimilar covariates (<0.15) only a single covariate (red) was chosen for modelling.

SI Tables

SI Table S1: Summary of covariates and their input data.

Covariate	**Unit**	**Description**	**Input**
MAT	°C	Temperature	WorldClim v2 [52]
RAD	kJ m^{-2} d^{-1}	Solar radiation	WorldClim v2 [52]
MAP	mm yr^{-1}	Precipitation	MSWEP v2.2 [97]
DRY	d	Number of consecutive dry days	PET, MSWEP v2.2 [97]
PET	mm d^{-1}	Potential evapotranspiration	*f*(MAT, RAD) [58]
PH	-	Soil pH	SoilGrids [56]
CEC	$cmol_c$ kg^{-1}	Cation exchange capacity	SoilGrids [56]
BLD	kg m^{-3}	Bulk density	SoilGrids [56]
ORC	g kg^{-1}	Organic carbon content	SoilGrids [56]
CLY	%	Clay content	SoilGrids [56]
SLT	%	Silt content	SoilGrids [56]
SND	%	Sand content	SoilGrids [56]
AWC	-	Available water-holding capacity	*f*(BLD, ORC, SLT, CLY) [55]
CWC	-	Climatic water content	*f*(PET, DRY, AWC)
NPP	g C m^{-2} yr^{-1}	Mean net primary productivity (2000-2015)	MODIS17 [50]

SI Table S2: Leave one out cross-validated test errors of the log ratio of bacterial richness with different (global) relative abundance cutoffs.

	Log ratio := $log(N_{rare}/N_{common})$		
Global relative abundance cutoff	***0.0005%***	***0.005%***	***0.05%***
MAT	15.8%	11.4%	12.2%
RAD	18.6%	23.8%	22.9%
MAP	12.5%	16.7%	12.2%
DRY	11.3%	19.5%	19.2%
PH	10.6%	21.2%	21.5%
CEC	17.9%	16.4%	15.6%
CLY	-1.0%	-0.3%	-0.9%
SLT	15.4%	20.3%	20.4%
CWC	11.7%	22.3%	24.3%
NPP	14.1%	14.6%	6.7%

SI Methods

Calculation of climatic water content

Climatic water content (CWC), was introduced to approximately describe the state of soil wetness specific to climate and soil storage capacity. It was calculated based on the assumption that the top one meter of soil (d_{soil} = 1 m) can be fully replenished up to field capacity (θ_{FC} defined as half porosity/AWC) during rainfall events, and drain exponentially in consecutive dry days (DRY). During this time, water mass is lost at a constant rate determined by (mean daily) potential evapotranspiration (PET) resulting in an exponential reduction of average water content. The MSWEP [97] precipitation records of 37 years (1979–2016) are used at daily resolution to derive average rainfall quantities per wetting-drying cycle. The precipitation time series is subjected to a threshold taken from estimates of PET to identify wetting events. The metric used is the mean time interval between rainfall events (an ensemble average) τ. This quantity combined with daily PET (m d^{-1}) lead to the following expression for climatic water content θ_τ:

$\theta_\tau = \theta_{FC} e^{-\alpha<\tau>}$ with $\alpha = \frac{PET}{d_{soil}\,\theta_{FC}}$

Diversity indices

Diversity of ecological communities can be quantified from different aspects, e.g. richness measures the number of unique types present in a community, while evenness compares the relative abundances that make up the local community[175]. Here, to measure how diverse a local community is, we opted for Hill's diversity ${}^qD$[62], defined as:

$${}^qD = \left(\sum_{i=1}^{N} p_i^q \right)^{1/(1-q)}$$

where p_i refers to the relative abundance (with $\sum p_i = 1$) of the i^{th} type and N is the total number of types in the population. The order q controls the weights given to species of different local abundance, i.e. fewer weights will be given to the rarities if $q > 0$, and vice versa. ${}^0D$ (richness) simply counts unique types in a population and thus gives equal weight to all species regardless of local abundance. ${}^1D$ and ${}^2D$ are closely related to the Shannon index[176], as a limiting case for q=1 and the Simpson index[177], respectively. Therefore, ${}^1D$ is less sensitive to low abundant species in local communities compared to ${}^0D$, while ${}^2D$ is the least sensitive and can be considered as a measure of dominant species[178,179]. Their counterparts, i.e. ${}^{-2}D$ and ${}^{-1}D$ were also included to get a full picture of the local abundance distribution.

A3 Supplementary Information for: The chosen few – variations in common and rare soil bacteria across biomes

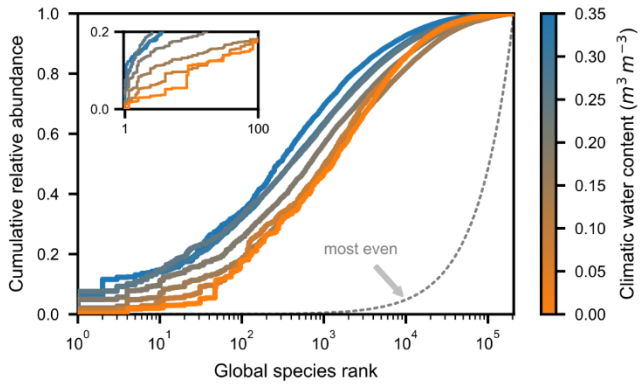


Fig. S1. Observed gradual shift in community composition with varying climatic water contents. The relative abundance distributions (RADs) of soil bacteria for groups of climatic water contents (bins of 0.05) and are shown as cumulative relative abundance using previously published data[17]. Values are sorted by global species rank that ranges from most (rank one) to least abundant. The distribution displays a systematic shift towards increased proportion of rare species under dry conditions. The distribution is shifted towards more even soil bacterial communities where each species would contribute equally to community composition regardless of their global ranks (dashed line). The inset figure on the left shows the 100 most abundant species on a linear scale. Additionally, a scenario is shown where every species would contribute equally to the composition of the community regardless of rank (dashed line).

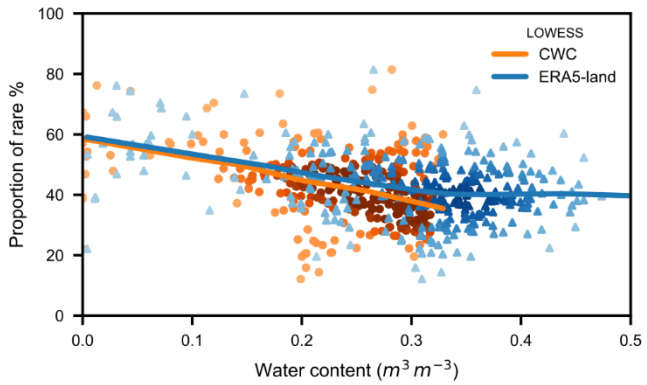


Fig. S2. Proportion of rare bacteria decrease with increasing water contents. Two long term estimates of soil water content show consistent trends of decreasing proportion of rare bacteria indicated by smoothened estimates (LOWESS) of the data. ERA5-land (https://doi.org/10.24381/cds.68d2bb30) derived estimates of mean water contents (0.1°x0.1°, monthly for 1981-2019) compare favorably with climatic water contents (CWC, based on rainfall frequency[9,17]).

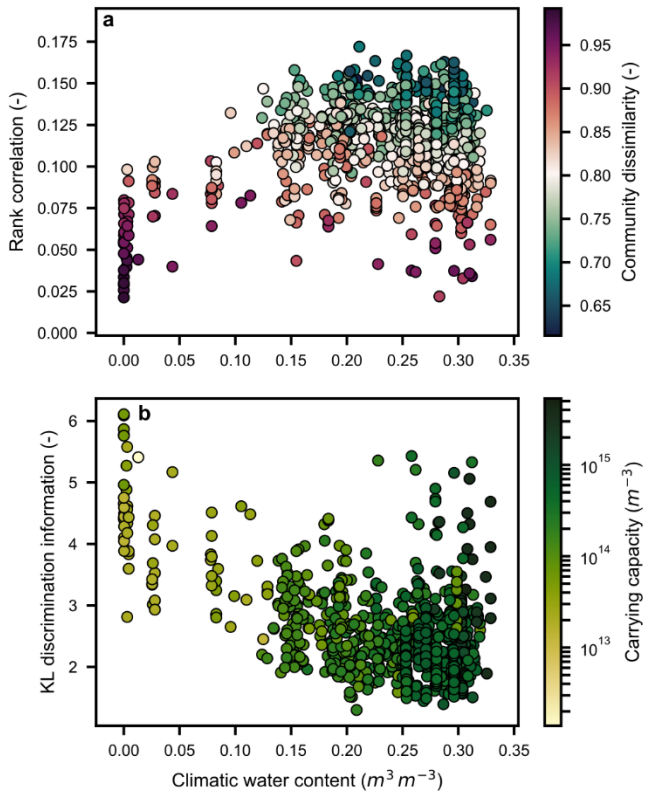


Fig. S3. Shifts in community composition and species ranking with climatic water contents relative to the global average relative abundance distribution (RAD). **a,** Spearman rank correlation between each sample's RAD and the global RAD. Colors indicate Bray-Curtis community dissimilarity. **b,** Discrimination information measured via Kullback-Leibler (KL) divergence between each samples RAD and the global RAD. Higher values indicate that more information is needed to describe the samples RAD relative to the global RAD. Colors indicate estimated cell density[9] (carrying capacity) that was calculated using previously published data[17]. Few exceptions to the clear tendency include mostly samples from nutrient rich environments as indicated by their high potential carrying capacity.

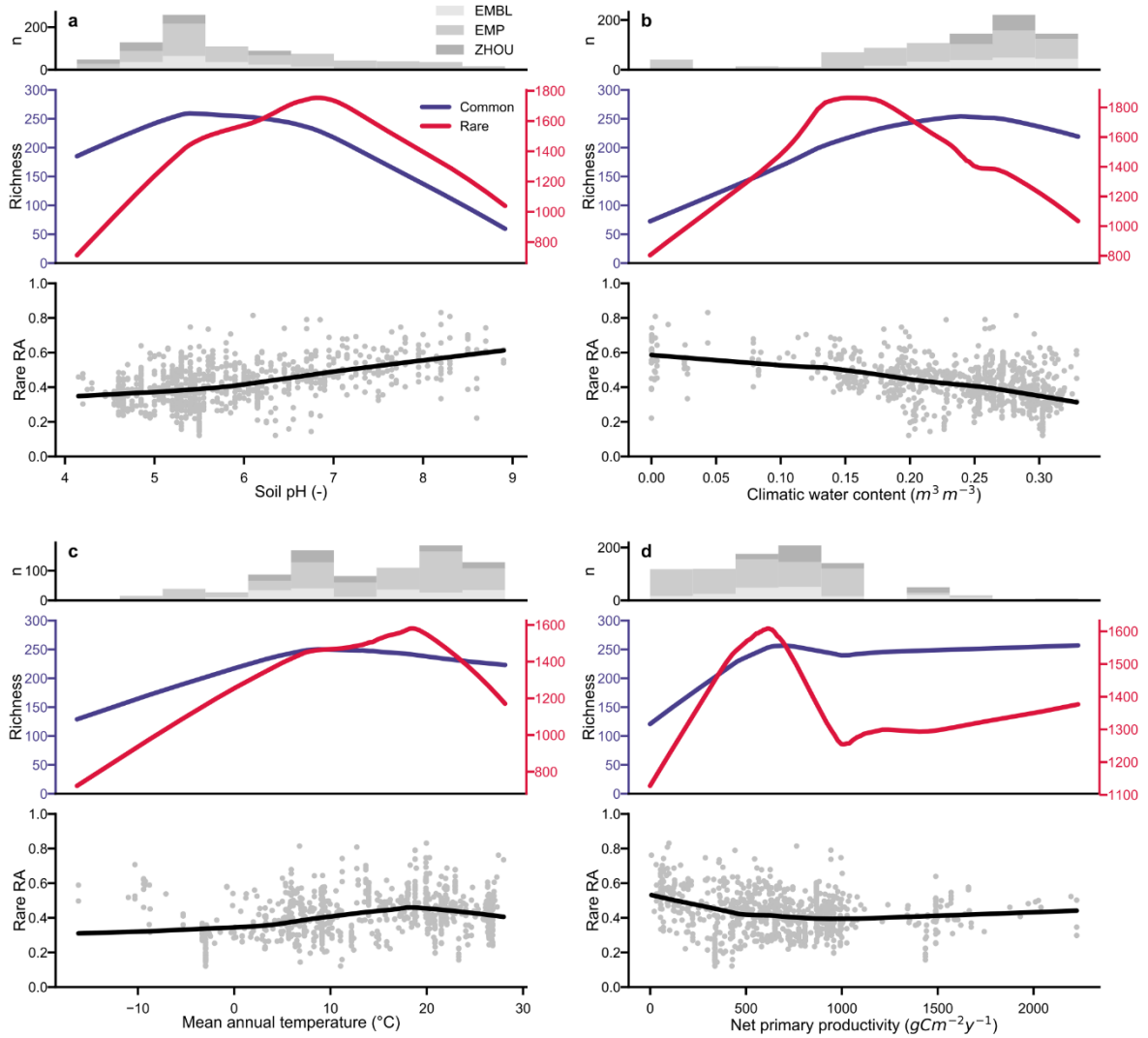


Fig. S4. The proportion of rare bacteria varies with environmental conditions. **a**-**d**, Empirical trends in relative abundance (RA) of rare bacteria compared for selected variables (grey symbols and solid black line, individual samples (n=844, from studies EMBL[5], EMP[18] and ZHOU[28]) and locally weighted scatter plot smooth - LOWESS). Colored lines indicate LOWESS of bacterial richness for common and rare bacteria (purple and red, respectively). Rare and common species are indicated on separate y-axis. **a**, soil pH, **b**, climatic water contents, **c**, mean annual temperature and **d**, net primary productivity.

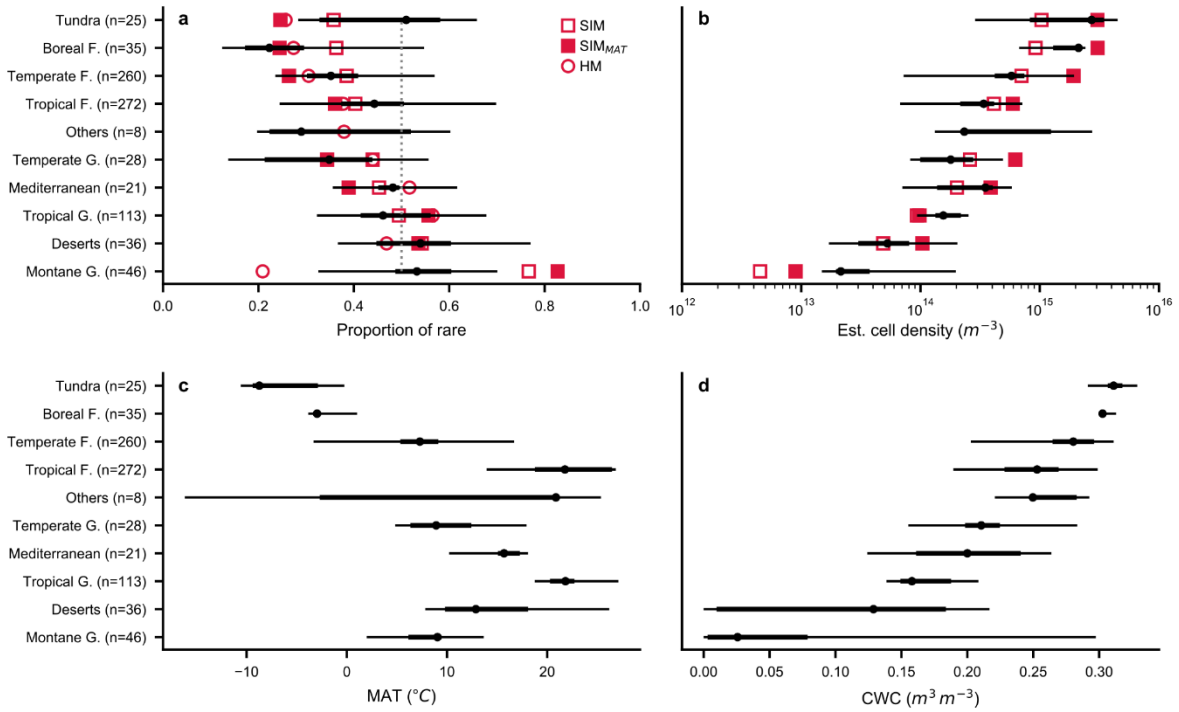


Fig. S5. Simulated bacterial community for different biomes. **a**-**d**, Median and central 50% and 95% of values are indicated for each biome. Predictions by a heuristic model[9] (HM, open circles) are shown together with model results of the SIM, and SIM with temperature dependency (SIM_{MAT}; open and closed squares, respectively). **a**, Proportion of rare species. **b**, Carrying capacity (estimated cell density). **c**, Mean annual temperature (MAT) and **d**, climatic water content (CWC) that provide average input values for the SIM.

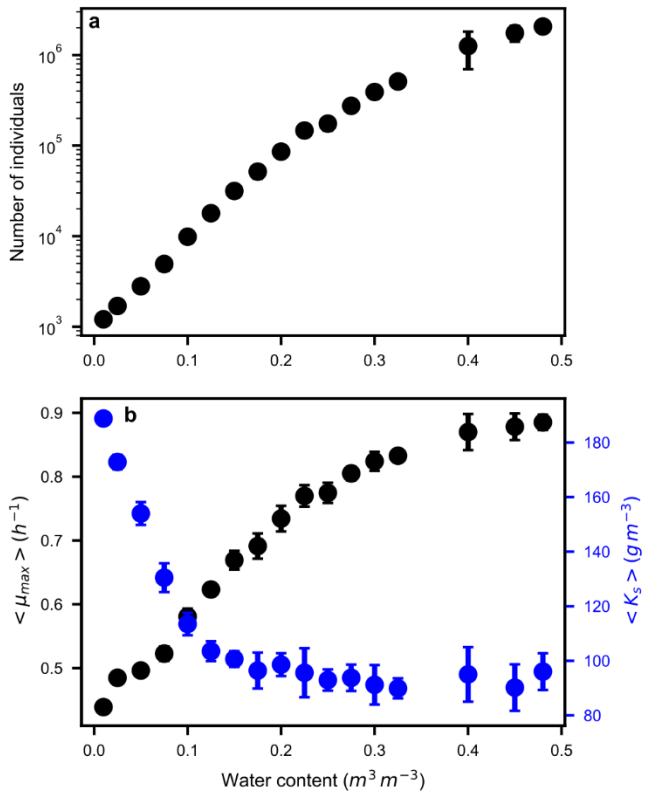


Fig. S6. Simulated bacterial community for varying water contents. **a**, The number of individuals increases exponentially with water contents although some saturation can be observed for very wet conditions (mean ± SD, n = 5 except for water contents > 0.35 where n = 3). The number of potentially maintained individuals is not prescribed and emerges from increased fluxes at high water contents, carbon input (boundary conditions) and physiological properties of individual species. **b**, Ensemble averages of Monod parameters (μ_{max} and K_s, maximal growth rate and half saturation constant, respectively) display consistent patterns with water contents.

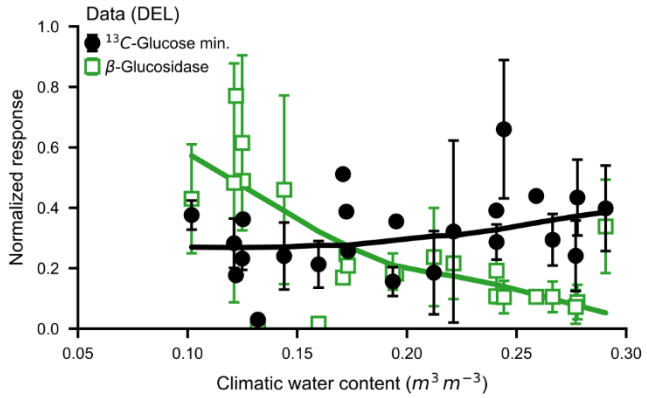


Fig. S7. Ecosystem functions vary with climatic water contents (CWC). Ecosystem functions, exemplified by Glucose mineralization and β-Glucosidase activity, can respond differently to CWC. The normalized response is adapted from a previous report (DEL[113], 81 samples aggregated spatially within 0.1° for 23 sites; mean ± SD) and solid lines are smoothened estimates (LOWESS).

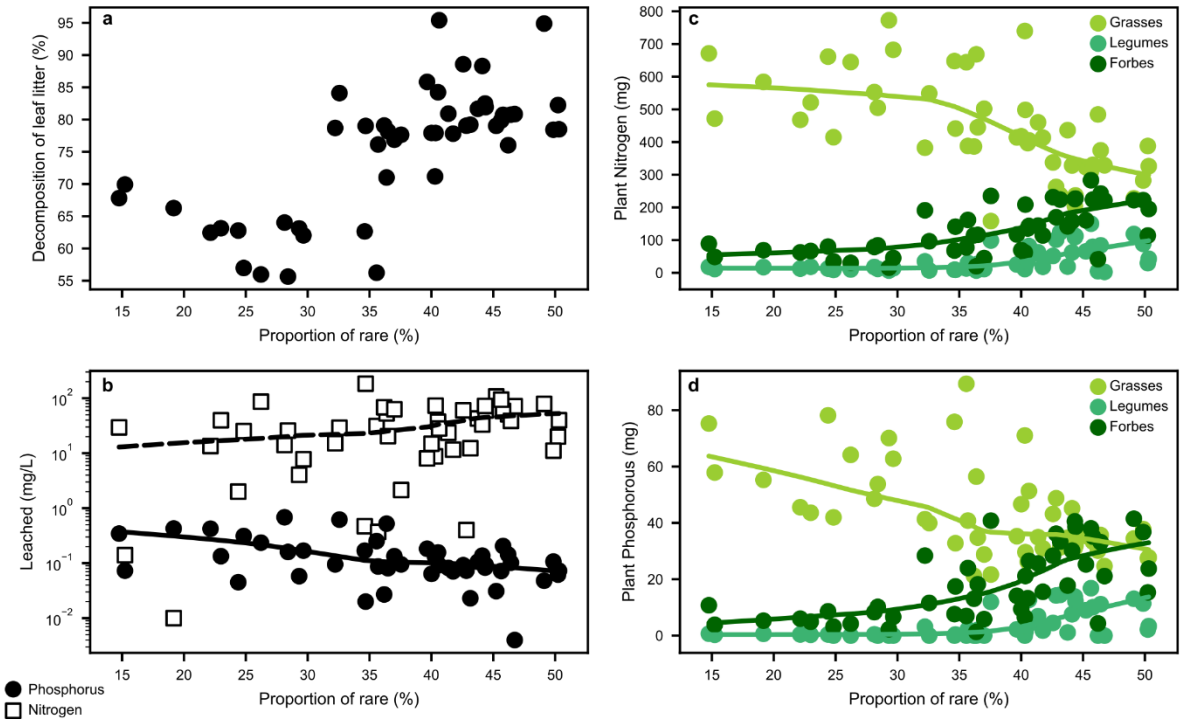


Fig. S8. Ecosystem functions related to microbial diversity that are potentially associated with rare soil bacteria. **a**-**d**, Our framework for delineating common and rare bacteria was applied to a recent study[111] that used microcosm experiments with an artificial diversity gradient. Reported species count data was rarefied to the minimum number of total counts across samples (1'438) and averaged across 15 independent realizations. Rare soil bacteria contributed largely to bacterial richness that was associated with multiple ecosystem functions[111]. Lines represent smoothed estimates (LOWESS). **a**, Leaf litter decomposition increased with proportion of rare bacteria. **b**, Leaching of Nitrogen and Phosphorus showed opposing trends with increased proportions of rare bacteria. **c**, Incorporation of Nitrogen and **d**, Phosphorus into plant tissue possibly associated with the proportion of rare species differed among plant type. Grasses incorporated less Nitrogen and Phosphorus with increasing proportion of rare species while Legumes and Forbes displayed opposite tendencies.

Table S1. Resampling of the dataset to test robustness of threshold selection. A subset of samples was selected randomly to perform resampling with replacement (n_{boot} = 1'000) from the previously published data[17] on community composition. The threshold *t* of relative abundance that is used to distinguish rare and common species and the resulting relative abundance of rare bacteria RA_r are shown for the resampled datasets. Mean and SD are reported.

Samples	t_{mean} **(%)**	t_{SD} **(%)**	$RA_{r,mean}$ **(%)**	$RA_{r,SD}$ **(%)**
52	0.021	0.003	39	11
105	0.021	0.002	41	12
211	0.021	0.002	42	12
422	0.020	0.002	42	12
844	0.019	0.002	42	12

A4 Supplementary Information for: How soil bacterial microgeography affects community interactions and soil functions

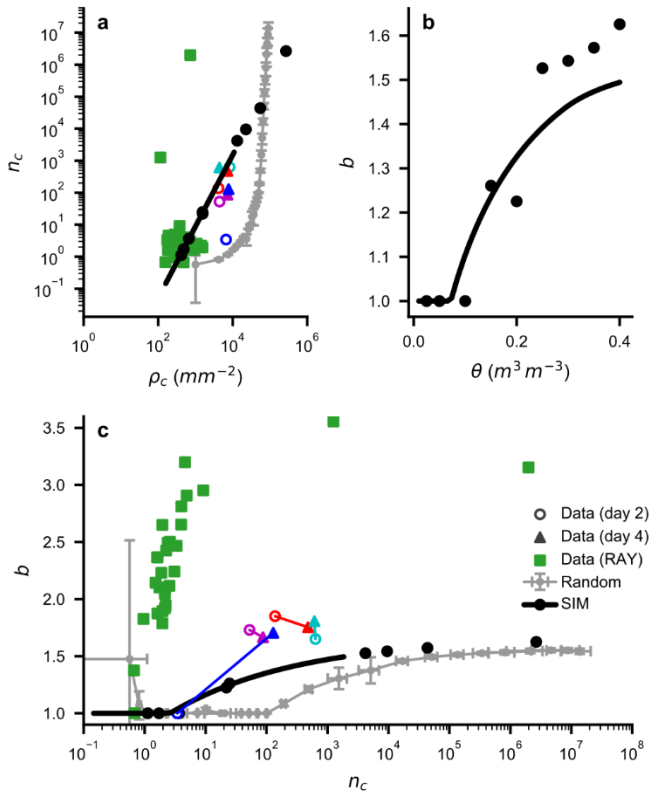


Figure S1. Modeled and observed community size distribution parameters. **a**-**c**, A power law with exponential cutoff is used to describe the soil bacterial community size distribution. Estimated parameters are shown for empirical data (microcosm, RAY) and the spatially-explicit individual-based model (SIM) with parametrization for typical cell densities (solid line). For comparison, parameters were also calculated for random cell distributions under varying density (grey symbols; mean ± SD, n = 24). **a**, Cutoff parameter n_c is related to cell density ρ_c. **b**, Exponent b is not independent from n_c and varies with water content θ mediated by ρ_c. **c**, Relation between n_c and b used for parametrization (see Methods).

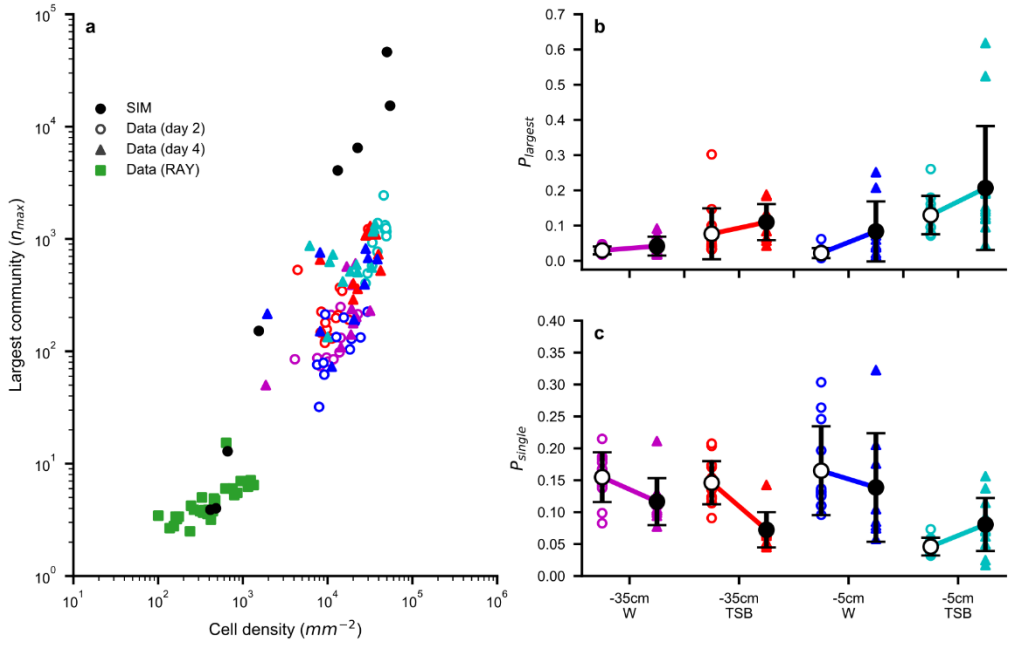


Figure S2. Soil bacterial communities observed in the soil microcosm experiment. **a**, The size of the largest community increases consistently with cell density for microcosm data, results of the spatially-explicit individual-based model (SIM) and data from an independent study[12] (RAY). **b**, In the microcosm experiment the proportion of cells in the largest community ($P_{largest}$) increased with increasing nutrient and hydration conditions. **c**, the proportion of isolated, single cells (P_{single}) varies across treatments.

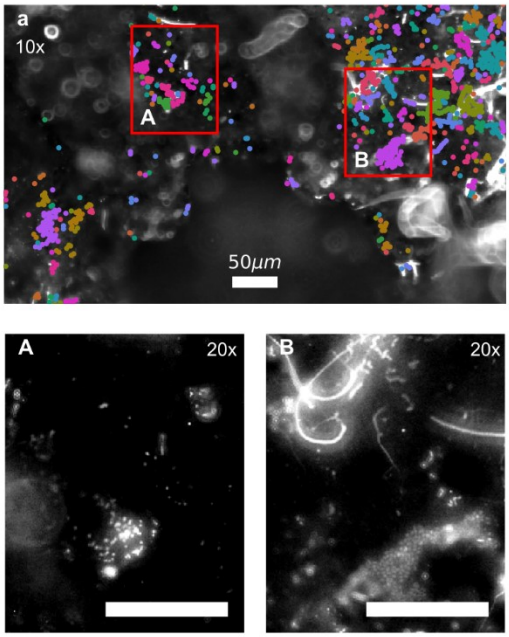


Figure S3. Soil bacterial communities observed in the soil microcosm experiment. **a**, Communities detected after 2 days of growth stained with SYTO9 (grey scale) and imaged at one micrometer resolution (10x) as used for measurement of community sizes. Cells are grouped into the same community if they are located within five micrometers (shown as different colors for each group of cells). Only cells in the focal plane are labeled and used in the analysis. Detailed view on soil microcolonies (boxes **A** and **B**) at 0.5 micrometer resolution (20x); for clarity truncated below median and above 99 percent intensity. Varying soil surface topology and growth morphologies challenge cell density estimation. **A**, Small settlement of round cells on a soil grain. **B**, Filamentous growth is only partially detected as 'chains' of individual cells. Densely packed colonies of round cells with low fluorescence intensity as visible in the lower half could cause additional uncertainties.

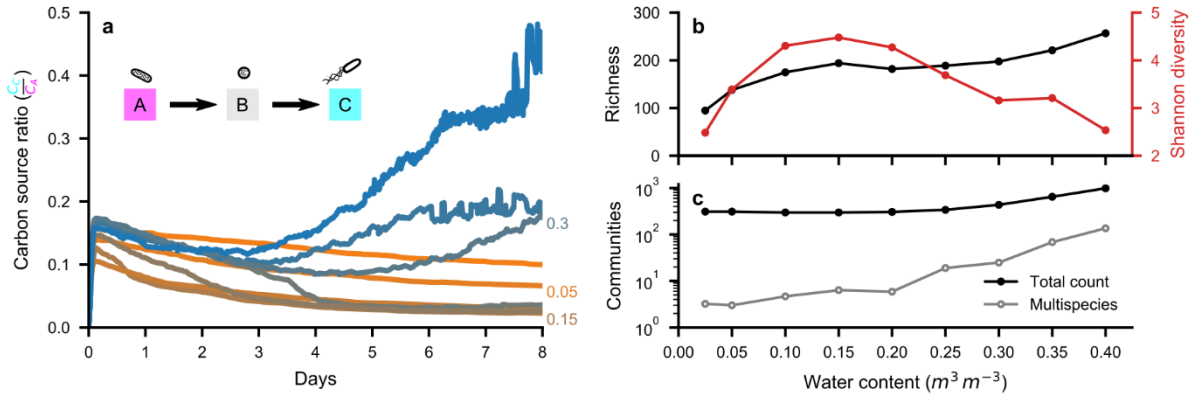


Figure S4. Trophic interactions are enhanced in wet soils with implications for species diversity. **a**, Dynamics of median (n=9) carbon source concentrations as obtained from the spatially-explicit individual-based model (SIM). The SIM considers a degradation pathway from carbon source A to C (consumption A of releases B to the aqueous phase, etc.). The concentration ratio of the end product C (cyan) to the source compound A (magenta) indicates trophic interactions via diffusion in the aqueous phase (C_C/C_A, colored lines from orange to blue indicate different water contents in m^3m^{-3}). **b**, Average bacterial diversity changes with hydration condition. Richness (black) increases with increasing water content. Shannon diversity (red) decreases towards wet conditions indicating reduced evenness. **c**, The total number of communities (black) and the number of multispecies communities (grey) increase with water content.

A5 Global earthworm distribution and activity windows determined by soil hydromechanical constraints

GLOBAL EARTHWORM DISTRIBUTION AND ACTIVITY WINDOWS DETERMINED BY SOIL HYDROMECHANICAL CONSTRAINTS

SIUL A. RUIZ[1,2*], SAMUEL BICKEL[1*], AND DANI OR[1,3]

[1]INSTITUTE OF BIOGEOCHEMISTRY AND POLLUTANT DYNAMICS,
SOIL AND TERRESTRIAL ENVIRONMENTAL PHYSICS, ETH, ZURICH SWITZERLAND

[2]FACULTY OF ENGINEERING AND PHYSICAL SCIENCES,
BIOENGINEERING GROUP, UNIVERSITY OF SOUTHAMPTON, SOUTHAMPTON

[3]DIVISION OF HYDROLOGIC SCIENCES, DESERT RESEARCH INSTITUTE, RENO, NV, USA

* THESE AUTHORS CONTRIBUTED EQUALLY TO THIS WORK

Name of corresponding author: Siul Ruiz
E-mail: s.a.ruiz@soton.ac.uk

Abstract

Earthworms activity modifies soil structure and promotes ecological and hydrological soil functioning. Earthworms use their flexible hydro-skeleton to burrow and expand biopores, hence their activity is constrained by soil hydromechanical conditions that permit deformation at earthworm's maximal hydro-skeletal pressure (≈ 200 kPa). A novel biophysical model links earthworms' physiological limits with bioturbation permitting soil conditions across biomes and climate regions. We inject additional constraints such as freezing temperatures, soil pH, and high sand content that exclude earthworm activity to develop the first predictive global map of earthworm habitats in good agreement with observations. Earthworm activity is strongly constrained by variable seasonal patterns across latitudes. The mechanistic model delineates potential for earthworm migration and regions sensitive to climate and land use changes.

Main

Subterranean activity by earthworms sustains soil structure and provides numerous ecosystem services[1]. Soil biopores formed by burrowing earthworms serve as preferential pathways for water flow and aeration[2]. They are hot spots of biological activity that can be reused by growing roots, improve groundwater recharge, soil water retention and support oxic conditions in soil profiles[3,4]. In locations with abundant plant-derived particulate organic carbon (POM), earthworms ingest POM-rich soil[5] and often line their burrows with secreted castings. Soil ingestion by earthworms can augment microbial activity and stimulates the formation of soil aggregates[6]. Overall, earthworm activity is attributed to significant enhancement in specific crop yields up to 25% [7]. Empirical evidence suggests that earthworms are efficient "ecosystem engineers"[8] and play a prominent role in remediating adverse soil compaction[9] that affects nearly 5% of the world's arable land (about 68 Mha)[10]. Soil bioturbation by earthworms is driven by subterranean resource exploration at rates and frequencies that are linked to the availability of soil organic carbon from decomposing plant residue[2] and their mechanical ability to move in the subsurface. The soil hydro-mechanical conditions[11] link soil strength with soil water content and regulate earthworms ability to burrow through soil. The kinematics of earthworm burrowing rely on locally extending the frontal segments of their body to mechanically penetrate the soil, followed by subsequent expansion of these segments to anchor and recollect extended segments, thereby pushing themselves through the soil[12,13]. The local pressures required by the earthworm's hydro-skeleton for expanding a new burrow are the primary determinants of penetration-cavity expansion[13], and vary widely with soil type and hydration conditions. Availability of spatially resolved soil properties and climatic records of soil hydration conditions offer opportunities for harnessing spatial and dynamic information to identify potential earthworm habitats at high resolution[14]. Ecological studies have provided insight into regional earthworm

distributions[15,16] along with earthworm seasonal activity windows[17,18]. In addition to innate ecological patterns, physical constraints may affect earthworms behaviors that include sensitivity to temperature, soil compaction, and soil moisture[19].

Physical bounds on earthworm bioturbation have been quantified recently by considering the interplay of soil hydro-mechanical constraints and biomechanical limit pressures that could be exerted by the earthworms' hydro-skeleton[11]. These insights allow delineation of regions that permit bioturbation activity and offer a biophysical and climatic context for global earthworm abundance and distribution[14,15,20]. Mechanistic models could predict consequences of agricultural intensification with potential for soil compaction while simultaneously considering climatic shifts that would affect future earthworm bioturbation activity windows (e.g. dormancy during dry seasons in Mediterranean climates) and associated ecosystem services.

Here we show that climatic conditions and highly dynamic soil mechanical states are the primary constraints for global earthworm occurrence and activity. The seasonal and dynamic nature of soil moisture conditions in many regions defines temporal activity windows that support bioturbation and shape biogeographic patterns[11]. The objectives of this study were: (i) to model soil hydro-mechanical conditions and derive temporal windows of potential earthworm burrowing activity, and (ii) to delineate regions where earthworm activity would be mechanically prohibited (iii) to compare predicted regions with earthworm presence data at the global scale.

We present a mechanistic soil bioturbation model[11] with associated soil mechanical properties and general biophysical traits of earthworms. Soil and climatic information is used to predict the global distribution of habitats and associated temporal windows of bioturbation activity. Although soil moisture and soil type dominate earthworm burrowing potential, other factors such as temperature[21], soil pH[22] and high sand contents[23] were taken into account.

Earthworm bioturbation - *cavity expansion model and soil mechanical properties*

Contrary to popular view, the primary mechanism for soil bioturbation by burrowing earthworms relies on their ability to penetrate and deform the wet soil matrix using their flexible hydro-skeleton rather than ingesting POM-rich soil[13]. A recent biophysical model quantifies earthworm soil penetration and cavity expansion pressures[11]. The model defines the mechanical stress required for radial cavity expansion in an elasto-viscoplastic soil[11] that is linked with radial stresses σ_r induced by the earthworm hydro-skeleton at the cavity wall (Fig. 1). The minimal stress for cavity expansion in a soil is given as:

$$\sigma_r(R_p) = P_L - 2s_u ln\left(\frac{R_p}{r_c}\right) = s_u \quad (1)$$

where r_c is the radius of the cavity, P_L is the pressure at the cavity interface, R_p is the radius of the elasto-viscoplastic interface (far field), and s_u is the soil shear strength. Solving for the cavity expansion pressure yields the following limiting pressure for soil deformation:

$$P_L = s_u\left(1 + 2\ln\left(\frac{R_p}{r_c}\right)\right) = s_u\left(1 + \ln\left(\frac{G}{s_u}\right)\right) \quad (2)$$

where G is the shear modulus of rigidity. The ratio between the cavity zone and the viscoplastic zone converge to the ratio between the shear modulus and shear soil strength ($\left(\frac{R_p}{r_c}\right)^2 \rightarrow \left(\frac{G}{s_u}\right)$) as the initial cavity radius approaches zero (*e.g.* when initiating creation of a new burrow). Soil mechanical properties and soil moisture affect the model parameter values and thus the conditions that permit bioturbation by earthworms. We adopt a macroscopic rheological description of soil deformation[24,25] and use simplified power law relations for linking soil mechanical properties to soil texture and water content similar to the work of Gerard[26] (Supplementary Information, Extended Data Fig. 1 and 2). The resulting expressions describe the minimum pressure an earthworm must exert to radially expand a cavity in soil (Fig. 1). Observations suggest that the earthworm hydro-skeleton[27] can apply a maximum pressure of P_w = 200 kPa [28,29]. In other words, earthworm bioturbation becomes mechanically

impeded by soil mechanical conditions when $P_L(\theta, n) \geq P_w$, where θ is the soil water content, and n is the summed fraction of silt and clay.

Results

Predicted earthworm hospitable regions

We calculated mean annual cavity expansion limit pressures globally (0.1° x 0.1°, monthly for 1981-2019) using the ERA5-land soil moisture reanalysis and SoilGrids[30] topsoil textural information (Fig. 2 a). Different averaging methods were compared (Extended Data Fig. 3) and the harmonic average annual pressures are reported (Fig. 2 a). Geographical regions indicated in green are, on average, below the earthworm's biomechanical pressure limits. Independent data from a recent study[20] indicated less than 10% of observed earthworm abundance above a limiting pressure of 200 kPa (Extended Data Fig. 4). Additional factors that might exclude earthworm activity were considered to further constrain the predictions of potential earthworm habitats (Fig. 2 b). Regions with low mean annual temperature (MAT), i.e. MAT < 0 °C, are marked in blue, red regions indicate where the soil pH is below 4.5, and yellow regions where the soil sand content exceeds 80%. For regions with pronounced seasonality, earthworms have developed ecological strategies to cope with periods during which soil mechanical conditions impede bioturbation (*e.g.* extended period of dormancy[18,31]). Considering the minimal time window for a reproductive cycle and survival of newly hatched earthworms (total 4-6 weeks)[31], we required two consecutive months of favorable, soil mechanical conditions for permissible habitation. This would ensure at least one reproductive cycle per year[31]. Regions with shorter time windows are shown in cyan (Fig. 2 b). Distributions of additional factors were compared to sites with earthworm occurrence from a recent study[14] (Extended Data Fig. 5). Comparing reported soil pH with values obtained from digital soil maps (SoilGrids[30]) revealed a narrowed range of values than

observed at the sample scale. Most occurrences of earthworms were reported for soil pH above 3.5 that mapped to SoilGrids[30] pH values above 4.5 (used for spatial mapping). These locations also received more than the previously reported[15] minimum mean annual precipitation (MAP) of 400 mm yr^{-1}.

Modeled and observed earthworm global distributions

Detailed comparison of regions with ample observations were used for model evaluation. For example, earthworm spatial distributions for Australia and North America are depicted in Fig. 3 a and b, respectively[32]. The large extent of arid regions in Australia limits earthworm activity to the coasts that receive sufficient rainfall to moisten the soil. This is in good agreement with model predictions as shown with the 400 mm yr^{-1} contour of MAP[15] (Fig. 3 a). For North America, the model predicts that earthworm activity is possible from the east coast to the Midwest followed by a sharp decrease in occurrence until the west coast (Fig. 3 b). These trends are similar to previously estimated earthworm distributions[16] with a sharp cutoff near arid regions. Around half of the terrestrial surface (>-60°N) permits earthworm activity but most observations of earthworm presence originate from Europe (Fig. 3 c). Reported earthworm presence agreed with model classification for 86% of the geographical occurrences (global within 0.1°x0.1°, n = 7346). Although there were 13% of false negatives, these were often associated with local geographical features (e.g. river banks, anomalous precipitation zones, etc.) as depicted in Fig 3. To test the robustness of classification and its sensitivity (hit-rate) we performed random re-sampling of occurrences with replacement (Extended Data Fig. 6).

Earthworm seasonal activity windows

The global map of average conditions conducive to earthworm burrowing activity conceals the nuanced dynamics associated with seasonal activity windows that are driven primarily by precipitation. To provide a succinct picture of this ingredient, temporal activity windows (seasonality or wet periods) for earthworms are illustrated in Fig. 4. The temporal variability of limiting soil pressures is described spatially by the coefficient of variation and highlights regions in which the impact of seasonality on earthworm activity is most pronounced (Fig 4 a). Fig. 4 b presents the median limiting pressure across latitudes for a climatic year to highlight the dynamic nature of soil conditions that constrain seasonal earthworm activity and delineates regions where soil conditions prohibit earthworm activity year-round (*i.e.*, arid regions). The required minimal cavity expansion pressures are compared for two contrasting biomes where MAT, sand content, and pH, were not limiting. A grassland located at 9.55^{o}N, 14.65^{o}E and a desert located at -22.95^{o}N, 132.95^{o}E are indicated in Fig. 4 a. Results suggest that soil moisture content mediated by precipitation facilitates mechanical activity for as much as 4.5 consecutive months in the grassland (Fig. 4 c) while the infrequent precipitation in the desert (Fig. 4 d) resulted in no appreciable temporal activity window for bioturbation or reproduction. Lastly, we compared species richness reported in Phillips et al.[14] to the fragmentation of habitats across latitudes (Fig. 4 e). Latitudinal habitat fragmentation was measured by counting the number of land fragments that are broken up by inhospitable zones and water bodies within a 0.1° wide strip around the globe. Results suggest higher species richness with increased number of fragmented habitats at the spatial resolution of ~10 km.

Discussion

A novel biomechanical model for earthworm bioturbation in combination with climatic and soil conditions enabled mapping of global habitat suitability (Fig. 2) and comparison with earthworm distributions (Fig. 3). Favorable soil moisture and mechanical conditions dominate the global distribution of earthworms. Additional constraints such as permafrost soil and subzero MAT[21] preclude earthworm activity in large parts of the world. Despite evidence for soil acidity limitations (soil pH < 4.5)[22], the global distribution of earthworm was not overly sensitive to low values of soil pH[16]. The primary mechanism[14] that shapes earthworm occurrence appears to be driven by soil physical (hydro-mechanical) conditions; determined by soil moisture and earthworm physiological limitations in unfrozen soils. The distributions of environmental conditions associated with earthworm occurrence compare favorably with the range of values reported in a recent global study[14] (Extended Data Fig. 5). The modeled soil limit-pressures appeared to also correspond strongly with observed earthworm abundance using independent data (Extended Data Fig. 4). However, modeled trends at ~10 km resolution preclude representation of many small-scale niches. For example, river corridors that cut across arid regions in the US Midwest reported presence of earthworms not represented by the model. Other examples were found along rivers in South-East Australia and Eurasia. Similarly, inhospitable regions with low soil pH may not be properly captured by the smoothed estimates of digital soil maps[30] as evident when comparing with values reported for soil samples (Extended Data Fig. 5 a and b). We note that many biological and chemical soil properties are also related to climatic hydration conditions[31,34] and our results represent average climatic tendencies manifested across biomes and spatial scales (~10 km resolution). Such global estimates might average out locally limiting factors (soil moisture, soil compaction, temperature and soil pH), thus contributing to model predicted false negatives. Furthermore, our estimation for maximal earthworm hydro-

skeletal pressures are based on earthworms residing in temperate regions[28]. Large earthworms found in the tropics or in Australia may exert greater pressures and could thus be less limited. However, this could be readily accounted for in future studies given more refined physiological information. Moreover, it remains challenging to address potential observational bias in the spatial patterns of reported earthworm occurrences. Most occurrences are reported for few countries in Europe (United Kingdom, Germany) resulting in strong spatial clustering of presence data that hampers the assessment of model sensitivity (hit-rate). By considering the observation density and performing weighted, random re-sampling we observe a minor reduction in hit-rate (from 86% to 84%) and find that average estimates are robust against variations in sample size (Extended Data Fig. 6). While this may not fully resolve the issue of observational bias, we can analyze possible tendencies of reduced sensitivity. Overall, the lowest hit-rate is still well above 50%, which would be expected by a coin toss and, coincidentally, by the fraction of terrestrial area that is predicted to be hospitable to earthworms.

In addition, the seasonality of limiting soil pressures defines temporal windows of earthworm activity and selects for particular ecological life strategies. Model predicted activity windows (Fig. 4) correspond closely to previously reported seasonal variations in earthworm communities[17,18]. This suggests that their ecological strategies (i.e. dormancy cycles, reproduction cycles, etc.) are mediated by soil hydro-mechanical factors. While the shortest possible temporal window that supports thriving earthworm communities is unknown, a sufficiently long window is required for earthworm annual reproduction[18]. Earthworms may live several years, but the fertilization and egg incubation take 3-4 weeks[18,31]. Additionally, young earthworms need a few weeks to build up biomass to survive dormancy[18,31]. We could assume 1-2 months of favorable conditions to be the minimum requirement for survival and reproduction[31]. Narrow windows would also limit earthworms' accessibility to plant-derived

POM, which could further preclude their activity in deserts with low net primary productivity (Fig. 4 c and d). Strong seasonal variation poses further constraints on earthworm activities linked to the variability of limit pressure (Fig. 4 a). Although we present harmonic averaging that provides more inclusive bounds for earthworm habitats in regions with strong seasonal variation (e.g. Spain, Fig. 3; for comparison of averaging methods see Extended Data Fig. 3), the mechanistic model allows for quantification of the seasonal variability in earthworm habitats (Fig. 4 a). Despite few regions of high volatility, climatic predictions are robust for most regions. For example, permissive regions of earthworm activity in Asian islands such as the Philippines[35] are predicted.

Furthermore, our results quantify the dynamics of latitudinal patterns (Fig. 4 b). While there are particular regions that remain stable (i.e. favorable or uninhabitable), there are several latitudes that exhibit strong fluctuations. One of the more striking features is observed between 20^{o}N and 30^{o}N. These zones are characterized by particularly harsh conditions. Interestingly, the highest number of earthworm species was reported for this range[14]. Compatibility between the two results would suggest that species richness is high under environmentally harsh conditions (Fig. 4 e). However, taking the latitudinal median might miss small regions that permit earthworm burrowing activity. The limited spatial extent of such “patches” would not allow for widespread migration and favor endemic (isolated) populations; resulting in high species richness over climatic timescales. Nonetheless, this is not to suggest that the short-term, anthropogenic fragmentation of earthworm habitats would promote species diversity.

The study provides a framework for prognosis of potential migration trends, climatic barriers, and the promotion of sustainable land use. Regions of North America with limited earthworm activity are predicted by our model in agreement with previously reported earthworm distributions (Fig. 3). Isolation of earthworm communities in North America could be

attributed to drier regions central-westward that act as geographic barriers. These regions obstruct earthworm migration and could explain why few native earthworm species returned to North America post glaciation[14].

The growing threat of soil compaction associated with increased land use intensification[36] is motivating a large push towards no-tillage practices[9,36]. Regions that indicate soil bioturbation potential by earthworms may be used to further prompt more sustainable agricultural practices[37], which would reduce the frequency and intensity of tillage machinery while maintaining soil structure suitable for crop growth[36]. The modeled regions of bioturbation potential are based on first principals that are independent of earthworm occurrence or abundance data and can serve as a reference for evaluating agricultural practices across biomes.

The modelling framework (Fig. 3) could be readily incorporated in climate models with minor computational costs to represent dynamics of global earthworm habitats and activity windows[38]. Unlike a static picture of global distributions[14,39], the model could be used to assess future trends in regions viable for agriculture and land use management (tillage vs. no-tillage) with respect to earthworm contribution to soil structure. Predictions of earthworm activity and migration patterns could be linked to future expansion of wetter (or drier) regions. Although the focus has been on hospitable regions for earthworm activity, soil water contents associated with limiting earthworm pressures have been shown to affect plant root growth for many soil types. Bengough *et al.*[40] reported that this lower bound in soil moisture provides favorable mechanical conditions and water availability for plant roots. This becomes evident when considering global gross primary production (GPP), which highlights very similar spatial patterns[41] compared to predicted earthworm habitats. Furthermore, plant roots could benefit from a mutualistic interactions with earthworms[5], thus finding benefits from regions where earthworms thrive and vice versa.

Although comparisons made in this study inspire confidence in our model, refinements would be needed to better predict bioturbation and foraging activity. We envision, development of population densities based on energetic considerations that include soil carbon input fluxes[34] (e.g. GPP). Reported earthworm populations range between 60 and 350 individuals per m^2 of soil surface[42] and it is likely that resource availability (i.e. soil organic carbon (SOC) or POM) could limit earthworm abundance in particular regions. Considering such factors in a mechanistic modeling framework would help disentangle the various effects of organic matter accumulation on soil mechanical properties (bulk density), soil water characteristics (water retention) and physiological (energetic) constraints. Such refinements would enable the model to generate estimates regarding earthworm abundance, which is beyond the scope of the current study.

Insights into the fundamental principles that shape earthworm ecological trends as reported in previous studies[15,16] [14] place such empirical observations on a mechanistic basis. This deepens our understanding of the processes relevant to predators, soil flora and microbes that interact with earthworms, and the general ecosystem services that earthworms provide [5]; all of which are built on the foundations of soil hydro-mechanical status.

Methods

Earthworm limiting pressure and activity windows

Using global soil moisture data combined with the critical soil hydro-mechanical states that limit earthworm burrowing, we determined climatic regions that could support potential earthworm bioturbation activity. Regions with high likelihood of permafrost are removed prior to calculations (with permafrost zonation index[43] exceeding 0.1). For each geographic location we then evaluate the parametrized model using soil textural information from SoilGrids digital soil maps[30] and monthly averaged soil moisture estimates from ERA5-land (https://doi.org/10.24381/cds.68d2bb30). All global raster data was harmonized to a common grid of 0.1° resolution (~11 km) using nearest neighbor interpolation of the upper most soil depth layer (0-5cm and 0-7 cm for SoilGrids and ERA-5 land, respectively). The limiting pressure (equation (2)) was calculated for the entire record of the ERA5-land dataset that ranges from 1981 to 2019 at a monthly resolution. Based on the limiting pressure time series, we estimate the number of consecutive months below 200 kPa and the ensemble average pressure for every grid cell. A comparison of averaging methods is reported in the Supplementary Information and we reported harmonic averages throughout the main text. Two specific regions were selected to illustrate temporal activity windows: a grassland located at 9.55^{o}N, 14.65^{o}E and a desert located at -22.95^{o}N, 132.95^{o}E. We aggregated the limiting pressure time series to climatic monthly values and compared with daily climatic precipitation estimates obtained from MSWEP[33]. Daily precipitation estimates were smoothened using a 30-day rolling average for comparison with monthly pressure values and to delineate time windows of earthworm burrowing activity.

Additional factors that impede earthworm activity

Climatic factors and soil properties were used to illustrate additional factors that could impede bioturbation activity by defining thresholds for earthworms' tolerance. Regions

where the mean annual temperatures (MAT) were below zero were considered zones of impedance. Besides the soil mechanical impedance becoming augmented in a manner not currently considered in our model, these low temperatures will decelerate earthworms' metabolic cycles to critical states[21], which may ultimately lead to earthworms freezing. Besides soil temperature, low soil pH is often cited as being critical for earthworm habitat suitability[14]. We outline global regions where soil pH is below 4.5[22,31]. Regions where sand content exceeded 80% were considered as regions of impedance. Although there are sandy soils where earthworms have been observed (e.g. sand dunes in the UK[44]), the abrasive nature of sand grains is typically obstructive[45]. We note that SOC and POM would also play a role in limiting earthworm abundance. However, as they are likely to co-occur in hydro-mechanically hospitable conditions, we focus our study on physical and chemical factors impeding potential earthworm activity.

Earthworm occurrence data

We compared our theoretically determined regions with previously published empirical maps that outline earthworm distributions for Australia[15] and North America[16] and with presence-only data of ten earthworm species (*Almidae, Eudrilidae, Glossoscolecidae, Hormogastridae, Lumbricidae, Microchaetidae, Moniligastridae, Ocnerodrilidae, Octochaetidae, Sparganophilidae*) as deposited in the Global Biodiversity Information Facility (GBIF) database (https://doi.org/10.15468/dl.xstqow, https://doi.org/10.15468/dl.wghggg, https://doi.org/10.15468/dl.3yj8pk, https://doi.org/10.15468/dl.lzuwlg, https://doi.org/10.15468/dl.vwqtsk, https://doi.org/10.15468/dl.brqmht, https://doi.org/10.15468/dl.ghccto, https://doi.org/10.15468/dl.dk97gk, https://doi.org/10.15468/dl.xjw6kc, https://doi.org/10.15468/dl.9a4ojx). The distribution of each species occurrence is shown in Extended Data Fig. 7.

Data availability

All data used in this study is available from public sources. Data underlying maps of potential earthworm habitats will be deposited in a public repository upon publication (meanwhile it is available from the corresponding author upon request).

References

1 Young, I. M. *et al.* The interaction of soil biota and soil structure under global change. *Global Change Biology* **4**, 703-712 (1998).

2 Lavelle, P. *et al.* Earthworms as key actors in self-organized soil systems. *Theoretical Ecology Series* **4**, 77-106 (2007).

3 Blakemore, R. & Hochkirch, A. Soil: Restore earthworms to rebuild topsoil. *Nature* **545**, 30-30 (2017).

4 Kuzyakov, Y. & Blagodatskaya, E. Microbial hotspots and hot moments in soil: Concept & review. *Soil Biology and Biochemistry* **83**, 184-199 (2015).

5 Brown, G. G., Barois, I. & Lavelle, P. Regulation of soil organic matter dynamics and microbial activityin the drilosphere and the role of interactionswith other edaphic functional domains. *European Journal of Soil Biology* **36**, 177-198 (2000).

6 Denef, K. *et al.* Influence of dry–wet cycles on the interrelationship between aggregate, particulate organic matter, and microbial community dynamics. *Soil Biology and Biochemistry* **33**, 1599-1611 (2001).

7 Van Groenigen, J. W. *et al.* Earthworms increase plant production: a meta-analysis. *Scientific reports* **4** (2014).

8 Blouin, M. *et al.* A review of earthworm impact on soil function and ecosystem services. *European Journal of Soil Science* **64**, 161-182 (2013).

9 Capowiez, Y. *et al.* Experimental evidence for the role of earthworms in compacted soil regeneration based on field observations and results from a semi-field experiment. *Soil Biology and Biochemistry* **41**, 711-717 (2009).

10 Wu, X. D., Guo, J. L., Han, M. & Chen, G. An overview of arable land use for the world economy: From source to sink via the global supply chain. *Land use policy* **76**, 201-214 (2018).

11 Ruiz, S., Schymanski, S. & Or, D. Mechanics and Energetics of Soil Penetration by Earthworms and Plant Roots - Higher Burrowing Rates Cost More. *Vadose Zone Journal* **16**, doi:10.2136/vzj2017.01.0021 (2017).

12 Quillin, K. J. Kinematic scaling of locomotion by hydrostatic animals: ontogeny of peristaltic crawling by the earthworm Lumbricus terrestris. *Journal of Experimental Biology* **202**, 661-674 (1999).

13 Ruiz, S., Or, D. & Schymanski, S. Soil Penetration by Earthworms and Plant Roots—Mechanical Energetics of Bioturbation of Compacted Soils. *PLOS ONE* **10.1371/journal.pone.0128914** (2015).

14 Phillips, H. R. *et al.* Global distribution of earthworm diversity. *Science* **366**, 480-485 (2019).

15 Abbott, I. Distribution of the native earthworm fauna of Australia-a continent-wide perspective. *Soil Research* **32**, 117-126 (1994).

16 Hendrix, P. F. & Bohlen, P. J. Exotic earthworm invasions in North America: ecological and policy implications: expanding global commerce may be increasing the likelihood of exotic earthworm invasions, which could have negative implications for soil processes, other animal and plant species, and importation of certain pathogens. *Bioscience* **52**, 801-811 (2002).

17 Nakamura, Y. Studies on the Ecology of Terrestrial Oligochaeta: I. Sesonal Variation in the Population Density of Earthworms in Alluvial Soil Grassland in Sapporo, Hokkaido. *Applied Entomology and Zoology* **3**, 89-95 (1968).

18 Edwards, C. A. & Bohlen, P. J. *Biology and ecology of earthworms*. Vol. 3 (Springer Science & Business Media, 1996).

19 Kretzschmar, A. Burrowing ability of the earthworm Aporrectodea longa limited by soil compaction and water potential. *Biology and Fertility of Soils* **11**, 48-51 (1991).

20 Johnston, A. S. Land management modulates the environmental controls on global earthworm communities. *Global Ecology and Biogeography* **28**, 1787-1795 (2019).

21 Rao, K. P. Physiology of low temperature acclimation in tropical poikilotherms. I. Ionic changes in the blood of the freshwater mussel, Lamellidens marginalis, and the earthworm, Lampito mauritii. *Proceedings of the Indian Academy of Sciences-Section B* **57**, 290-295 (1963).

22 Baker, G. H. & Whitby, W. A. Soil pH preferences and the influences of soil type and temperature on the survival and growth of Aporrectodea longa (Lumbricidae): The 7th international symposium on earthworm ecology· Cardiff· Wales· 2002. *Pedobiologia* **47**, 745-753 (2003).

23 El-Duweini, A. K. & Ghabbour, S. I. Population density and biomass of earthworms in different types of Egyptian soils. *Journal of Applied Ecology*, 271-287 (1965).

24 Ghezzehei, T. A. & Or, D. Rheological properties of wet soils and clays under steady and oscillatory stresses. *Soil Science Society of America Journal* **65**, 624-637 (2001).

25 Ghezzehei, T. A. & Or, D. Dynamics of soil aggregate coalescence governed by capillary and rheological processes. *Water Resources Research* **36**, 367-379 (2000).

26 Gerard, C. The Influence of Soil Moisture, Soil Texture, Drying Conditions, and Exchangeable Cations on Soil Strength. *Soil Science Society of America Journal* **29**, 641-645 (1965).

27 Quillin, K. J. Ontogenetic scaling of burrowing forces in the earthworm Lumbricus terrestris. *Journal of Experimental Biology* **203**, 2757-2770 (2000).

28 Ruiz, S. A. & Or, D. Biomechanical limits to soil penetration by earthworms: direct measurements of hydroskeletal pressures and peristaltic motions. *Journal of The Royal Society Interface* **15**, 20180127 (2018).

29 McKenzie, B. M. & Dexter, A. R. Radial pressures generated by the earthworm Aporrectodea rosea. *Biology and Fertility of Soils* **5**, 328-332 (1988).

30 Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one* **12**, e0169748 (2017).

31 Burges, A. *Soil biology*. (Elsevier, 2012).

32 Ruiz, S. A. Mechanics and Energetics of Soil Bioturbation by Earthworms and Growing Plant Roots. doi:10.3929/ethz-b-000280625 (2018).

33 Beck, H. E. *et al.* MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment. *Bulletin of the American Meteorological Society* **100**, 473-500 (2019).

34 Beer, C., Reichstein, M., Ciais, P., Farquhar, G. & Papale, D. Mean annual GPP of Europe derived from its water balance. *Geophysical Research Letters* **34** (2007).

35 Heaney, L. R., Balete, D. S., Rickart, E. A. & Niedzielski, A. *The mammals of Luzon Island: biogeography and natural history of a Philippine fauna*. (Johns Hopkins University Press, 2016).

36 Keller, T. *et al.* Long-Term Soil Structure Observatory for Monitoring Post-Compaction Evolution of Soil Structure. *Vadose Zone Journal* **16** (2017).

37 Lacoste, M., Ruiz, S. & Or, D. Listening to earthworms burrowing and roots growing-acoustic signatures of soil biological activity. *Scientific reports* **8**, 10236 (2018).

38 Change, I. C. the Physical Science Basis: Working Group Ⅰ Contribution to the Fifth Assessment Report of the Intergovernment Panel on Climate Change. (2013).

39 Van Den Hoogen, J. *et al.* Soil nematode abundance and functional group composition at a global scale. *Nature* **572**, 194-198 (2019).

40 Bengough, A. G. *et al.* Root responses to soil physical conditions; growth dynamics from field to cell. *Journal of Experimental Botany* **57**, 437-447 (2005).

41 Beer, C. *et al.* Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science* **329**, 834-838 (2010).

42 Paoletti, M. G. The role of earthworms for assessment of sustainability and as bioindicators. *Agriculture, Ecosystems & Environment* **74**, 137-155 (1999).

43 Gruber, S. Derivation and analysis of a high-resolution estimate of global permafrost zonation. *The Cryosphere* **6**, 221 (2012).
44 Chamberlain, E. J. & Butt, K. R. Distribution of earthworms and influence of soil properties across a successional sand dune ecosystem in NW England. *european journal of soil biology* **44**, 554-558 (2008).
45 Booth, L. H., Heppelthwaite, V. & McGlinchy, A. The effect of environmental parameters on growth, cholinesterase activity and glutathione S-transferase activity in the earthworm (Apporectodea caliginosa). *Biomarkers* **5**, 46-55 (2000).
46 Lu, N. & Kaya, M. Power law for elastic moduli of unsaturated soil. *Journal of Geotechnical and Geoenvironmental Engineering* **140**, 46-56 (2014).
47 Fan, L., Lehmann, P. & Or, D. Load redistribution rules for progressive failure in shallow landslides: Threshold mechanical models. *Geophysical Research Letters* **44**, 228-235 (2017).
48 Alramahi, B., Alshibli, K. A. & Fratta, D. Effect of fine particle migration on the small-strain stiffness of unsaturated soils. *Journal of geotechnical and geoenvironmental engineering* **136**, 620-628 (2010).

Acknowledgements

This research was carried out at ETH Zürich and the University of Southampton. Authors would like to acknowledge the help from Dr. Peter Lehmann for preliminary soil moisture maps, which were crucial to motivating this study. Authors acknowledge helpful discussions with Prof. Ning Lu regarding soil mechanical properties and thank Dr. Katherine Williams for proof reading the document.

Competing interests

Authors declare no competing interest.

Figures:

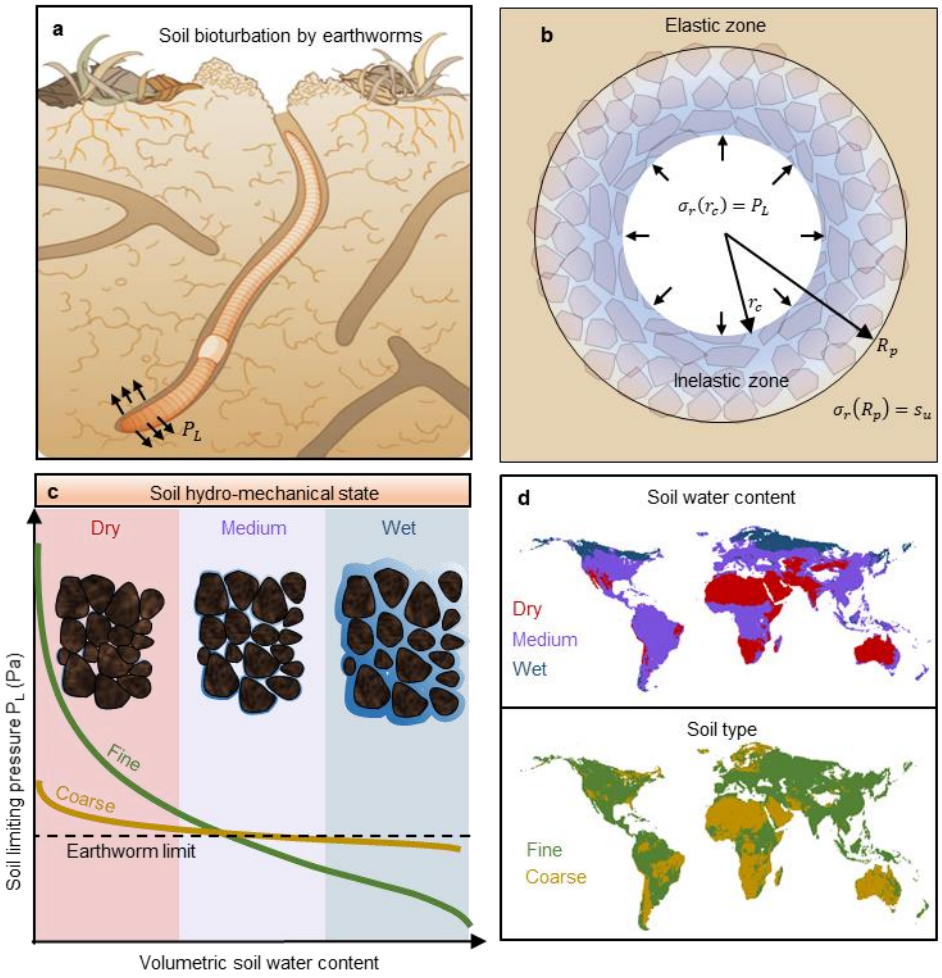
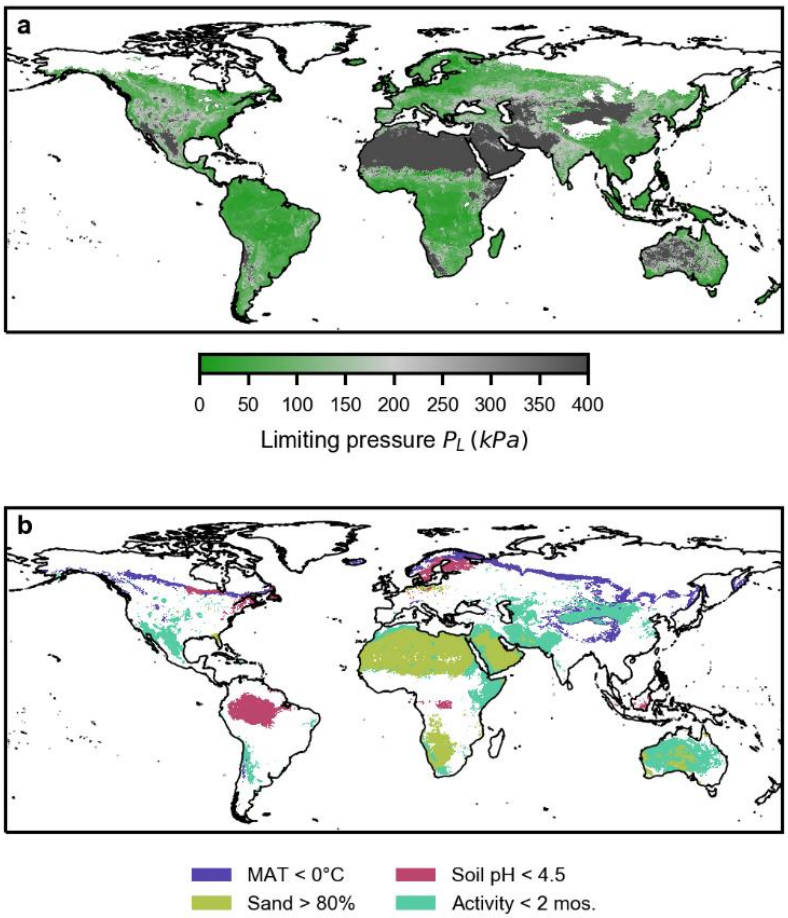


Fig. 1: Earthworm bioturbation activity in structured soil. a Subterranean bioturbation relies on earthworms' ability to mechanically penetrate and deform the soil using their flexible hydroskeleton, which is **b** modeled by means of penetration and cavity expansion transverse to the earthworm body where radial stresses σ_r exerted by the earthworm form the local cavity of size r_c. Yielding soil material is bounded by a remote elastic zone at a distance R_P from the center of the cavity is dependent on **c** soil hydro-mechanical conditions that enable their hydro-skeleton to form cavities. **d** Hydro-mechanical soil states can be mapped globally depending on soil texture, enabling inferences to earthworm distributions.



473

474 **Fig. 2: Global map of earthworm hospitable zones. a**, Green regions indicate that annual
475 average pressures required for cavity expansion are below the earthworm's hydrostatic
476 pressure limit (200 kPa). Pressures are truncated to values below 400 kPa for visualization
477 (dark grey) and permafrost regions were removed (white). **b**, Other factors that may impede
478 earthworm activity. Blue regions indicate subzero mean annual temperature (MAT), red
479 regions mark soil pH<4.5, green regions indicate coarse soil texture (sand content > 80%),
480 and cyan regions indicate that there are fewer than two consecutive months during which the
481 soil mechanical properties permit cavity expansion. Regions of different limiting factors may
482 overlap and were ordered for visibility.

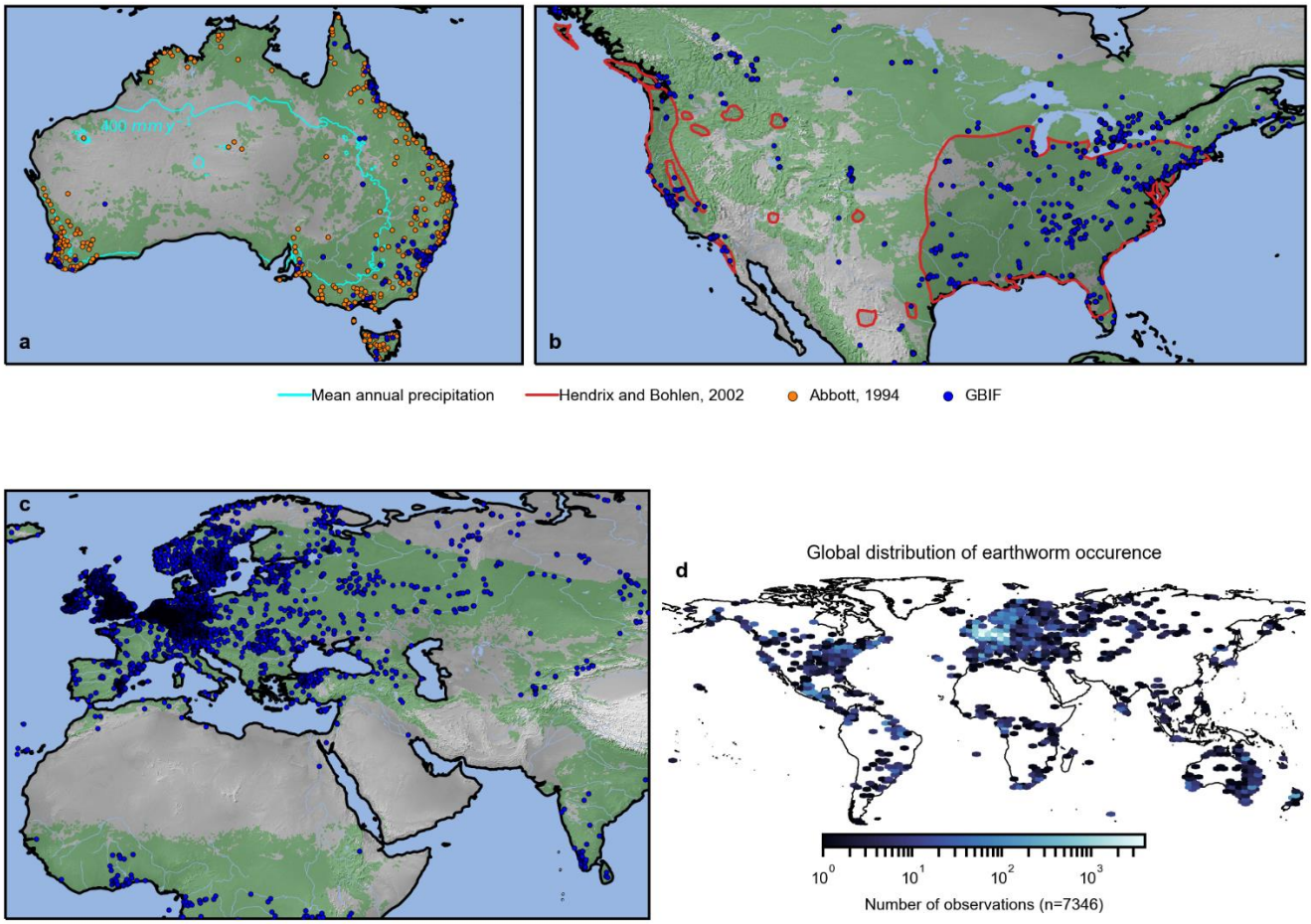


Fig. 3: Comparison of predicted hospitable zones and reported earthworm distribution. **a**, Potential earthworm habitats (green) including soil hydro-mechanical limitations for Australia. Locations with reported presence of earthworms from two datasets are displayed; GBIF (blue points) and Abbott[15] (orange points). Regional limitation of earthworm activity is delineated by 400 mm yr^{-1} of mean annual precipitation[33] (cyan contour) as previously reported[15]. **b**, Predicted earthworm habitats for North America. Observed occurrences (Global Biodiversity Information Facility, GBIF) are in good agreement with regional extents of earthworm communities (redrawn from Hendrix and Bohlen[16], red). **c**, Regions in East Eurasia and Northern Africa that could support earthworm soil bioturbation. **d,** Global distribution of earthworm occurrence.

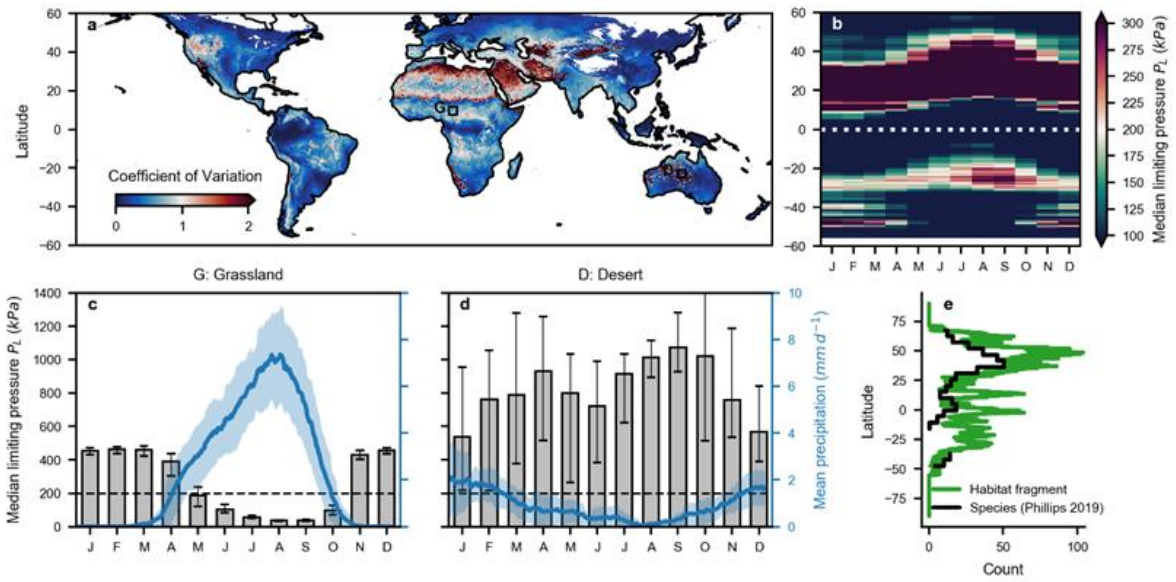


Fig. 4: Temporal windows of potential earthworm burrowing activity. a, Global map of temporal hydro-mechanical variations (coefficient of variation of limiting pressures). **b,** Median earthworm limit pressures across latitudes for a climatic year. **c-d,** Median climatic limiting pressures (bars ± IQR) required to burrow through soil are associated with mean daily precipitation[33] (blue line and shading; 30 day running mean and SD) for **c,** a grassland (**G**: 9.55°N, 14.65°E) and **d**, a desert (**D**: -22.95°N, 132.95°E) as indicated in **a**. **e.** Habitat fragmentation based on habitable regions is plotted in comparison with species richness[14] results for different latitudes. The maximum radial earthworm pressures P_w (dashed line) are shown. Soil limit pressures are reported for the topsoil (0-7 cm) and are assumed to represent the driest part of the soil profile.

Supplementary Information

In this supplement, we provide additional information concerning the description of the biophysical, cavity-expansion model (SI.1), the functional dependency of hydro-mechanical properties with their parametrization (SI.2) and the assessment of averaging methodology for summarizing earthworm limiting pressures (SI.3).

SI.1 – Cavity expansion mechanical model – an overview

The mechanics of soil bioturbation by burrowing earthworms relies on their ability to deform the soil. The biophysical model considers soil penetration-cavity expansion sequences by the earthworm similar to cone penetration [11]. The model provides the minimal mechanical stress required to radially expand a cavity in an elasto-viscoplastic soil [11]. To quantify the magnitude of radial pressure required by an earthworm to expand in wet elasto-viscoplastic soils, we first consider the force balance at equilibrium:

$$\frac{\partial \sigma_r}{\partial r} + \frac{\sigma_r - \sigma_\theta}{r} = 0 \tag{S1}$$

where r [m] is the distance from the center of the cavity, σ_r [Pa] is the radial stress and σ_θ [Pa] is the hoop (circumferential) stress. The deformation behavior is expressed by the Von-Mises criterion considering viscous deformation (i.e. Bingham model [24]), relating the difference between the radial and hoop stresses to the summation of the undrained soil strength and the viscoplastic strain rate:

$$\sigma_r - \sigma_\theta = 2s_u + \frac{4}{3}\eta\dot{\epsilon}_r \tag{S2}$$

where η [Pa s] is the soil plastic viscosity, s_u [Pa] is the undrained soil strength, and $\dot{\epsilon}_r$ [m m^{-1}s^{-1}] is the radial strain rate.. Substitution of Eq. (S2) into (S1) yields the following expression:

$$\frac{\partial \sigma_r}{\partial r} = -\frac{2s_u}{r} - \frac{4}{3}\eta\frac{\dot{\epsilon}_r}{r} \tag{S3}$$

By integration, we determine the radial stresses as a function of the radius (and the strain rate):

$$\sigma_r(t,r) = P_L - 2s_u ln(\frac{r}{r_c}) - \frac{4}{3}\eta\int \frac{\dot{\epsilon}_r}{r}dr \tag{S4}$$

where r_c [m] is the minimum cavity size and P_L [Pa] is the time independent limit pressure to which the static cavity pressure converges. Under static conditions, the strain rate term in the integral

vanishes. We solve for the limit pressure by equating the change in the cavity zone to the change in the plastic region local to the cavity:

$$\left(\frac{R_p}{r_c}\right)^2 \rightarrow \frac{G}{s_u} \tag{S5}$$

Where G [Pa] is the soils shear modulus, and R_p is the elasto-plastic interfacial radius. Under static conditions, the radial stress by the earthworm at the cavity wall is expressed as:

$$\sigma_r(R_p) = P_L - 2s_u ln\left(\frac{R_p}{r_c}\right) = s_u \tag{S6}$$

Leading to the minimum radial pressure required to expand a cavity in soil:

$$P_L = s_u\left(1 + 2\ln\left(\frac{R_p}{r_c}\right)\right) = s_u\left(1 + \ln\left(\frac{G}{s_u}\right)\right) \tag{S7}$$

The resulting expression would be the minimum amount of pressure an earthworm would have to exert with its hydroskeleton in order to expand a cavity radially in soil. However, earthworms hydroskeleton is made up of soft flexible muscle fibers [27] that are mechanically limited to a maximum pressure of P_w = 200 kPa [28,29]. Thus, earthworms are mechanically impeded by soil conditions when $P_L \geq P_w$. These constraining soil mechanical conditions are linked to soil's hydration status and soil texture[45].

SI.2 – Functional relationship between soil hydration status, textural class and mechanical properties

Soil mechanical properties are linked to multiscale physical phenomena, which are sensitive to soil textural class and soil moisture content. At the submicron scale, soil clay particles are tightly bound by electrical forces, and their ability to yield depends on their alignment, liquid lubrication, and platelet spacing [24]. The soil moisture plays a prominent role in binding together soil aggregates via capillarity under drier condition [24, 25] or reduce soil friction under wetter conditions. Ultimately, these forces acting on different scales jointly increase soils shear strength and shear modulus of rigidity under drier conditions [11,26]. While these different processes warrant more rigorous analysis, these details extend beyond the scope of our current study.

Instead, we adopt simplified power law relations for linking soil mechanical properties to soil texture and water content similar to the work of Gerard et al. [26]. We collected experimentally determined values for soil shear strength and shear modulus of rigidity and interpolated their behavior for a range of soil textures and soil water contents. Data were collected and consolidated from Gerard et al. 1965[26], Lu and Kaya 2013 [46], Fan et al. 2017[47], Alramahi et al. 2010 [48] , Ruiz 2017 [11], and Ghezzehei and Or 2001[24]. Soil shear strength and shear modulus of rigidity are related to soil water contents via a power law [26]:

$$s_u = a_y(f)\theta_v^{-b_y(f)} \quad \text{(S8)}$$

$$G = a_G(f)\theta_v^{-b_G(f)} \quad \text{(S9)}$$

where θ_v [% m^3 m^{-3}] is the soil water content, the pre-factors $a_{y,G}(f)$ [Pa] and the exponent $b_{y,G}(f)$ [-] are functions of the soil fine fraction f [% Silt+Clay]. Coefficients take the functional forms:

$$a_{y,G}(f) = \alpha_{y,G}\exp(\beta_{y,G}f) \quad \text{(S10)}$$

$$b_{y,G}(f) = \zeta_{y,G}f + \xi_{y,G} \quad \text{(S11)}$$

where α, β, ζ, and ξ are fitting coefficients that relate the mechanical properties to the soil texture. The relations of coefficient to soil fine fraction. Lastly, the focus lies on the hydro-mechanical properties of relatively fine textured soils, as coarse soils have been reported too abrasive and are

561 often too dry for earthworm activity due to their frictional nature and low water retention properties
562 respectively[45].

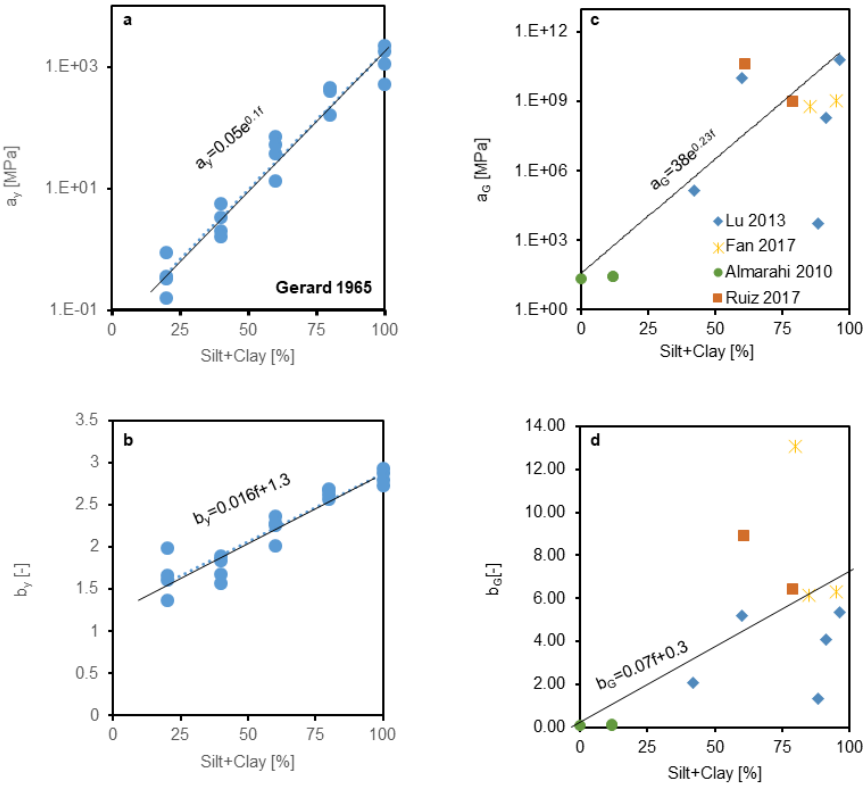
563 **Linking soil mechanical properties to soil texture and hydration state**

564 Coefficients and exponential pre-factors as related to soil fine texture content was plotted in Extended
565 Data Fig. 1. A comprehensive span of soil shear strengths were taken from Gerard et al. 1965 [26] and
566 were parametrized using equations (S10) and (S11) (Extended Data Fig. 1 a and b). Equations (S10)
567 and (S11) were also fit to soil shear modulus data (Extended Data Fig. 1 c and d). These soil
568 mechanical relations allowed us to determine minimal cavity expansion pressures via equation (S7) as
569 a function of soil moisture contents and fine texture percentages (Extended Data Fig. 2). The red
570 contour path in Extended Data Fig. 2 highlights soil conditions that inhibit earthworm mechanical
571 activity. This functional relation is used to relate earthworm limiting pressures to soil moisture and
572 soil texture.

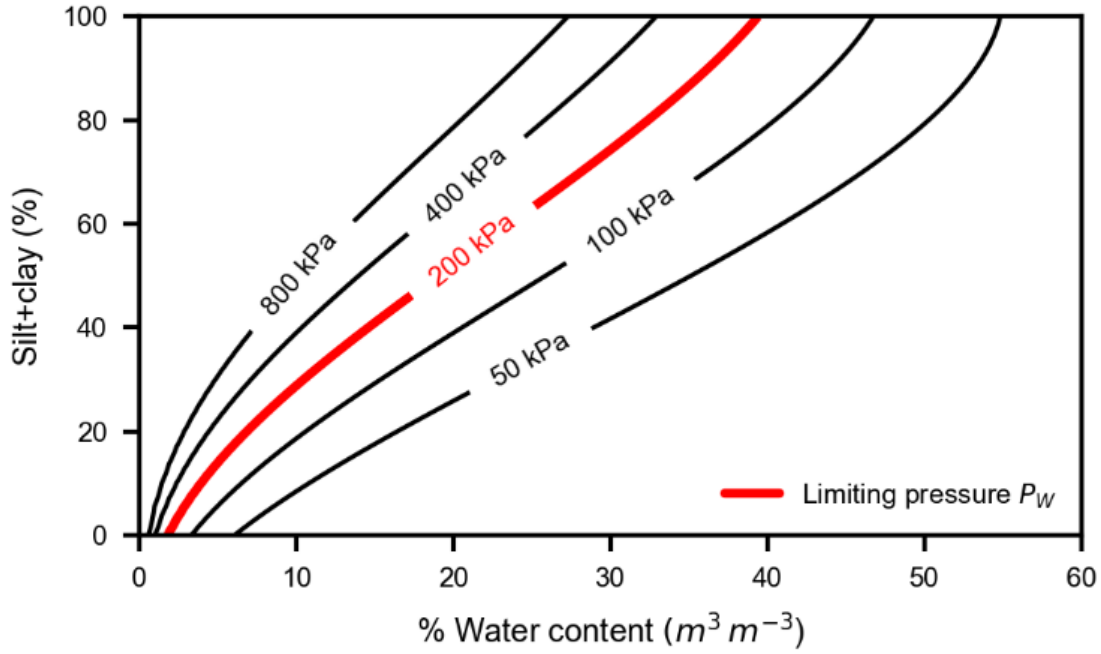
SI.3 Assessing averaging methodologies for limiting pressures

Soil moisture status is highly dynamic and limiting pressures respond non-linearly, thus it is not clear as to what averaging methods provide the most representative estimates for promoting potential earthworm habitats. We compared the model results considering arithmetic averaging, harmonic averaging, and median values of the global limit pressures that would support earthworm activity (Extended Data Fig. 3). To systematically compare the effect of averaging method on predicted regions below the earthworm limiting pressure (P_w = 200 kPa) we overlay masks for each method (Extended Data Fig. 3 a-c) and count the number of times predictions agree (Extended Data Fig. 3 d). Most regions are considered permissible to earthworm bioturbation by all three averaging methods. The arithmetic average results in more restricted regions, while the harmonic mean classifies a larger proportion of the terrestrial surface as suitable habitats based on limit pressure. The difference between averaging methods are most pronounced in regions with larger monthly variability in soil moisture (e.g. Mediterranean, India, Sahel).

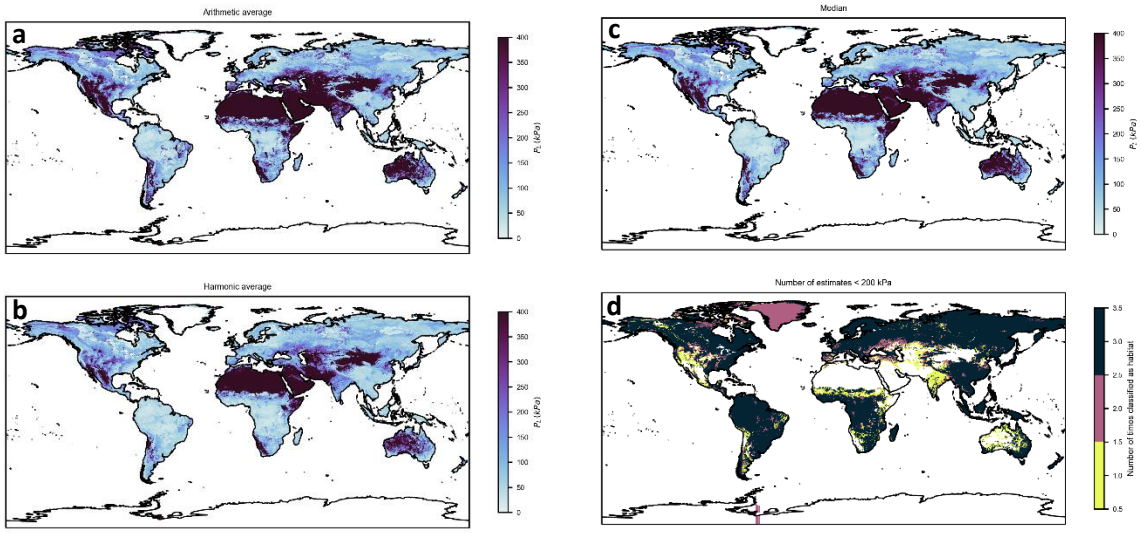
Extended Data



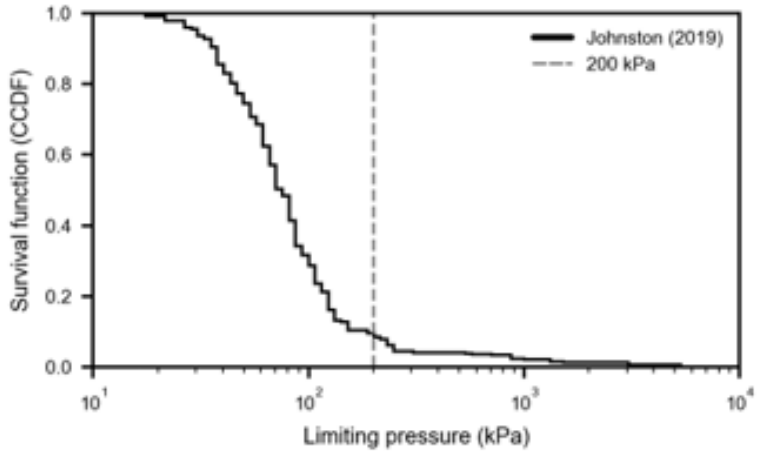
Extended Data Fig. 1: Soil hydro-mechanical coefficients as a function of soil fine texture. Soil shear strength relationships (**a**, **b**) were derived from Gerard et al. 1965[26]. Parameter relationship for shear modulus (**c, d**) to fine texture were derived from data points taken from literature[11, 46-48].



Extended Data Fig. 2: Profiles of cavity expansion pressures required for given soil texture and soil water contents. The red curve indicates the earthworm limiting pressures that would hinder bioturbation activity under different texture classes and soil moistures.



Extended Data Fig. 3: **Assessing different averaging techniques. a**, Arithmetic mean for limit pressures, **b**, harmonic mean of limit pressures, and **c**, median values for limit pressures. **d**, provides an estimate for how many times the regions are considered potential earthworm habitats amongst the different techniques (black indicating agreement of all averaging methods considered).



602

603 **Extended Data Fig. 4** : **Normalized cumulative distribution of earthworm abundance with**

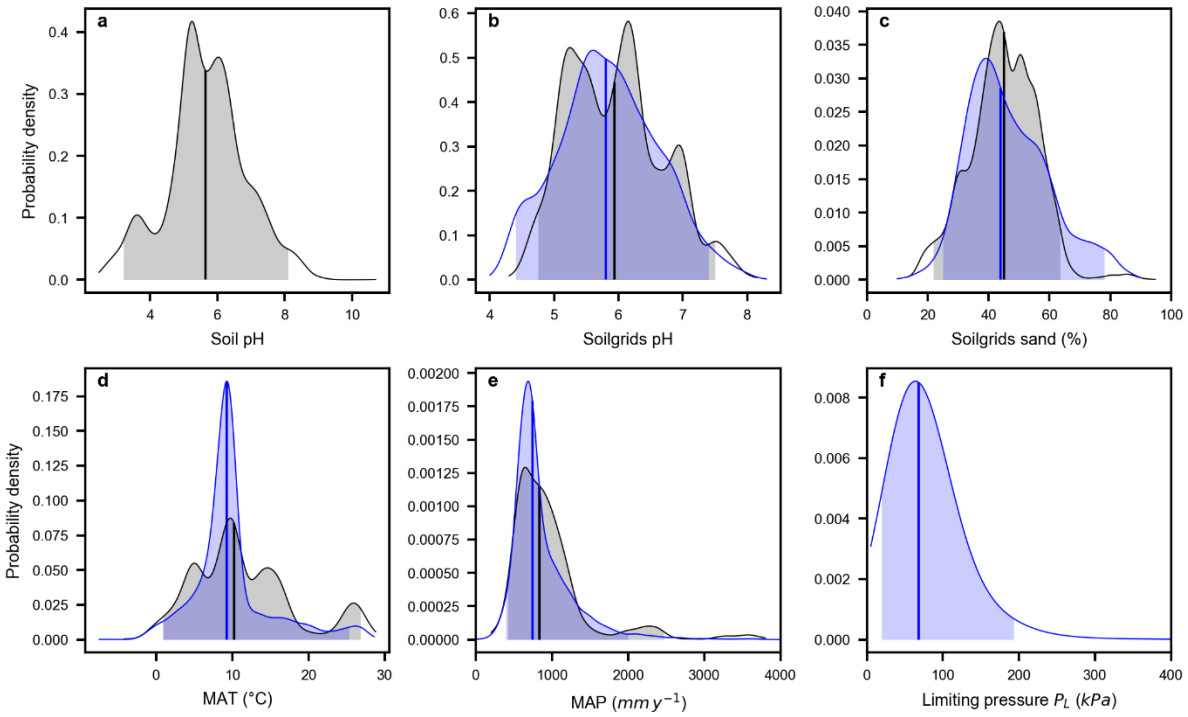
604 **average soil limiting pressure magnitudes.** Abundance data was taken from Johnston (2019)[20] and

605 mapped to limiting pressures using reported geographical coordinates. Over 90% of the earthworms

606 were located in regions with pressures below 200 kPa, which is consistent with earthworm's

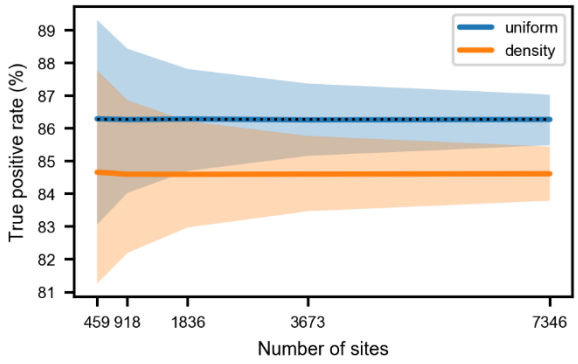
607 physiological hydro-skeletal pressure limit.

608



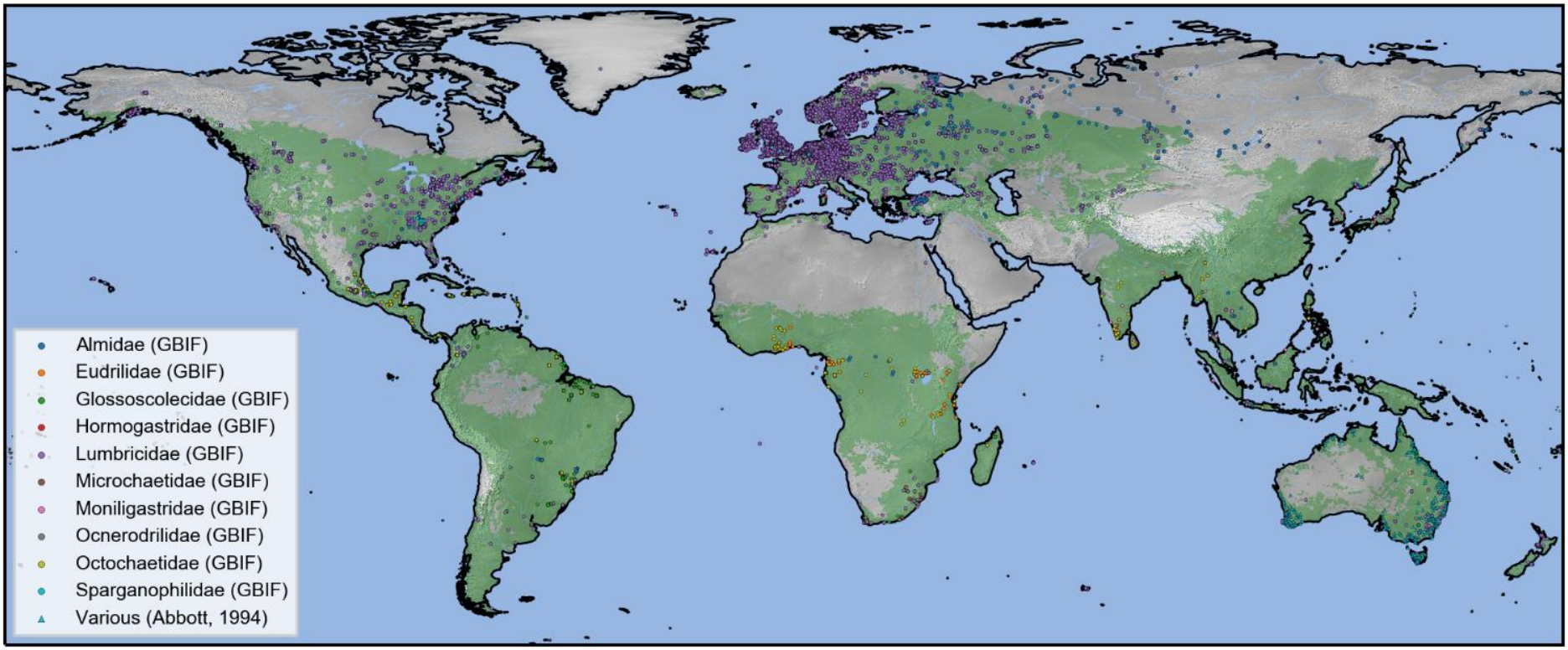
609

610 **Extended Data Fig. 5**: **Distribution of additional factors associated with sites of earthworm**
611 **occurrences. a-e**, kernel density estimates for selected variables used to delineate regions of potential
612 earthworm activity. Distributions of occurrences from a recent study[14] (black) and values at sites used
613 in the current study[15,43] (blue) are compared. Shaded areas contain 95% of values and the vertical line
614 indicates the median. The recent study[14] enables comparison of **a**, Soil pH measured on site with **b**,
615 soil pH from SoilGrids[32] as used in our study. The range of SoilGrids[30] pH values is narrower and
616 most of the occurrences were reported for sites with SoilGrids[30] pH > 4.5. **c**, Sand content from
617 SoilGrids[30] at which occurrences were reported. **d**, distribution of mean annual temperature (MAT)
618 and **e**, mean annual precipitation (MAP). **f,** Limiting pressure for earthworm cavity expansion as
619 estimated in this study.



620

621 **Extended Data Fig. 6**: **Robustness of true positive rate (hit-rate, sensitivity) under variation of**
622 **sample size for two sampling schemes.** Random re-sampling with replacement (n_{boot} = 5000) of sites
623 with earthworm occurrences are shown as solid lines and shading (representing median and central
624 95%) for two sampling schemes. Sites were selected with uniform probability (blue) or with
625 probabilities inverse to the density of reported occurrences in a five-point neighborhood (orange)
626 thereby penalizing sites with many reported occurrences nearby (attempting to address observational
627 bias). The dashed line represents the hit rate using the full dataset.



Extended Data Fig. 7: **Global distribution of occurrences for ten species (obtained from the Global Biodiversity Information Facility, GBIF) and an additional study from Australia [15].** The green shading indicates the modelled regions that are hospitable to earthworms based on soil mechanics and additional factors.