Global mammalian zooregions reveal a signal of past human impacts

Marta Rueda[1*], Manuela González-Suárez[2], Eloy Revilla[1]

1. Department of Conservation Biology, Estación Biológica de Doñana (EBD-CSIC), Seville, Spain

2. Ecology and Evolutionary Biology, School of Biological Sciences, University of Reading, Reading, UK.

**Corresponding author*

Understanding how the world's biodiversity is organized and how it changes across geographic regions is critical to predicting the effects of global change[1]. Ecologists have long documented that the world's terrestrial fauna is organized hierarchically in large regions - or realms - and continental scale subregions[2-6], with boundaries shaped by geographic and climatic factors[2,7]. However, little is known about how global biodiversity is assembled below the continental level and the factors, including the potential role of human impacts, triggering faunistic differences as the biogeographical scale becomes smaller. Here we show that the hierarchical organization of global zoogeographical regions extends coherently below the region level to reach a local scale, and that multiple determinants act across varying spatial and temporal scales. Among these determinants, anthropogenic land use during the Late Holocene stands out showing a footprint across biogeographical scales and explaining 22% of the faunistic differences among the larger bioregions. The Late Holocene coincided with the development of large cities and substantial transformation of ecosystems into agricultural land[8,9]. Our results show that past human activity has played a role in the global organization of present-day animal assemblages, leaving a detectable signal that warns us about significant time-lag effects of human-mediated impacts on biodiversity.

The questions of how the world's biodiversity is organized, and why large-scale patterns of taxonomic diversity change through natural geographic regions have attracted the attention of naturalists since the early 19th century[2,10-15]. The answers to these questions are important to satisfy our curiosity about the natural world, but have also become critical to forecast the future of biodiversity in the face of global change[1]. A key step in understanding the organization of biodiversity is the assemblage of regions based on their shared elements[14]. Alfred R. Wallace was among the first to propose that the world's fauna is organized hierarchically in broad regions shaped by geographic and climatic factors[2]. About 150 years later, the development of multivariate analytical techniques has led to the revaluation of Wallace's proposal[3-6] and refining of the extrinsic factors explaining the major dissimilarities among zooregions[7]. However, biogeographic boundaries and the signal of evolutionary processes associated to species isolation are not so evident at smaller scales, and importantly, still remain globally unexplored. Smaller regions, which are generally the units of conservation actions, contain more similar biota and thus, the factors determining faunistic dissimilarities among them are likely to be more diverse and include spatial and taxonomic idiosyncracies[7,16,17].

We hypothesize that global biodiversity patterns can be characterized by a hierarchical system of biogeographic regions extending from global to local scales, with regions at different levels explained by determinants that represent varying temporal scales. To test this, we considered determinants already identified as important: plate tectonics, climate - including Quaternary climate changes, orography and changes in habitat type[7,16,17]. But critically, we also explore the role of largely overlooked predictors associated with present and past anthropogenic global impacts. The effects of recent human actions on current species distributions are undeniable[18], already affecting current biogeographic patterns[19,20], but past human actions are generally portrayed as localized and insignificant in comparison[21]. The increasing evidence that Quaternary human impacts induced shifts in the plant and animal communities we see today[22,23] challenges this view and poses the question of whether past anthropogenic land impacts may have been large enough to induce changes detectable today at biogeographical scales.

To address these questions we first applied an affinity propagation clustering algorithm[24] to a co-occurring species matrix of global terrestrial mammals generating a hierarchical bioregionalization upscaling from the smallest detectable bioregions to the largest realms. We then used random forest classification models to identify the determinants that best predict taxonomic differences among bioregions within the framework of two hypothesized scenarios (Fig. 1). These scenarios always consider remote past, recent past and contemporary determinants but assume their influence will differ across the hierarchical levels. Differentiation between large realms should require longer evolutionary times, and therefore, both scenarios assume that factors related to historical and macroevolutionary processes of speciation and extinction will be most important to explain taxonomic dissimilarities of the largest realms. As bioregions decrease in size we predict processes related to tolerances to given habitats or climates (which are also forged over evolutionary time) and human impact would gain importance. The scenarios differ in how we suggest this process may occur: linearly (Fig. 1A) or with a nested structure (Fig. 1B).

The clustering algorithm generated a hierarchical system of biogeographic regions with four levels showing that global biodiversity patterns can be cohesively shaped from local (area of the smallest bioregions detected is ~93 km^2) to regional and to realm scales

(Extended Data Fig. 1 and Supplementary Fig. 1). The broadest delineation of nine large bioregions was strikingly similar to the six zoogeographical regions and boundaries proposed by Wallace[2], showing our method is a suitable approach to define bioregions. Particular differences include the delineation of Madagascar and Chilean subregions (*sensu* Wallace) as regions[2,6], differences in the limits of the Palearctic also detected in previous analytical regionalizations[3,5,6], and an extension towards the arid steppes of Mongolia of the 'Saharo-Arabian' realm[3,6].

We found a nested effect of temporal determinants of bioregion assemblages (Fig. 1C), similar to our proposed nested scenario (Fig. 1B). The signal of events occurring millions of years ago, such as tectonic movements or orographic barriers, remained apparent from the largest to the smallest bioregions, while recent past and contemporaneous determinants acquired importance at smaller scales. Overall, results for the two broadest scales (nine and 27 bioregions respectively) supported findings from prior work[7]. Plate tectonics drove the main taxonomic dissimilarities between large landmasses in interplay with variability in climate and orographic barriers, the latter with less weight in the global model but important for determining differences between specific regions (Figs. 1C, 2, Extended Data Fig. 3 and Supplementary Table 1 and 2). For the smaller scales detected (141 and 1128 bioregions), we found that a combination of multiple determinants, that varied spatially in importance, was critical to predict assemblages (Fig. 1C, Extended Data Fig. 4, and Supplementary Table 3 and 4). Among them, the association of geological factors, past climate change and current variability in temperature resulted decisive (Fig. 1C, Extended Data Table 1).

Interestingly, we identified a prevalent footprint of past anthropogenic impacts, particularly human land use 2000 years ago, across the hierarchical bioregionalization (Figs. 1C, 3). Human land use 2000 years ago was the second most important determinant for the largest bioregions showing a moderate but non-negligible predictive value (~22%, Fig. 2). The importance of human land use 2000 years ago increased in the global model for smaller bioregions but its individual predictive power was reduced because differentiating smaller bioregions required many more determinants. Differences in human land use 2000 years ago contribute to set boundaries separating bioregions where human land use was noticeable (e.g. among the Neotropic/Nearctic, Neotropic/Chilean, and the Oriental/Palaearctic) and bioregions where human land use was negligible at that time (e.g. Madagascar, and the arctic boundary between the Palearctic-Nearctic) (Fig. 3 and Extended Data Fig. 2). Current human land use was a less relevant determinant, which is consistent with the hypothesis that anthropogenic transformation of ecosystems has been extensive and started longer ago than is often recognized[21-22].

What happened 2000 years ago that resulted in such as noticeable footprint in the Earth's bioregions? This period coincides with the development of major cities (populations > 100,000) in the Near East, Europe, and Asia[8]. At this time, human populations already inhabited ecosystems reshaped by their ancestors to enhance agricultural productivity[9]. Long-term impacts, including forest clearing, increased fire frequencies, megafaunal extinctions, species invasions, and soil erosion, were already apparent in some regions[8,21,22]. Indeed, reconstructions suggest that >20% of Earth's temperate woodlands had been already impacted by humans by the late-Holocene[21]. Our results join recent studies in revealing that anthropogenic impact during the Holocene played an important role in how modern mammals' communities are assembled[22-23]. Here we show for the first time that this signal can be detected in the configuration of realms, which are traditionally assumed to reflect the natural organization of biodiversity that resulted from ecological, historical, and evolutionary processes acting over millions of years.

Our results also highlight the increased importance of Quaternary climate for the smaller bioregional levels (Fig. 1C). This is in line with previous work showing how strongly climatic shifts during the Quaternary have determined changes on modern species distributions[25-26]. Among all the measures of past climate considered here, we found a notable increase in importance of temperature change in the mid-Holocene. This is remarkable as generally mid-Holocene climates are considered to be relatively stable and similar to modern climate[26]. Paleodata indicate however, that around 6000 years ago there were rapid climatic changes that triggered the northward expansion of boreal forest and the greening of the Sahara[27], and accompanied by human societal collapses at local and regional scales[28]. Our results suggest these rapid climatic changes could have also affected the distribution of modern species, resulting in a detectable signal of taxonomic differentiation in the smaller bioregions.

Taken together, our results indicate that the world's biodiversity is organized with a hierarchical structure of global biogeographical patterns that include a local basis, which is determined by multiple and spatially heterogeneous factors. Geological events that occurred over millions of years permeate from the largest to the smaller scales, yet understanding biogeographical delineations at more local scales requires considering determinants acting at multiple temporal and spatial scales. At more local levels, we also observed a reduced predictive capacity, which may reflect the importance of biotic determinants, species traits and interactions, not considered in our analyses but known to play a role in structuring community composition[29]. Previous studies have documented lasting effects of human land changes during the last millennia on current biodiversity patterns[30], but this is the first time the signal has been recognized on the taxonomic differentiation of the largest realms. If human impacts over the Late Holocene can result in such long-lasting and widely spread signals, we should be concerned about the effects of the much more widespread and severe changes that have occurred since the beginning of the industrial revolution. Current human impact was not a strong determinant of current bioregions, but this likely reflects a time-lag effect. The signal of current human land use will likely be detected by the future generations of biogeographers. It is in our hands to ensure those future bioregions are not solely determined by our impacts.

References

1. Botkin, D.B. *et al.* Forecasting the effects of global warming on biodiversity. *Bioscience* **57,** 227-236 (2007).

2. Wallace, A. R. 1876. *The geographical distribution of animals*. Harper & Brothers, New York (1876).

3. Holt, B. *et al.* An update of Wallace's zoogeographic regions of the world. *Science* **339,** 74–78 (2013).

4. Proches, S. & Ramdhani, S. The world's zoogeographical regions confirmed by cross-taxon analyses. *Bioscience* **62,** 260–270 (2012).

5. Kreft, H. & Walter, J. A framework for delineating biogeographical regions based on species distributions. *J. Biogeogr.* **37,** 2029-2053 (2010).

6. Rueda, M., Rodriguez, M. A. & Hawkins, B. A. Identifying global zoogeographical regions: lessons from Wallace. *J. Biogeogr.* **40,** 2215–2225 (2013).

7. Ficetola, G. F., Mazel, F. & Thuiller, W. Global determinants of zoogeographical boundaries. *Nature Ecology & Evolution* **1,** 0089 (2017).

8. Kirch, P.V. Archaeology and global change: The Holocene record. *Annu Rev Environm Resour* **30**, 409 (2005).

9. Ruddiman, W. & Ellis, E.C. Effect of per-capita land use changes on Holocene forest clearance and CO_2 emissions. *Quat. Sci. Rev*. **28,** 3011-3015 (2009).

10. von Humboldt, A. *Essai sur la geographie des plantes; accompagné d'un tableau physique des régions équinoxales, accompagné d'un tableau physique des régions équinoctiales*. Schoel & Co., Paris (1806).

11. de Candolle, A. *Geographie botanique raisonnée*. Librairie de Victor Masson, Paris (1855).

12. Sclater, P. L. On the general geographical distribution of the members of the class Aves. *J. Proc. Linn. Soc. Lond. Zool.* **2,** 130–145 (1858).

13. Ricklefs, R.E. A comprehensive framework for global patterns in biodiversity. *Ecol. Lett.* **7,** 1-15 (2004).

14. Wiens, J.J. The niche, biogeography and species interactions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **366,** 2336-2350 (2011).

15. Daru, B.H., Elliot T.L., Park, D.S. & Davies, T.J. Understanding the processes underpinning patterns of phylogenetic regionalization. *Trends Ecol. Evolut.* **32,** 845-860 (2017).

16. Glor, R.E. & Warren, D. Testing ecological explanations for biogeographic boundaries. *Evolution* **65,** 673-683 (2010).

17. Rueda, M., Rodríguez, M. A. & Hawkins, B. A. Towards a biogeographic regionalization of the European biota. *J. Biogeogr.* **37,** 2067–2076 (2010).

18. Di Marco, M. & Santini, L. Human pressures predict species´ geographical range size better than biological traits. *Glob. Change Biol.* **21,** 2169-2178 (2015).

19. Capinha, C. *et al.* The dispersal of alien species redefines biogeography in the Anthropocene. *Science* **348,** 1248-1251 (2015).

20. Bernardo-Madrid, R. *et al*. Human activity is altering the world´s zoogeographical regions. Preprint at https://www.biorxiv.org/content/biorxiv/early/2018/03/23/287300.full.pdf (2018)

21. Ellis, E.C. *et al.* 2013. Used planet: a global history. *Proc. Natl. Acad. Sci. USA* **110,** 7978-7985 (2013).

22. Lyons *et al.* Holocene shifts in the assembly of plant and animal communities implicate human impacts. *Nature* **529,** 80-85 (2016).

23. Sandom, C., Faurby, S., Sandel, B. & Svenning, J-C. Global late Quaternary megafauna extinctions linked to humans, not climate change. *Proc. Royal Soc. B* **281,** 20133254 (2014).

24. Frey, B.J. & Dueck, D. Clustering by passing messages between data points. *Science* **315,** 972-976 (2007).

25. Sandel, B. *et al.* The influence of late Quaternary climate-change velocity on species endemism. *Science* **334,** 660-664 (2011).

26. Graham, R.W. *et al.* Spatial response of mammals to Late Quaternary environmental fluctuations. *Science* **272,** 1601-1606 (1996).

27. Ganopolski, A. *et al.* The influence of vegetation-atmosphere-ocean interaction on climate during the mid-Holocene. *Science* **280,** 1916-1919 (1998).

28. Wanner, H. *et al.* Mid- to late Holocene climate change: an overview. *Quat. Scie. Rev.* **27,** 1791-1828 (2008).

29. Wisz, M.S. *et al.* The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev.* **88,** 15-30 (2013).

30. Dambrine, E. *et al.* Present forest biodiversity patterns in France related to former Roman agriculture. *Ecology* **88,** 1430-1439 (2007).

Acknowledgements

Support for this research was provided by the EU H2020 Framework Programme under grant agreement number 707587 (DRIVER) to MR, and by CGL2017-83045-R AEI/FEDER EU (Agencia Estatal de Investigación from the Ministry of Economy, Industry and Competitiveness, Spain co-financed with FEDER) to ER.

Author contributions

MR, MG-S and ER designed the study. MR analysed the data. All authors discussed the results and contributed to the final manuscript.

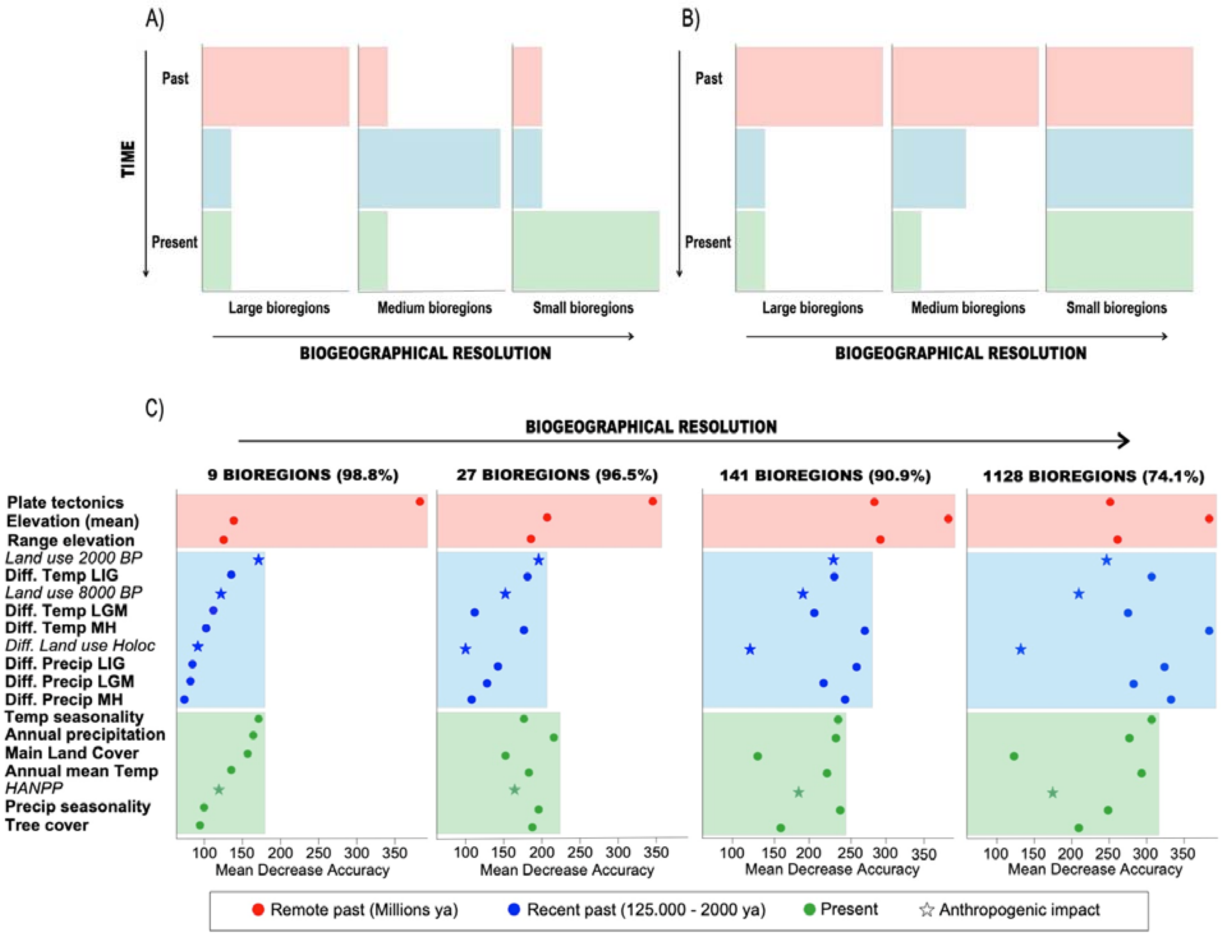


Figure 1 | Hypothesized scenarios (A, B) and results describing variable importance from random forest classification models (C). A) This scenario proposes that remote past determinants are largely irrelevant to explain the assembly of species in smaller bioregions. The signal of the past has been diluted as the biogeographic resolution increases, such that large bioregions can be described by remote past determinants, medium-sized bioregions by recent past determinants, and smaller bioregions by contemporary determinants. B) This scenario proposes nested importance of determinants, with remote past determinants being important at all scales, recent past determinants being important at medium and small scales, and contemporary determinants being only relevant at local scale. C) Observed global variable importance of geological, historical and environmental determinants of taxonomic differentiation among bioregions and across hierarchical resolutions. Determinants are ordered by decreased importance at the largest bioregion scale. Importance was measured by the drop in classification accuracy after predictor randomization in random forests of 5000 trees. Higher values of mean decreased in accuracy indicate variables that are more important to the classification. Percentages indicate prediction accuracy (percent correctly classified, 1-OBB) of global models. BP = Before present; LIG = Last interglacial; LGM = Last Glacial Maximum; MH = Mid Holocene; HANPP = Human appropriation of the net primary productivity.

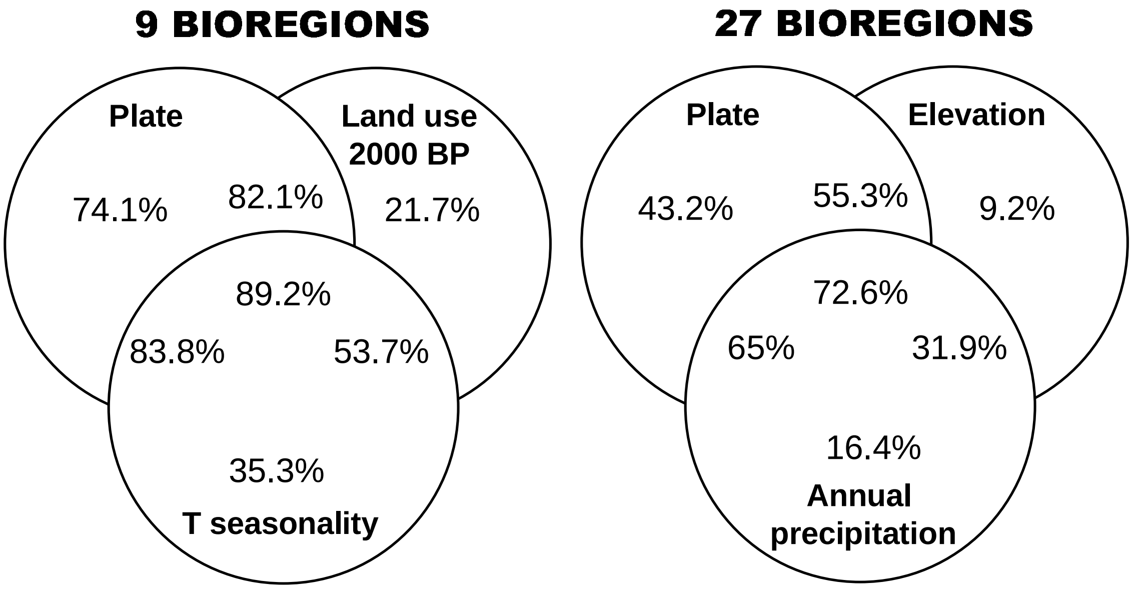


Figure 2 | Venn diagrams describing the predictive power of the three most important determinants of the two largest bioregion levels. Values indicate percentages of samples correctly classified (1-OBB error) and are obtained from running RF classification models for the individual determinants, by pairs and for the three most important determinants together. Venn diagrams for the smaller bioregion levels (141 and 1128) are not shown because the predictive value of RF models using only the three most important determinants was low (32.4% for the 141 bioregions model and 7.3% for the 1128 bioregions). But see Extended Data Table 1 for more complete information, where the determinants needed to reach ≥ 50% of samples correctly classified for each biogeographical level is shown.

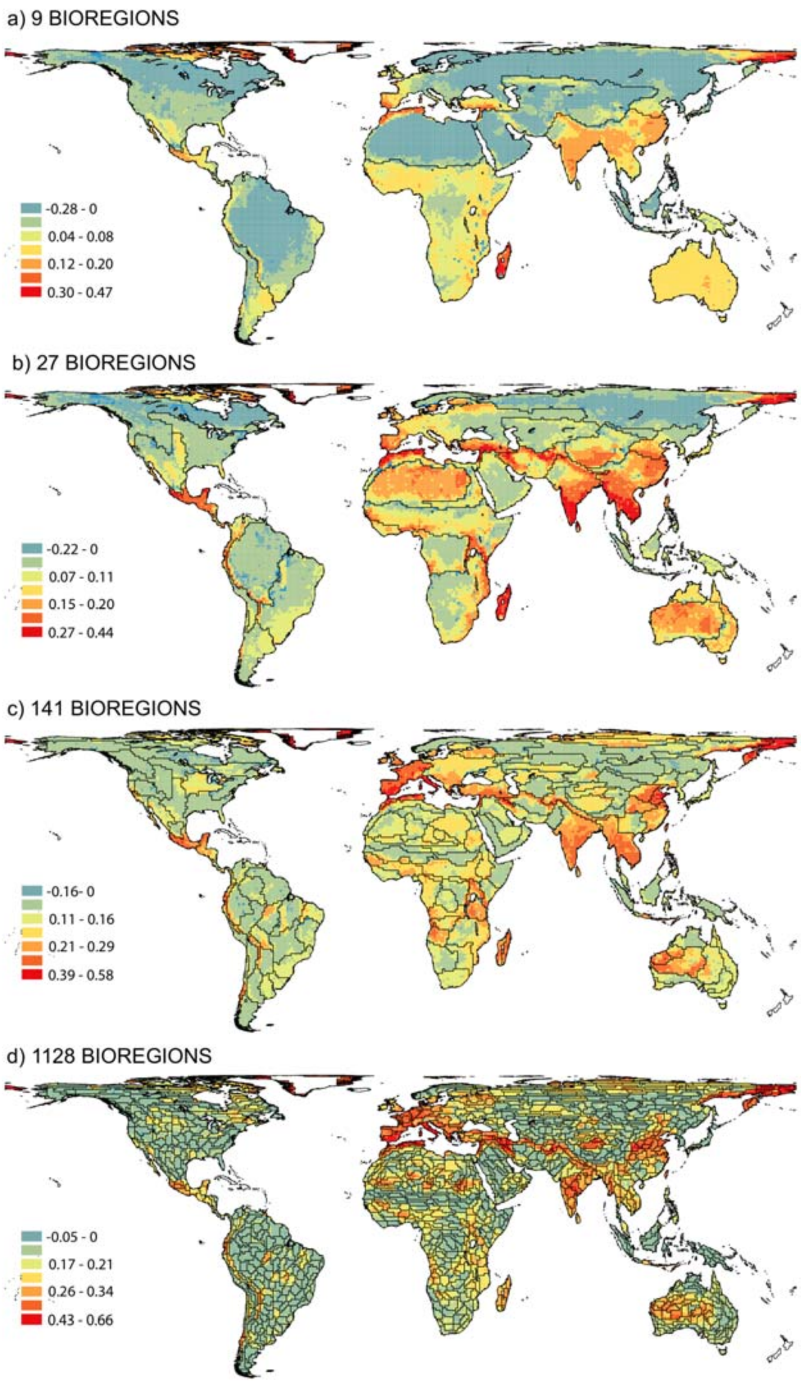


Figure 3 | Local importance values of the human impact 2000 years ago for the four biogeographical scales. The legends show the impact on correct classification of single samples (grid cells): negative, 0 (the variable is neutral) and positive. See Extended Data Fig. 2 to compare these results with the map showing human land use 2000 years ago.

METHODS

Data. Range maps for non-volant terrestrial mammals were obtained from the IUCN (IUCN 2015). We extracted species occurrences into a global grid with a 9309 km^2 grain size – based on Behrmann's projection – to generate a presence–absence matrix in which every row represents a grid cell and every column a species. We then excluded (1) cells containing less than 50% of land area to approximate equal-size samples, and as a result small islands and coastal cells were not included in the analyses, and (2) cells containing fewer than five species to reduce potential distortions caused by having few taxa[5,6]. These exclusion criteria rendered a total of 14,097 cells representing 3960 terrestrial mammal species.

Predictors. To each of the above 14,097 cells we assigned a mean value of the following predictors, which were grouped in contemporary (~ present day), recent past (~ from 130,000 until 2000 years ago) and remote past predictors (~65 Millions of years ago). Contemporary predictors included current climate (annual total precipitation, mean annual temperature, seasonality in temperature and seasonality in precipitation), habitat-related predictors (primary productivity, main land cover type and tree cover), and human impact related predictors (human appropriation of net primary production, and the human influence index). The four climatic variables were extracted from the WorldClim dataset[31]. These variables represent both average conditions and variability within years and have been shown to be determinants of vertebrate distributions[32]. Primary productivity was represented by the net primary productivity in terrestrial ecosystems[33]. Main land-cover type was obtained from Globcover 2009[34] and tree cover from the

Hansen et al. (2013) dataset[35]. These variables are surrogates of the vegetation structure and composition and can be considered the habitat template on which animal life-history strategies are shaped[36]. We acknowledge that the extent of tree cover may not directly affect the distribution of non-forest mammals, but here we assume that forest directly or indirectly affect the number and species composition in mammals. The human appropriation of net primary production from earth's terrestrial ecosystems (HANPP) was obtained from Haberl et al. dataset[33], and is considered an integrated socio-ecological indicator quantifying effects of human-induced changes in productivity and harvest on ecological biomass flows. The global human influence index (HII) is a composite predictor covering population density, human land use and infrastructure, and human access[37]. Net primary productivity and annual total precipitation, and HANPP and HII were highly correlated (Supplementary Table 5), so only annual total precipitation and HANPP were included in the final models.

Recent past predictors were represented by Quaternary climate changes and past anthropogenic land use. Both groups of determinants have been shown to play a role in the present-day assemblages or biodiversity patterns of plants and animals[22,23,25,38]. We selected six variables reflecting Quaternary climate changes, which represented the anomalies (i.e., absolute differences) of temperature and precipitations between the Mid Holocene (MH; ~ 6000 years ago), the Last Glacial Maximum (LGM; ~22,000 years ago), and the Last Interglacial (LIG; ~130,000 years ago) with the present. Mean annual temperature and annual total precipitation for the MH and the LGM were calculated using the Model for Interdisciplinary Research on Climate (MIROC-ESM)[39], whereas for the LIG we used the model of Otto-Bliesner et al. (2006)[40]. Past anthropogenic land use was obtained from Ellis et al. (2013) dataset[21]. We chose the more realistic KK_{10} model, which assumes that humans use land more intensively when population density is high and land scarce[41]. In counterpoint, the most popular HYDE model omits land-use intensification and predicts that except for the developed regions of Europe, human use of land was insignificant in every biome and region before A.D. 1750. We calculated human pressures at four different time spans (8000, 5000, 2000 years ago, and the present time) and the differences between past human land use (8000, 5000 and 2000 years ago) and the present time. These variables were, however, largely correlated among them (Supplementary Table 5), and finally only human land use 8000 years ago, 2000 years ago and the difference in the percentage of land used between 8000 years ago and the present were included in the models.

Finally, remote past predictors were represented by orographic barriers and plate tectonics. Mountain ranges represent major barriers to dispersal for most mammals, whereas plate tectonics are responsible of the long-term isolation of the biotas on some plates[42]. We use the GTOPO30 to calculate the mean elevation and the range in elevation per grid cell. Plate tectonics were obtained from Bird (2003) dataset[43]. Each grid cell was assigned the tectonic plate to which it belongs. When a cell was represented by more than one tectonic plate, it was assigned the one that occupied a greater percentage of the cell.

Building Biogeographical regionalizations. We applied a machine-learning algorithm referred to as affinity propagation[24] to build zoogeographical regionalizations at different biogeographical resolution. Affinity propagation (AP hereafter) is a powerful clustering algorithm extensively used in bioinformatics and astrophysics and is making inroads in ecology[44] and biogeography[6]. One of its main advantages is that it can compress massive data sets very efficiently (i.e. with lower error). The algorithm detects special data points called exemplars, and by a message-passing procedure it iteratively connects every data point to the exemplar that best represent it until an optimal set of exemplars and clusters emerges. Contrary to algorithms in which exemplars are found by randomly choosing an initial subset of data points, AP takes as input measures of 'similarities' between pairs of data points (grid cells here) and simultaneously considers all the points as potential exemplars. The optimal set of exemplars is the one for which the sum of similarities of each point to its exemplar is maximized. Hence, detecting exemplars goes beyond simple clustering because the exemplars themselves store compressed information[45].

We devised a protocol based on a successive application of AP to obtain a biogeographical upscaling from the smallest possible bioregions (i.e. the highest biogeographical resolution) to the largest ones. We first used the mammals' presence-absence matrix to calculate pairwise similarities between pairs of cells. As with clustering algorithms, there are many similarities indices or taxonomic/phylogenetic turnover metrics to choose from, and none can be said to be perfect from a biogeographical perspective. Here we selected Hellinger distance[46], which is calculated by first modifying the species-presence data and then computing the Euclidean distance among pairs of cells based on the modified data[47]. The Hellinger distance is used to avoid the 'double-zero' problem, i.e. when two sites or grid cells that have no species in common are assigned the same distance as two sites that share species; and the ‘species-abundance paradox’, which frequently occurs when two sites share only a small fraction of all the species in the same regional pool[48,49]. We performed an initial AP analysis involving all grid cells of the similarity matrix. This first AP run generates the optimal solution of the highest resolution bioregions, while also identifying its exemplars. We obtained 1128 clusters/exemplars. Then, using the exemplars (grid cells) as the new units of analysis we conducted again an AP, i.e., we calculated a new similarity matrix and re-run a new AP. This process was repeated until a small and coherent number of large clusters emerged. Finally, to obtain maps of each clustering result, we classified each grid cell (row) of every presence-absence matrix according to the cluster to which they were assigned in its corresponding AP analysis. AP analyses were performed using the '*APCluster*' package in R[50].

Statistical analyses. We used random forest[51,52] models of 5000 classification trees, implemented in the R package '*RandomForest*'[53], to assess the factors that may predict the classification of grid cells in bioregions at different biogeographical resolutions and to estimate the relative importance of the predictors. Random forest is a machine learning method based on a combination of a large set of decision trees. Each tree is trained by selecting a random set of variables and a random sample from the training dataset (i.e., the calibration data set). Out-of-bag (OOB) samples are used to calculate an unbiased error rate of the model and predictor importance, eliminating the need for a test set or cross-validation. The number of variables randomly sampled as candidates at each split, *mtry*, is a tuning parameter of random forest models. The default value for random forest classification is the squared root of the number of variables (i.e., the squared root of 19 resulting in a *mtry* of 4). We used the *tuneRF* code (available within the *RandomForest* package) to look for the optimal *mtry* parameter. We ran the code 5 times for each biogeographic scale and the results varied between 4 and 8, although the trend was always higher towards 4, which consequently was the *mtry* we finally chose.

Global and local importance of variables for classification: random forest also calculates estimates of global variable importance for classification, which are very useful to interpret the relevance of variables for the dataset under study. We measured variable importance using the mean decrease in classification. This measure is obtained by permuting randomly each predictor and assessing the decrease in classification accuracy of the model. The local variable importance is an estimate of the importance of a variable for the classification of a single sample (grid cell here) and shows a direct link between variables and samples[54]. It may therefore reveal specific variable importance patterns within groups of samples that may not be evident from the global importance values. The local importance score is derived from all trees for which the sample was not used to train the tree (i.e. its value is OBB). The percentage of correct votes for the correct class in the permuted OBB data is subtracted from the percentage of votes for the correct class in the original OOB data to assign a local importance score for the variable for which the values were permuted. The score reflects the impact on correct classification of a given sample: negative, 0 (the variable is neutral) and positive. Given that local importances are noisier than global importances we run the same classification

5 times (5 per biogeographical scale) and averaged the local importance scores to obtain a robust estimation of local importance values[54].

Random forests are able to disentangle interacting effects and identify nonlinear and scale-dependent relationships that often occur at the scale of the analysis performed here among multiple correlated predictors[52]. Although random forests are generally assumed to not be affected by highly correlated predictor variables, we eliminate some predictors showing a high correlation (r > 0.70, Supplementary Table 5) as some evidence from genomic studies suggests that variable importance measures may show a bias towards correlated predictor variables[55]. We made an exception with annual mean temperature and temperature seasonality. Both variables are highly correlated (r = -0.85), however, their ecological significance when it comes to explaining regional taxonomic differences can be very different.

References

31. Fick, S.E. & Hijmans, R.J. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climat.* **37,** 4302-4315 (2017).

32. Sexton, J.P., McIntyre, P.J., Angert, A.L. & Rice, K.J. Evolution and ecology of species range limits. *Ann. Rev. Ecol. Evol. Syst.* **40,** 415-436 (2009).

33. Haberl, H. *et al.* Quantifying and mapping the human appropriation of net primary production in earth´s terrestrial ecosystems. *Proc. Natl. Acad. Sci USA* **104,** 12942-12947 (2007).

34. European Space Agency. GlobCover 2.2: GlobCover Land Cover, European Space

Agency, Paris (2009).

35. Hansen, M.C. *et al.* High-resolution global maps of 21st-century forest cover change. *Science* **342,** 850-853 (2013).

36. Southwood, T.R.E. Habitat, the templet fo ecological stratsgies? *J. Anim. Ecol.* **46,** 337-365 (1977).

37. Sanderson, E.W. *et al.* The human footprint and the last of the wild. *BioScience* **52,** 891-904 (2002)

38. Araujo, M.B. *et al.* Quaternary climate changes explain diversity among reptiles and amphibians. *Ecography* **31,** 8-15 (2008).

39. Watanabe, S. *et al.* MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments. *Geosci. Model Dev.* **4,** 845–872 (2011).

40. Otto-Bliesner, B.L. *et al.* Simulating Arctic Climate Warmth and Icefield Retreat in the Last Interglaciation. *Science* **311,** 1751-1753 (2006).

41. Kaplan, J.O. *et al.* Holocene carbon emissions as a result of anthropogenic land cover change. *Holocene* **21,** 775–791 (2011).

42. Lomolino, M.V., Riddle, B.R. & Whittaker, J. *Biogeography*. 4th edn, Sinauer Associates, (2010).

43. Bird, P. An updated digital model of plate boundaries. *Geochemistry, Geophysics, Geosystems* **4,** 1027, doi:10.1029/2001GC000252 (2003)

44. Cardille, J.A. & Lambois, M. From the redwood forest to the Gulf Stream waters: human signature nearly ubiquitous in representative US landscapes. Front. Ecol. Environ. **8,** 130–134 (2009).

45. Mézard, M. Where are the exemplars? *Science* **315,** 949-951 (2007).

46. Rao, C.R. A review of canonical coordinates and an alternative to correspondences analysis using Hellinger distance. *Qüestiió* **19,** 23–63 (1995).

47. Legendre, P. & Gallagher, E.D. Ecological meaningful transformations for ordination of species data. *Oecologia* **129,** 271-280 (2001).

48. Legendre, P. & Legendre, L. *Numerical ecology*. 2nd edn. Elsevier, Amsterdam (1998).

49. Gagné, S. & Proulx, R. Accurate delineation of biogeographical regions depends on the use of an appropriate distance measure. *J. Biogeogr*. **36,** 561–567 (2009).

50. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27,** 2463–2464 (2011).

51. Breiman, L. Random forests. *Machine Learning* ***45,*** 5–32 (2001).

52. Cutler, D.R. *et al*. Random forests for classification in ecology. *Ecology* **88,** 2783–2792 (2007).

53. Liaw, A. & Wiener, M. Classification and regression by RandomForest. *R News* **2,** 18–22 (2002).

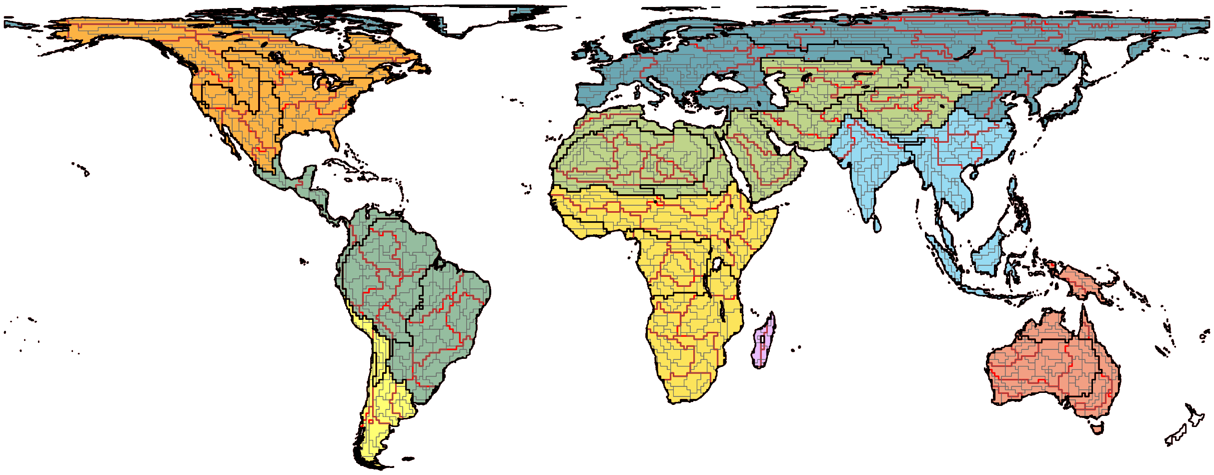
54. Touw, W.G. et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle. *Briefings in Bioinformatics* **14**, 315-326 (2012).

55. Nicodemus, K.K. *et al.* The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* **11**, 110 (2010).

Extended Data Table 1: Results in terms of percentage of samples (grid cells) correctly classified (1-OOB) of step by step RF models. The most important determinants, according to global RF models, were added one by one to the models until a 1-OOB ≥ 50% was reached.

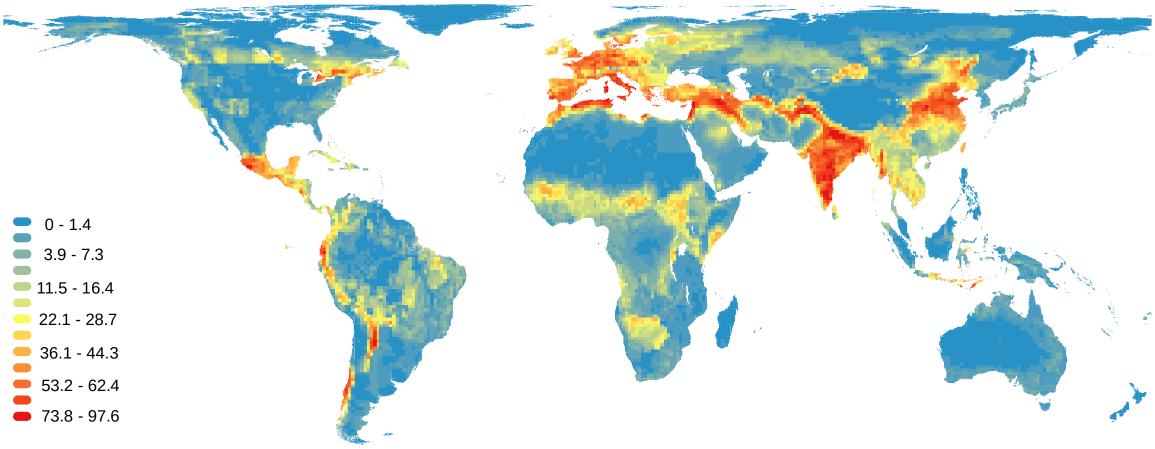
	1-OOB
9 Bioregions	
Plate tectonics	74.1
27 Bioregions	
Plate tectonics	43.2
Plate tectonics + Ann Prec	65
141 Bioregions	
Elevation	3.2
Elevation + Range Elev	22.1
Elevation + Range Elev + Plate tectonics	32.4
Elevation + Range Elev + Plate tectonics + Dif T MH	40.9
Elevation + Range Elev + Plate tectonics + Dif T MH + Dif P LIG	52.4
1128 Bioregions	
Elevation	0.6
Elevation + Dif T MH	2.1
Elevation + Dif T MH + Dif P MH	7.3
Elevation + Dif T MH + Dif P MH + Dif P LIG	17.6
Elevation + Dif T MH + Dif P MH + Dif P LIG + Dif T LIG	28.2
Elevation + Dif T MH + Dif P MH + Dif P LIG + Dif T LIG + T season	49.1
Elevation + Dif T MH + Dif P MH + Dif P LIG + Dif T LIG + T season + Ann mean T	58.2

Ann Prec = Annual precipitation; Range Elev = Range in elevation; Dif T and Dif P = Differences in annual mean temperature and annual precipitation; MH = Mid-Holocene; LIG = Last Interglacial; T season = Temperature seasonality; Ann mean T = Annual mean temperature.

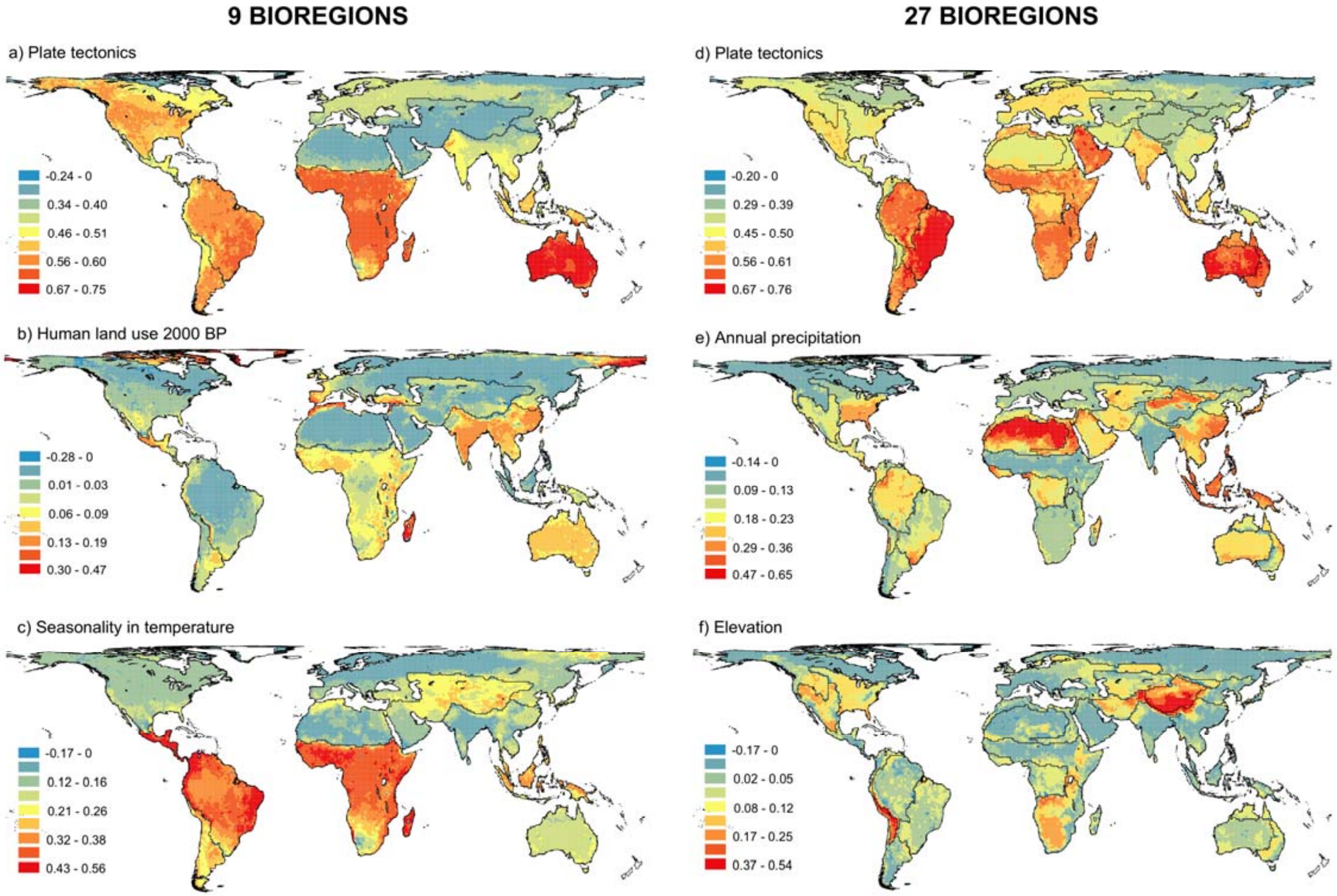


Extended Data Figure 1 | Nested global bioregionalizations for terrestrial mammals.

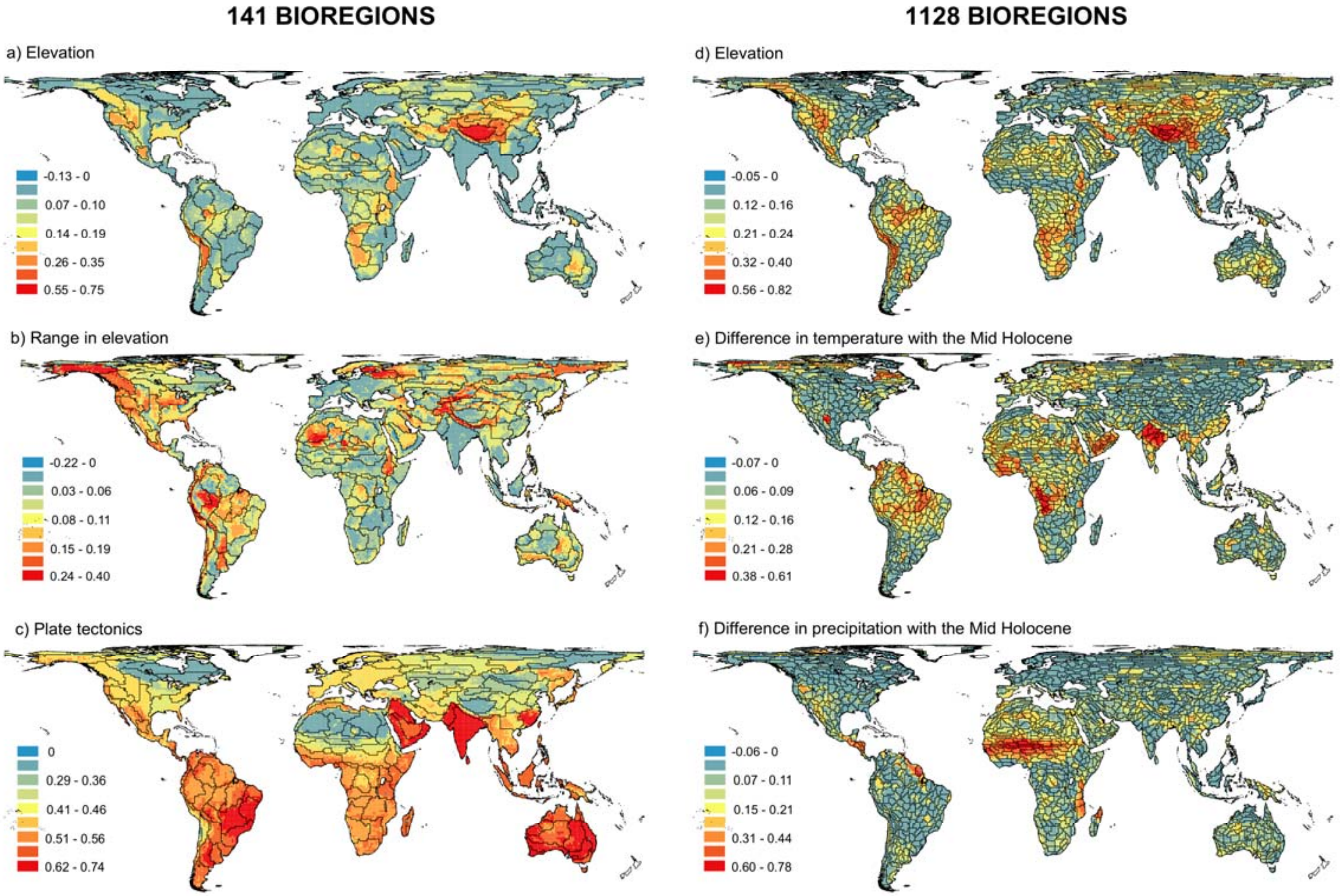
The hierarchical application of the affinity propagation algorithm resulted in a nested global bioregionalization containing 9 (filled colours), 27 (black boundaries), 141 (red boundaries) and 1128 (grey boundaries) bioregions from the largest extant realms to the smallest obtained bioregions.



Extended Data Figure 2 | Map showing anthropogenic land use 2000 years ago. The legend shows % total land in use for crops and pastures.



Extended Data Figure 3 | Local importance values for the three most important determinants of the broadest bioregions. The legends show the impact on correct classification of single samples: negative, 0 (the variable is neutral) and positive.



Extended Data Figure 4 | | Local importance values for the three most important determinants of the smaller bioregions. The legends show the impact on correct classification of single samples: negative, 0 (the variable is neutral) and positive.