Climatic heterogeneity and low precipitation dictate the global distribution of range-edge hotspots of major tree species

Supplementary Information

David Lerner[1], Marcos Fernandez Martinez[2], Stav Livne-Luzon[1], Jonathan Belmaker[3,4], Josep Peñuelas[2,5], Tamir Klein[1]

[1]Department of Plant & Environmental Sciences, Weizmann Institute of Science. Rehovot 76100, Israel.

[2]CREAF, Bellaterra (Cerdanyola de Valles), Catalonia E08193, Spain.

[3]School of Zoology, Tel Aviv University, Tel Aviv, Israel.

[4]Steinhardt Museum of Natural History, Tel Aviv University, Tel Aviv, Israel.

[5]CSIC, Global Ecology Unit CREAF-CSIC-UAB, Bellaterra, Catalonia E08193, Spain

*David Lerner

Supplementary methods

1. Polygon formation

In order to cluster occurrences into polygons we used the DBSCAN package. We used the kNNdistplot function (DBSCAN package) to determine a suitable epsilon size given the size of the distribution (number of occurrences), as recommended[1]. We used epsilon sizes of 2, 3.8, and 8.41 degrees for species with >20000 occurrences, <20000 and >7500 occurrences, and <7400 occurrences, respectively. The variable *minPts* was selected to balance several objectives: remove occurrences that were distant enough to the population mean (probably suggesting false positives), identify and isolate populations that seemed distant enough to avoid overestimating false negatives, and use the same value across all species.

We nonetheless identified cases where DBSCAN may have removed true occurrences. We decided to implement an algorithm that would identify these possible *false* outliers and reintroduce them into the clusters due to the importance of accurately identifying REs. The algorithm standardizes distances of every occurrence given the centroid of every cluster (using the z-score of the medians from every point relative to the centroid). We re-introduced occurrences based on two criteria: 1) the percentage increase of outlier z-score to the most distant occurrence (highest z-score) within the cluster (reintroduce any occurrence with a z-score <1.5-fold higher), and 2) the percentage increase in difference (Δ_1) in z-scores between the most distant occurrence (highest z-score) in the cluster and the outlier to the difference (Δ_2) between the most distant occurrence (highest z-score) in the cluster and the two previous most distant occurrences (reintroduced occurrences where $\frac{\Delta 1}{\Delta 2} < 3$). Both conditions needed to be true for an occurrence to be reintroduced.

We calculated the geographical edge perimetry of each species to fit a concave-hull on the previously determined clusters (using a concavity value of 2.5). We chose this value by manually curating a random set of samples for obtaining a perimeter that did not overrepresent the area of land when encapsulating the occurrence areas, did not underrepresent the area by overconstricting the perimeter, and used the same value that would work accordingly throughout all randomly selected species distributions.

2. RE determination

The REs for each quartile were determined as the two most outward coordinates of the corresponding cardinal directions (e.g. north and east cardinal coordinates for the NE quartile). Eight REs were thus determined for each population (two for each cardinal direction). As a filtering step, we accounted for both REs from the same cardinal direction if they were >20 arc-degrees apart, otherwise we only accounted for the farthest point from the centroid.

Coastline REs were determined by creating a semicircle (buffer of 3 arc-degrees) around each RE in the direction of its cardinal coordinate and measured the percentage overlap with water. Cardinal coordinates with ≥50% overlap were considered a "coastal RE". We estimated kernel densities to identify the similarities of inland RE distributions between continents. Bootstrapping was carried out with the global population (except for the Australian population) and was used to determine if any of the cardinal REs from the different continents had a significantly higher percentage of internal than coastline REs than what we would find by chance.

3. Statistical Analyses - Linear regression models**:**

We tested the dependence of the REs on each climatic variable independently using generalized linear models (linear equation: RE density ~ $climate_x$) given the number of REs at each hexagon due to the strong correlations between the absolute climatic variables and between their *SH* equivalents (Fig. S7a). We used two alternative methods that accounted for the bias of the number of species in each hexagon. We used a binomial-regression model (Fig. 3a), assuming a binomial distribution for the response variable, split into two columns: *successes* (# REs) and *failures* (# species - # REs). The second method normalized the number of REs by the number of species in that hexagon (Fig. S5a).We approximated the distribution of normalized RE densities per hexagon to be a Tweedie distribution[2,3], as described by Hasan et. al. (2012), where the mass of the distribution was set to zero but otherwise continuous on the positive real numbers. We set var.power to 1 and link.power to zero. Predictor variables were standardized for the regression models (mean of zero, standard deviation of 1).

Dominance analysis[4,5] was used to determine the total and relative contribution of each absolute climatic variable and its spatial heterogeneous counterpart. Apart from accounting for multicollinearity between predictor variables, the method accounts for the relative contribution of each predictor by averaging R^2 from the three effects. Computational power increases exponentially as the number of variables increases, so we grouped all climatic variables (n = 40) into four groups (Fig. S6): absolute vs spatially heterogeneous climates, divided into temperature-associated (BioClim 1-11) vs precipitation-associated + elevation (BioClim 12-19 + elevation). We used the most dominant predictors from the dominance analysis to run a regression model using a binomial distribution (see above) with the strongest predictors.

Predictor pairs with strong multicolinearity (R>0.75) were disregarded, and only the most dominant pairs were selected (Figs. 4d and S7c).

Figures

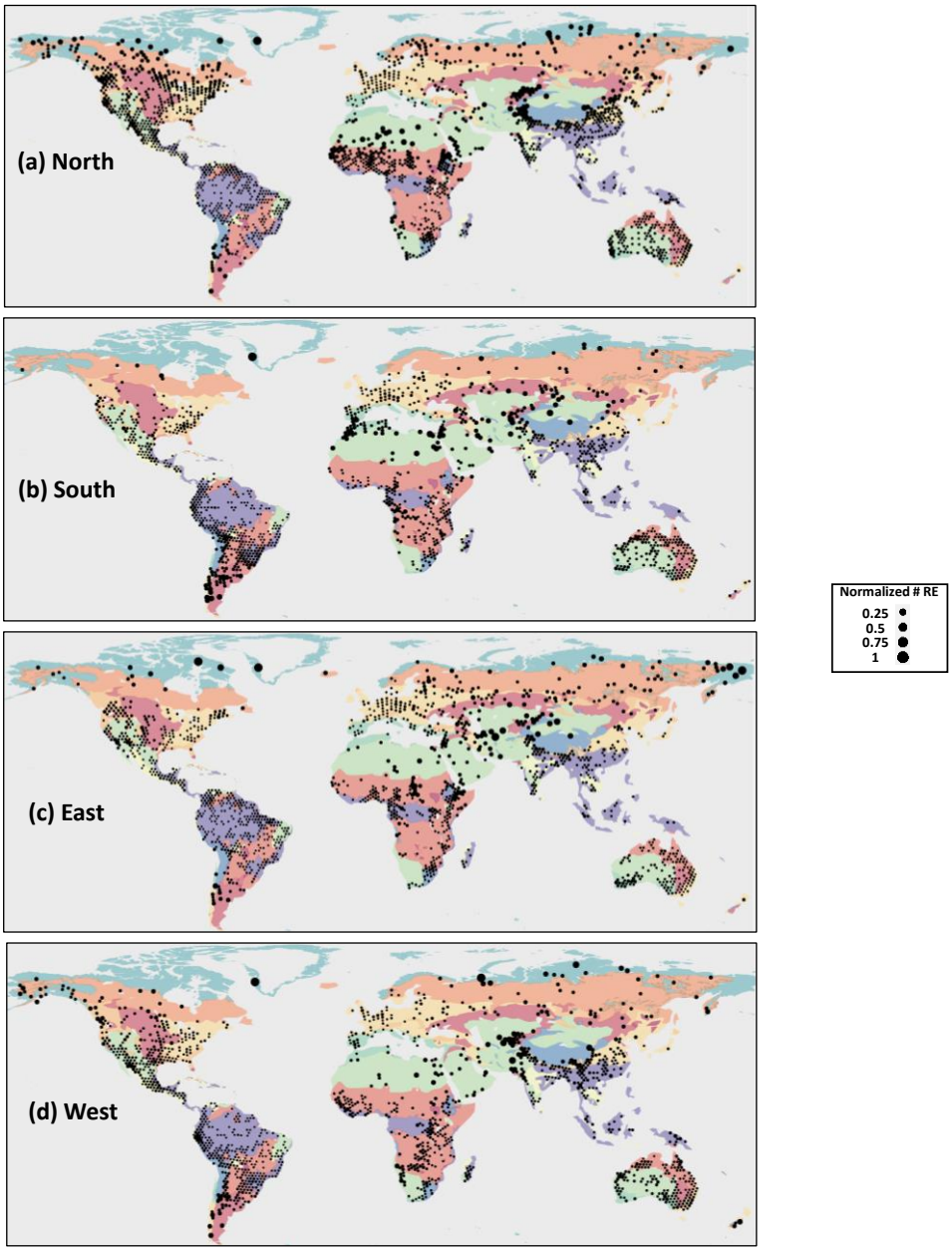


Figure S1: Normalized Range edges using the 'cardinal coordinate system' overlaid with the distribution of biomes, as in Fig 1.

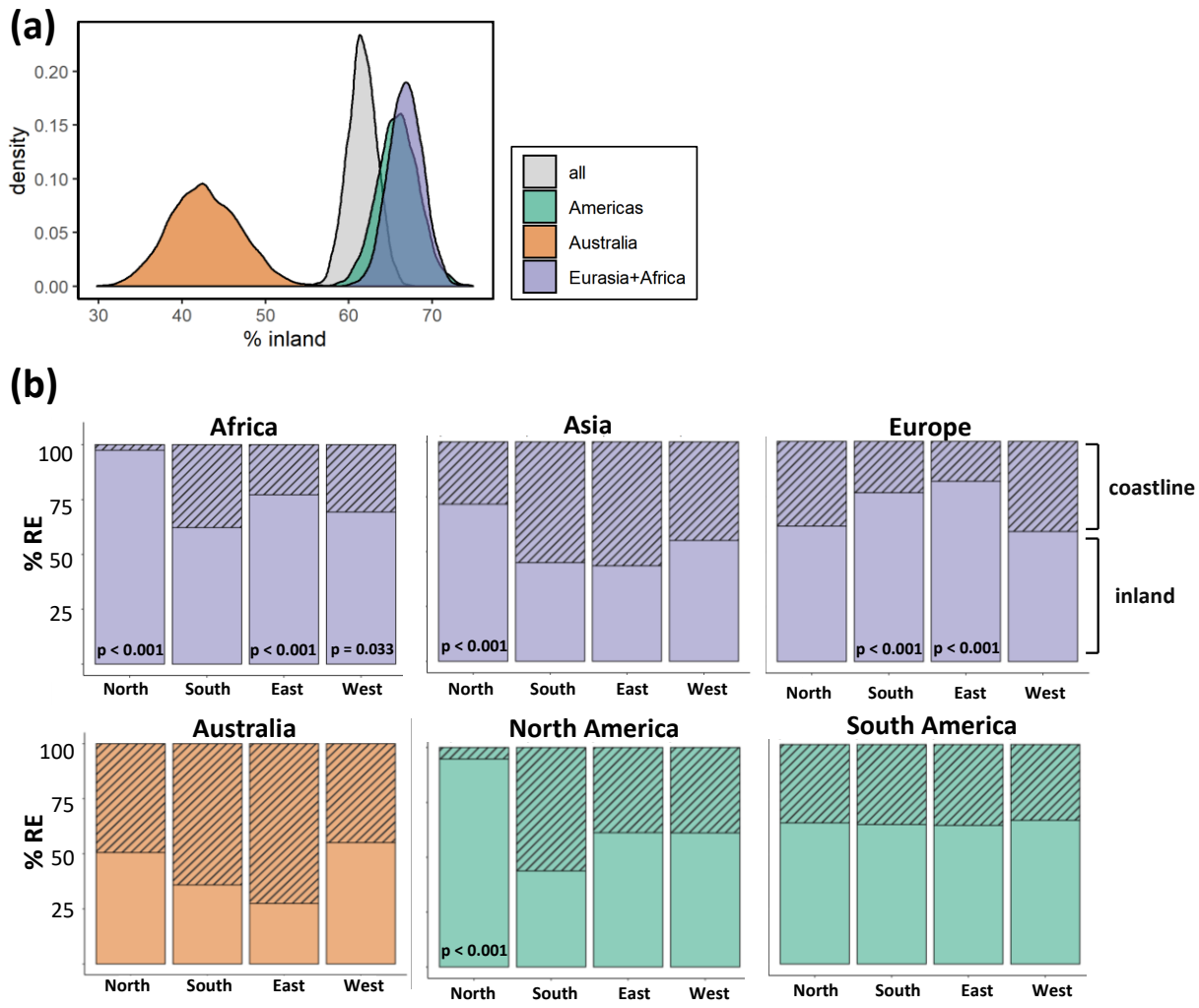


Figure S2. Global distribution of inland range edges calculated in percentage between inland and coastline RE. (a) Modeled distribution using kernel-density estimation of the distinct continental regions. The grey distribution represents the a global modelled distribution. (b) % inland & coastline RE (smooth and striped colors, respectively) divided into the distinct continents and further subdivided into the four cardinal directions . Statistical significance of high % internal RE was calculated using a modeled distribution of inland RE for Americas and Eurasia + Africa. Situations with high significance are marked with their respective p-value. See Methods for further information on the methodology used to determine inland RE and to calculate the modeled distributions.

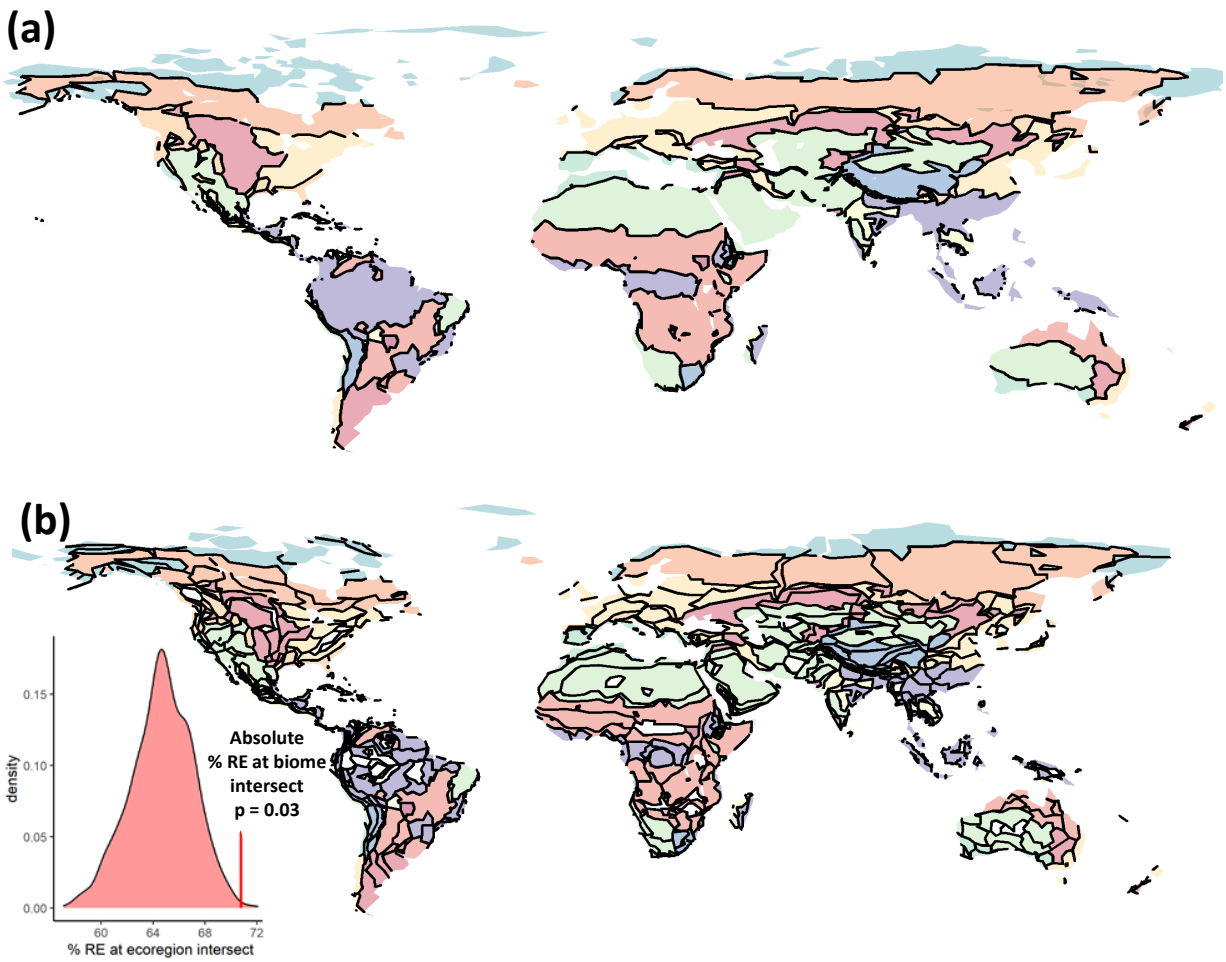


Figure S3: Biome and ecoregion intersections. The black line illustrates the buffer zone delineated at the intersection between (a) biomes (14) and (b) ecoregions (867). The insert in (b) represents the modeled distribution of % hotspots at the edge of ecoregions from a permutated (randomized) global distribution of hotspots. The arrow marks the % RE hotspots at biome intersections from the actual dataset (p = 0.03)

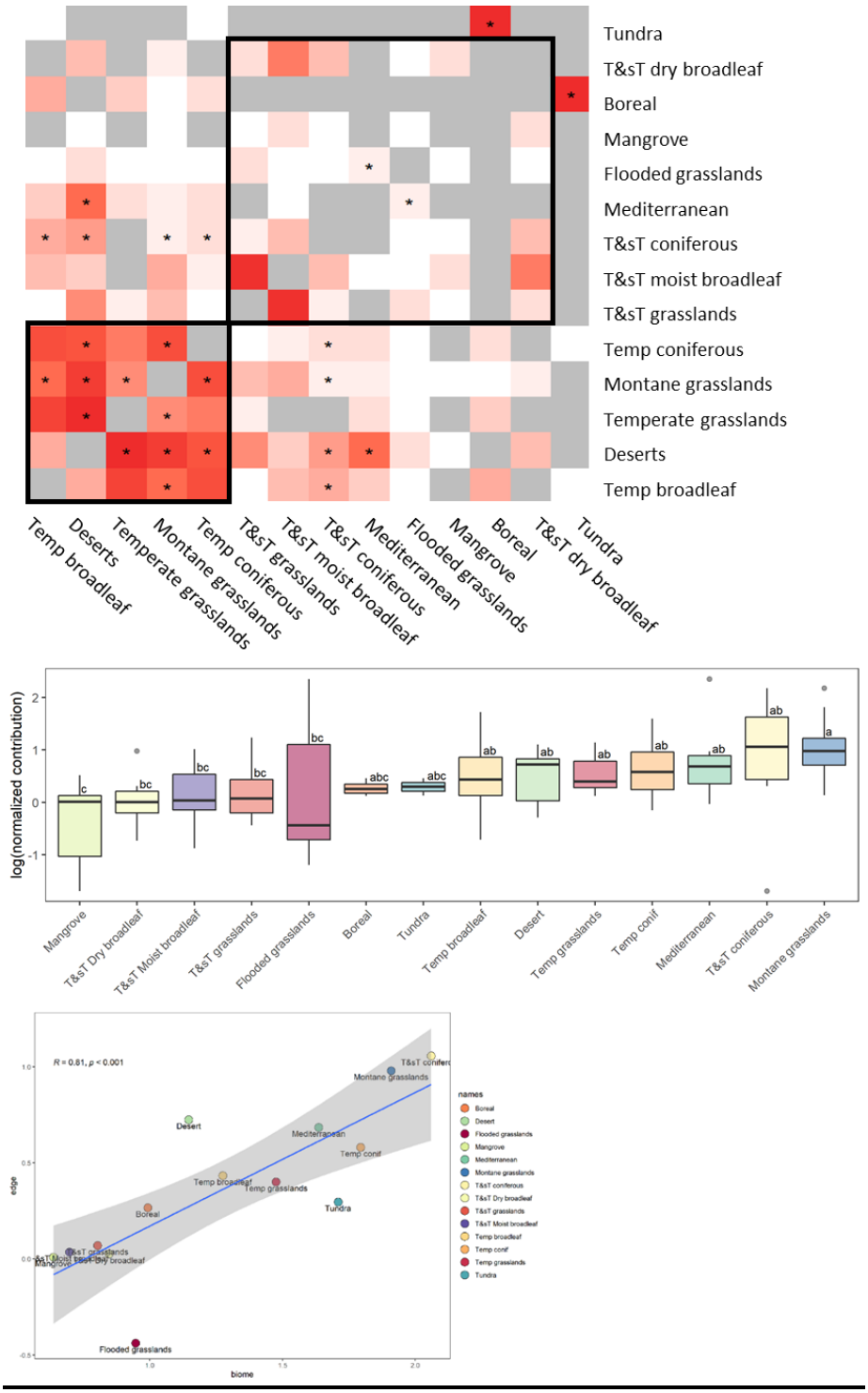


Figure S4: Range edges at the intersection of biomes. Parallel analyses as those carried out in Fig 2. Rather than analyzing the number of hotspots at the edge of biomes, we have analyzed here the total number of (normalized) RE.

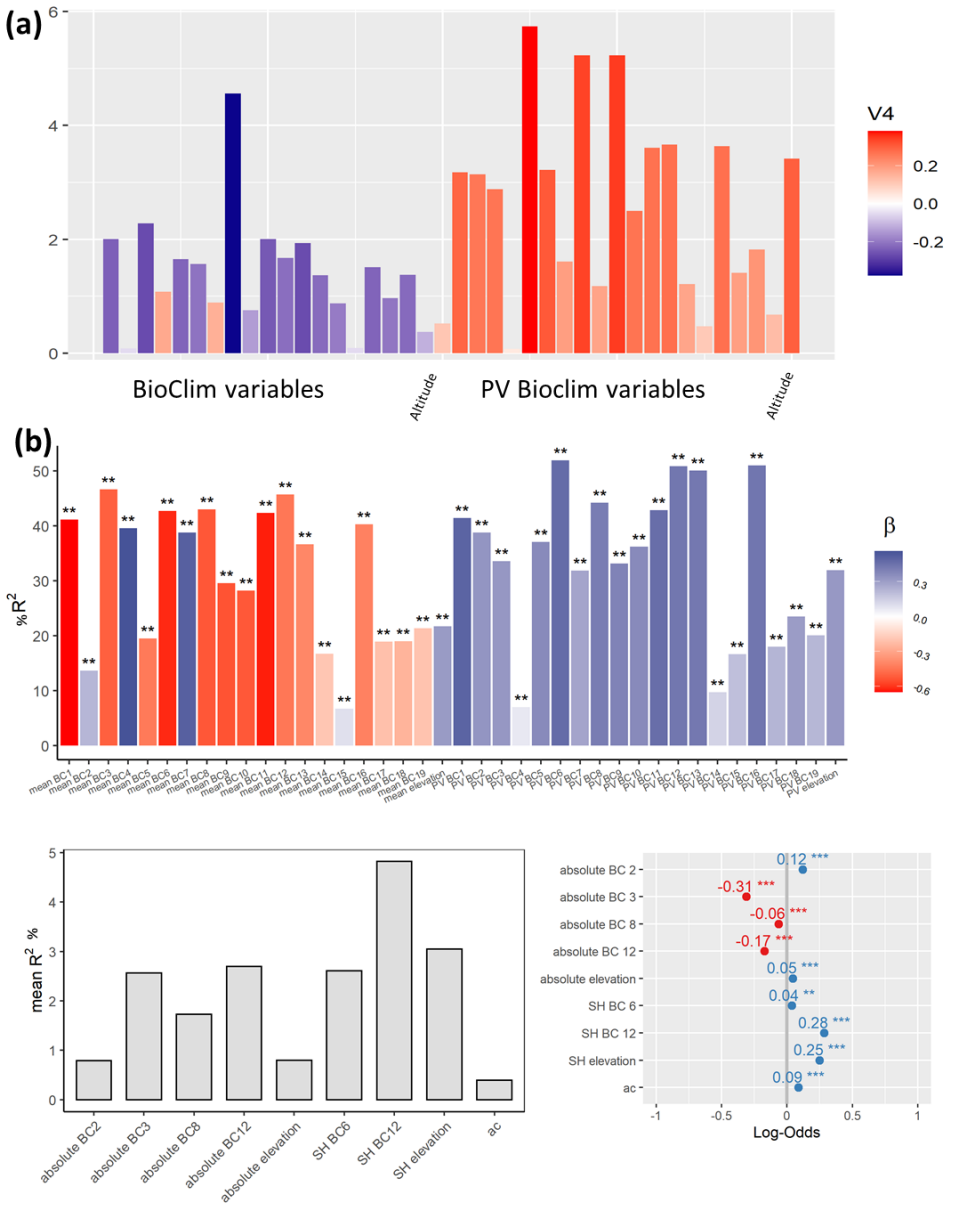


Figure S5: GLM model results a) Individual regression models between each of the predictor variables and number of RE. (1-19 are BioClim (BC) variables). In this instance, the dependent variable RE is the normalized number of range edges by the total number of species. A Tweedie distribution model was used; b) carried out when considering for ac: the spatial autocovariate to account for and remove spatial autocorrelation from the linear regression

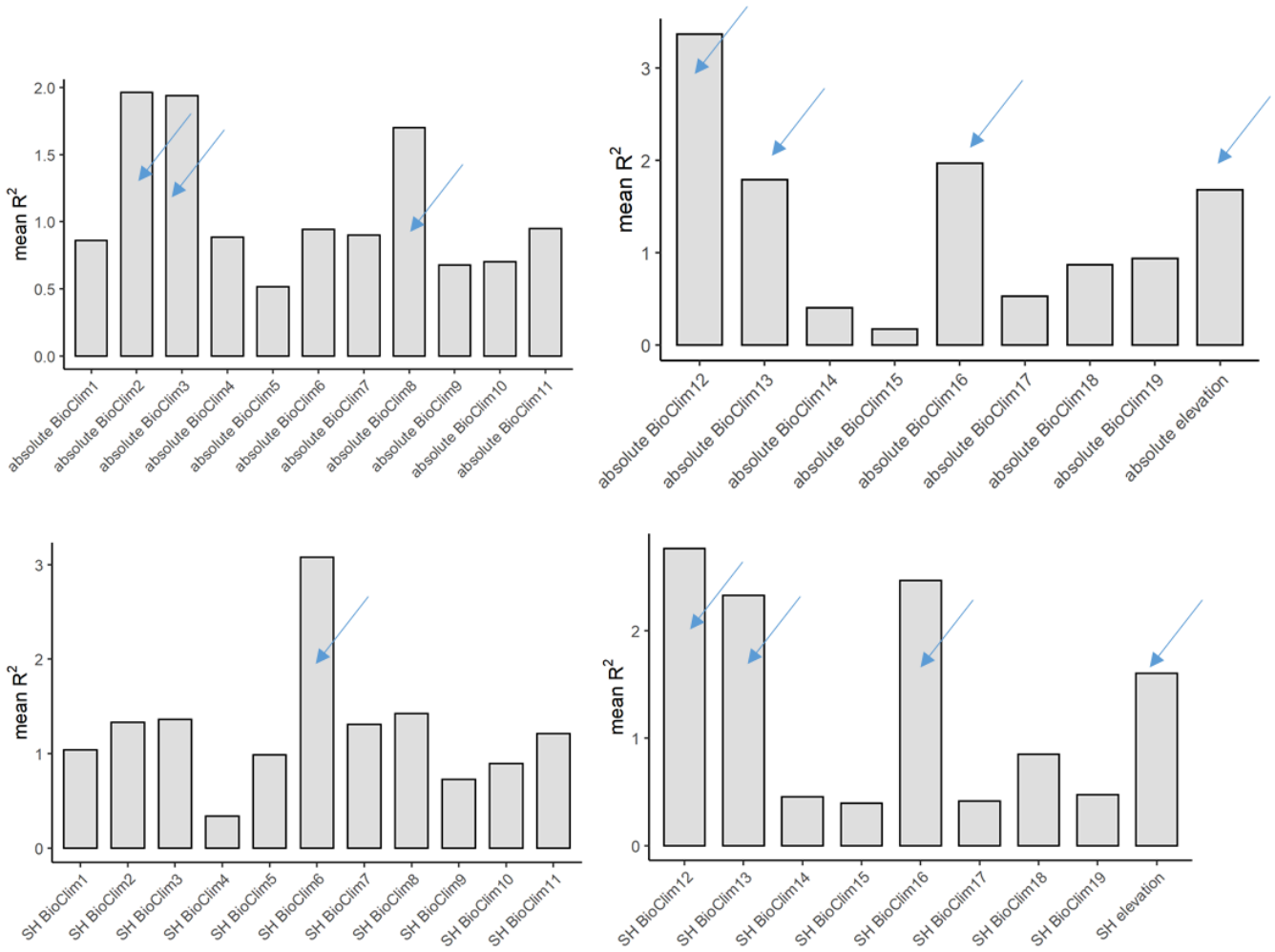


Figure S6: Dominance Analyses on the four sub-groups of climatic predictors: Absolute temperature climates, absolute precipitation, spatial heterogenous climate and spatial heterogeneous precipitation

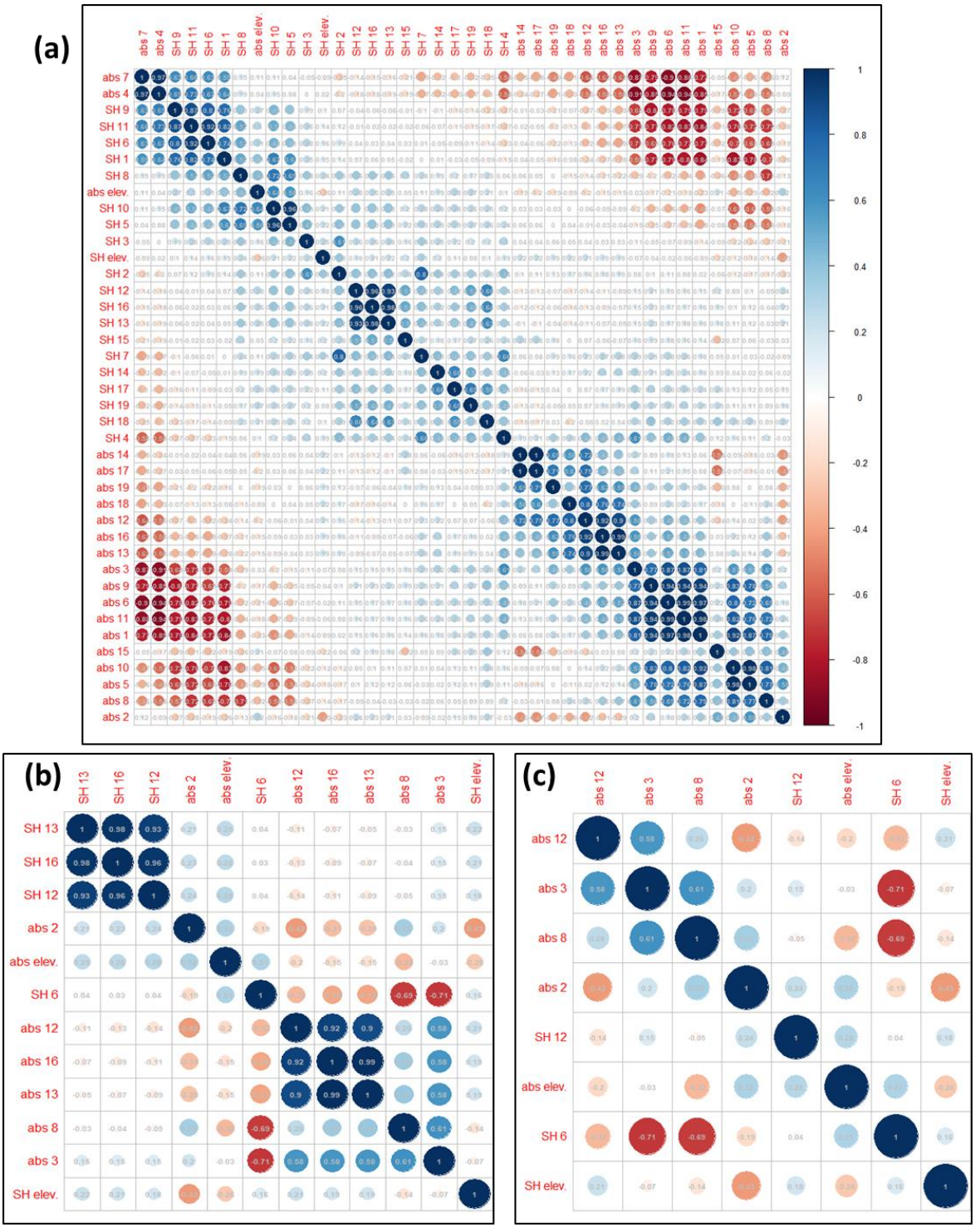


Figure S7: Correlations between all the climatic variables. The red and blue indicate negative and positive correlations, respectively.

Cleaning Pipeline

1 Download occurrences (60,000 species) fom GBIF

2 Filter occurences:
- Uncertainty > 100km
- Records from: Fossil, Unknown
- Using *CoordinateCleaner*

3 Filter species by occurences: 300>n>50,000

Polygon formation

1 DBSCAN:
1. Remove outliers
2. Create clusters

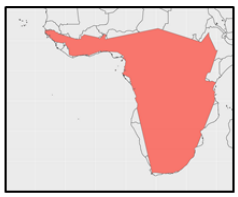
2 Reinsert *inliers*

3 Concave hulls

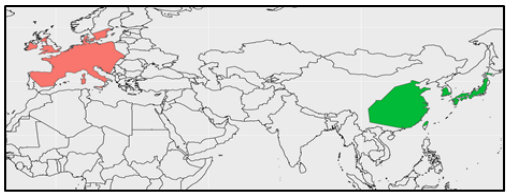
4 Cluster specie into populations (d<500km)

E.g.

Ficus sur

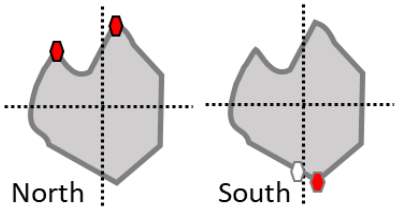


Ginkgo biloba

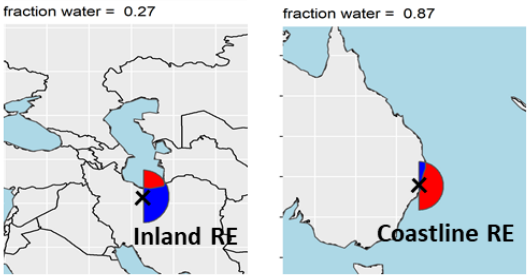


RE distribution

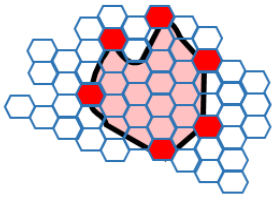
1 RE identification:
Cardinal coordinate system



2 Inland vs. Coastile RE determination



3 Range edge global distribution



4 Hotspot determination

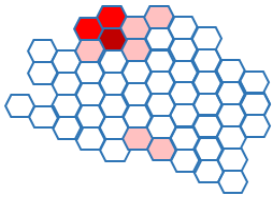


Figure S8. Summary and visual illustration of the methodology to obtain global RE distribution.

A. Cleaning pipeline – (1) Data for 60,000 tree species from GBIF. (2) Filter occurrence on the basis of different criteria, and using the R package *CoordinateCleaner*. (3) Filter species on the basis of the final number of occurrences.

B. Polygon Formation – (1) Spatial clustering using DBSCAN allowing for outlier removal. (2) *Inliers* reinserted using an algorithm on the basis of their relative distance to rest of the cluster occurrences. (3) Concave hulls delineated from the final clusters. (4) Polygons clustered into populations on the basis of absolute distance from one another (500km).

C. RE distribution – (1) REs identified by subdividing polygons into four quadrants based on the distribution's centroid. The two most extreme coordinates for each quadrant are obtained (two coordinates in each cardinal direction). If absolute distance is large, both are considered (e.g. North in illustration), otherwise only the most extreme (e.g. South in illustration). (2) Inland REs determined by forming a semicircle in the direction of the RE (e.g. East in the illustration). If the fraction of the semicircle at the edge of the water source is >0.5 then RE classified as coastline, otherwise inland. (3) Inland REs logged onto a rastered world (hexagon grid). (4) Hotspots determined on the basis of the hexagon's RE density and proximity to other dense hexagons. In the illustration, the red patch represents a hotspot area (darker color represents stronger hotspot hexagons, i.e. Z-score.)

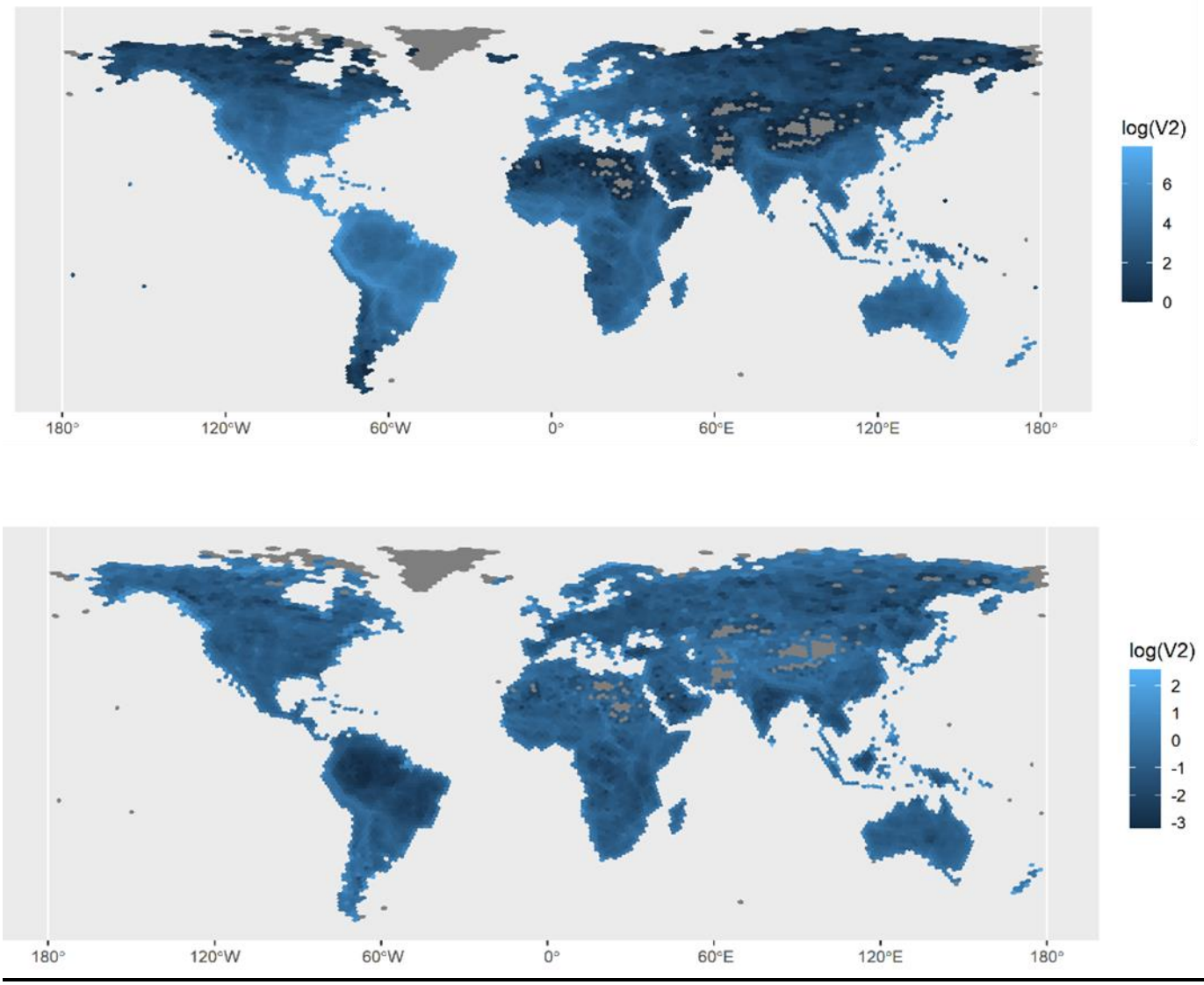


Figure S9: Range edges through the 'perimeter system' (a) Absolute density of range edges per hexagon (b) Normalized density of range edges in each hexagon (RE intersecting/species intersecting). The intensity of each hexagon's density is represented by the intensity of color. Hexagons with no intersections are colored grey.

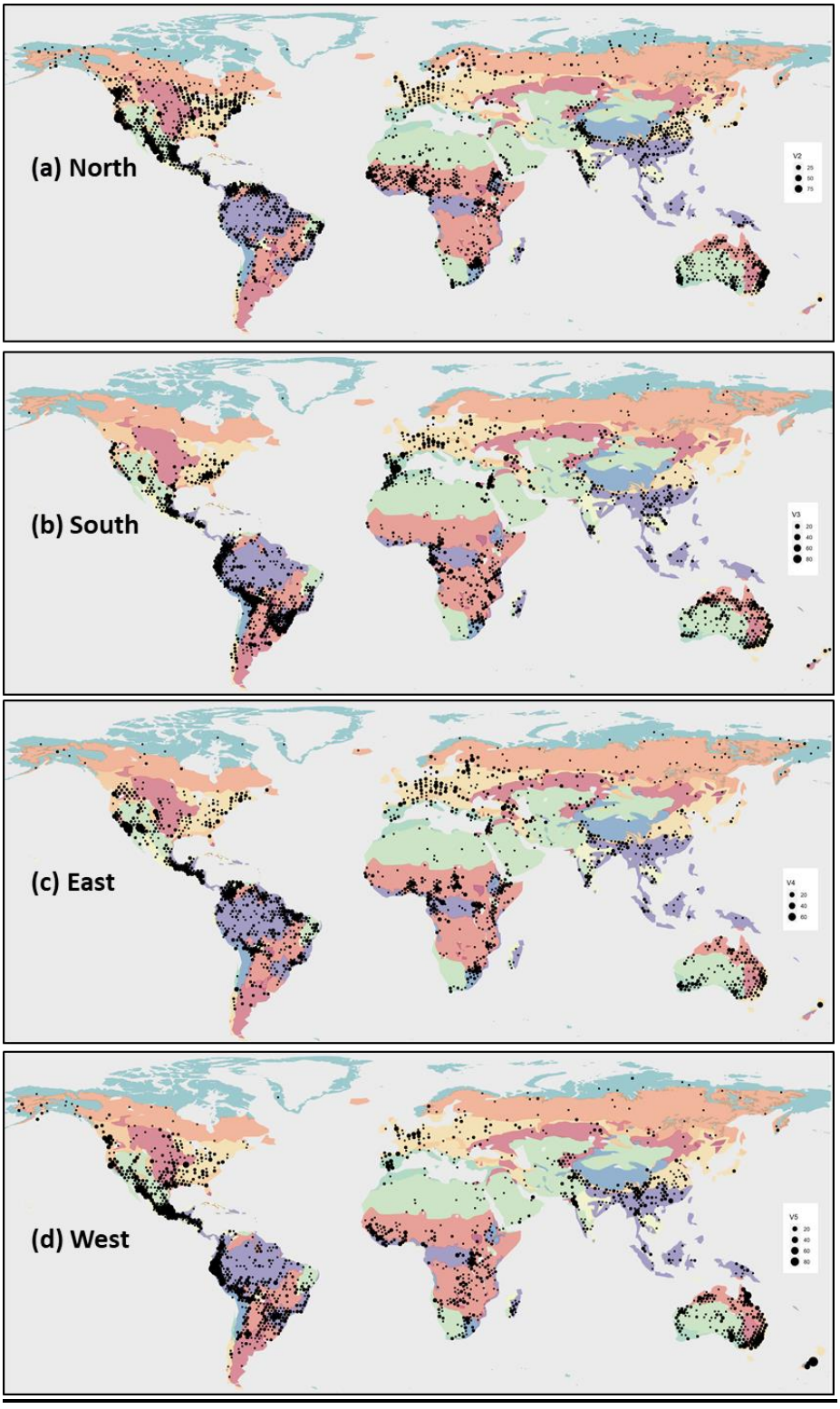


Figure S10: Absolute number of Range edges (before normalization) using the 'cardinal coordinate system' overlaid with the distribution of biomes, as in Fig 1.

References

1. Hahsler, M., Piekenbrock, M. & Doran, D. Dbscan: Fast density-based clustering with R. *Journal of Statistical Software* **91**, (2019).

2. Tweedie, M. C. K. An index which distinguishes between some important exponential families. In Statistics: Applications and New Directions. *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference.* 579–604 (1984).

3. Dunn, P. K. & Smyth, G. K. *Generalized Linear Models with Examples in R*. (Springer-Verlag., 2018).

4. Budescu, D. v. *Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression*. *Psychological Bulletin* vol. 114 (1993).

5. Luo, W. & Azen, R. Determining Predictor Importance in Hierarchical Linear Models Using Dominance Analysis. *Journal of Educational and Behavioral Statistics* **38**, 3–31 (2013).