

# **Paleoclimate explains a unique proportion of the global variation in soil bacterial communities**

Manuel Delgado-Baquerizo<sup>\*1,2</sup>, Andrew Bissett<sup>3</sup>, David J. Eldridge<sup>4</sup>, Fernando T. Maestre<sup>5</sup>, Ji-Zheng He<sup>6,7</sup>, Jun-Tao Wang<sup>6</sup>, Kelly Hamonts<sup>2</sup>, Yu-Rong Liu<sup>6</sup>, Brajesh K. Singh<sup>2,8</sup>, Noah Fierer<sup>1,9</sup>.

1. Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309.
2. Hawkesbury Institute for the Environment, Western Sydney University, Penrith, 2751, New South Wales, Australia.
3. CSIRO, Oceans and Atmosphere, Hobart, Tasmania, 7000, Australia.
4. Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia.
5. Departamento de Biología y Geología, Física y Química Inorgánica, Escuela Superior de Ciencias Experimentales y Tecnología, Universidad Rey Juan Carlos, Calle Tulipán Sin Número, Móstoles 28933, Spain.
6. State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China.
7. Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, Victoria 3010, Australia.
8. Global Centre for Land Based Innovation, Western Sydney University, Building L9, Locked Bag 1797, Penrith South, New South Wales 2751, Australia.
9. Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309.

## **\*Author for correspondence:**

Manuel Delgado-Baquerizo. Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309. E-mail: M.DelgadoBaquerizo@gmail.com

### **Summary paragraph**

The legacy impacts of past climates on the current distribution of soil microbial communities are largely unknown. Here, we used data from >1000 sites from five separate global and regional datasets to identify the importance of paleoclimatic conditions (Last Glacial Maximum and Mid-Holocene) in shaping the current structure of soil bacterial communities in natural and agricultural soils. We show that paleoclimate explained a greater amount of the variation in the richness and composition of bacterial communities than current climate. Moreover, paleoclimate accounted for a unique fraction of this variation that could not be predicted by geographic location, current climate, soil properties, or plant diversity. Climatic legacies (temperature and precipitation anomalies from the present to ~20k years ago) likely shape soil bacterial communities both directly and indirectly via shifts in soil properties and plant communities. The ability of both paleo- and current climate to predict the distribution of soil bacteria declined dramatically in agricultural soils, highlighting the fact that anthropogenic activities have a strong influence on soil bacterial diversity. We illustrate how climatic legacies can help explain the current distribution of soil bacteria in natural ecosystems, and advocate that climate legacies should be considered when predicting microbial responses to climate change.

## Introduction

The climate of a particular region varies over time, often resulting in large-scale biome migrations that drive the current distribution of plant communities<sup>1-4</sup>. For example, long-term climatic legacies have shaped the distribution and diversity of plant communities in terrestrial ecosystems via dispersal-limited recolonization and environmental filtering<sup>2-4</sup>. Similarly, long-term regional climate history could conceivably explain significant proportions of the variation found in the current richness and composition of soil microbial communities. For example, a recent study provides indirect evidence that the last glaciation may have influenced the current distribution of strains of the soil bacterial genus *Streptomyces* across the United States<sup>5-6</sup>. However, the broader role of past climatic conditions in regulating the current distribution of microbial communities remains largely unexplored<sup>5</sup>. If past climates help explain the current distribution of microbial communities, careful consideration of climatic legacies could improve our capacity to predict how soil microbial communities will respond to forecasted climate changes, and how this response will affect the myriad ecosystem services that they provide (e.g., decomposition, nutrient cycling and climate regulation)<sup>7-9</sup>.

In theory, paleoclimate could explain the current distribution of soil microbial communities directly, via differential changes in temperature and precipitation patterns across millennia<sup>5-6</sup>. Soil bacteria are known to have short generation times leading to a fast turnover rate, but they have also been found to be highly sensitive to changes in temperature<sup>10</sup>. For example, a recent study demonstrated that a wide range of soil bacterial taxa exhibit predictable and consistent preferences for particular temperature conditions<sup>10</sup>. These intrinsic characteristics of microbial communities surely influence their direct response to paleoclimate. For instance, Atkinson et al.<sup>11</sup> found that the community composition of fast-growing invertebrates responded immediately to large and abrupt changes in temperature after the most recent glaciation, a response that left a strong signature in their contemporary distribution. Likewise, abrupt changes in climate, which may have occurred prior to 10000 years ago<sup>12</sup>, might have also left a strong signature on the structure of soil bacterial communities. In this respect, a direct effect from paleoclimate on soil microbial communities might have occurred in the past (e.g. in response to a particular drastic climatic event), but the consequences of this rapid compositional shift might still be detectable today.

Paleoclimate can also influence the current structure of bacterial communities indirectly, via its influence on soil properties and plant community structure<sup>13-15</sup>. Thus, variations in soil

properties such as pH and total organic carbon, which can have strong effects on microbial distributions<sup>13-15</sup> and change slowly during ecosystem development<sup>16-17</sup>, could drive the effects of paleoclimate on the contemporary patterns of microbial community composition and richness. Likewise, paleoclimate effects on plant communities<sup>2-4</sup> may be associated with corresponding changes in the composition of soil microbial communities<sup>18</sup>. Although the growing literature focuses on the main drivers of soil microbial communities in terrestrial ecosystems, we do not know whether climatic legacies contribute to their current richness and composition patterns at regional or global scales.

Additionally, if climatic legacies play an important role in regulating current soil microbial distribution, agricultural practices may reduce or remove any potential effects of paleoclimate on microbial community composition and richness. Agricultural practices are known to alter soil microbial communities directly, e.g. via introduction of new bacterial taxa associated with crop rhizospheres<sup>19</sup> or via fertilization<sup>20</sup>, and indirectly, via changes in soil properties (soil carbon, soil pH, and microbial communities)<sup>21</sup>. Drastic changes in the composition and richness of soil bacteria derived from agricultural practices might potentially mitigate the direct and indirect influences of climatic legacies on soil microbial communities via soil properties. Soil disturbance is expected to increase exponentially this century due to the increasing intensification of agricultural production needed to meet an increase in demand for food by 70 to 100% by 2050<sup>22</sup>. Thus, understanding how agricultural intensification will shift the signature of climatic legacies on microbial communities could improve our ability to evaluate and manage anthropogenic soil disturbances.

Here, we evaluated the relative importance of paleoclimate and current climate as predictors of the richness (number of phylotypes observed per sample) and composition (relative abundance of phylotypes) of soil bacteria at global and regional scales after accounting for key drivers of bacterial distribution such as geographical location, soil properties and plant diversity. We did so using data from five separate regional and global datasets including information on the structure of bacterial communities assessed via 16S rRNA gene sequencing (see Methods). Together, these datasets included more than 1000 sites from all continents except Antarctica, covering a broad range of ecosystem types (see Methods and Supplementary Fig. 1). We tested the following hypotheses: i) paleoclimate predicts a unique portion of the variation in the current richness and composition of soil bacterial communities in terrestrial ecosystems, ii) climatic legacies (measured as the temperature and precipitation anomalies<sup>12</sup> between an estimate of



climate 20,000 year ago and another estimate for the present day) affect the structure of current bacterial communities both directly and indirectly via soil properties and plant diversity and iii) soil disturbances linked to agricultural production reduce the relative importance of paleoclimate as a predictor of current microbial community richness and composition. It was not our intention to merge the five datasets used, which vary in sampling design and experimental methods (e.g. primer sets), but to test our hypotheses using five independent regional and global data sets from ecosystems that differed in their vegetation, climate and soil attributes (Methods; Supplementary Fig. 1).

## Results

We first used Variation Partitioning<sup>23</sup> to quantify the relative contribution of past and current climates as predictors of the richness and composition of soil bacterial communities. We also included soil properties and spatial variables<sup>15</sup> in our models. This approach allowed us to quantify the unique contribution of climate from a particular period to explain the current distribution of soil bacteria, and to differentiate this contribution from that shared among all predictors. Environmental drivers such as plant diversity, soil properties and geographic location explained unique portions of the variation in soil bacterial richness and composition in all datasets (Fig. 1, Supplementary Figs. 2 and 3). Most importantly, climatic variables from mid-Holocene and Last Glacial Maximum climates explained a unique percentage of the variation in the richness and composition of soil bacterial communities (Fig. 1, Supplementary Figs. 2 and 3). Overall, paleoclimate was a better predictor of soil bacterial richness and composition than current climate in all five datasets (Fig. 1), suggesting that models using current climate alone have a limited predictive power at regional to global scales. Paleoclimate also shared a large part of the variance explaining bacterial community richness and composition with plant richness and/or soil properties, suggesting that a large fraction of the apparent effects of paleoclimate on soil bacterial communities may be driven by its direct and indirect effects on these ecosystem variables.

We then used structural equation modeling (SEM, see Methods) to assess the role of climatic legacies in driving bacterial community composition and richness, and to separate direct (i.e., temperature and precipitation anomalies between an estimate of climate 20,000 year ago and another estimate for the present day) and indirect (via soil properties and plant diversity) effects of such legacies on soil microbial communities. Unlike regression analyses, SEM offers the ability to separate multiple pathways of influence and to investigate the complex relationships among

environmental predictors commonly found in terrestrial ecosystems (Methods). As SEM works on single response variables, we collapsed the bacterial community compositional data using non-metric multidimensional scaling (NMDS) for each dataset independently, and retained the first two axes from a 2D solution (Bacterial comm. 1 and 2; stress ~ 0.1 in all cases). Prior to conducting SEM, we used a Random Forest<sup>8</sup> procedure (Methods) to reduce the number of predictors to those that significantly explained the variation found in bacterial community richness and composition (i.e. geographical location, climatic legacies, soil properties and plant diversity) for each dataset (Supplementary Table 4). Random Forest procedures are recommended for identifying the main significant predictors of environmental response variables (Methods). Finally, after conducting Random Forest but prior to the final SEM analyses, we ran preliminary SEMs to further evaluate whether the effects of climatic legacies on soil microbial community composition and richness were independent of those of current climate. We included in these analyses the selected climatic legacies from Random Forest analyses, but also included their corresponding current climate variables. In general, climatic legacies were as important as, or more important than, current climate in directly driving the richness and composition of bacteria across all datasets (Supplementary Fig. 4).

Our final SEM analyses provided solid evidence that climatic legacies had both direct (four of five cases) and indirect (four of five cases) effects on bacterial richness (Fig. 2; Supplementary Tables 5 and 6) across the five datasets used. Annual mean temperature and precipitation in the driest month showed the largest total effects (sum of direct and indirect effects from SEM) on bacterial richness in three of five datasets (Supplementary Fig. 5). Similarly, we also found both direct (four of five cases) and indirect (all cases) effects of climatic legacies on the composition of bacterial communities (for both NMDS axes; Supplementary Figs. 6-9; Supplementary Tables 5 and 6). In this case, direct effects were driven both by changes in temperature and precipitation, with particular importance of annual mean temperature and isothermality and precipitation in the driest month (Supplementary Figs. 6 and 7). Indirect effects of precipitation and temperature legacies on microbial richness were largely driven by soil properties such as pH (three of five databases for bacterial richness and all databases cases for bacterial community composition), organic carbon concentration and texture (two of five cases for bacterial community composition, respectively; Fig. 2 and Supplementary Figs. 6 and 7). Other soil properties such as available phosphorus and micronutrients indirectly drove part of the effects of climatic legacies on microbial

community structure (Fig. 2 and Supplementary Figs. 6 and 7). We also found strong indirect effects of climatic legacies on soil bacterial richness (three of three databases) via changes in plant diversity. On the contrary, the effects of climatic legacies on bacterial community composition were only indirectly driven via plant species richness in China (Supplementary Fig. 6).

Additional Random Forest analyses (see Methods and Supplementary Data Table 7) allowed us to identify some of the bacterial taxa that were consistently (i.e. in more than half of the datasets) good predictors of major climatic legacies (i.e., AMT and PDM, which were selected using standardized total effects from SEM; Methods). For example, we found that the relative abundance of both *Planctomycetes* and candidate phylum WS3 (recently renamed *Latescibacteria*) consistently increased with increasing precipitation in the driest month from paleoclimatic to current climates (Supplementary Data Table 7). In addition, phylum *Actinobacteria* was found to be an indicator of changes in temperature over millennia (Supplementary Data Table 7).

We repeated our Variation Partitioning models for a subset of data from the Australia dataset where we were able to partition sites between croplands and natural ecosystems located close to these croplands (66 sites each). This allowed us to evaluate whether agriculture might alter the predictive power of paleo- and current climates. We found that paleoclimate (mid-Holocene + Last Glacial Maximum) still predicted a unique part of the variation in bacterial diversity within croplands (Fig. 3 and Supplementary Fig. 10). However, paleoclimate always had a significantly lower capacity to predict bacterial diversity in croplands than in natural ecosystems (Fig. 3 and Supplementary Fig. 10). When the SEMs were repeated using data from only natural and croplands sites in Australia (Fig. 3 and Supplementary Fig. 11), we found a strong reduction in the importance of soil properties as predictors of microbial community richness and composition due to the extreme disturbance caused by cotton and wheat farming (Fig. 3 and Supplementary Fig. 11).

## **Discussion**

Together, our work provides, to our knowledge, the first empirical evidence that paleoclimate and climatic legacies (climate anomaly between 20,000 years ago and today) can leave a strong signature on soil bacterial communities, which may have influenced the contemporary distribution of bacterial richness and composition from regional to global scale. The importance of these results lies in the fact that climatic legacies can be used to better understand and predict the response of microbial communities to ongoing climate changes, including rising temperatures and changes in

precipitation patterns<sup>28</sup>. For example, in arid environments (Global drylands and New South Wales datasets; average of 338/334 and 417/398mm of current/Last Glacial Maximum annual precipitation, respectively), increasing precipitation in the driest month from paleoclimates to current climates resulted in a net increase (sum of direct and indirect effects) in bacterial richness (Supplementary Fig. 5). This result is supported by a recent study highlighting that aridity, a proxy of water availability, is a key driver of bacterial diversity in global drylands<sup>15</sup>. However, in more humid environments such as those of the Americas and China (mostly temperate and tropical ecosystems; average of 948/894mm and 903/1020mm of current/Last Glacial Maximum annual precipitation, respectively), increases in precipitation (Americas) or precipitation in the driest month (China) from paleo to current climates led to reductions in bacterial richness. This response is likely driven by increases in the relative abundance of specific microbial taxa under the wettest conditions<sup>15</sup> (Fig. 2 and Supplementary Fig. 5). For example, the relative abundance of both *Planctomycetes* and phylum *Latescibacteria* was positively related to the precipitation in the driest month anomaly from paleoclimatic to current climates (Supplementary Data Table 7). This result agrees with expectations that members of these phyla typically prefer wetter environments<sup>29-30</sup>. Interestingly, phylum *Actinobacteria* was also found to be an indicator of changes in temperature over millennia, suggesting that this taxon may have been highly influenced by the last glaciation at the continental scale<sup>5-6</sup>. The effects of increasing precipitation on bacterial richness observed in China may be indirectly driven via soil acidification as a consequence of soil weathering<sup>31</sup> (Fig. 2d), as reductions in soil pH are known to reduce soil bacterial diversity in terrestrial ecosystems<sup>13</sup>. Annual mean temperature showed the highest total (sum of direct and indirect effects) positive and negative effect on bacterial richness for the China and Australia datasets, respectively (Supplementary Fig. 5). This contrasting result might be related to the fact that Australia showed the lowest increase in temperature from paleo to current climates in this study (3.5°C), which resulted in a total positive effect on the diversity of bacteria, compared to the increases found in China (5.6°C), Global drylands (5.2°C) and the Americas (10.1°C), where annual temperature legacies had a total negative effect on the diversity of soil bacterial communities.

Climatic legacies (measured as the temperature and precipitation anomalies<sup>12</sup> between an estimate of climate 20,000 year ago and another estimate for the present day) drove the richness and composition of bacterial communities both directly and indirectly via changes in soil properties and plant diversity. Direct effects were driven both by changes in temperature and

precipitation, with particular importance of annual mean temperature, isothermality and precipitation in the driest month (Supplementary Data Figs. 6 and 7). This finding is supported by recent studies that have identified temperature and precipitation as key global and continental-scale predictors of bacterial community richness and composition<sup>15</sup>. Direct effects include the impacts derived from rapid climatic changes in the past –which mostly occurred prior to 10000 years ago<sup>12</sup> and that have left a strong signature on the contemporary structure of soil bacterial communities (see Appendix S3 for further rationale on direct effects from paleoclimates on soil bacterial communities). Indirect effects of climatic legacies on bacterial richness and composition were largely driven by soil properties such as pH and, to a lesser extent, by organic carbon concentration and texture. These soil variables, which were included in all datasets, are known to determine changes in bacterial communities in terrestrial ecosystems<sup>8,13-15</sup>. Other soil properties such as soil P and micronutrients were also found to indirectly drive part of the effects of climatic legacies on bacterial community structure. In general, we found strong indirect effects of climatic legacies on soil bacterial richness via plant diversity; an indirect effect that was only observed for bacterial composition in China. Highly diverse plant communities may promote the richness of soil bacteria by supporting a wider variety of niches (e.g. litter qualities and rhizosphere products)<sup>17</sup>. Climatic legacies in these datasets always had direct and indirect (via soil properties) effects on plant richness, providing further support for the notion that climatic legacies play important roles in driving current plant diversity in terrestrial ecosystems<sup>2-4</sup>.

As expected, geographic location, soil properties and plant richness accounted for significant variation in microbial community richness and composition in our model<sup>8,13-15,18,24,26</sup>. However, a unique and significant portion of variation was explained by climate legacies, which suggests that geographical location, soil properties and plant diversity cannot account solely (via direct effects) for most of the effects of climatic legacies on soil bacterial communities (Figs. 1-2 and Supplementary Figs. 2-3 and 5-9). We acknowledge that other soil properties not included in our models could improve the predictive power attributed to climatic legacies. Alternatively, part of the direct effects from climatic legacies on bacterial communities may still be indirectly driven by processes that we did not explicitly account for in the SEMs, such as dispersal-limited recolonization, which have been traditionally considered as main drivers of paleoclimate effects on plant diversity<sup>2-4</sup>.

Our analyses also provide evidence that paleoclimate is also influencing the contemporary distribution of bacterial communities in croplands. However, paleoclimate always had a significantly lower capacity to predict the richness and composition of bacterial communities in croplands than in natural ecosystems. This suggests that agricultural practices have reduced the unique contribution of paleoclimate as a predictor of contemporary soil microbial distribution. Our SEM results suggested that the influence of direct and indirect effects from climatic legacies on the richness and composition of soil bacterial communities are largely reduced in croplands (cotton and wheat farming) compared to natural ecosystems. Direct effects of climatic legacies on soil microbial communities can be potentially erased or lessened by agricultural practices via introducing new taxa of bacteria associated with the rhizosphere of particular crop species<sup>19</sup> and/or by artificially promoting particular bacterial species responsive to watering or fertilization<sup>20</sup>. Indirect effects from paleoclimate on soil bacterial communities can be erased (richness) or lessened (composition) via rapid changes in soil carbon and pH as a result of agricultural practices. By drastically changing microbial communities in soil, agricultural practices hinder our capacity to predict the richness and composition of these bacterial communities using paleoclimatic information. This result suggests that agriculture not only removes paleoclimatic legacies on microbial communities, but also leads to predictions with a lower level of accuracy than those obtained using data from natural ecosystems. Our results, coupled with those from previous studies<sup>21</sup>, highlight the fact that agricultural intensification markedly alters soil microbial community composition and diversity in terrestrial ecosystems in ways that can be difficult to predict.

Together our results provide strong evidence that past climates have left their signature on current bacterial diversity patterns across the globe, and that agricultural practices may significantly reduce the unique signature of climatic legacies on soil bacteria. Overall, our findings indicate that using paleoclimatic data can improve our ability to predict the global distribution of soil bacterial communities in natural ecosystems. Thus, paleoclimatic data should be used when assessing the responses of these communities, and the ecosystem services they provide, to global environmental change.

## **Methods**

### **Study sites and data collection**

**Drylands** (Global scale). We used data from Maestre et al.<sup>15</sup>. This dataset is focused on drylands (i.e. regions with an aridity index [precipitation/potential evapotranspiration] < 0.65)<sup>32</sup> and includes a wide variety of ecosystem types, including grasslands, shrublands and open woodlands, and environmental conditions across “natural” dryland ecosystems. Field samples were collected between 2006 and 2010 from 80 sites located in 12 countries from all continents except Antarctica (Supplementary Fig. 1), under the most representative vegetation of each plot, according to a standardized sampling protocol<sup>15</sup>. A composite sample (i.e. from five soil samples; top 7.5 cm) was randomly taken under the canopy of the dominant perennial plant species. Each sample was separated into two portions. One portion was air-dried and used for chemical analyses. The other was immediately frozen at -20 °C for molecular analyses.

Soil DNA was extracted using the Powersoil® DNA Isolation Kit (Mo Bio Laboratories, Carlsbad, CA, USA) according to the manufacturer’s instructions. A portion of the bacterial 16S rRNA gene was sequenced using the Illumina MiSeq platform and the 341F/805R primer sets<sup>33</sup>. Bioinformatic analyses were conducted using the QIIME package (Caporaso et al. 2010) as describes in ref. 15. Operational Taxonomic Units (OTU) were picked at 97% sequence similarity. The resultant OTU abundance tables from these analyses were rarefied to an even number of sequences per samples to ensure equal sampling depth (11789). In addition, we removed OTUs that had only one read per OTU across all samples. Plant species richness and soil properties, including texture (% of sand), pH, electrical conductivity, total organic C, C:N ratio, total P, available P, available N (sum of inorganic and organic N), dissolved phenols, aromatic compounds, proteins, amino acids, carbohydrates and N mineralization, were measured as described elsewhere<sup>34</sup> (Supplementary Table 3).

**Americas** (Cross-continental scale). We used data from Ramirez et al.<sup>24</sup>. This dataset includes 48 “natural” sites across North and South America that cover a wide range of biomes and environmental conditions going from Arctic to tropical forests (Supplementary Fig. 1). Composite soil samples (top 5cm) were collected under the most representative vegetation of each study site. Each sample was separated into two portions. One portion was kept fresh and used for chemical analyses, the other was stored at -80°C until DNA extraction.

Soil DNA was extracted using the Powersoil® DNA Isolation Kit (Mo Bio Laboratories, Carlsbad, CA, USA) following the manufacturer’s instructions with the modifications described previously<sup>13</sup>. A portion of the bacterial 16S rRNA gene was sequenced using the Illumina MiSeq

platform and the 515F/806R primer sets<sup>35</sup>. Bioinformatic analyses were conducted using the QIIME package<sup>36</sup> as described in ref. 24. Operational Taxonomic Units (OTU) were picked at 97% sequence similarity. All samples were rarified to 40000 randomly selected reads per sample. In addition, we removed OTUs that had only one read per OTU across all samples.

Soil properties including texture (% of sand), pH, total organic C, C:N ratio and C mineralization were measured as described in reference 13 (Supplementary Table 3).

**Australia** (Continental scale). We used a subset of sample locations from the Biome of Australia Soil Environments (BASE) project<sup>25</sup> (Supplementary Fig. 1). This dataset include 531 soil samples belonging to “natural” (465) and agricultural (66) (cropping by cotton and wheat) ecosystems from Australia. Samples were collected between 2011 and 2014. Soil samples were collected according to the methods described in ref. 25. In brief, at each plot, a composite soil sample (nine discrete soil samples) from the top 0-10cm<sup>25</sup> was collected and separated into two portions. One portion was air-dried for chemical analyses, the other was frozen (-80°C) until DNA extraction.

All soil DNA was extracted in triplicate, according to the methods employed by the Earth Microbiome Project<sup>25</sup>, at the Australian Genome Research Facility. Amplicons targeting the bacterial 16S rRNA gene were sequenced using the Illumina Miseq platform and the 27F – 519R<sup>37</sup> primer set (see reference 25 for details on these analyses). Bioinformatic analyses were performed as explained using MOTHUR (v1.34.1)<sup>38</sup> as explained in ref. 25. Operational Taxonomic Units (OTU) were picked at 97% sequence similarity. The OTU abundance tables were rarefied at 14237 sequences/sample to ensure even sampling depth. In addition, we removed OTUs that had only one read per OTU across all samples.

Soil properties including texture (% of sand), pH, electrical conductivity, total organic C, available N (sum of ammonium and nitrate), available P, available K, and total K, S, Cu, Fe, Mn, Zn, Al, Ca, Mg, Na and B were measured as described in reference 25 (Supplementary Table 3). **China** (Continental scale). This dataset focuses on forest ecosystems (i.e., boreal, temperate mixed coniferous, temperate deciduous, subtropical evergreen and tropical forests) and includes 300 plots across a wide latitudinal gradient (approximately 4,000 kilometers) in Eastern China<sup>26</sup> (Supplementary Fig. 1). In each plot, a composite soil sample from 15 soil cores was collected from the top 0-10cm. Each sample was separated into two portions. One portion was air-dried and used for soil chemical analyses and the other was stored at -80°C until DNA extraction.



Soil DNA was extracted using the Powersoil® DNA Isolation Kit (Mo Bio Laboratories, Carlsbad, CA, USA) with a slight modification as explained in reference 26. A portion of the 16S rRNA gene (515F/806R primer set)<sup>38</sup> was sequenced using the Illumina Miseq platform. Bioinformatic analyses were completed using the QIIME pipeline<sup>36</sup> (see ref. 26 for details on these analyses). Operational Taxonomic Units (OTU) were picked at 97% sequence similarity. All samples were rarefied to 40000 randomly selected reads per sample. In addition, we removed OTUs that had only one read per OTU across all samples.

Plant species richness and soil properties, including texture (% of sand), pH, total organic C, C:N ratio and available P, were measured as described in reference 26 (Supplementary Table 3).

**New South Wales** (Regional scale). We used a data from Eldridge et al.<sup>27</sup>. This dataset includes data from 54 sites scattered across a 500 km<sup>2</sup> area of semi-arid eastern Australia (Supplementary Fig. 1). This survey was undertaken in three woodland communities dominated by blackbox (*Eucalyptus largiflorens*), white cypress pine (*Callitris glaucophylla*) and river red gum (*Eucalyptus camaldulensis*). This dataset includes sites extensively used for livestock grazing, large areas dedicated for conservation (national parks, nature reserves) and smaller areas devoted to native forestry. At each site, a soil sample was collected in 2014 from the top 5cm of soil. For this study, we used the subset of samples collected under tree microsites<sup>27</sup>. Each sample was separated into two portions. One portion was air-dried and used for soil chemical analyses, the other was stored at -20°C until DNA extraction.

Soil DNA was extracted using the Powersoil® DNA Isolation Kit (Mo Bio Laboratories, Carlsbad, CA, USA) according to the manufacturer's instructions. Amplicons targeting the bacterial 16S rRNA gene were sequenced using the Illumina MiSeq platform and the 341F-805R primer set<sup>33</sup>. Bioinformatic analyses were done using MOTHUR<sup>38</sup> and UPARSE<sup>39</sup>. Operational Taxonomic Units (OTU) were picked at 97% sequence similarity. We removed OTUs that had only one read per OTU across all samples and the resulting OTU abundance tables were rarefied to 10851 sequences per sample (the fewest sequences obtained in a single soil sample).

At each of three positions along a 100 m transect (0 m, 50 m, 100 m) we selected the nearest tree (perennial plant > 4 m in height). Two small (0.5 m x 0.5 m) quadrats were placed midway between the trunk and the canopy edge on opposite sides of the canopy. Within these small quadrats we assessed the cover of all vascular plants by species and used these data to obtain a value of total

plant species richness for each site. Soil properties including texture (% of sand), bulk density, and available P were measured as described in reference 27 (Supplementary Table 3).

### **Climate data**

In all cases, a total of 19 standardized climatic variables (Supplementary Table 1) were obtained for all the sites surveyed from the Worldclim database ([www.worldclim.org](http://www.worldclim.org)). In the case of mid-Holocene and Last Glacial Maximum climates, we used estimates provided by the Community Climate System Model (CCSM4; [www.worldclim.org](http://www.worldclim.org))<sup>40-43</sup>. We used data at a 2.5 minutes resolution (~4.5km at equator), as this is the highest resolution available for the Last Glacial Maximum period. Bioclimatic data are also available for this resolution for current and mid-Holocene climates, allowing the direct comparison among bioclimatic data at different periods. In all cases, climatic data at 30 seconds resolution for current and mid-Holocene, which allowed us to compare 2.5 minutes and 30 seconds resolution data for these two periods. Values calculated using 2.5 minutes were identical to those calculated using a resolution of 30 seconds in all cases (Pearson's  $r > 0.99$ ;  $P < 0.001$ ) (See Climatic data cross-validation in Appendix S1). We acknowledge that paleoclimatic data from islands may be inaccurate as a consequence of their spatial location, which influences the accuracy of the available climate data. This should not bias, however, the conclusions from this study –which was conducted at the global scale– given the low number of data points coming from islands<sup>12</sup>.

### **Pre-selection of multicollinearity free climatic variables**

We decided to use only those climatic variables (i.e., from the original 19 climatic variables available from Worldclim) that were free of multicollinearity within each period of time (i.e. current, Last Glacial Maximum and mid-Holocene). For example, the inclusion of strongly positively correlated ( $r > 0.8$ )<sup>44</sup> variables within a particular group of predictors is not recommended for Variation Partitioning modeling as they may cause multicollinearity problems in the analyses (see below). To preselect multicollinearity-free climatic variables from the original list, we collapsed the climatic information from all datasets and conducted correlation analyses (Pearson) within each period of time (i.e., group of predictors in our Variation partitioning) for the original 19 climatic variables available (Supplementary Tables 1 and 2). Based on these analyses, we selected for our analyses the same eight out of 19 climatic variables for each period of time that were not strongly correlated with the rest ( $r < 0.8$ )<sup>44</sup>: annual mean temperature (AMT), mean diurnal temperature range (MDR), Isothermality (ISO), temperature in the wettest quarter

(TWETQ), annual precipitation (AP), precipitation in the driest month (PDM), precipitation seasonality (PSEA) and precipitation in the colder quarter (PCQ). These eight variables include variables that are highly correlated to multiple non-selected variables and also variables that were unrelated to any other climatic variable, hence which could only be explained by themselves. Together, these variables are a good representation of the rest of non-selected climatic variables (Supplementary Table 2). We retained these eight variables for the remainder of statistical analyses presented in this manuscript.

### ***Statistical modeling***

We used a combined approach including multiple statistical models to address our different hypotheses. In particular, we used (1) Variation Partitioning modeling to identify whether paleoclimate can explain a unique portion of the variation in bacterial community richness and composition that cannot be accounted for other key predictors of soil microbial communities; (2) Random Forest analysis to identify the main individual predictors of bacterial community richness and composition including spatial predictors, climatic legacies, soil properties and plant richness; and (3) Structural Equation Modeling (SEM) to identify the direct and indirect (via soil properties and plant richness) effects of climatic legacies on bacterial community richness and composition. All these statistical models address a particular part of our research question that cannot be addressed using each approach on its own.

### ***Variation partitioning modeling.***

We used Variation Partitioning<sup>23</sup> to quantify the relative importance of four groups of predictors: 1) climatic variables from the Last Glacial Maximum, 2) climatic variables from the mid-Holocene, 3) climatic variables from current climates and 4) other key environmental drivers of microbial communities, including plant species richness (available for Drylands, China and New South Wales survey) and/or soil properties (14 for Drylands, 5 for Americas, 18 for Australia, 5 for the China and 6 for the New South Wales survey; total organic carbon, texture and pH were included in all models) and space (sites location as defined by latitude and longitude) as predictors of the bacterial community composition (number of reads / OTU), richness (i.e., number of OTUs per sample). Climate includes the eight multicollinearity-free variables described above (Supplementary Tables 1 and 2). Geographic location (i.e. latitude and longitude) was included in all models to account for spatial autocorrelation (see reference 15 for a similar approach). The complete list of predictors available for each database is presented in Supplementary Table 3.

Variation partitioning is a method specifically recommended to deal with between-group multicollinearity, as it partitions the variance in a given response variable that is attributed to a particular group of predictors from that variance shared among all predictors<sup>23</sup>. Thus, this analysis allow us to identify whether climatic variables from current, mid-Holocene and Last Glacial Maximum periods can explain a unique portion of the variance that is not explained by climate in other periods<sup>23</sup>. Note that adjusted coefficients of determination in multiple regression and canonical analysis can, on occasion, take negative values<sup>23</sup>. Negative values in the variance explained for a group of predictors on a group of response variable are interpreted as zeros and correspond to cases in which the explanatory variables explain less variation than random normal variables would<sup>23</sup>. In all cases, Variation Partitioning analyses were conducted with the R package Vegan<sup>45</sup>. The complete list of predictors available for each database is presented in Supplementary Table 3.

***Assessing comprehensive indices of climatic legacies.*** To obtain a greater mechanistic understanding of the role of paleoclimate in regulating current microbial richness and composition, we calculated specific climatic legacy indices for each of the preselected eight climatic variables. To do so, we calculated the mathematical difference in the values for each climatic variable from Last Glacial Maximum to current climates (e.g. Annual precipitation<sub>Current climate</sub> - Annual precipitation<sub>Last Glacial Maximum</sub>) for each site. Therefore, climatic legacies represent the temperature and precipitation anomalies between an estimate of climate 20,000 year ago and another estimate for the present day<sup>12</sup>. This difference informs us about the climatic legacies –increases, decreases or lack of changes for a particular climatic condition with time- in each of the sites surveyed from the different datasets (See Climatic legacy indexes cross-validation in Appendix S2). We used paleoclimatic information from the Last Glacial Maximum rather than mid-Holocene conditions because 1) the former is included in the period between Last Glacial Maximum and the current climate and 2) in general, Last Glacial Maximum conditions were a better predictor of bacterial richness and composition than climatic conditions in mid-Holocene (Supplementary Figs. 2 and 3) (Appendix S2). Note that the climatic legacy index used here is based on the differences between two single snapshots in time (Current vs. Last Glacial Maximum climates), thus calculation of climate legacy comes with a number of inherent and important assumptions<sup>12</sup>, some of which are accounted for in Appendix S2. Also, note that most abrupt changes in climate occurred prior to 10000 years ago<sup>12</sup> (e.g., Supplementary Fig. S12-14). Even so, our climatic legacy index still

allowed us to address our research question on whether the difference between climate today and 20000 years ago affects the structure of current bacterial communities, and whether these effects were directly mediated by climate legacies or indirectly via soil properties and plant diversity.

***Random Forest modeling I: pre-selection of main microbial drivers used in structural equation modeling.*** Due to the large amount of predictors used, we conducted a classification Random Forest analysis<sup>46</sup> as described in ref. 8 to identify the major statistically significant predictors of the composition and richness of bacteria to be included in our structural equation models (next section). Contrary to the Variation Partitioning model described above, both Random Forest and structural equation modeling take one response variable at each time. In this respect, in the case of bacterial community composition at the OTU level, we conducted Random Forest analyses on the two scores from the 2D solution of a non-metric multidimensional ordination (nMDS) using the Bray-Curtis dissimilarity metric (i.e. Bacterial comm. 1 and 2). The complete list of predictors available for each database is presented in Supplementary Table 3. These analyses were conducted using the rfPermute package<sup>47</sup> of the R statistical software (<http://cran.r-project.org/>).

***Structural equation modeling.*** We used structural equation modeling (SEM)<sup>48</sup> to evaluate the direct (i.e., changes in temperature and precipitation variables with time) and indirect (i.e. plant diversity and/or soil properties and space) effects of climatic legacies on the richness and composition of bacterial communities. The use of SEM is particularly useful in large scale correlative studies, as it allows the partitioning of causal influences among multiple variables, and separation of the direct and indirect effects of model predictors<sup>48</sup>. The main structure of our *a priori* model was shared across all datasets and response variables (Supplementary Fig. 15). We only included in these models those variables that were identified as major statistically significant predictors of the richness and composition of bacterial communities from Random Forest analyses (Supplementary Table 4). Therefore, SEM models conducted for the different datasets contain different predictors and were independently constructed. The only exception to this was latitude and longitude, which were included in all the models to account for spatial autocorrelation in our models (as done in reference 15). By simplifying our models with such approach, we acknowledge that we may be missing some indirect effects from excluded variables on bacterial community richness or composition. However, we also reduce the complexity of our models, providing a more comprehensive understanding on the major direct and indirect effects from climatic legacies on the richness and composition of bacterial communities; therefore allowing us to properly address

our research question. A direct consequence of this approach is that direct effects between climatic legacies and soil properties may not include some of the major climatic legacies controlling soil properties, obscuring the interpretation of these parts of the models. It was not the goal of this study to identify the major direct effects of climatic legacies on soil properties. Consequently, we only included in our models those soil properties that directly influenced soil bacterial community richness or composition, and that ultimately could hold indirect effects of climatic legacy on bacterial communities.

In our models, all variables are included as independent observed variables. We grouped the different categories of predictors (climate legacies, soil properties and spatial) in the same box in the model for graphical simplicity, however these boxes do not represent latent variables. The climatic legacies box includes all selected individual climatic legacies identified as significant predictors of bacterial community richness or composition from Random Forest. The spatial box includes latitude and longitude. The soil properties box include all individual soil properties identified as significant predictors of bacterial community richness or composition by our Random Forest analyses. Note that if none of the variables within a particular box (e.g. plant richness or soil properties) were selected by our Random Forest analyses as significant predictors of a particular microbial variables and for a particular dataset, that box is excluded in that specific model. We included both direct and indirect (via soil properties and plant richness) effects of climatic legacies on the richness and composition of bacterial communities in our models (see rationale for direct impacts of climatic legacies on soil properties, microbial communities and plant diversity in Appendix S3).

All variables within a particular box were allow to covary. Because of this, all models were originally saturated (zero degrees of freedom). In order to release a degree of freedom and make possible for us to test the goodness of fit of our models, we conducted the following *a priori* analyses: (1) we conducted partial correlations (Pearson) between all predictors within a particular model and (2) we removed the weakest *a priori* correlation (Supplementary Table 9) between two predictors in our models; therefore releasing a degree of freedom and making our models testable. The goodness of fit of SEM models was checked following ref. 49. There is no single universally accepted test of overall goodness of fit for SEM, applicable in all situations regardless of sample size or data distribution<sup>49</sup>. We used five measures of goodness of fit of our models including the (1) Comparative Fit Index (CFI) (the model has a good fit when  $0.97 \leq CFI \leq 1.00$  and acceptable

fit when  $0.95 \leq \text{CFI} < 0.97$ ), (2) Goodness-of-Fit Index (GFI) (the model has a good fit when  $0.95 \leq \text{GFI} \leq 1.00$  and acceptable fit when  $0.90 \leq \text{GFI} < 0.95$ ), (3) Normed Fit Index (NFI) (the model has a good fit when  $0.95 \leq \text{NFI} \leq 1.00$  and acceptable fit when  $0.90 \leq \text{NFI} < 0.95$ ), (4)  $\chi^2$  test ( $\chi^2$ ; the model has a good fit when  $0 \leq \chi^2/\text{df} \leq 2$  and  $0.05 < P \leq 1.00$  and acceptable fit when  $2 < \chi^2/\text{df} \leq 3$  and  $0.01 \leq P \leq 0.05$ ), and (5) the root mean square error of approximation (RMSEA; the model has a good fit when  $0 \leq \text{RMSEA} \leq 0.05$  and  $0.10 < P \leq 1.00$  and acceptable fit when  $0.05 < \text{RMSEA} \leq 0.08$  and  $0.05 \leq P \leq 0.10$ )<sup>49</sup>. In general, our *a priori* models attained an acceptable/good fit. In particular, 16/21 cases showed a good/acceptable fit by all criteria (Supplementary Table 10). The remaining 5/21 cases still showed a good fit in three of the five indexes used here (CFI, GFI and NFI) (Supplementary Table 10). No post-hoc alterations were made. With an acceptable/good model fit, we were free to interpret the path coefficients of the model and their associated *P* values. SEM models were conducted with the software AMOS 20 (IBM SPSS Inc, Chicago, IL, USA).

We also calculated the standardized total effects of plant diversity and/or soil properties, space and climatic legacies on the richness and composition of bacteria. The net influence that one variable has upon another is calculated by summing all direct and indirect pathways between the two variables. If the SEM model fits the data well, the total effect should approximate be the bivariate correlation coefficient for that pair of variables<sup>48</sup>.

***Random Forest modeling II: identifying the main phyla characterizing particular climatic legacies.*** We also used Random Forest analysis<sup>46</sup> as described in ref. 8 to identify the main bacterial phyla predicting a particular climatic legacy. We focused on the major climatic legacies driving bacterial community composition, which were identified using the standardized total effects from SEM: AMT (in China and Australia) and PDM (in Global-drylands and New South Wales) (Supplementary Figs. 8 and 9). AMT was also a major predictor of the bacterial community composition in the Americas dataset (Supplementary Fig. 8). In these analyses, the relative abundance of bacterial phyla acts as a predictor of a particular climatic legacy variable (AMT or PDM). These phyla include: *Thermi*, *Acidobacteria*, *Actinobacteria*, *AD3*, *Armatimonadetes*, *Bacteroidetes*, *BRC1*, *Chlorobi*, *Chloroflexi*, *Cyanobacteria*, *Elusimicrobia*, *FBP*, *FCPU426*, *Fibrobacteres*, *Firmicutes*, *Gemmatimonadetes*, *GN02*, *Nitrospirae*, *NKB19*, *OD1*, *OP11*, *OP3*, *Planctomycetes*, *Proteobacteria*, *Tenericutes*, *TM6*, *TM7*, *Verrucomicrobia*, *WPS-2*, *WS2*, *WS3* and *WS4*. These analyses were conducted using the rfPermute package<sup>47</sup> of the R statistical

software (<http://cran.r-project.org/>). The main goal of these analyses is to identify the main taxa characterizing a particular climatic legacy from AMT or PDM.

We first identified the main (i.e. significant,  $P < 0.05$ , according to Random Forest results) microbial phyla accounting for the variation of particular climatic legacies and that are highly related to a particular climatic legacy in a consistent way (i.e., those microbial taxa that are selected from a Random Forest model as important drivers of either AMT or PDM in more than half of the datasets; Supplementary Table 7). We then identified the shape of the relationship between climatic legacies and the relative abundance of selected taxa. All statistical analyses were independently performed with each dataset. To identify the best shape describing the relationship between climatic legacies and microbial taxa, we fitted two different functions that involve different biological interpretations: linear (positively or negatively affected by precipitation and temperature legacies) and quadratic (microbial taxa that are positively or negatively affected by intermediate levels of precipitation and temperature legacies). We selected the best model fits by following Akaike Information Criteria (AIC)<sup>50</sup>. The lower the AIC index, the better the model. Here, we consider a  $\Delta AIC > 2$  threshold to differentiate between two different models and then select the best of those models (see reference 50 for a similar approach). When both quadratic and linear models were similar (i.e.,  $\Delta AIC < 2$ ) we selected the linear (most parsimonious) model.

### **Acknowledgments**

M.D-B. acknowledge support from the Marie Skłodowska-Curie Actions of the Horizon 2020 Framework Programme H2020-MSCA-IF-2016 under REA grant agreement n° 702057. We would like to acknowledge the contribution of the BASE project partners DOI 10.4227/71/561c9bc670099, an initiative supported by Bioplatforms Australia with funds provided by the Australian Commonwealth Government through the National Collaborative Research Infrastructure Strategy. B.K.S., M.D-B are supported by the Australian Research Council projects (DP13010484 and DP170104634). D.J.E. was supported by the Hermon Slade Foundation. N.F was supported by grants from the U.S. National Science Foundation (PLR 1241629 and DEB 1542653). The work from J-Z.H. and the China dataset were supported by the Natural Science Foundation of China (Grant No. 41230857) and the Chinese Academy of Sciences (Grant No. XDB15020200). The work of FTM and the global drylands database were supported by the European Research Council (ERC Grant Agreements 242658 [BIOCOM] and 647038 [BIODESERT]).



**Competing financial interests.**

The authors declare no conflict of interest.

**Data accessibility**

Data associated with this paper has been deposited in figshare: <https://figshare.com/s/e3e47dc51ac2090f38bb> (DOI: 10.6084/m9.figshare.5048311).

**Author contribution**

M.D-B. conceived the idea of this study in consultation with N.F. The microbial datasets of the Global-Drylands, Americas, Australia, China and New South Wales datasets were originally compiled by F.T.M/B.K.S./M.D-B, N.F., A.B., J-Z.H./Y-R.L./J-T.W. and D.J.E./B.K.S./K.H., respectively. M.D-B. conducted statistical modeling. The manuscript was written by M.D-B with contributions from all co-authors.

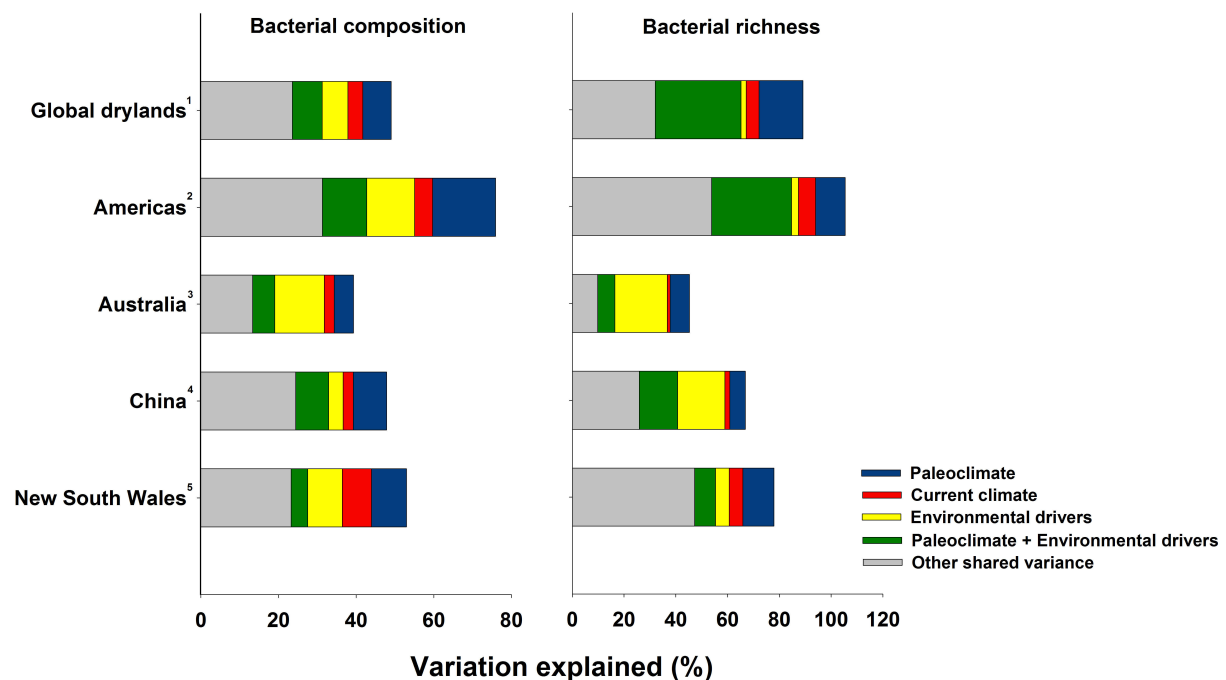
**References**

1. Monger, C. et al. Legacy effects in linked ecological–soil–geomorphic systems of drylands. *Front Ecol Environ* 13, 13–19 (2015).
2. Gajewski K. Impact of Holocene climate variability on Arctic vegetation. *Glob Planet Chang* 133, 272–287 (2015).
3. Svenning J-C. et al. The Influence of Paleoclimate on Present-Day Patterns in Biodiversity and Ecosystems. *Ann Rev Ecol Evol* 46, 551–572 (2015).
4. Lyons, S.K. et al. Holocene shifts in the assembly of plant and animal communities implicate human impacts. *Nature* 529, 80–83 (2016).
5. Martiny, J.B.H. History Leaves Its Mark on Soil Bacterial Diversity. *mBio* 7, e00784-16 (2016).
6. Andam C.P. et al. A Latitudinal Diversity Gradient in Terrestrial Bacteria of the Genus *Streptomyces*. *mBio* 7, e02200-15 (2016).
7. Wagg C. et al. Soil biodiversity and soil community composition determine ecosystem multifunctionality. *Proc Natl Acad Sci U.S.A* 111, 5266–5270 (2014).
8. Delgado-Baquerizo, M. et al. Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nat Commun* 28, 10541 (2016).
9. van der Heijden M.G. et al. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol Lett* 11, 296–310 (2008).

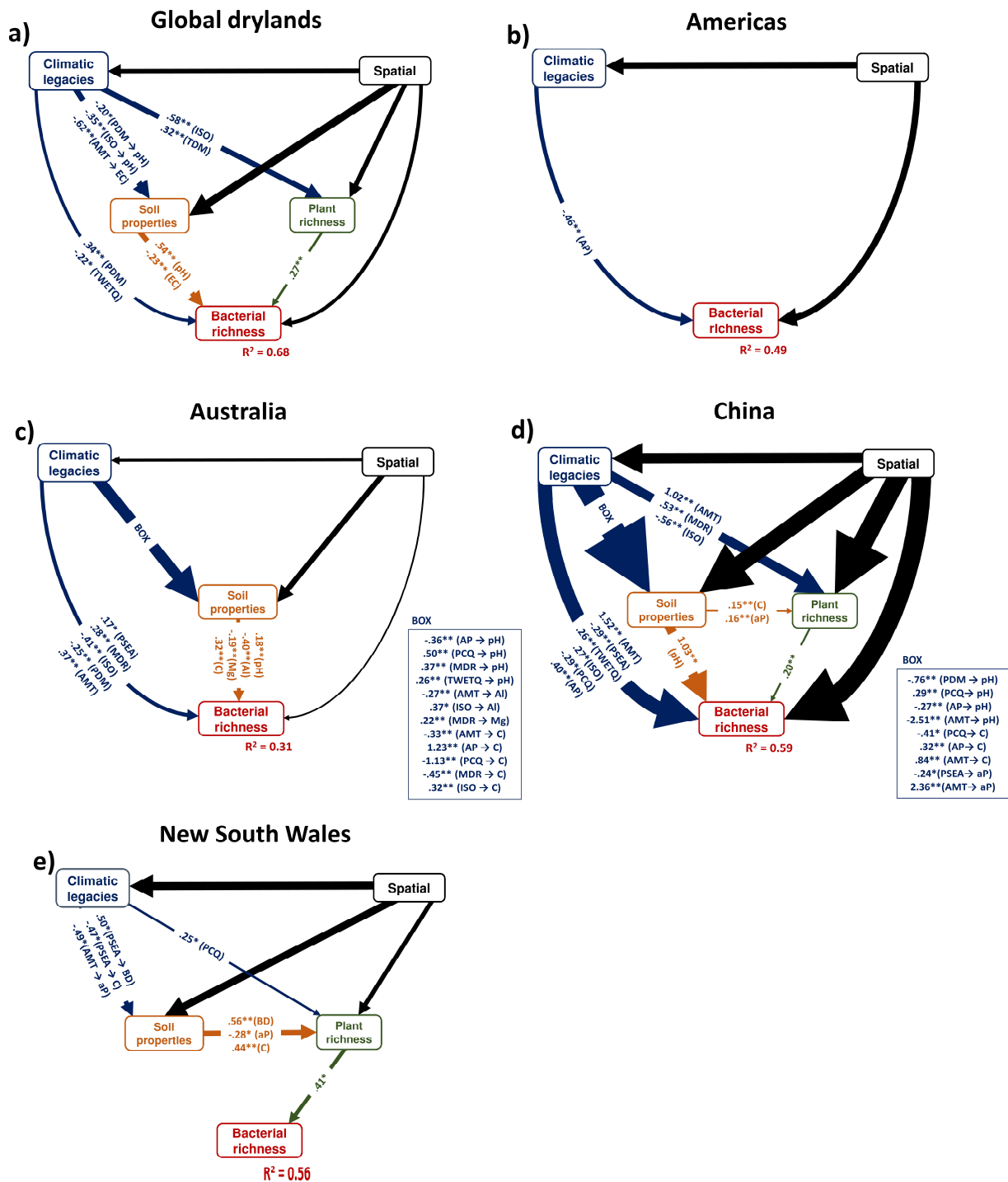
10. Oliverio, A. et al. Identifying the microbial taxa that consistently respond to soil warming across time and space. *Glob Chang Biol* doi: 10.1111/gcb.13557 (2017).
11. Atkinson, T.C. et al. Seasonal temperatures in Britain during the past 22,000 years, reconstructed using beetle remains. *Nature* 325, 587 - 592 (1987).
12. Fordham, D.A. et al. PaleoView: a tool for generating continuous climate projections spanning the last 21000 years at regional and global scales. *Ecography* DOI: 10.1111/ecog.03031 (2017).
13. Lauber, C.L. et al. Soil pH as a predictor of soil bacterial community structure at the continental scale: a pyrosequencing-based assessment. *Appl Environ Microbiol* 75, 5111-5120 (2009).
14. Delgado-Baquerizo, M. et al. Carbon content and climate variability drive global soil bacterial diversity patterns. *Ecol Monogr* 86, 373–390 (2016).
15. Maestre F.T. et al. Increasing aridity reduces soil microbial diversity and abundance in global drylands. *Proc Natl Acad Sci U.S.A* 112, 15684–15689 (2015).
16. Vitousek, P.M. *Nutrient Cycling and Limitation: Hawai'i as a Model System* (Princeton University Press, New Jersey, NY, 2004).
17. Wardle, D.A. et al. Ecosystem properties and forest decline in contrasting long-term chronosequences. *Science* 305, 509-513 (2004).
18. Prober, S.M. et al. Plant diversity predicts beta but not alpha diversity of soil microbes across grasslands worldwide. *Ecol Lett* 18, 85–95 (2015).
19. Leff, J.W. et al. Plant domestication and the assembly of bacterial and fungal communities associated with strains of the common sunflower, *Helianthus annuus*. *New Phytologist* 214, 412–423 (2017).
20. Leff, J.W. et al. Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10967-72 (2015).
21. Trivedi, P. et al. Response of Soil Properties and Microbial Communities to Agriculture: Implications for Primary Productivity and Soil Health Indicators. *Front Plant Sci* 7, 990 (2016).
22. World Bank. *World Development Report, Agriculture for Development* (World Bank, Washington, DC 2008).
23. Legendre, P. et al. Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis. *J Plant Ecol* 1, 3-8 (2008).

24. Ramirez, K.S. et al. Biogeographic patterns in belowground diversity in New York City's Central Park are similar to those observed globally. *Proc R Soc Lond [Biol]* 281, 20141988 (2014).
25. Bissett, A. et al. Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database. *GigaScience* 20165, 21 (2016).
26. Wang, J.-T. et al. Coupling of soil prokaryotic diversity and plant diversity across latitudinal forest ecosystems. *Sci Rep* 5, 19561 (2015).
27. Eldridge, D.J. et al. Competition drives the response of soil microbial diversity to increased grazing by vertebrate herbivores. *Ecology* DOI: 10.1002/ecy.1879 (2017).
28. Nogués-Bravo, D. et al. Amplified plant turnover in response to climate change forecast by Late Quaternary records. *Nat Clim Chang* 6, 1115-1119 (2016).
29. Fieseler L. et al. Discovery of the novel candidate phylum "Poribacteria" in marine sponges. *Appl Environ Microbiol* 70, 3724-32 (2004).
30. Youssef, N.H. et al. In Silico Analysis of the Metabolic Potential and Niche Specialization of Candidate Phylum "Latescibacteria" (WS3). *PLoS ONE* 10, e0127499 (2015).
31. Schlesinger, W.H. *Biogeochemistry, an analysis of global change* (Academic Press, San Diego, CA, USA, 1996).
32. United Nations Environment Programme *World Atlas of Desertification* (Edward Arnold, London, UK) (2002)
33. Herlemann, D. P. et al. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* 5, 1571–1579 (2011).
34. Maestre, F.T. et al. Plant Species Richness and Ecosystem Multifunctionality in Global Drylands. *Science* 335, 214-218 (2012).
35. Bates, S. et al. Bacterial communities associated with the lichen symbiosis. *Applied Environ. Microbiol.* 77, 1309-1314 (2011).
36. Caporaso, J.G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336 (2010).
37. Lane, DJ (1991) *Nucleic Acid Techniques in Bacterial Systematics* (John Wiley and Sons, NY, USA).

38. Schloss, P.D. et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microb.* 75, 7537-7541.
39. Edgar R.G. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* 10, 996-998 (2013).
40. Gent, P.R. et al. The Community Climate System Model Version 4. *J. Clim.* 24, 4973-4991 (2011).
41. Bystriakova, N. et al. Present, past and future of the European rock fern *Asplenium fontanum*: combining distribution modelling and population genetics to study the effect of climate change on geographic range and genetic diversity. *Ann Bot* doi: 10.1093/aob/mct274 (2013)
42. Tallavaara, M. et al. Human population dynamics in Europe over the Last Glacial Maximum. *Proc. Natl. Acad. Sci. USA* 112, 8232–8237 (2015).
43. Delgado-Baquerizo, M. et al. Biogeographic bases for a shift in crop C:N:P stoichiometries during domestication. *Ecol. Lett.* 19, 564-575.
44. Katz, M.H. *Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers* (Cambridge University Press, Cambridge, UK, 2006).
45. Oksanen, J. et al. *vegan: Community Ecology Package*. R package version 2.3-0 (2015).
46. Breiman, L. *Machine Learning*, 45, 5 (2001).
47. Archer, E. *rfPermute: Estimate Permutation p-Values for Random Forest Importance Metrics*. R package version 1.5.2 (2016).
48. Grace J.B. *Structural Equation Modeling Natural Systems* (Cambridge Univ. Press, Cambridge) (2006).
49. Schermelleh-Engel, K. et al. Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods Psychol. Res.* 8, 23–74 (2003).
50. Delgado-Baquerizo, M. et al. Lack of functional redundancy in the relationship between microbial diversity and ecosystem functioning. *J. Ecol.* 104, 936–946 (2016).

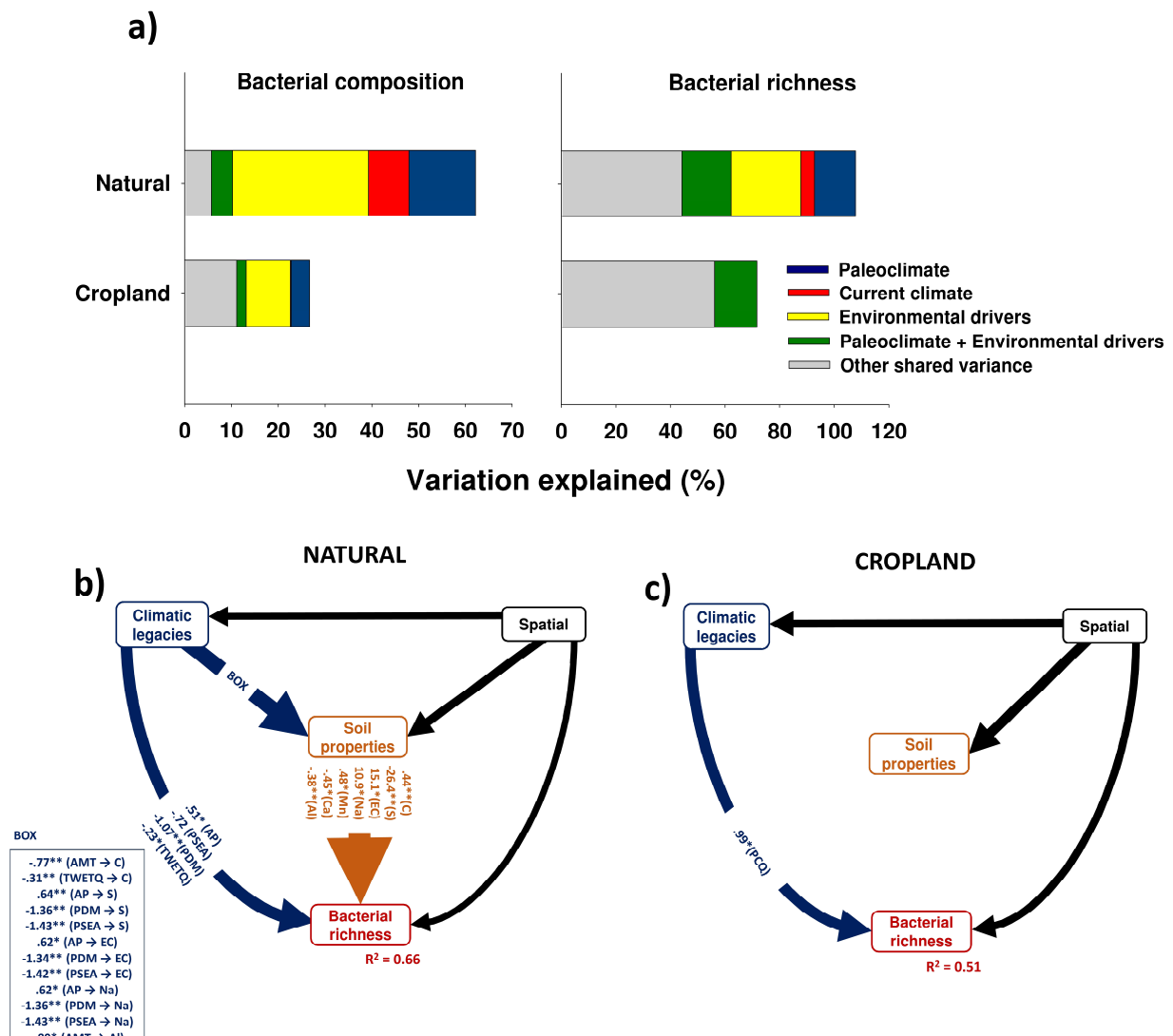


**Figure 1.** Relative contribution of the different predictors used to model bacterial composition and diversity. Panels represent results from Variation Partitioning modelling aiming to identify the percentage variance of bacterial community composition explained by past and current climate variables across five independent large scale datasets. Unique and shared variance from Last Glacial Maximum and mid-Holocene in predicting bacterial community composition and richness were merged in this figure for simplicity. An alternative version of this figure showing the unique and shared variance of each group of predictors can be found in Supplementary Figs. 2 and 3. Further information on the datasets used in these analyses can be found in: 1) Maestre et al.<sup>15</sup>, 2) Ramirez et al.<sup>24</sup>, 3) Bissett et al.<sup>25</sup>, 4) Wang et al.<sup>26</sup> and 5) Eldridge et al.<sup>27</sup>.



**Figure 2.** Structural equation model accounting for the direct and indirect (plant diversity and/or soil properties) effects of climatic legacies on the diversity of bacteria across the five datasets used. Numbers adjacent to arrows are path coefficients ( $P$  values), and indicative of the standardized effect size of the relationship. Spatial influence (latitude and longitude) were included to control

spatial autocorrelation; however, in this case, path coefficients were not included for simplicity. The size of the arrow represents the strength of the relationship when significant. All variables are included as independent observable variables. We grouped the different categories of predictors (soil properties, climate legacies and spatial) in the same box in the model for graphical simplicity. For the same reason, we only included those direct effects from climate legacies on soil properties that could indirectly affect the diversity of bacteria. The rest of the effects from climate legacies on soil properties are available in Supplementary Tables 5 and 6.  $R^2$  = the proportion of variance explained. Acronyms of climatic variables are shown in Supplementary Table 1. Lower case letters adjacent to a particular chemical element indicates that the element is in an “available” (a) or “occluded” (o) form.  $R^2$  = the proportion of variance explained. Significance levels of each predictor are  $*P < 0.05$ ,  $**P < 0.01$ . Environmental drivers include plant diversity (Global drylands, China and New South Wales) and/or soil properties and geographical location.



**Figure 3.** Relative contribution of the different predictors used to model bacterial composition and diversity in croplands ( $n = 66$ ) and natural ( $n = 66$ ) ecosystems from Australia (a). Unique and shared variance from Last Glacial Maximum and mid-Holocene in predicting bacterial community composition and richness were merged in this figure for simplicity. An alternative version of this figure showing the unique and shared variance of each group of predictors is available in Supplementary Fig. 10. Panels (b-c) represent structural equation models accounting for the direct and indirect effects (plant diversity and/or soil properties) of climatic legacies on the diversity of soil bacteria in this database. Rest of caption as in Fig. 2.



## **Supplementary Information**

### **Paleoclimate explains a unique proportion of the global variation in soil bacterial communities**

Manuel Delgado-Baquerizo, Andrew Bissett, David J. Eldridge, Fernando T. Maestre, Ji-Zheng He, Jun-Tao Wang, Kelly Hamonts, Yu-Rong Liu, Brajesh K. Singh, Noah Fierer.

**Correspondence to:** [M.delgadobaquerizo@gmail.com](mailto:M.delgadobaquerizo@gmail.com)

#### **This PDF file includes:**

Appendices S1-S3

Supplementary Tables 1-12

Supplementary Figures 1-15

## **Appendix S1. Climatic data cross-validation**

*It can be argued that the paleoclimatic data used here do not have enough temporal resolution to represent the paleoclimate of a particular region. Because of this, besides using CCSM4 ([www.worldclim.org](http://www.worldclim.org)) to calculate paleoclimate, we also calculated three climatic variables (average precipitation and minimum and maximum temperature) for 49 unique locations from North America (from the Global-drylands and Americas datasets) using the paleoclimatic data available from references 51 and 52. In these references, authors summarized information for North America for several climatic variables (including average precipitation and minimum and maximum temperature) into decadal averages (from ~22000y before present<sup>12</sup>). Paleoclimatic data used in this manuscript were identical to those calculated using references 51 and 52 (Supplementary Table 11) supporting the paleoclimatic indexes used here.*

## **Appendix S2. Climatic legacy indexes cross-validation**

It could be argued that the approach used here to calculate climatic legacies is not representative for changes in climates over last ~20000 years. For the climatic variables used here (several representations of temperature and precipitation), climatic changes during the last ~20000 years (Supplementary Figs. 12, 13 and 14) were largely unidirectional and many time linear; including for example large –but location dependent– increases in temperature. These analyses support our approach and suggest that calculating differences between paleo and current climate can actually account for major changes in temperature and precipitation over the last ~20000y<sup>12</sup>.

To further support our approach, we correlated information on precipitation, and maximum or minimum temperature legacies used here (e.g.  $\text{Annual precipitation}_{\text{Current climate}} - \text{Annual precipitation}_{\text{Last Glacial Maximum}}$  for each site; [www.worldclim.org](http://www.worldclim.org)) with the rate of change of precipitation, and maximum or minimum temperature calculated using the slope of a linear regression between each of these climatic parameters and time (years) using data from references 51 and 52. These analyses aim to explicitly evaluate whether the approach used here –based on differences between two snapshots: one ~20000y before the present and another one from current climate– can account for long-term changes in precipitation and temperature over thousands of years. Interestingly, values from the climatic legacy index used here were strongly correlated to the rate of change in climate over thousands of years (Supplementary Table 12). Therefore, our approach is proven to be a good approach to account for climatic legacies that occurred for the period of time between 0 and 20000 years before the present.

***Appendix S3. Rationale on the direct effects from climatic legacies on soil properties and plant and microbial communities.***

The direct effect of climatic legacies (measured as the temperature and precipitation anomalies<sup>12</sup> between an estimate of climate 20,000 year ago and another estimate for the present day) on soil properties and plant communities is strongly supported by current literature<sup>2-4,15-16</sup>. One of the most significant developments in the field of soil sciences in recent years involves the increased recognition that most soils are polygenetic, that is, archival products of pedogenic processes that vary widely over time (ref. 53). Empirical support for this point comes from soil age chronosequence studies. For example, Schlesinger et al.<sup>54</sup>, has shown that C accumulates over millennia at different rates across different ecosystem types from tundra to tropical forest when comparing similar soil age ranges. Thus, the climatic changes over millennia (climatic legacies) can easily alter the rates of C sequestration in soil. A more recent study<sup>55</sup> provided further evidence that soil properties such as soil C are strongly influenced by paleoclimate compared with current climate, and that the direct effects from past climates are stronger than those from current climates (Fig. S8 in ref. 55). From a mechanistic point of view, the direct effects of climatic legacies on soil properties, plant diversity and microbial communities should follow a similar logic to those from contemporary climate. However, unlike contemporary climate, the direct effects of climatic legacies on soil properties, microbial community and plant diversity come from the long-term climate history of a region. Some examples follow. Let's imagine a forest under a current wet climate, which was previously a grassland ecosystem under a drier paleoclimate. This forest (compared to a similar forest with wet paleo and wet current climates) may now have a lower amount of soil C, plant and microbial diversity than expected based on its current climate as a consequence of the changing climate over millennia. Having a dry climate over long periods (centuries to thousands of years) should both have promoted strong reductions in biological activity and C sequestration and limited the number of species being able to survive under these conditions. Consequently, the precipitation history of a region can directly influence soil properties, microbial communities and plant diversity. Similarly, very low temperature in a glaciated area over prolonged periods *per se* may directly explain low levels plant and microbial diversity in a region (i.e., not many species can survive in such as harsh climate), compared to other location on Earth with higher paleo-temperatures. Note that all datasets included in this manuscript are large-scale datasets going from regional to global spatial scales –meaning that

multiple paleoclimatic conditions, therefore different climatic legacies, occurred in these locations simultaneously. Also, not all terrestrial ecosystems were under ice during last glaciation. For examples, many places from the tropics, Australia, Africa or South America did not suffer such glaciation effects, rather they experienced a wide variety of paleoclimates, including both extreme and more benign conditions. Another example for a direct effect of climatic legacies on soil properties would be a region with a particular paleoclimate in terms of precipitation in the driest month or isothermality, which may have strongly influenced soil texture by promoting high/low rates of bedrock weathering –a major driver of texture from centuries to millennia. Direct effects from climatic legacies on microbial communities also include the impacts derived from rapid climatic changes in the past –which mostly occurred prior to 10000 years ago<sup>12</sup>– and that have left a strong signature on the contemporary structure of soil bacterial communities. In this respect, a direct effect from paleoclimate on soil microbial communities might have occurred in the past (e.g. in response to a particularly drastic climatic event), but their consequences might still be detectable today.

**Supplementary Table 1.** Bioclimatic variables included in this study.

Worldclim number	Bioclimatic variable	Acronym
1	Annual Mean Temperature	AMT
2	Mean Diurnal Range	MDR
3	Isothermality	ISO
4	Temperature Seasonality	TSEA
5	Max Temperature of Warmest Month	MAXTWM
6	Min Temperature of Coldest Month	MINTCM
7	Temperature Annual Range	TRANGE
8	Mean Temperature of Wettest Quarter	TWETQ
9	Mean Temperature of Driest Quarter	TDQ
10	Mean Temperature of Warmest Quarter	TWARQ
11	Mean Temperature of Coldest Quarter	TCQ
12	Annual Precipitation	AP
13	Precipitation of Wettest Month	PWETM
14	Precipitation of Driest Month	PDM
15	Precipitation Seasonality	PSEA
16	Precipitation of Wettest Quarter	PWETQ
17	Precipitation of Driest Quarter	PDQ
18	Precipitation of Warmest Quarter	PWARQ
19	Precipitation of Coldest Quarter	PCQ

**Supplementary Table 2.** Correlation (Pearson) among bioclimatic variables across different time periods. Worldclim number of climatic variables are shown in Supplementary Table 1.

*Supplementary Table 2 is available online as a Separate Excel file under the Supporting Information for this article.*

**Supplementary Table 3.** Complete list of predictors available for each database and used for the Variation Partitioning and Random Forest modeling. Acronyms of climatic variables are shown in Supplementary Table 1.

Dataset	Variable	Group	Predictor abbreviation	Predictor spelled name
Global drylands	1	Climatic legacy	AMT legacy	Annual Mean Temperature legacy
	2		MDR legacy	Mean Diurnal Range legacy
	3		ISO legacy	Isothermality legacy
	4		TWETQ legacy	Mean Temperature of Wettest Quarter legacy
	5		AP legacy	Annual temperature legacy
	6		PDM Legacy	Precipitation of Driest Month
	7		PSEA Legacy	Precipitation Seasonality legacy
	8		PCQ Legacy	Precipitation of Coldest Quarter legacy
	9	Geographical location	Latitude	Latitude
	10		Longitude	Longitude
	11	Soil properties	pH	pH
	12		C	Soil carbon
	13		Sand	Texture - sand content (%)
	14		C:N	Soil C:N ratio
	15		Total P	Soil total phosphorus
	16		EC	Soil electrical conductivity
	17		N mineralization	Potential net nitrogen mineralization rate
	18		Phenols	Dissolved phenols
	19		Proteins (PRO)	Dissolved proteins
	20		Carbohydrates (aC)	Carbohydrates
	21		Aromatic c.	Aromatic compounds
	22		Available N (aN)	Available nitrogen
	23		Available P (aP)	Available phosphorus
	24		Amino acids	Amino acids
	25	Plant diversity	Plant richness	Plant richness
Americas	1	Climatic legacy	AMT legacy	Annual Mean Temperature legacy
	2		MDR legacy	Mean Diurnal Range legacy
	3		ISO legacy	Isothermality legacy
	4		TWETQ legacy	Mean Temperature of Wettest Quarter legacy
	5		AP legacy	Annual temperature legacy
	6		PDM Legacy	Precipitation of Driest Month
	7		PSEA Legacy	Precipitation Seasonality legacy

	8		PCQ Legacy	Precipitation of Coldest Quarter legacy
	9	Geographical location	Latitude	Latitude
	10		Longitude	Longitude
	11	Soil properties	pH	pH
	12		C	Soil carbon
	13		Sand	Texture - sand content (%)
	14		C:N	Soil C:N ratio
	15		C mineralization (C min)	Carbon mineralization
Australia	1	Climatic legacy	AMT legacy	Annual Mean Temperature legacy
	2		MDR legacy	Mean Diurnal Range legacy
	3		ISO legacy	Isothermality legacy
	4		TWETQ legacy	Mean Temperature of Wettest Quarter legacy
	5		AP legacy	Annual temperature legacy
	6		PDM Legacy	Precipitation of Driest Month
	7		PSEA Legacy	Precipitation Seasonality legacy
	8		PCQ Legacy	Precipitation of Coldest Quarter legacy
	9	Geographical location	Latitude	Latitude
	10		Longitude	Longitude
	11	Soil properties	pH	pH
	12		C	Soil carbon
	13		Sand	Texture - sand content (%)
	14		EC	Soil electrical conductivity
	15		S	Sulfur
	16		Zn	Zinc
	17		K	Potassium
	18		Available K (aK)	Available potassium
	19		B	Boron
	20		Ca	Calcium
	21		Cu	Copper
	22		Na	Sodium
	23		Al	Aluminum
	24		Fe	Iron
	25		Mg	Magnesium
	26		Mn	Manganese
	27		Available P (aP)	Available phosphorus
	28		DIN	Dissolved inorganic N
China	1	Climatic legacy	AMT legacy	Annual Mean Temperature legacy
	2		MDR legacy	Mean Diurnal Range legacy
	3		ISO legacy	Isothermality legacy

	4		TWETQ legacy	Mean Temperature of Wettest Quarter legacy
	5		AP legacy	Annual temperature legacy
	6		PDM Legacy	Precipitation of Driest Month
	7		PSEA Legacy	Precipitation Seasonality legacy
	8		PCQ Legacy	Precipitation of Coldest Quarter legacy
	9	Geographical location	Latitude	Latitude
	10		Longitude	Longitude
	11	Soil properties	pH	pH
	12		C	Soil carbon
	13		Sand	Texture - sand content (%)
	14		C:N	Soil C:N ratio
	15		Available P (aP)	Available phosphorus
	16	Plant diversity	Plant richness	Plant richness
New South Wales	1	Climatic legacy	AMT legacy	Annual Mean Temperature legacy
	2		MDR legacy	Mean Diurnal Range legacy
	3		ISO legacy	Isothermality legacy
	4		TWETQ legacy	Mean Temperature of Wettest Quarter legacy
	5		AP legacy	Annual temperature legacy
	6		PDM Legacy	Precipitation of Driest Month
	7		PSEA Legacy	Precipitation Seasonality legacy
	8		PCQ Legacy	Precipitation of Coldest Quarter legacy
	9	Geographical location	Latitude	Latitude
	10		Longitude	Longitude
	11	Soil properties	pH	pH
	12		C	Soil carbon
	13		Sand	Texture - sand content (%)
	14		C:N	Soil C:N ratio
	15		Bulk density	Bulk density
	16		Available P (aP)	Available phosphorus
	17	Plant diversity	Plant richness	Plant richness



**Supplementary Table 4.** Results from Random Forest analyses aiming to identify the most important predictors of soil bacterial richness and composition. Acronyms of climatic variables are shown in Supplementary Table 1. Importance is calculated as the % of increase in the mean square error in our models. Importance (Variation explained by Random Forest models).

Dataset	Diversity			Composition					
	Richness			Microbial comm.1			Microbial comm.2		
	Variable	Importance (42.32%)	P-value	Variable	Importance (72.48%)	P-value	Variable	Importance (60.01%)	P-value
<b>Global drylands</b>	PDM Legacy	17.065	0.001	pH	32.821	0.001	C	24.168	0.001
	Plant richness	15.354	0.002	Latitude	14.520	0.001	PDM Legacy	15.879	0.001
	Longitude	11.344	0.004	PSEA Legacy	11.902	0.002	Latitude	14.063	0.001
	Proteins	10.253	0.005	AP legacy	10.843	0.002	Longitude	13.415	0.001
	Sand	9.256	0.007	PCQ Legacy	9.882	0.005	Carbohydrates	11.932	0.002
	C	8.895	0.020	Proteins	9.380	0.009	Sand	11.431	0.004
	ISO legacy	7.530	0.013	Aromatic c.	8.853	0.012	Available P	9.983	0.007
	Aromatic c.	7.187	0.029	Longitude	7.620	0.030	AP legacy	6.988	0.026
	Latitude	6.646	0.034	Amino acids	6.677	0.033	ISO legacy	6.617	0.011
	AMT legacy	6.453	0.024	ISO legacy	5.905	0.030	Plant richness	6.369	0.023
	TWETQ legacy	6.315	0.040	N mineralization	5.883	0.052	PSEA Legacy	6.057	0.034
	PCQ Legacy	6.239	0.035	Phenols	5.466	0.081	PCQ Legacy	5.796	0.057
	EC	6.192	0.034	MDR legacy	4.964	0.077	AMT legacy	5.593	0.049
	pH	5.981	0.039	TWETQ legacy	4.858	0.080	Amino acids	3.887	0.137
	Carbohydrates	5.094	0.090	Available N	4.779	0.064	TWETQ legacy	3.119	0.161
	Phenols	4.966	0.131	EC	4.638	0.071	Aromatic c.	3.061	0.302
	Available P	4.834	0.052	AMT legacy	4.188	0.072	pH	2.706	0.235
	AP legacy	4.816	0.092	Carbohydrates	3.992	0.134	Phenols	2.508	0.353
	N mineralization	2.946	0.195	PDM Legacy	3.665	0.069	Proteins	2.374	0.357
	Total P	2.019	0.275	Total P	3.616	0.125	Total P	2.056	0.277
	PSEA Legacy	1.848	0.364	Sand	3.101	0.182	EC	1.419	0.300
	MDR legacy	1.650	0.291	C	2.366	0.338	MDR legacy	1.220	0.314
	Available N	1.611	0.350	Plant richness	2.274	0.169	N mineralization	0.617	0.442
	C:N	1.084	0.292	Available P	1.931	0.209	C:N	-0.589	0.507
	Amino acids	0.832	0.514	C:N	1.330	0.248	Available N	-1.211	0.748
<b>Americas</b>	AP legacy	4.332	0.020	pH	72.608	0.000	Latitude	21.831	0.001
	Longitude	3.884	0.030	C	33.853	0.002	Longitude	20.261	0.001
	ISO legacy	3.587	0.020	C:N	32.644	0.001	TWETQ legacy	13.166	0.003

	Latitude	3.082	0.010	Longitude	22.400	0.010	AMT legacy	11.582	0.005
	PCQ Legacy	2.630	0.080	C mineralization	20.862	0.025	Sand	11.095	0.007
	C:N	2.272	0.090	AMT legacy	17.545	0.029	pH	10.613	0.008
	PDM Legacy	2.015	0.030	Latitude	15.343	0.055	AP legacy	10.099	0.012
	TWETQ legacy	1.893	0.130	AP legacy	14.733	0.047	ISO legacy	10.036	0.011
	AMT legacy	1.796	0.100	PCQ Legacy	10.189	0.125	C:N	7.916	0.051
	C	1.152	0.270	PDM Legacy	9.912	0.099	C	6.951	0.054
	MDR legacy	0.906	0.360	TWETQ legacy	9.761	0.129	PCQ Legacy	6.812	0.066
	C mineralization	0.853	0.350	MDR legacy	7.345	0.189	MDR legacy	4.049	0.152
	PSEA Legacy	0.779	0.340	Sand	3.456	0.298	PSEA Legacy	3.653	0.194
	pH	0.141	0.710	PSEA Legacy	2.442	0.367	PDM Legacy	3.548	0.187
	Sand	-0.470	0.640	ISO legacy	1.965	0.362	C mineralization	-0.982	0.688
Dataset	Variable	Importance (45.78%)	P-value	Variable	Importance (82.36%)	P-value	Variable	Importance (68.67%)	P-value
Australia	Al	30.368	0.001	pH	55.153	0.001	Latitude	29.322	0.002
	pH	28.485	0.001	Al	30.926	0.001	Longitude	19.685	0.002
	Ca	25.211	0.001	Fe	28.776	0.001	Ca	18.427	0.002
	Longitude	23.571	0.001	C	23.177	0.001	Available P	13.882	0.002
	Mn	20.931	0.001	Ca	19.224	0.001	PCQ Legacy	12.984	0.002
	Latitude	19.423	0.001	Longitude	19.140	0.001	AP legacy	12.569	0.002
	AP legacy	18.373	0.001	Available K	18.455	0.001	MDR legacy	12.504	0.002
	Fe	18.127	0.001	PCQ Legacy	17.442	0.001	C	11.812	0.008
	C	16.515	0.005	AP legacy	16.448	0.001	AMT legacy	11.243	0.002
	PSEA Legacy	16.408	0.001	Latitude	15.671	0.002	pH	11.212	0.010
	Na	15.655	0.009	K	15.645	0.003	Fe	10.249	0.024
	EC	15.594	0.028	MDR legacy	15.627	0.001	B	9.599	0.150
	PDM Legacy	15.403	0.001	EC	15.347	0.026	ISO legacy	9.454	0.002
	Mg	15.094	0.033	AMT legacy	14.764	0.001	Mn	9.012	0.102
	PCQ Legacy	14.619	0.002	Mn	14.226	0.005	Cu	8.984	0.102
	MDR legacy	14.345	0.001	TWETQ legacy	14.203	0.001	PDM Legacy	8.851	0.006
	S	14.257	0.033	Available P	13.717	0.012	Zn	8.843	0.088
	ISO legacy	13.718	0.001	Sand	13.324	0.024	PSEA Legacy	8.584	0.010
	AMT legacy	12.608	0.002	PDM Legacy	12.573	0.001	Al	8.289	0.094
	B	12.402	0.231	Zn	11.240	0.049	Available K	8.250	0.281
	Sand	11.954	0.142	B	10.336	0.262	DIN	7.917	0.170
	Available K	11.523	0.238	ISO legacy	10.209	0.002	Mg	7.448	0.547
	K	11.490	0.265	S	10.175	0.143	EC	7.368	0.665
	DIN	10.448	0.145	Cu	10.117	0.111	K	6.754	0.659
	Zn	9.551	0.221	Na	9.395	0.185	Sand	6.158	0.669

	TWETQ legacy	8.132	0.049	PSEA Legacy	9.132	0.019	TWETQ legacy	5.221	0.178
	Cu	7.394	0.492	Mg	8.995	0.401	Na	3.141	0.990
	Available P	6.209	0.679	DIN	8.249	0.177	S	1.650	0.988
Dataset	Variable	Importance (62.28%)	P-value	Variable	Importance (87.85%)	P-value	Variable	Importance (85.09%)	P-value
China	pH	49.286	0.001	pH	52.865	0.001	Latitude	24.599	0.001
	Latitude	20.295	0.002	PDM Legacy	18.905	0.001	Longitude	24.478	0.001
	Longitude	19.569	0.002	Latitude	18.847	0.001	AMT legacy	16.576	0.001
	MDR legacy	19.422	0.001	Longitude	17.256	0.001	Available P	15.070	0.020
	PSEA Legacy	19.243	0.002	AMT legacy	17.171	0.001	MDR legacy	14.910	0.002
	PDM Legacy	19.002	0.001	PSEA Legacy	14.729	0.003	AP legacy	14.857	0.003
	C	17.921	0.004	TWETQ legacy	14.702	0.001	TWETQ legacy	14.151	0.003
	AP legacy	17.208	0.002	C:N	13.979	0.011	Sand	14.130	0.026
	AMT legacy	15.537	0.002	C	13.367	0.012	ISO legacy	13.217	0.002
	Available P	14.402	0.016	PCQ Legacy	11.938	0.007	PSEA Legacy	12.862	0.012
	ISO legacy	14.116	0.001	MDR legacy	11.802	0.005	pH	12.311	0.072
	TWETQ legacy	13.524	0.004	Available P	11.691	0.020	C:N	12.215	0.112
	Plant richness	12.623	0.049	AP legacy	11.455	0.014	Plant richness	9.505	0.274
	C:N	11.069	0.181	Plant richness	10.973	0.037	PDM Legacy	8.704	0.047
	PCQ Legacy	10.893	0.025	ISO legacy	8.782	0.006	PCQ Legacy	8.536	0.113
	Sand	3.173	0.702	Sand	7.386	0.146	C	6.964	0.450
Dataset	Variable	Importance (46.24%)	P-value	Variable	Importance (57.18%)	P-value	Variable	Importance (41.84%)	P-value
New South Wales	Available P	14.467	0.002	pH	21.366	0.001	Bulk density	19.385	0.001
	Bulk density	11.885	0.009	PDM Legacy	15.609	0.002	Available P	19.220	0.001
	C	11.341	0.010	Sand	11.364	0.009	Plant richness	14.061	0.003
	Latitude	9.780	0.019	Bulk density	10.006	0.022	C	9.392	0.017
	PDM Legacy	9.622	0.008	Latitude	9.937	0.020	Latitude	9.189	0.030
	Plant richness	9.249	0.022	C	8.795	0.045	AP legacy	6.840	0.039
	PSEA Legacy	8.846	0.012	Available P	8.565	0.040	PDM Legacy	5.586	0.047
	PCQ Legacy	7.579	0.022	PSEA Legacy	8.060	0.020	Sand	4.445	0.140
	Sand	7.107	0.074	Longitude	7.486	0.050	pH	4.012	0.167
	AMT legacy	5.507	0.032	AP legacy	6.864	0.036	ISO legacy	3.892	0.047
	AP legacy	5.477	0.104	ISO legacy	5.004	0.048	PSEA Legacy	3.799	0.139
	pH	5.410	0.091	Plant richness	3.593	0.198	MDR legacy	3.771	0.084
	Longitude	5.198	0.105	MDR legacy	3.421	0.124	C:N	3.231	0.234
	ISO legacy	2.756	0.140	PCQ Legacy	3.115	0.162	Longitude	2.544	0.253
	TWETQ legacy	2.400	0.225	C:N	2.517	0.321	AMT legacy	1.996	0.266
	MDR legacy	1.871	0.303	TWETQ legacy	2.473	0.206	PCQ Legacy	0.649	0.411
	C:N	1.629	0.403	AMT legacy	2.264	0.230	TWETQ legacy	-2.121	0.755

**Supplementary Table 5.** Standardized direct effects from SEMs in Fig. 2 and Supplementary Figs 6 and 7.

*Supplementary Table 5 is available online as a Separate Excel file under the Supporting Information for this article.*

**Supplementary Table 6.** Correlations (standardized effects) from SEMs in Fig. 2 and Supplementary Figs 6 and 7.

*Supplementary Table 6 is available online as a Separate Excel file under the Supporting Information for this article.*

**Supplementary Table 7.** Results from Random Forest analyses aiming to identify the most important bacterial composition predictors of selected paleoclimatic legacies (AMT or PDM). Acronyms of climatic variables are shown in Supplementary Table 1. Importance is calculated as the % of increase in the mean square error in our models.

*Supplementary Table 7 is available online as a Separate Excel file under the Supporting Information for this article.*

**Supplementary Table 8.** Shape of the relationship between climatic legacies and main bacterial taxaSupplementary. To identify the best shape describing the relationship between climatic legacies and bacterial taxa, we fitted two different functions: linear and quadratic. We selected the best model fits by following Akaike Information Criteria. The lower the AIC index the better the model. We considered a  $\Delta AIC > 2$  threshold to differentiate between two different models and then select the best of those models. When both quadratic and linear models were similar between them (i.e.  $\Delta AIC < 2$ ) we selected the most parsimonious model (i.e. the linear model). P = Positive relationship; N = Negative relationship; U = U-shaped relationship;  $\cap$  = hump-shaped relationship. CC = Current climate; LGM = Last Glacial Maximum. AMT = Annual Mean Temperature; TSEA = Temperature Seasonality; AP = Annual Precipitation; PDM = Precipitation in the Driest Month. N = number of datasets. CC/LGM indicates the range in climatic variables for these two periods in each dataset. Acronyms of climatic variables are shown in Supplementary Table 1.

Climatic variable	Microbial taxa	N	Global drylands	Americas	Australia	China	New South Wales
AMT (°C)	Actinobacteria	3			U (<0.001)		U (<0.001)
	Chloroflexi	3				N (<0.001)	
	FBP	3	U (<0.001)			P (<0.001)	
PDM (mm)	Nitrospirae	4	P (0.021)		$\cap$ (<0.001)	N (<0.001)	$\cap$ (<0.001)
	AD3	3			$\cap$ (0.045)	P (<0.001)	$\cap$ (0.045)
	Armatimonadetes	3	N (<0.001)		$\cap$ (<0.001)	U (0.007)	$\cap$ (<0.001)
	FBP	3	N (<0.001)		N (0.021)	N (<0.001)	N (0.021)
	FCPU426	3		N (<0.001)	P (0.003)	U (<0.001)	P (<0.001)
	Firmicutes	3			N (<0.001)		N (<0.001)
	Planctomycetes	3	P (0.022)		$\cap$ (0.040)	P (<0.001)	$\cap$ (0.040)
	WS3	3	P (<0.001)		P (<0.001)	U (<0.001)	P (<0.001)

**Supplementary Table 9.** Selected *a priori* degrees of freedom for our SEMs based on partial correlation (Pearson) analyses. Acronyms of climatic variables are shown in Supplementary Table 1.

Dataset	Microbial variable	Predictors		Pearson's r	P-value
<b>Drylands</b>	<b>Richness</b>	EC	ISO legacy	0.000	1.000
	<b>Composition 1</b>	Latitude	Proteins	-0.009	0.943
	<b>Composition 2</b>	Longitude	ISO legacy	0.003	0.984
<b>Americas</b>	<b>Richness</b>	Longitude	AP legacy	0.328	0.026
	<b>Composition 1</b>	Longitude	AP legacy	-0.021	0.893
	<b>Composition 2</b>	AP legacy	Sand	-0.024	0.880
<b>Australia</b>	<b>Richness</b>	Latitude	S	0.000	0.994
	<b>Composition 1</b>	TWETQ legacy	Sand	0.000	0.993
	<b>Composition 2</b>	AI	aP	0.000	0.996
<b>China</b>	<b>Richness</b>	MDR legacy	aP	-0.006	0.919
	<b>Composition 1</b>	TWETQ legacy	C	0.002	0.970
	<b>Composition 2</b>	ISO legacy	PSEA legacy	-0.006	0.924
<b>New South Wales</b>	<b>Richness</b>	AMT legacy	PSEA legacy	0.000	0.999
	<b>Composition 1</b>	Latitude	ISO legacy	-0.005	0.976
	<b>Composition 2</b>	PDM legacy	C	0.016	0.917

**Supplementary Table 10.** Measures of goodness of fit of the SEMs included in in Fig. 2 and Supplementary Figs 6 and 7.

	Microbial variable	CFI	GFI	NFI	$\chi^2$	P	RMSEA	P	df
Global-Drylands	Richness	1.000	1.000	1.000	0.014	0.907	0.000	0.915	1.000
	Composition 1	0.999	0.996	0.998	1.572	0.210	0.086	0.253	1.000
	Composition 2	0.994	0.992	0.994	4.075	0.044	0.200	0.064	1.000
Americas	Richness	1.000	0.982	0.998	0.940	0.332	0.000	0.360	1.000
	Composition 1	1.000	0.996	0.996	0.940	0.332	0.000	0.360	1.000
	Composition 2	1.000	1.000	1.000	0.003	0.959	0.000	0.961	1.000
Australia	Richness	1.000	1.000	1.000	1.248	0.264	0.000	0.525	1.000
	Composition 1	1.000	1.000	1.000	2.435	0.119	0.052	0.344	1.000
	Composition 2				5.110	0.024	0.080	0.134	1.000
China	Richness	1.000	1.000	1.000	0.176	0.675	0.000	0.771	1.000
	Composition 1	1.000	1.000	1.000	0.012	0.914	0.000	0.941	1.000
	Composition 2	1.000	1.000	1.000	0.148	0.701	0.000	0.790	1.000
New South Wales	Richness	1.000	0.999	1.000	0.212	0.645	0.000	0.999	1.000
	Composition 1	0.998	0.994	0.996	2.081	0.149	0.143	0.176	1.000
	Composition 2	1.000	1.000	1.000	0.000	0.989	0.000	0.990	1.000
Australia - Natural	Richness	1.000	0.999	1.000	0.760	0.383	0.000	0.421	1.000
	Composition 1	1.000	0.999	1.000	0.572	0.449	0.000	0.485	1.000
	Composition 2	1.000	0.998	0.999	1.171	0.279	0.338	0.317	1.000
Australia - Croplands	Richness	0.992	0.984	0.992	11.898	0.001	0.409	0.001	1.000
	Composition 1	0.983	0.969	0.985	30.194	0.000	0.670	0.000	1.000
	Composition 2	0.957	0.944	0.961	41.960	0.000	0.794	0.000	1.000

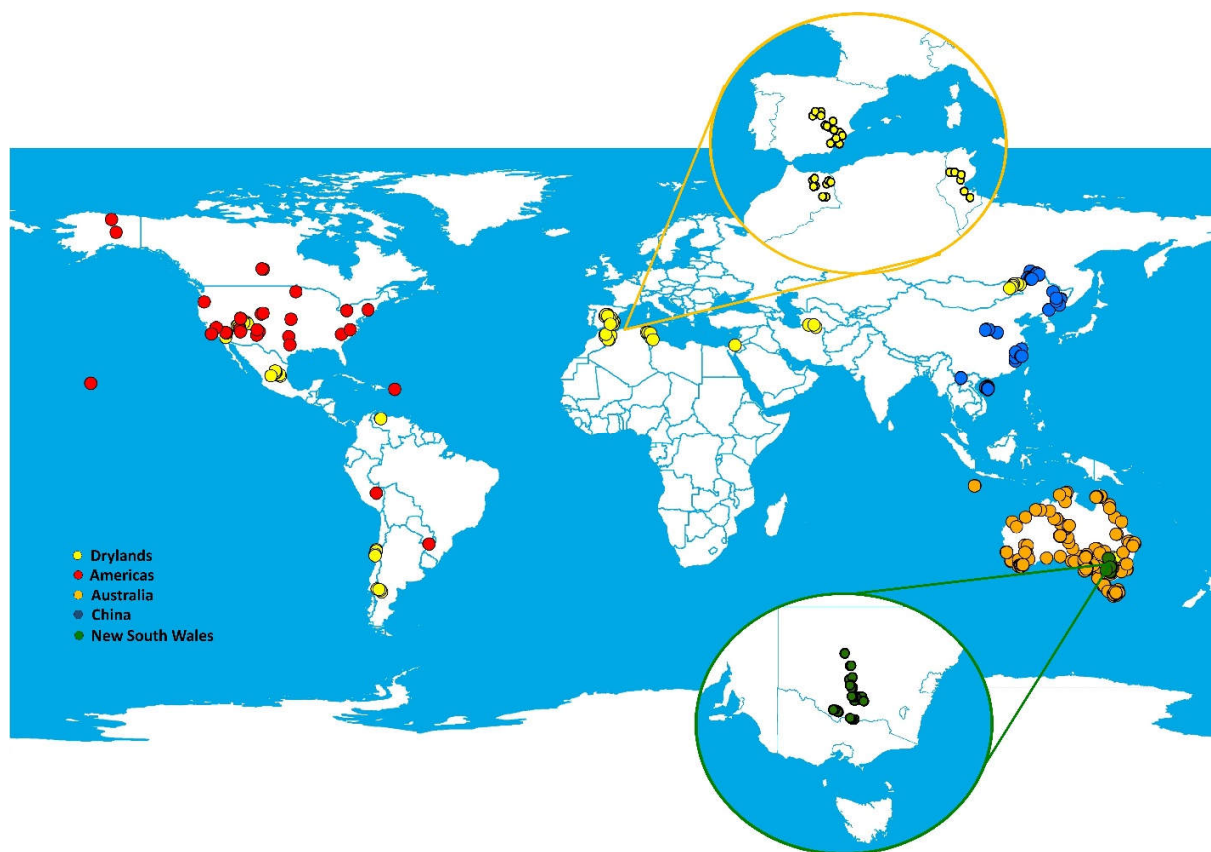
**Supplementary Table 11.** Pearson correlations (*P*-values in brackets) between precipitation, and maximum and minimum temperature used here ([www.worldclim.org](http://www.worldclim.org); resolution: 2.5minutes or 4.5km in the equator) and similar indexes from references 51 and 52 (resolution: 0.5° or 55km in the equator) summarizing information into decadal averages. Acronyms of climatic variables are shown in Supplementary Table 1.

	0y BP	6000y BP	20000y BP
<b>AP</b>	0.949 (<0.001)	0.957 (<0.001)	0.921 (<0.001)
<b>MAXTWM</b>	0.720 (<0.001)	0.633 (<0.001)	0.889 (<0.001)
<b>MINTCM</b>	0.926 (<0.001)	0.924 (<0.001)	0.903 (<0.001)

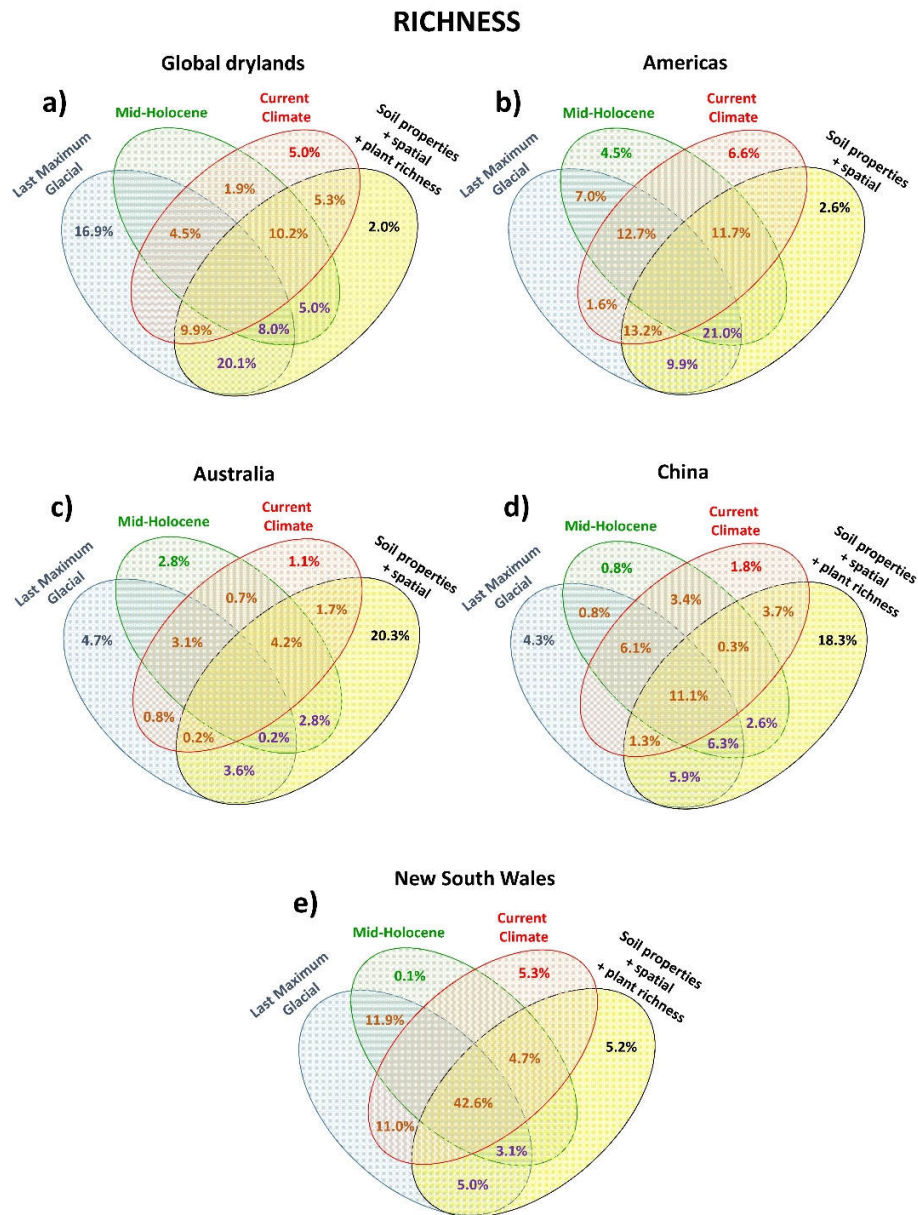


**Supplementary Table 12.** Pearson correlations (*P*-values in brackets) between precipitation, and maximum or minimum temperature legacies used here (e.g. Annual precipitation<sub>Current climate</sub> - Annual precipitation<sub>Last Glacial Maximum</sub> for each site; [www.worldclim.org](http://www.worldclim.org)) with the rate of change of precipitation, and maximum or minimum temperature calculated using the slope of a linear regression between each of these climatic parameters and time (years) using data from references 51 and 52. Acronyms of climatic variables are shown in Supplementary Table 1.

	AP Legacy (used)
Slope AP (Lorenz et al. 2016)	0.780 (<0.001)
	MAXTWM Legacy (used)
Slope MAXT (Lorenz et al. 2016)	0.864 (<0.001)
	MINTCM Legacy (used)
Slope MINT (Lorenz et al. 2016)	0.805 (<0.001)

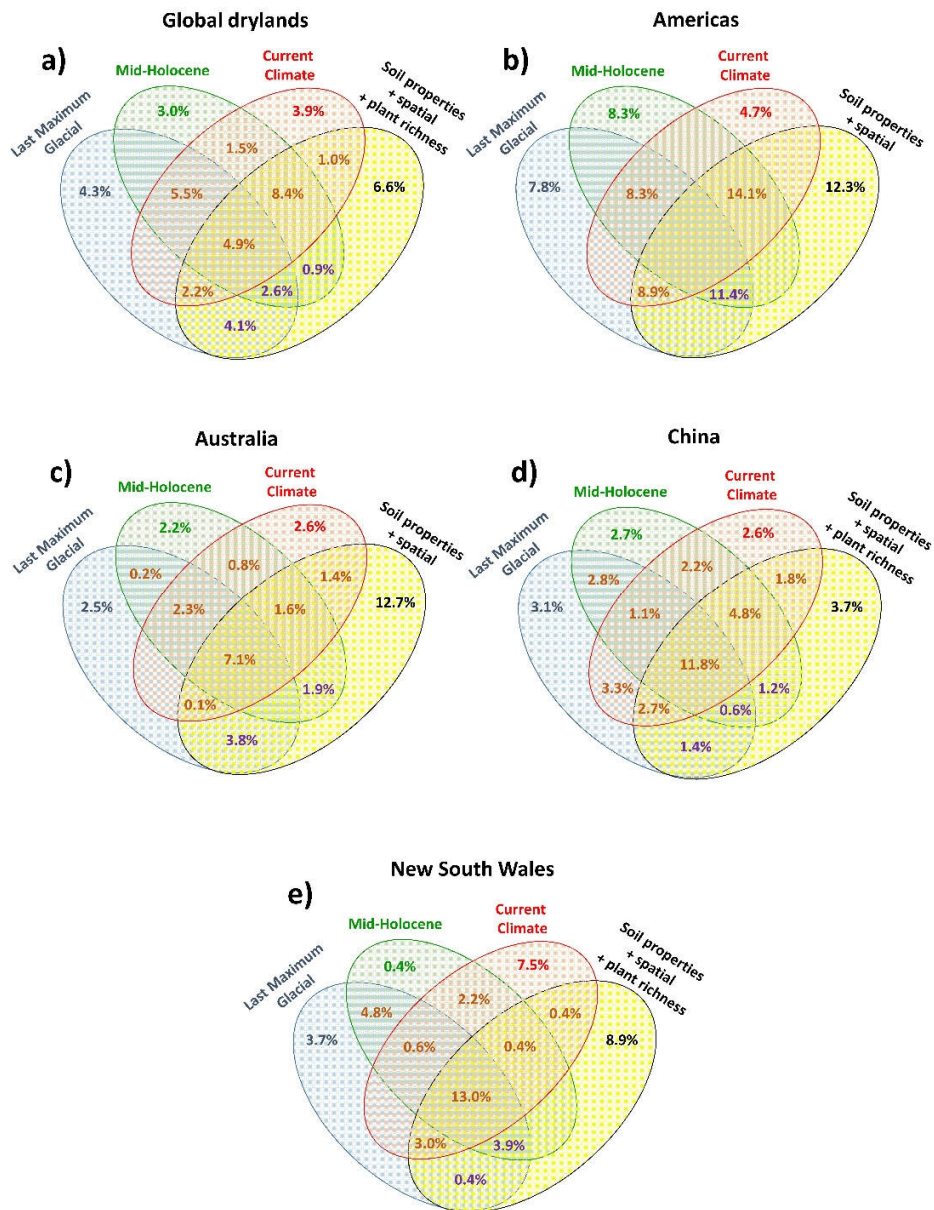


**Supplementary Figure 1.** Location of the sites included in the Drylands ( $n = 78$ ), Americas ( $n = 48$ ), Australia ( $n = 531$ ), China ( $n = 300$ ) and New South Wales ( $n = 54$ ) datasets.

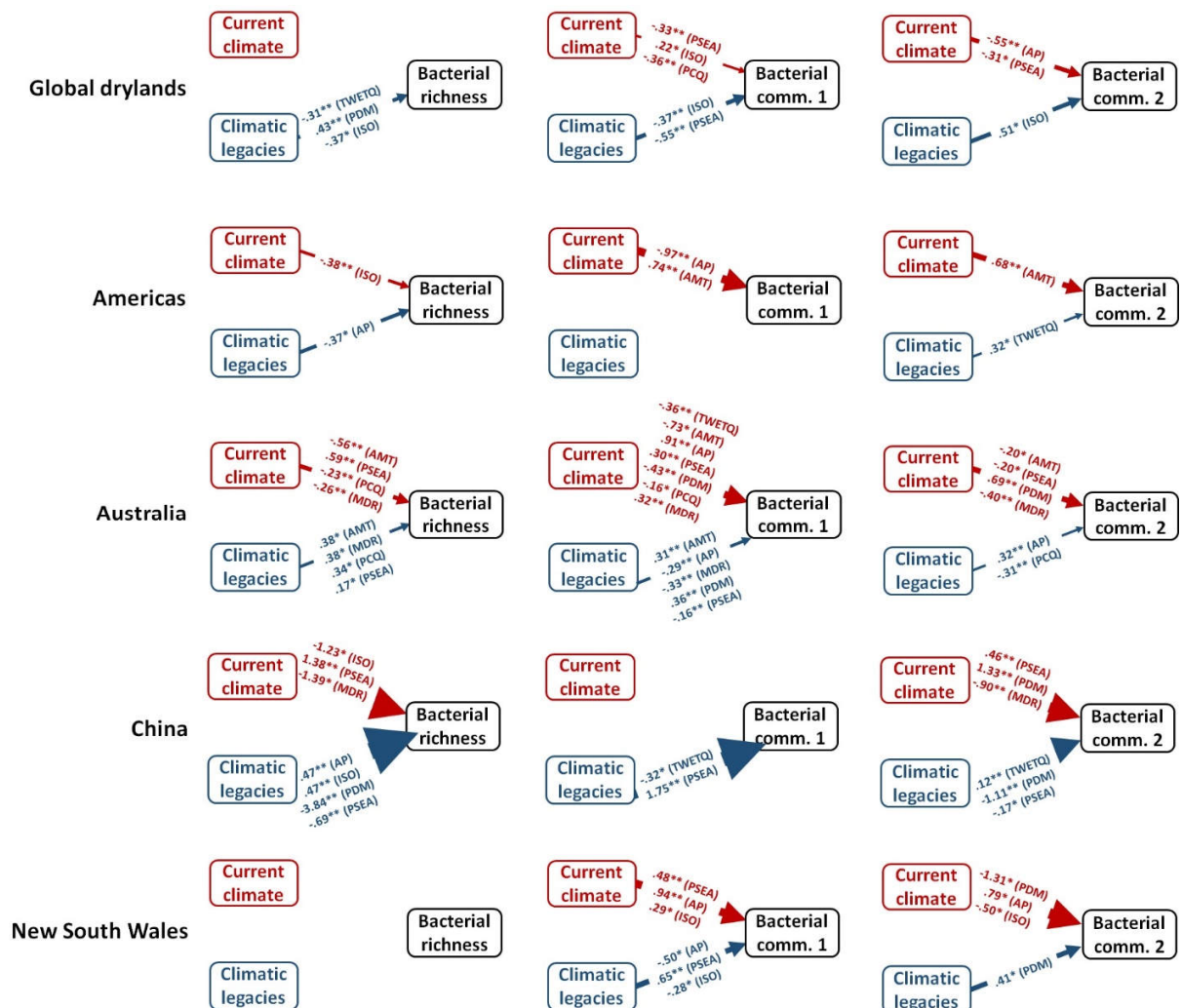


**Supplementary Figure 2.** Relative contribution of paleo (mid-Holocene and Last Glacial Maximum), current climate and other predictors (plant diversity and/or space – latitude and longitude – and soil properties) as drivers of bacterial richness at the OTU level. Panels represent results from Variation Partitioning modeling aiming to identify the percentage variance of bacterial diversity explained by past and current climate variables across five independent large scale datasets. Shared effects of these variable groups are indicated by the overlap of circles.

## COMPOSITION

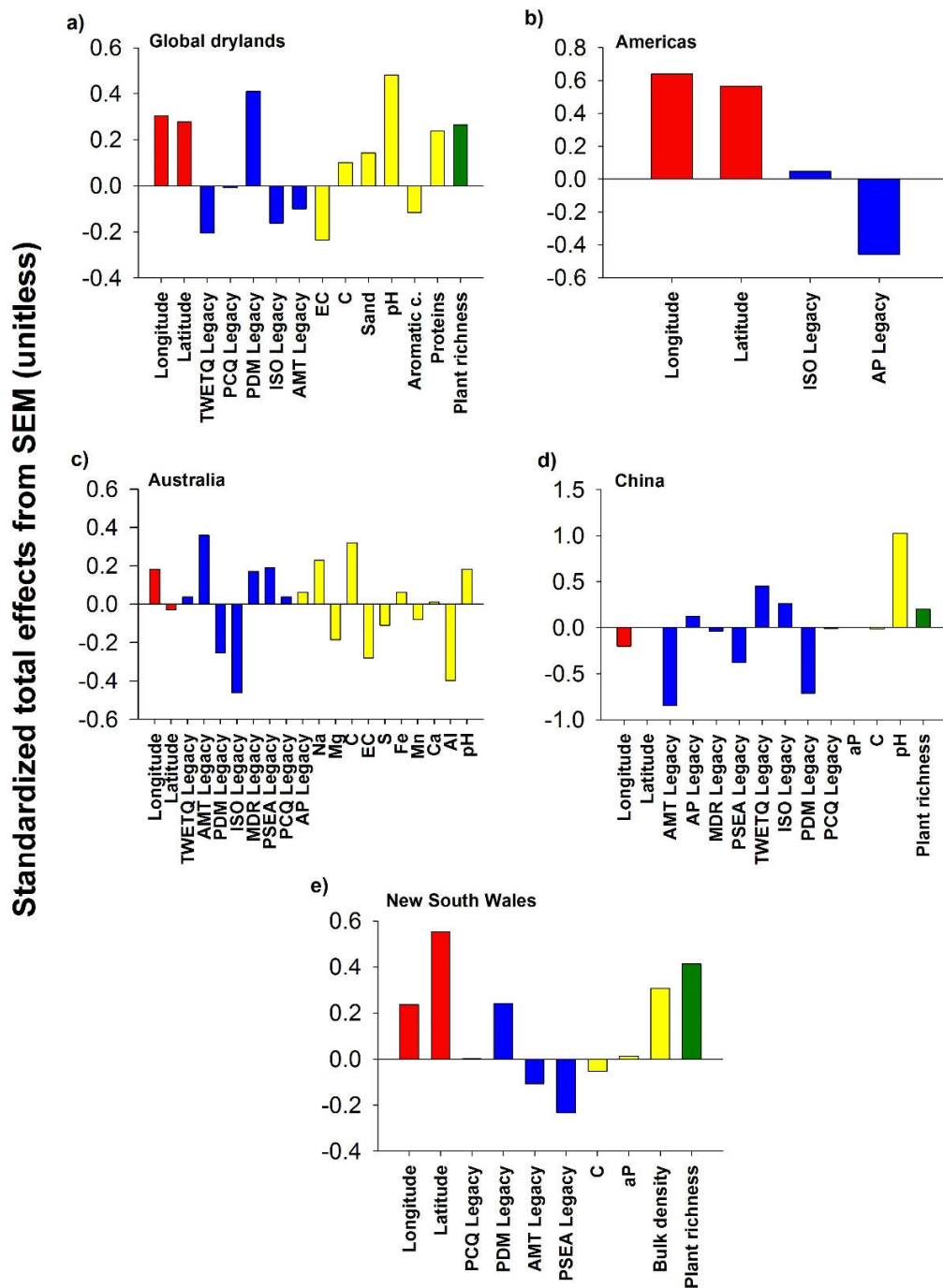


**Supplementary Figure 3.** Relative contribution of paleo (mid-Holocene and Last Glacial Maximum), current climate and other predictors (plant diversity and/or space – latitude and longitude – and soil properties) as drivers of bacterial community composition at the OTU level. Panels represent results from Variation Partitioning modeling aiming to identify the percentage variance of bacterial community composition explained by past and current climate variables across five independent large scale datasets. Shared effects of these variable groups are indicated by the overlap of circles.

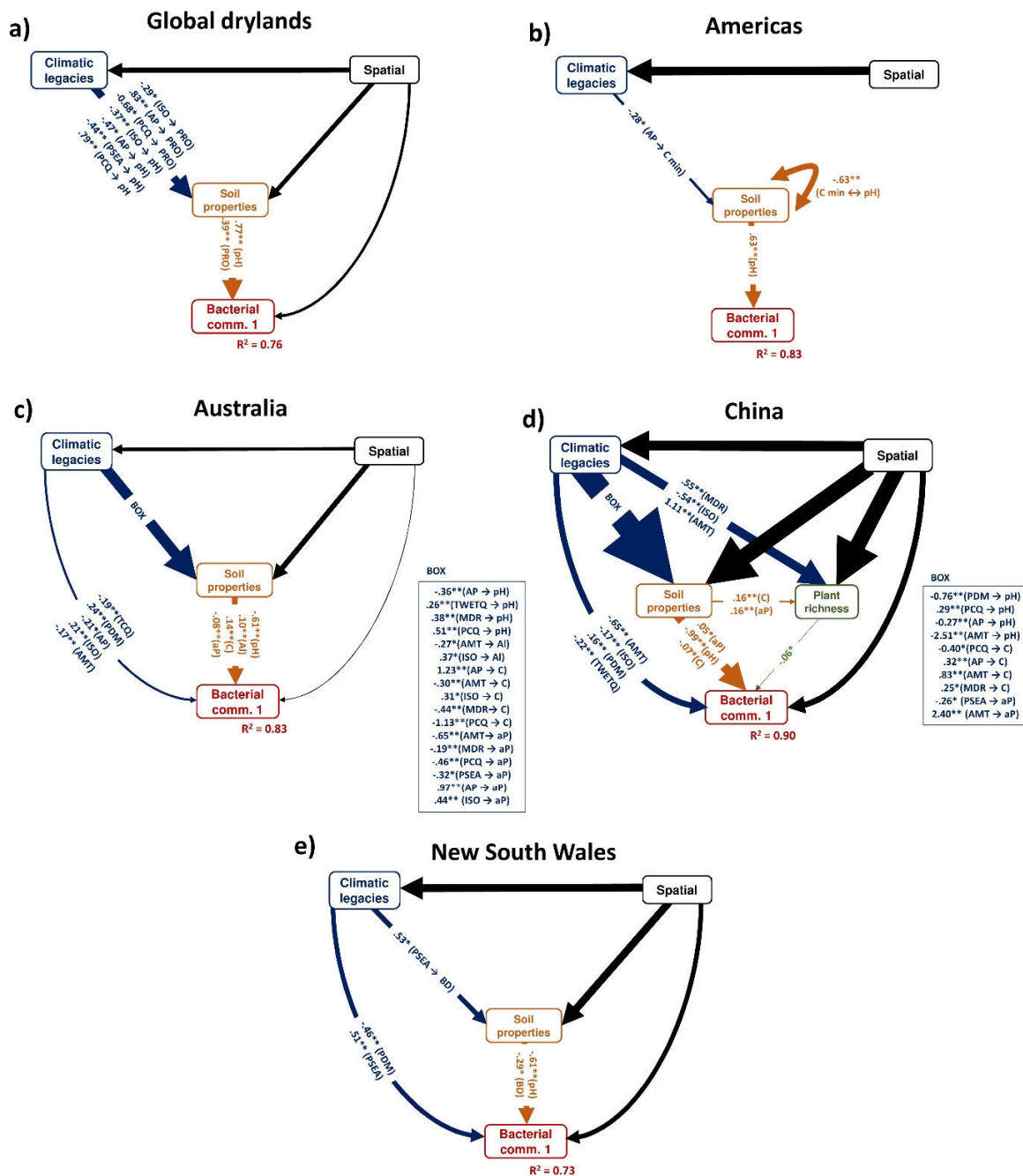


**Supplementary Figure 4.** Structural Equation Modeling aiming to identify the relative influence of the main bioclimatic variables from climatic legacies and current climate (as identified by Random Forest analyses) on soil bacterial community richness and composition. Current climate and climatic legacies are allowed to covary in these analyses, however, we did not include that information in this example. Acronyms of climatic variables are shown in Supplementary Table 1.





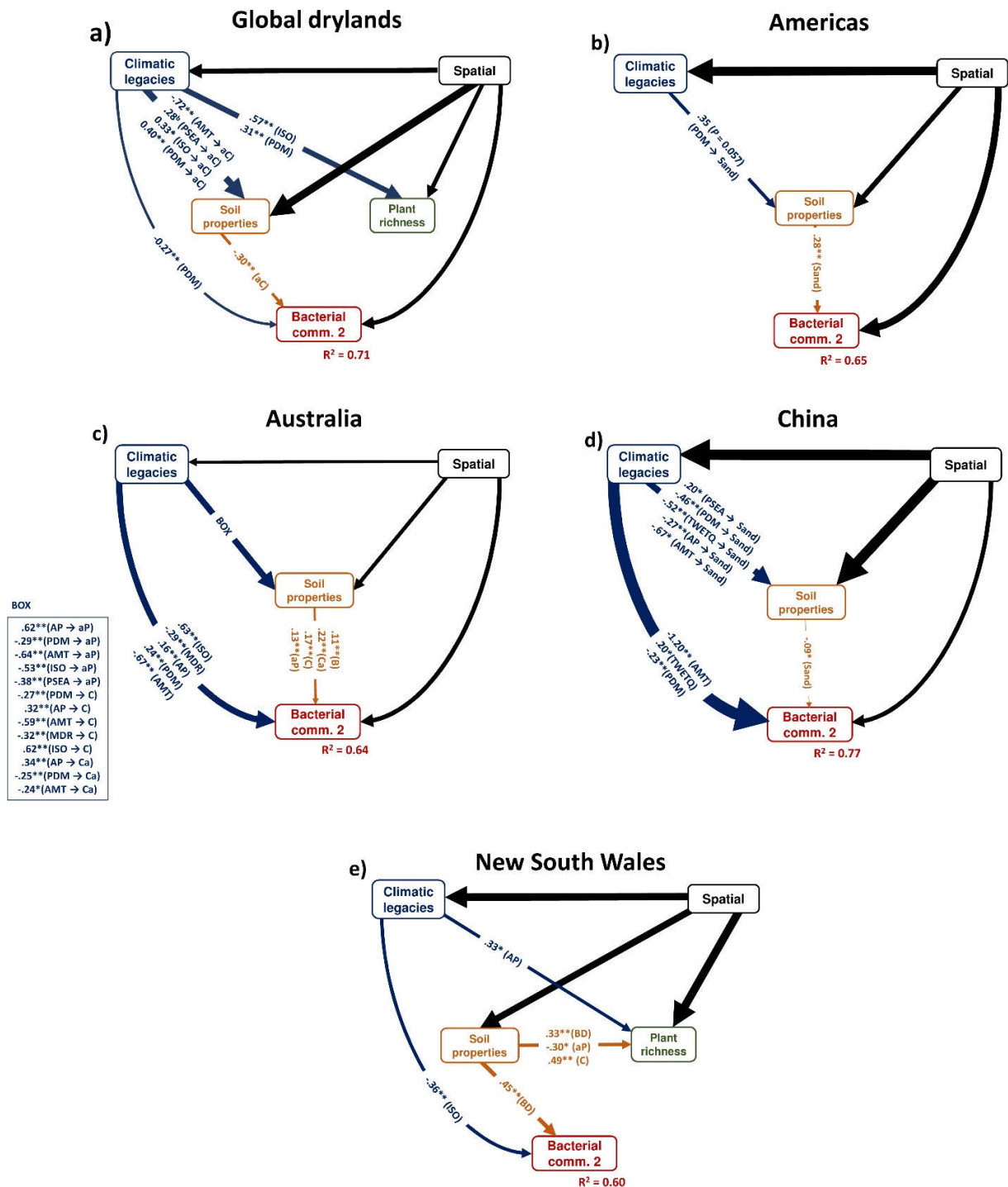
**Supplementary Figure 5.** Standardized total effects from structural equation models – this is the sum of direct and indirect effects from plant diversity and/or climatic legacies, soil properties and space on the richness of soil bacteria across five independent regional and global datasets. Acronyms of climatic variables are shown in Supplementary Table 1.



**Supplementary Figure 6.** Structural equation model accounting for the direct and indirect effects (plant diversity and/or soil properties) of climatic legacies on bacterial composition (Bacterial comm. 1) across five large scale datasets. Numbers adjacent to arrows are path coefficients (P values), and indicative of the standardized effect size of the relationship. Spatial influence (latitude and longitude) were included to control spatial autocorrelation; however, in this case, path coefficients were not included for simplicity and the size of the arrow represent the strength of the

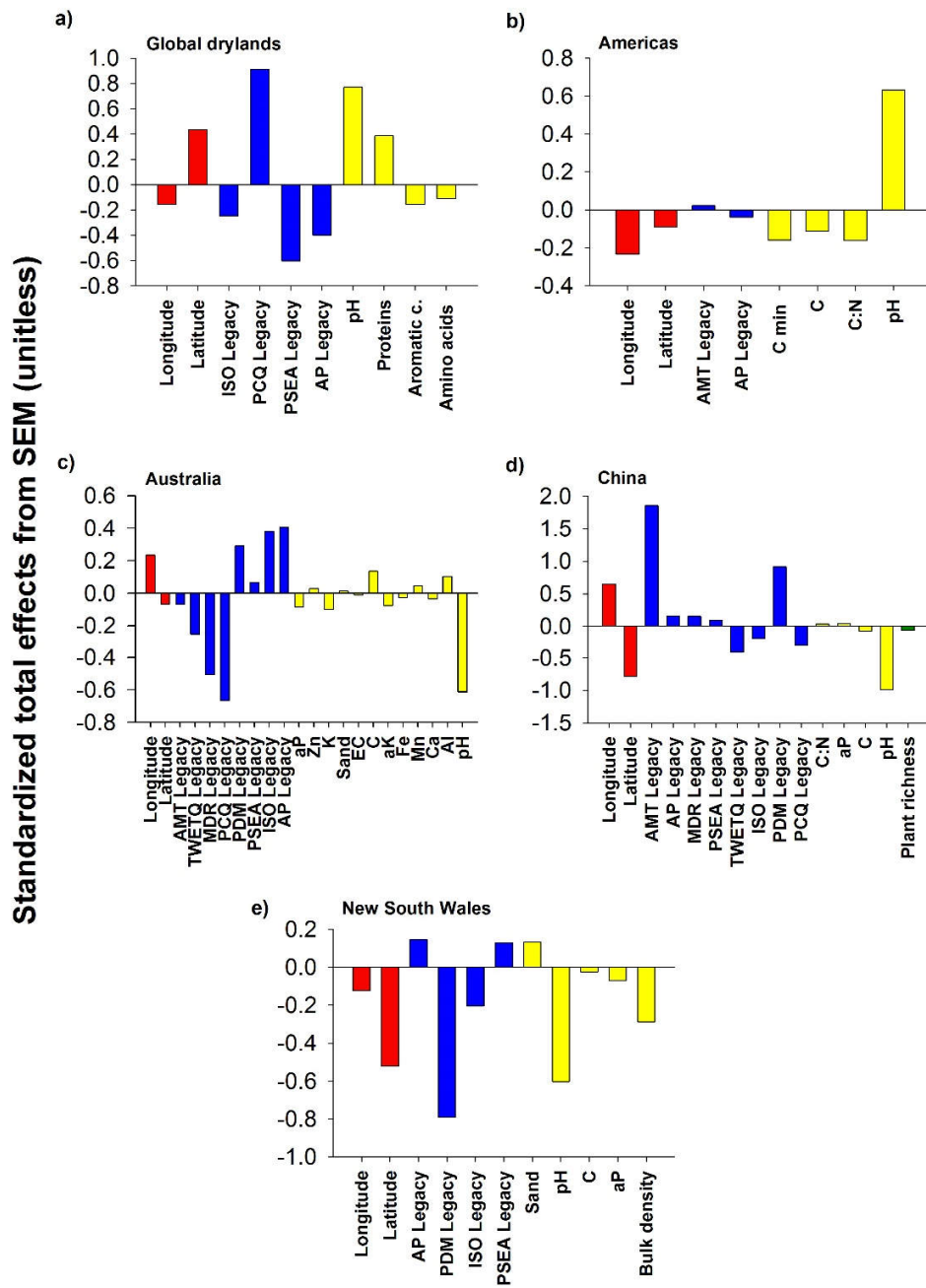
relationship when significant. All variables are included as independent observable variables. We grouped the different categories of predictors (soil properties, climate legacies and spatial) in the same box in the model for graphical simplicity. Also for simplicity, we only included those direct effects from climate legacies on soil properties that could indirectly affect the diversity of bacteria. The rest of the effects from climate legacies on soil properties are available in Supplementary Tables 5 and 6.  $R^2$  = the proportion of variance explained. Significance levels of each predictor are \* $P < 0.05$ , \*\* $P < 0.01$ . A small capital letter (a) and (o) adjacent to a particular chemical element indicate that element is in an “available” or “occluded” form. Acronyms of climatic variables are shown in Supplementary Table 1.



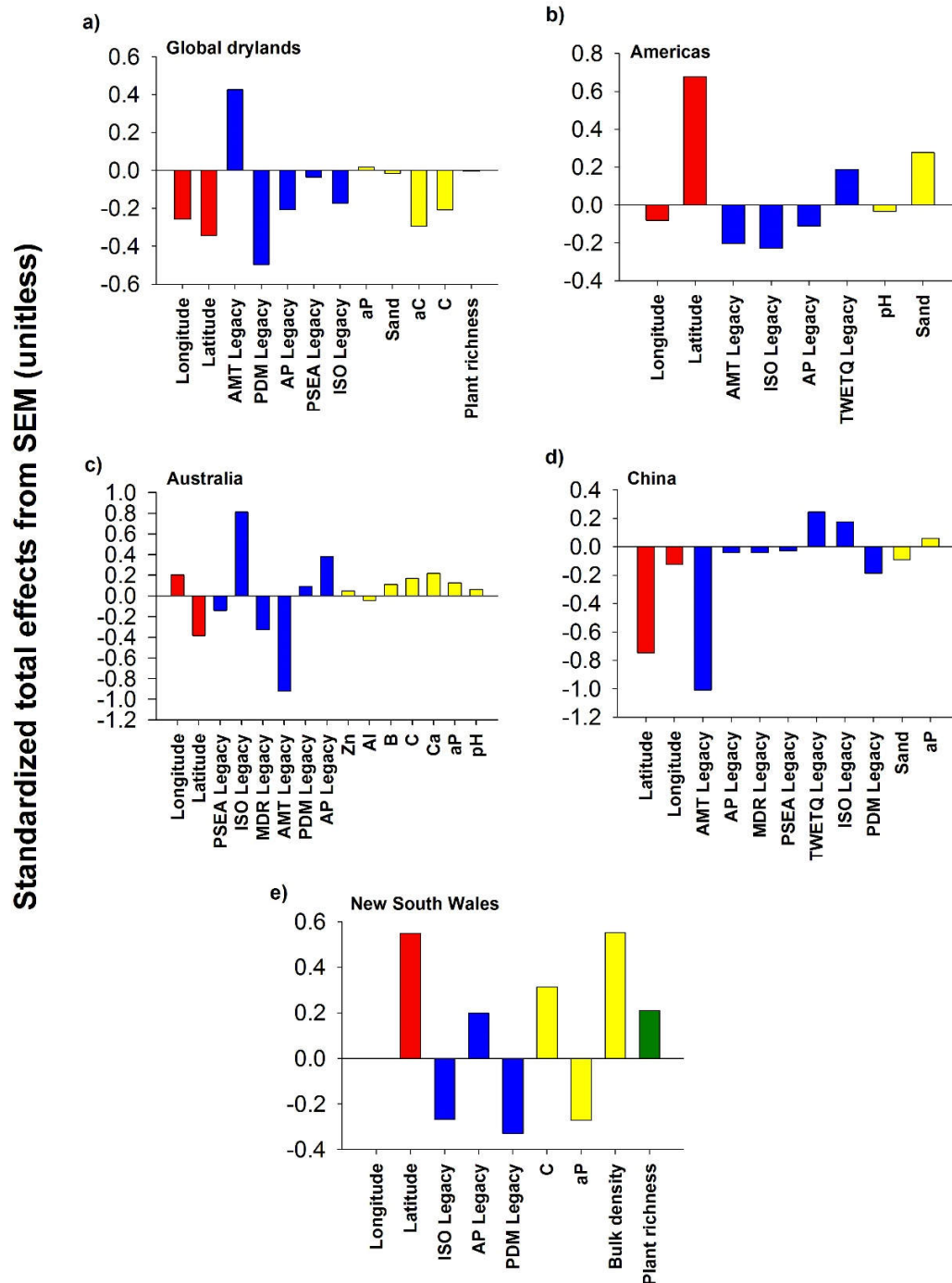


**Supplementary Figure 7.** Structural equation model accounting for the direct and indirect effects (plant diversity and/or soil properties) of climatic legacies on bacterial composition (Bacterial comm. 2) across five large scale datasets. \*P < 0.05, \*\*P < 0.01.

Rest of caption as in Supplementary Fig. 6.

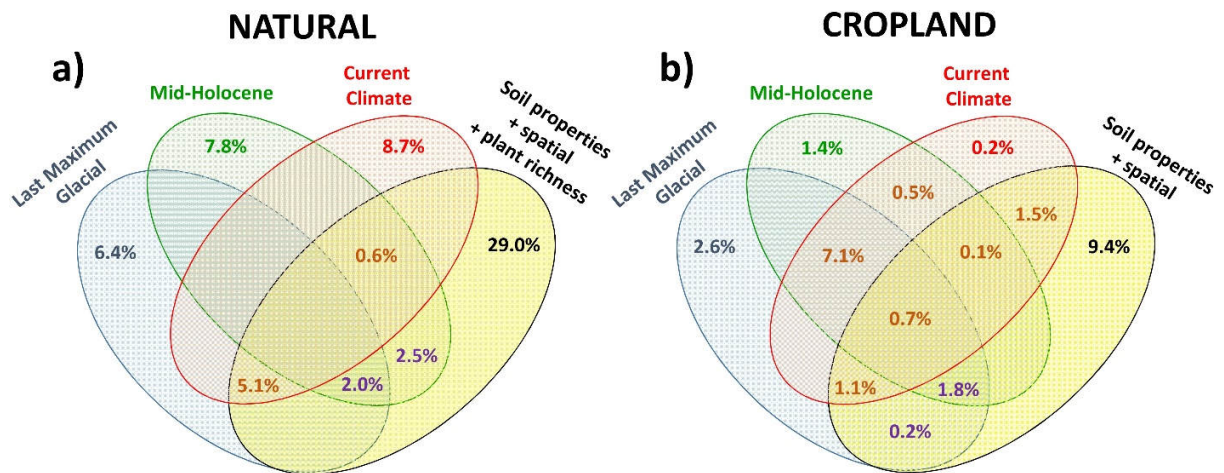


**Supplementary Figure 8.** Standardized total effects from SEM – this is the sum of direct and indirect effects from plant diversity and/or climatic legacies, soil properties and space on Bacterial comm. 1 (first axis of an NMDS analysis collapsing information of bacterial composition at the OTU level) across five independent regional and global datasets.

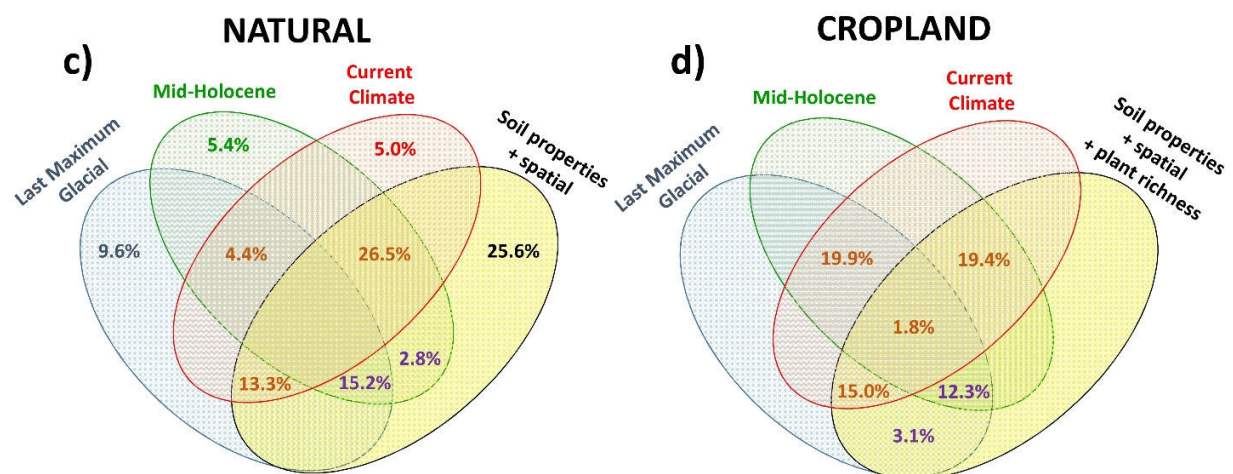


**Supplementary Figure 9.** Standardized total effects from SEM – this is the sum of direct and indirect effects from plant diversity and/or climatic legacies, soil properties and space on Bacterial comm. 2 (second axis of an NMDS analysis collapsing information of bacterial composition at the OTU level) across five independent regional and global datasets.

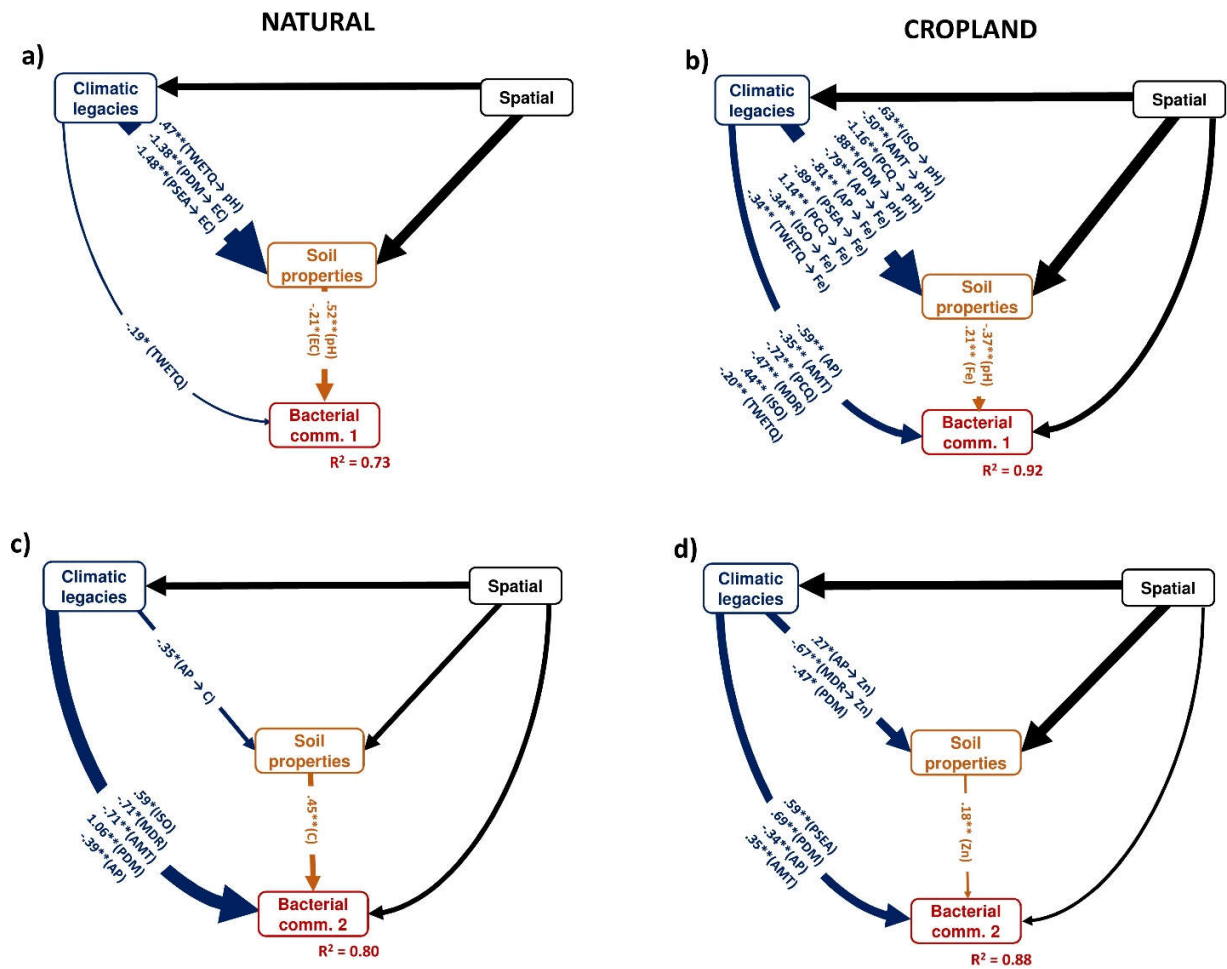
## COMPOSITION



## RICHNESS



**Supplementary Figure 10.** Relative contribution of paleo (mid-Holocene and Last Glacial Maximum), current climate and other predictors (plant diversity and/or space – latitude and longitude – and soil properties) as drivers of the composition of bacteria (reads/OTUs) for croplands and natural ecosystems of the Australia dataset. Shared effects of these variable groups are indicated by the overlap of circles.



**Supplementary Figure 11.** Structural equation model accounting for the direct and indirect effects (plant diversity and/or soil properties) of climatic legacies on bacterial composition (Bacterial comm. 1 and 2) for croplands and natural ecosystems of the Australia dataset. The variables included in this SE model are identical to that one presented in Figure3c. Rest of caption as in Supplementary Fig. 6.

**Supplementary Figure 12.** Changes in precipitation over the last 20000y in 49 unique location from North America using climatic information from references 51 and 52.

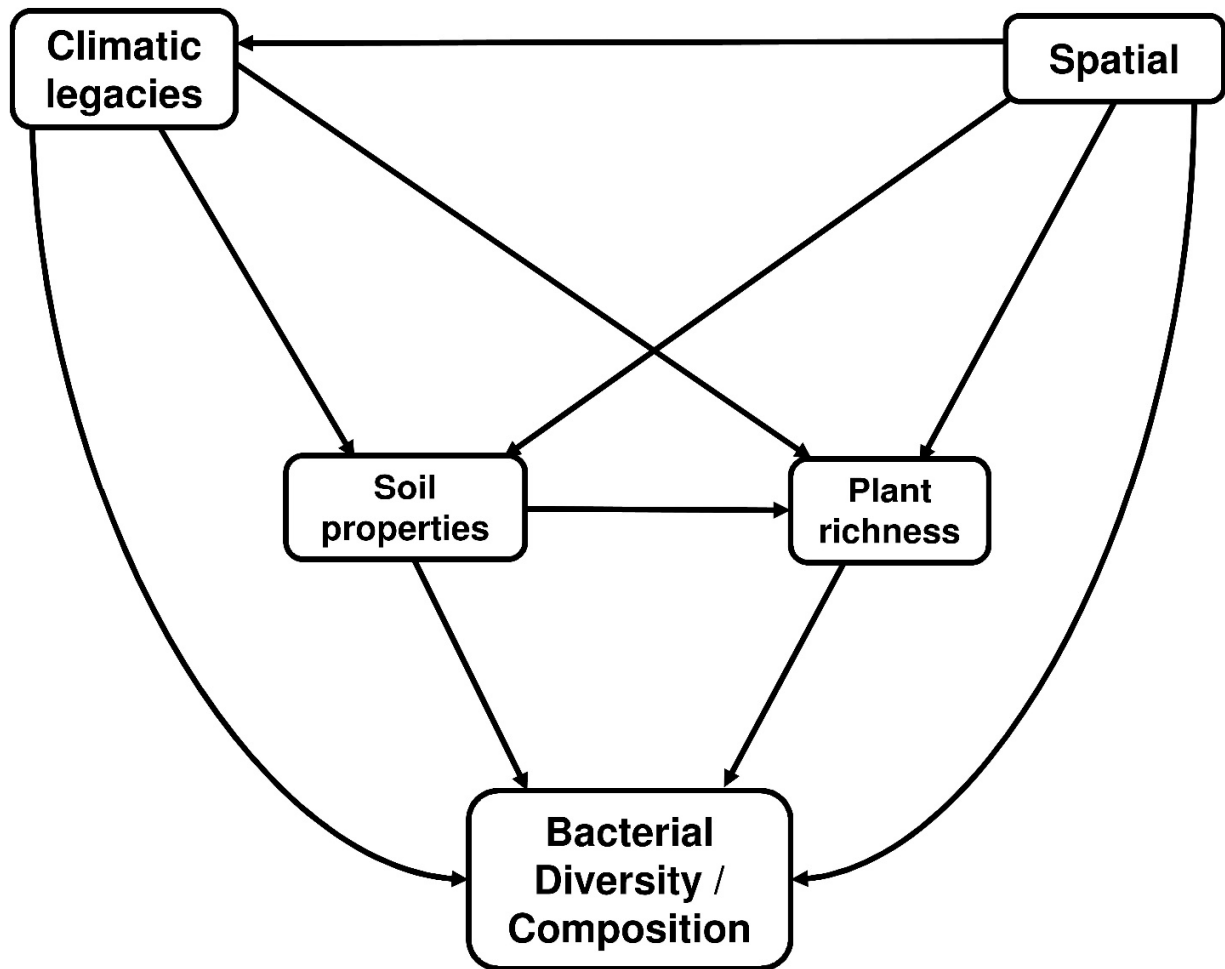
*Supplementary Figure 12 is available online as a Separate Excel file under the Supporting Information for this article.*

**Supplementary Figure 13.** Changes in maximum temperature over the last 20000y in 49 unique location from North America using climatic information from references 51 and 52.

*Supplementary Figure 13 is available online as a Separate Excel file under the Supporting Information for this article.*

**Supplementary Figure 14.** Changes in minimum temperature over the last 20000y in 49 unique location from North America using climatic information from references 51 and 52.

*Supplementary Figure 14 is available online as a Separate Excel file under the Supporting Information for this article.*



**Supplementary Figure 15.** *A priori* structural equation model accounting for the direct and indirect effects (plant diversity and/or soil properties) of climatic legacies on bacterial composition and diversity (richness).

### **References (not listed in the main text)**

51. Lorenz, D. J., Nieto-Lugilde, D., Blois, J. L., Fitzpatrick, M. C. & Williams, J. W. Dryad Digital Repository <http://dx.doi.org/10.5061/dryad.1597g> (2016).
52. Lorenz, D.J. et al. Downscaled and debiased climate simulations for North America from 21,000 years ago to 2100AD. *Scientific Data* 3, 160048 (2016).
53. Richter, D.B., Yaalon, D.H. "The Changing Model of Soil" Revisited. *Soil Sci. Soc. Am. J.* 76, 766-778 (2012).
54. Schlesinger, W.H. Evidence from chronosequence studies for a low carbon-storage potential of soils. *Nature* 348, 232-234 (1990).
55. Delgado-Baquerizo M. et al. Climate legacies drive global soil carbon stocks in terrestrial ecosystems. *Sci. Adv.* 3, e1602008 (2017).