Supplementary Information for

Drivers of geographic patterns of North American language diversity

Marco Túlio Pacheco Coelho, Elisa Barreto Pereira, Hannah Haynie,Thiago F. Rangel, Patrick Kavanagh,Kathryn R. Kirby, Simon J. Greenhill, Claire Bowern, Russell D. Gray, Robert K. Colwell, Nicholas Evans, Michael C. Gavin

Marco Túlio Pacheco Coelho

Email: *marcotpcoelho@gmail.com*

Michael C. Gavin

Email: *Michael.Gavin@colostate.edu*

**Supplementary Methods**

**Population Density Estimates**

Kavanagh et al. [41] used a piecewise structural equation modelling (SEM) that assumes the direct and indirect effects of environmental and cultural variables to estimate the population density of foraging societies. The model assumes the effect of productivity, topography, precipitation seasonality, distance to coast (i.e. access to marine resources), resource ownership (i.e. whether resources are owned or not) and

residential mobility (i.e. average distance travelled per residential move). All these variables were previously hypothesized to affect population density of foraging societies due to their influence on the availability of resources or the foraging practices of human groups. Kavanagh et al. [41] fitted a piecewise-SEM to empirical societies of hunters and gatherers, and they showed that the model explained 77% of the variation in population density among observed foraging societies. With the fitted model, the authors estimated population density at 0.5x0.5° cells for the world. In our study we used the estimations of population density for North America in order to explore the effect of population density on language diversity as well as to define the carrying capacity for the simulation model (see *Simulation Model* section).

**Simulation Model (Carrying capacity with group size limits)**

We estimated the variable “carrying capacity with group size limits” based on a recent simulation model developed to better understand the effects of climate and demography on language diversity [49]. Here we present a detailed description of the simulation model.

*Hexagon cell resolution*

We defined a hexagon and equal area gridded map of 0.5x0.5° for North America, and we extracted population density (people per km2) [based on 41, see population density section above] for each geographical cell. We used the hexagon cell to simulate the expansion of language ranges over space (see *Model Algorithm* section). Two criteria define the resolution of the hexagon gridded map: (i) cells must be large enough to encompass a group of individuals, but smaller than most observed language ranges in North America, and (ii) population density needs to be extracted without interpolating the data to finer grid resolutions. Population Density is only available in 0.5 degrees of resolution. The use of a finer resolution (< 0.5°) would require us to interpolate population density, which would generate uncertainty in the data used as an input to the simulation model (see *Carrying Capacity* section). However, coarser resolutions (> 0.5°) would produce fewer total languages because languages with smallest ranges generated by the simulation would have ranges larger than the smallest observed language ranges in North America. Therefore, using a coarsest resolution would generate fewer languages over space because of a spatial constraint in the definition of the grid. As we show latter, this grid resolution can produce a total number of languages that precisely resembles the observed data.

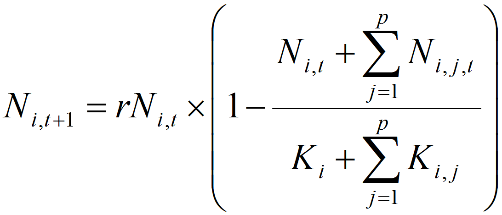
*Carrying Capacity*

Unlike in Gavin et al. [49], we did not use parameter estimation to define the best mathematical function that describes carrying capacity. Instead of estimating carrying capacity based on different mathematical functions [see 49], we calculated carrying capacity using estimates of population density [41] and the area of each hexagonal cell.

*Model Algorithm*

Here we summarize the steps of the simulation model, which was implemented following similar procedures as Gavin et al. [49]:

1. Ten individuals of a language group occupy a randomly chosen hexagon cell (*i*);
2. A maximum group size is defined for the language group by sampling the empirical distribution of hunters, gatherers and fishers group sizes [43].
3. At each algorithm time step (*t*), a regional carrying capacity is defined (K*i,j*) for each occupied cell (*i*) by summing the carrying capacity of the cell *i* (K*i*) and the carrying capacity of all its *p* neighboring cells (i.e. cells that share an edge with the focal cell *i*).
4. The increase in population size (*N*) between time step *t* and the next time step (t + 1) is given by:

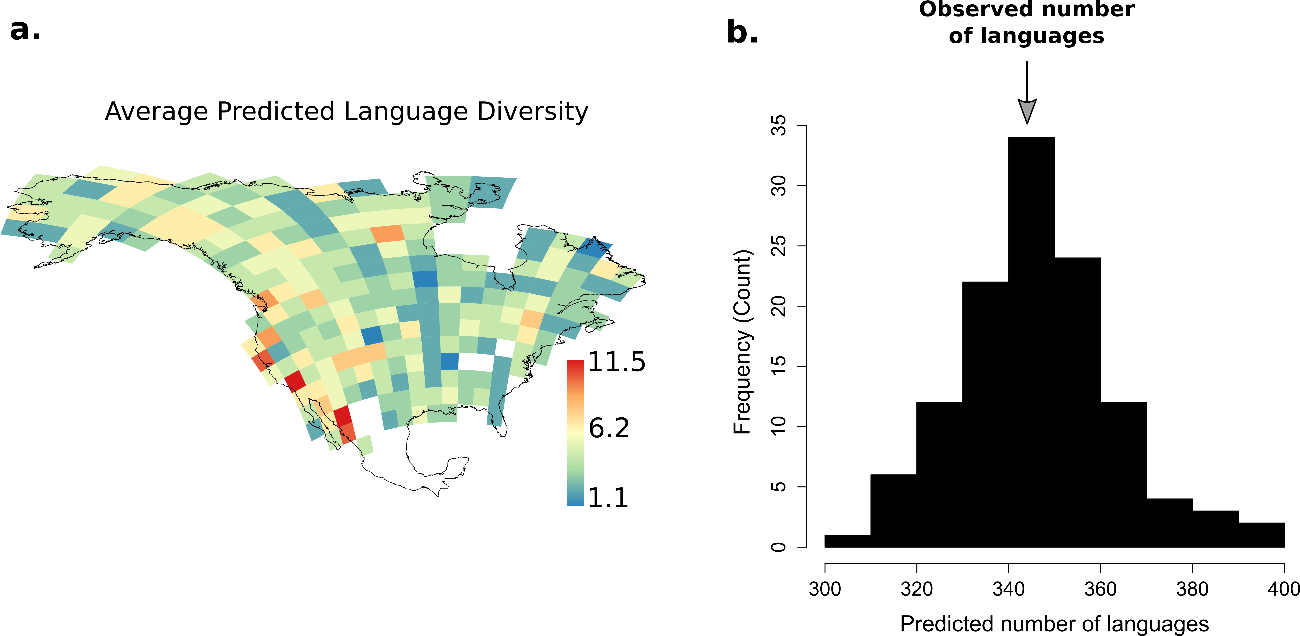


where *r* = 1.01 (i.e. per capita intrinsic rate in population growth), Σ*Ni,j,t* is the number of individuals at time *t* in all *p* cells, indexed by j, that are adjacent to cell *i*, and Σ*Ki,j* is the regional carrying capacity. This equation takes into consideration the potential population growth of individuals that are present in the cell *i,* but also the opportunity for colonization of the adjacent cells. Changes in *r* do not affect the outcome of the model, only the rates of expansion of each language.

1. The population of the focal language grows until it reaches the maximum group size that was defined based on sampling the empirical HGF distribution (step 2).
2. As soon as the population of the language reaches its maximum population, it stops growing and an empty cell is randomly chosen at the edge of the previous growing language (if it is the first language), or from a randomly selected language (if richness > 1).
3. As the new language emerge in the simulation the same procedures from 1 to 6 are repeated. The new population can colonize any empty cell but do not colonize any occupied geographical cell.
4. The simulation stops when all cells are colonized.

*Model prediction*

Because we randomly selected the first cell to be colonized by any language group (see steps 1 and 6 of model algorithm), the simulation model is stochastic. Thus, we replicated the model 120 times [49] and recorded the spatial pattern of language diversity (number of languages per 300x300km cells) and the total number of languages. The predictions extracted from the model and used in our path analysis was a ratio between the number of languages predicted in each cell and the total number of languages predicted for the geographic domain (see Fig S7). We used the average among 120 model replicates as our “carrying capacity with group size limits” variable in the path models. The average richness map and the distribution of the total number of languages predicted by the model are represented in Fig S1.



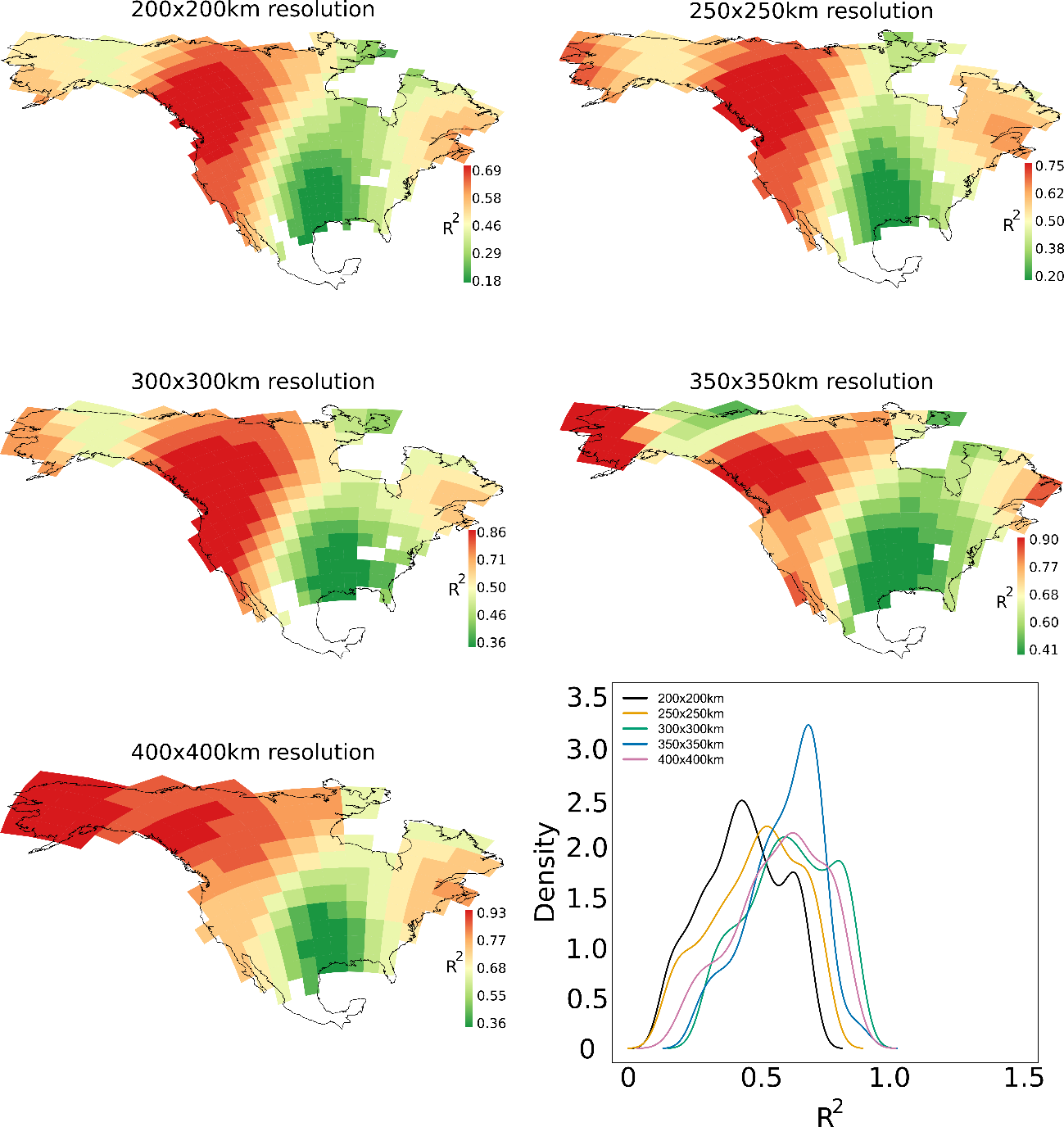
**Fig S1**. Average predicted language diversity in North America (a) and total predicted language diversity (b) based on 120 model replicates. North America presents 344 aboriginal languages. The simulation predicts and average of 346.49 languages.

Statistical Analysis – Additional details

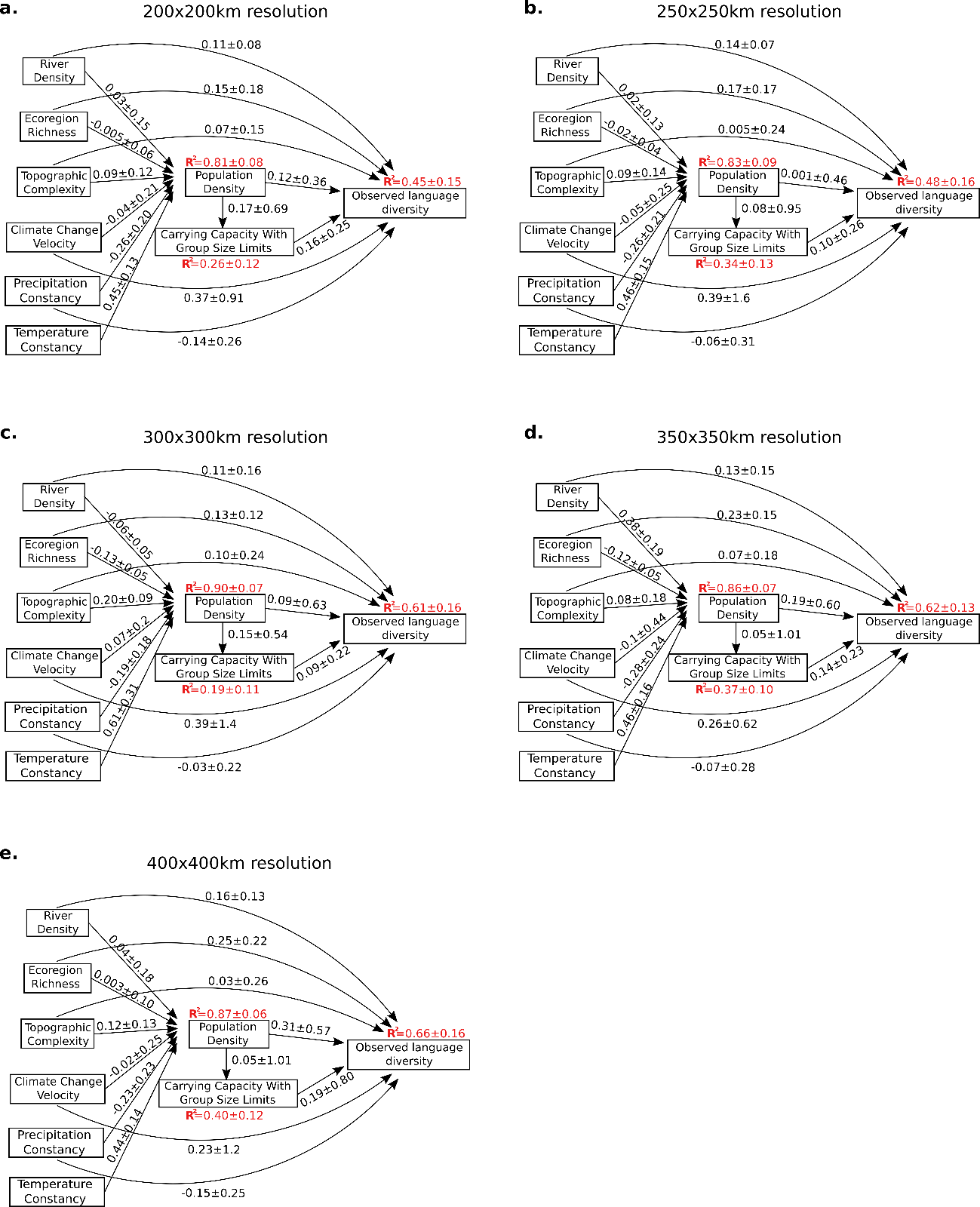
We Z-transformed all variables to allow for direct comparisons between path coefficients. Therefore, because variables are standardized, we can examine which variable presents the highest or lowest coefficient as in partial regression coefficients. To avoid multicollinearity issues, we tested the association among predictors and followed the standard statistical interpretation that correlations > |0.70| should be avoided between predictors [70]. Because population density and temperature constancy were highly correlated (r = 0.86, Table S1) and the direct effect of population density on language diversity already captures the effect of temperature constancy on language diversity, we removed the direct effect of temperature constancy from our analysis. Therefore, the final path model is composed of three linear regressions: (1) population density ~ river density + ecoregion richness + topographic complexity + climate change velocity + precipitation constancy + temperature constancy, (2) language diversity ~ river density + ecoregion richness + topographic complexity + climate change velocity + precipitation constancy + population density + carrying capacity with group size limits and (3) carrying capacity with group size limits ~population density (Code and data to perform the analysis are available as supplementary material).

**Sensitivity Analysis**

To explore the sensitivity of our analysis to grid resolution we defined five different grid resolutions: 200x200km, 250x250km,300x300km, 350x350km and 400x400km. However, as we show here, our results are qualitatively insensitive to grid resolution. Despite a slight increase in R2 at coarse resolution grids (>300x300km), the spatial pattern in R2 remains similar across different grid resolutions (Fig S2). Similarly, although the mean coefficient of each variable varies with different resolutions, the coefficients of all variables still vary over space (Fig S3). In our paper we present the results only for the 300x300km2 to ensure that grid cells were small enough to capture the variation in language diversity across space and because the same grid resolution has been used to characterize language diversity on other continents [49].



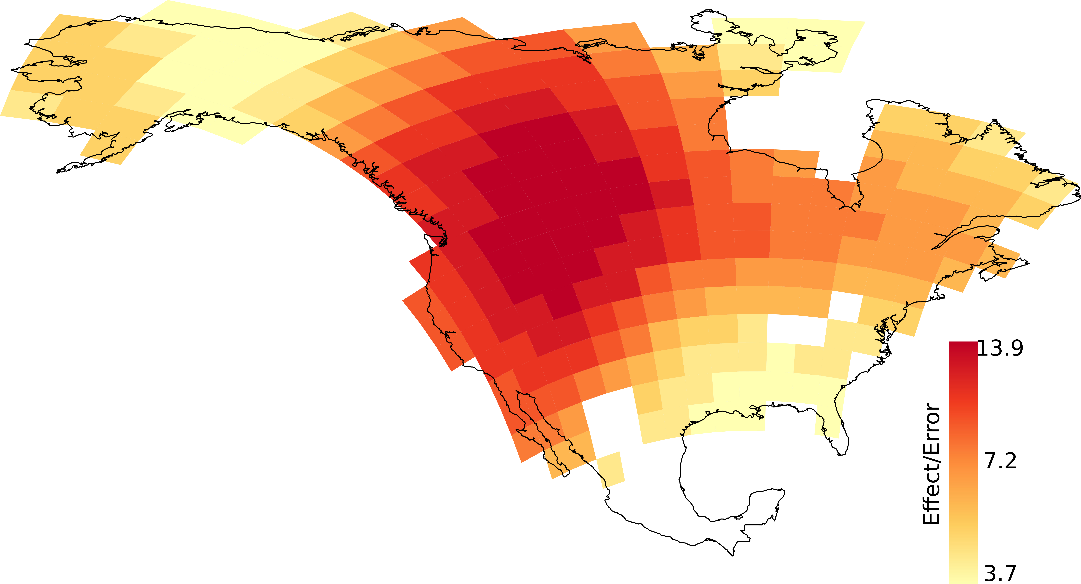
**Fig S2.** Local R2 for different grid resolutions and the distribution of R2 for each resolution.



**Fig S3.** Mean coefficients and their standard deviation in different grid resolution.

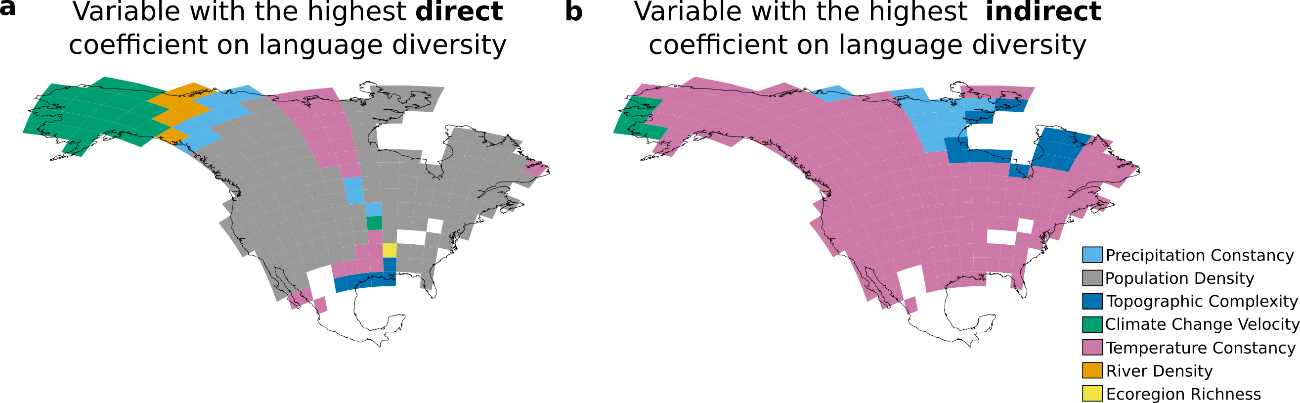
**Contrast against a null model**

We compared the predictions of our model against the expectations of a null model, which randomized language diversity in North America among grid cells, effectively removing the spatial pattern in language diversity. We replicated the null model 1000 times and ran the GWR path analysis for each replicate of randomized language diversity, recording the R2 for each grid cell (i.e. *local null R2*). We then calculated a map of the statistical *effect-error* ratio [71], in which the statistical effect is represented by the R2 for a grid cell (i.e. *local R2*) obtained by the analysis of empirical language richness, and the statistical error is the standard error of local R2, estimated as the standard deviation of local null R2 for a grid cell [72]. The ratio between the local R2 (effect) and the standard error of local R2 (error) follows a z-distribution and is a standardized measure of how much the observed effect is greater than the statistical error. Standard statistical interpretation argues that an effect at least two times larger than the error (i.e., an effect-error ratio of 2) represents substantial evidence that this effect would not have been obtained by sampling error with a 95% confidence level [73]. In our stationary analysis, the effect-error ratio is 28.430. In our geographically weighted model, the minimum *effect-error* ratio is 3.7, indicating that observed R2 depart substantially from the null expectation. In areas where our model explains more than 50% of the variation in language diversity (Fig 3) we estimate an even larger *effect-error* ratio (7 - 13.9; Fig S4).

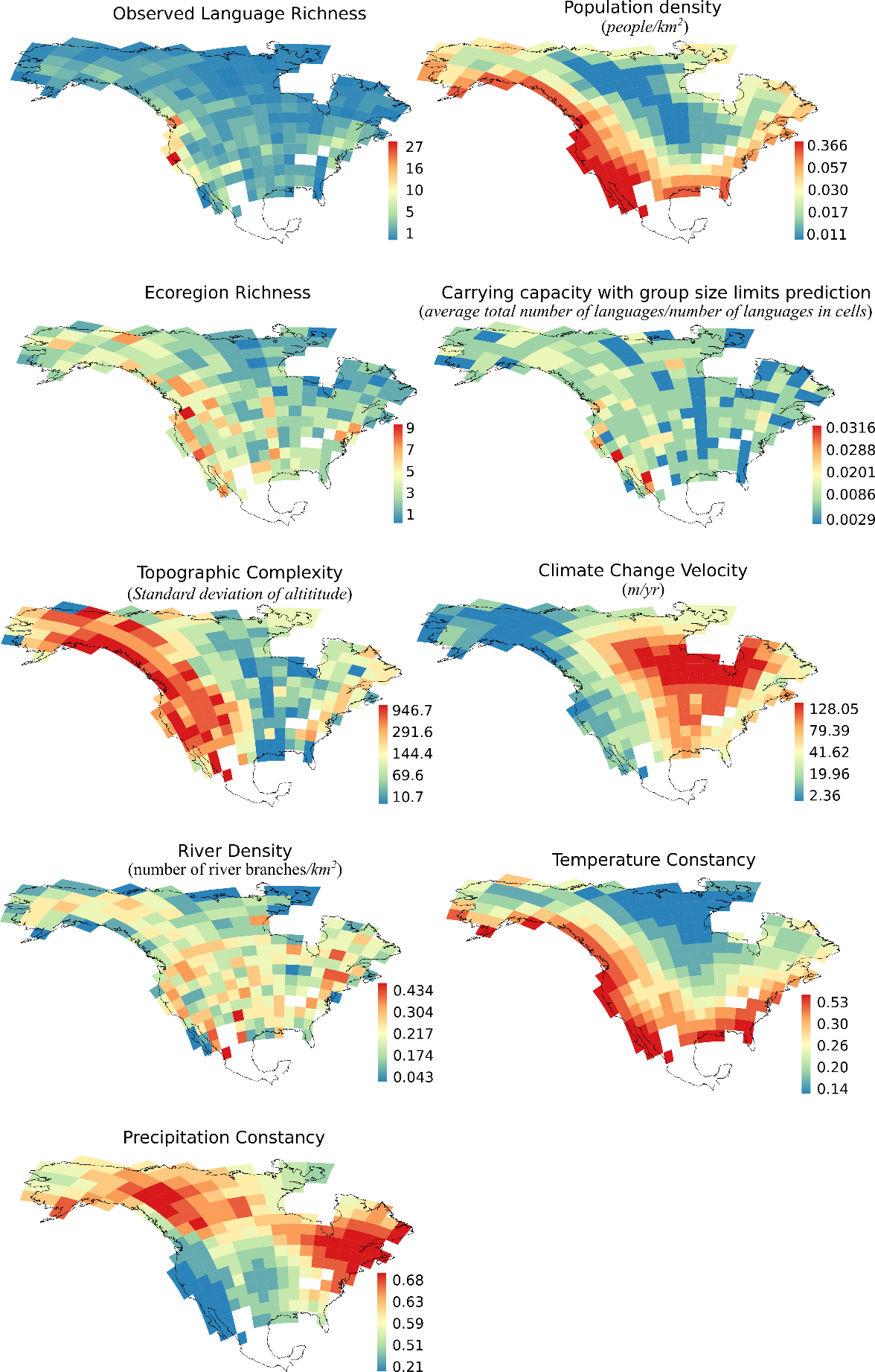


**Fig S4**. Effect-Error ratio. The effect is represented by the local R2 obtained by the analysis with the empirical language richness and the error is represented by the standard deviation of the local null R2 obtained by 1000 randomizations of language diversity in the gridded map of North America.

**Supplementary Figures**



**Fig S5.** Variables with the highest (a) direct and (b) indirect coefficients.



**Fig S6**.Spatial pattern of language diversity and all predictors tested in this study.

**Supplementary Tables**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Carrying capacity with group size limits** | **River**  **Density** | **Population**  **Density** | **Topographic**  **Complexity** | **Ecoregion**  **Richness** | **Climate Change**  **Velocity** | **Precipitation**  **Constancy** | **Temperature Constancy** |
| **Carrying capacity with group size limits** | 1.00 | 0.07 | 0.34 | 0.41 | 0.26 | -0.27 | -0.30 | 0.17 |
| **River**  **Density** | 0.07 | 1.00 | -0.01 | 0.26 | 0.26 | 0.01 | -0.10 | 0.04 |
| **Population**  **Density** | 0.34 | -0.01 | 1.00 | 0.54 | 0.37 | -0.51 | -0.68 | 0.87 |
| **Topographic**  **Complexity** | 0.41 | 0.26 | 0.54 | 1.00 | 0.53 | -0.68 | -0.42 | 0.38 |
| **Ecoregion**  **Richness** | 0.26 | 0.26 | 0.37 | 0.53 | 1.00 | -0.44 | -0.41 | 0.39 |
| **Climate Change**  **Velocity** | -0.27 | 0.01 | -0.51 | -0.68 | -0.44 | 1.00 | 0.38 | -0.43 |
| **Precipitation**  **Constancy** | -0.30 | -0.10 | -0.68 | -0.42 | -0.41 | 0.38 | 1.00 | -0.63 |
| **Temperature Constancy** | 0.17 | 0.04 | 0.87 | 0.38 | 0.39 | -0.43 | -0.63 | 1.00 |

Table S1 – Pairwise correlation between predictor variables. Temperature constancy and population density are the only pair of variables that are strongly correlated [see 71]. The direct effect of population density captures the effect of temperature constancy on language diversity. Thus, the direct effect of temperature constancy on language diversity is not assumed in the path models.