PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Pacheco Coelho MT *et al*. 2019 Drivers of geographical patterns of North American language diversity. *Proc. R. Soc. B* **286**: 20190242.
http://dx.doi.org/10.1098/rspb.2019.0242

Received: 29 January 2019
Accepted: 6 March 2019

Subject Category:
Evolution

Subject Areas:
evolution, ecology, environmental science

Keywords:
language diversity, path analysis, geographically weighted regression

Authors for correspondence:
Marco Túlio Pacheco Coelho
e-mail: marcotpcoelho@gmail.com
Michael C. Gavin
e-mail: michael.gavin@colostate.edu

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4440170.

THE ROYAL SOCIETY PUBLISHING

Drivers of geographical patterns of North American language diversity

Marco Túlio Pacheco Coelho[1,2], Elisa Barreto Pereira[2], Hannah J. Haynie[1], Thiago F. Rangel[2], Patrick Kavanagh[1], Kathryn R. Kirby[3,4], Simon J. Greenhill[4,8], Claire Bowern[5], Russell D. Gray[4], Robert K. Colwell[2,6,7], Nicholas Evans[8] and Michael C. Gavin[1,4]

[1]Department of Human Dimensions of Natural Resources, Colorado State University, Fort Collins, CO, USA
[2]Departamento de Ecologia, ICB, Universidade Federal de Goiás, 74.690-900 Goiânia, Goiás, Brazil
[3]Department of Ecology and Evolutionary Biology and Department of Geography and Planning, University of Toronto, Ontario, Canada
[4]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany
[5]Department of Linguistics, Yale University, New Haven, CT, USA
[6]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA
[7]University of Colorado Museum of Natural History, Boulder, CO 80309, USA
[8]CoEDL (ARC Centre of Excellence for the Dynamics of Language), Australian National University, Canberra, Australia

MTPC, 0000-0002-7831-3053; EBP, 0000-0002-3372-7295; MCG, 0000-0002-2169-4668

Although many hypotheses have been proposed to explain why humans speak so many languages and why languages are unevenly distributed across the globe, the factors that shape geographical patterns of cultural and linguistic diversity remain poorly understood. Prior research has tended to focus on identifying universal predictors of language diversity, without accounting for how local factors and multiple predictors interact. Here, we use a unique combination of path analysis, mechanistic simulation modelling, and geographically weighted regression to investigate the broadly described, but poorly understood, spatial pattern of language diversity in North America. We show that the ecological drivers of language diversity are not universal or entirely direct. The strongest associations imply a role for previously developed hypothesized drivers such as population density, resource diversity, and carrying capacity with group size limits. The predictive power of this web of factors varies over space from regions where our model predicts approximately 86% of the variation in diversity, to areas where less than 40% is explained.

1. Introduction

Humans collectively speak over 7000 distinct languages, and these languages are unevenly distributed across the globe [1,2]. Surprisingly, we still know little about the complex web of processes that shape these geographical patterns of language diversity (i.e. the number of languages spoken in a given region). Linguists distinguish three types of diversity—the number of languages (*language diversity*), the number of language families (*phylogenetic diversity*), and the amount of structural difference between languages (*typological diversity* or *disparity*). Here, we focus only on the number of languages, using the term *language diversity*, which in contrast to the more ambiguous term *linguistic diversity* indicates that languages are the unit of our diversity measures.

One barrier to our prior understanding has been contradictory results from the limited number of empirical studies that have investigated the relationship between environmental and/or sociocultural variables and language diversity [1,3–8]. Prior studies have found mixed results for the effect of environmental variables, spatial heterogeneity, and isolation on language diversity [8–12]. For

example, human populations may expand social networks to cope with higher levels of ecological risk, resulting in larger language ranges and lower levels of language diversity per unit area [13]. Although some prior studies have concluded that the most commonly used measure of ecological risk in linguistics—mean growing season—correlates with language diversity (e.g. [10,11]), others have found little support for this relationship (e.g. [4,8,12]).

Two methodological challenges contribute to the inconsistencies in these results: first, previous studies have tried to identify universal predictors of language diversity, but it is possible that no universal predictor exists. Research in macroecology has shown that the drivers of observed spatial patterns in biodiversity tend to be spatially variable [14–16]. We might assume that the mechanisms driving language diversification also vary from one location to another, but the methods used to date cannot capture this potential non-stationarity. Second, contradictory results may also reflect the complexity of the pattern being studied, which can be generated by a web of both direct and indirect pathways. For example, environmental drivers of language density vary across subsistence types [17]; the adoption of agriculture, or new boat and fishing technology, may transform the number of people a given ecoregion can support; or political centralization, the product of a particular historical trajectory, may homogenize a previously disparate linguistic mosaic.

Surprisingly, only a limited number of statistical techniques have been used to explore the direct and indirect associations between multiple predictors underlying the heterogeneous spatial patterns of language diversity [1]. To the best of our knowledge, only one previous study briefly explores a simple structural equation modelling approach that considers the direct and indirect effect of three variables on the distributional range size of languages [12]. Here, we overcome prior methodological limitations by designing a path analysis model that assumes direct and indirect effects of environmental and sociocultural variables on language diversity, while exploring spatial variation in the predictors' effects. Our study is the first to use a geographically weighted path analysis (GWPath) to examine possible drivers of human diversity patterns.

(a) Factors contributing to language diversity patterns

Because languages are markers of social boundaries within and between groups [18–20], group boundary formation is a critical step in language diversification. The formation or dissolution of group boundaries can be influenced by many different environmental and social factors [1]. Variation within a language can lead to new language formation (i.e. cladogenesis) if these group boundaries are stable and socially important, amplifying the degree of linguistic difference between groups to the point that erstwhile dialects become distinct languages. Here, our aim is to demonstrate the importance of complex paths and non-stationarity by examining a subset of variables that have been widely discussed in the literature and may contribute to group boundary formation, thus affecting spatial patterns of language diversity. We do not focus on the internal factors contributing to individual language variation [21–26], rather we focus on a subset of the large-scale processes that may shape language diversity patterns in a broader ecological context.

We examine the direct and indirect effects of eight factors hypothesized to influence group boundary formation and language diversity patterns: river density, topographic complexity, ecoregion richness, climate (i.e. temperature and precipitation constancy, and climate change velocity), population density, and carrying capacity with group size limits. Rivers and topography have recently been proposed as universal predictors of language diversity at a global scale [7]. Movement and isolation are both critical processes for the formation of group boundaries [26,27]. When groups of people move to the other side of physical barriers, the costs of interacting with neighbouring groups can increase, leading to social isolation and group boundary formation [7,28,29]. Rivers and complex topography may act as barriers to contact among groups, promoting isolation and driving diversification, in a mechanism similar to models of allopatric speciation developed in ecology and evolutionary biology to explain biodiversity patterns [29]. This mechanistic link implies that both river density and topographic complexity should be positively correlated with language diversity. Alternatively, rivers may also improve transportation, which can increase contact among groups and undermine group boundary formation leading to less language diversity in a region [7,30,31]. In addition, in regions such as Southern New Guinea [32,33] complex linguistic differentiation has occurred despite the absence of any complex topography, suggesting linguistic differentiation in circumstances of ethnic intermarriage and multilingualism can sometimes be accelerated by easily traversed terrain.

Many prior studies discuss possible links between language diversity and biological diversity [4,11,34,35]. One possible explanation for the association between biological and language diversity is that biodiversity facilitates group boundary formation through resource partitioning [11]. The development of unique subsistence strategies and technologies may allow different groups to thrive within different ecoregions, each of which represents a distinct assemblage of species [36]. Therefore, ecoregion richness (i.e. number of ecoregions) might be expected to associate positively with language diversity.

Climate may influence group boundary formation and geographical patterns in language diversity via multiple pathways [17]. For example, unstable and extreme climatic conditions of temperature and precipitation contribute to higher ecological risk for human groups, which can lead to the growth of larger social networks that provide a source of alternative resources and manage risk [9,13,32]. Larger social networks limit group boundary formation and promote linguistic homogenization [10,37]. Therefore, we would expect fewer languages in areas that experience greater fluctuation in climatic conditions of temperature and precipitation. We propose that the velocity at which the climate has changed may also be a proxy for long-term ecological risk, because higher velocity of climate change indicates more instability of climate in a region over longer periods of time. In addition, the velocity of climate change over longer periods of time played an important role in the human colonization of the globe, opening pathways and territories for settlement where climatic conditions were suitable for humans (e.g. warming of northern regions) [38].

Climate may also influence language diversity through its effects on human population densities. When climatic

conditions are favourable (i.e. warm and wet) and predictable, human groups can be more assured of rich and stable sources of resources that may support higher population densities [39–41]. Several other environmental and sociocultural variables also shape potential population densities. For example, population densities may increase in coastal regions, given greater access to marine resources; in topographically complex areas due to access to a range of nearby ecosystems and restrictions on available level surfaces for settlement [41,42]; and in areas of higher river density, where rivers provide services such as food and water that directly affect the establishment of human groups [7]. In addition, less mobile groups and those with established land ownership norms tend to have higher population densities [41,43,44].

Multiple possible mechanisms link higher population densities with greater language diversity per unit area. As has been suggested in ecological theory, regions that support more individuals may also accumulate more diversity over time due to stochastic diversification events [44,45]. If more individuals exist in a given location, the probability of high linguistic variation also increases, and therefore we expect higher rates of diversification. Similarly, Bromham *et al*. [46] found that larger populations have faster rates of innovation, which could lead to more languages as changes accumulate. Another possible link involves the effects of group size on boundary formation. Large groups provide more opportunities to cooperate in resource acquisition, but also increase the costs associated with maintaining social ties [10,47,48]. Limits on the size of human groups imply that regions that can support higher population densities will tend to have greater language diversity [49]. However, these limits are not fixed—for example, increases in food production per unit area (e.g. as a result of the development of intensive agriculture) as well as the evolution of centralized political institutions have both been associated with increases in maximum group sizes and linguistic homogenization [50,51].

Prior studies seeking to identify factors linked to language diversification have been almost exclusively based on correlative analyses [1], in which no causal story is modelled [52]. Recently, a relatively simple mechanistic simulation model explored causal explanations for language diversity in Australia [49]. The model reproduced the spatial pattern of language diversity in Australia assuming only that carrying capacity varies over space as a function of the environment, and groups have maximum size limits (i.e. carrying capacity with group size limits) [49]. However, the carrying capacity with group size limits mechanism remains untested in other regions of the world.

Here, we test the hypothesized effect of each of the eight factors discussed above (river density, topographic complexity, ecoregion richness, temperature and precipitation constancy, climate change velocity, population density, and carrying capacity with group size limits) using a path analysis that models the multiple paths through which predictors could be associated with language diversity. Each pathway implies a different set of mechanisms that may shape language diversity. River density, number of ecoregions, topographic complexity, and climate may directly shape language diversity, or influence diversity indirectly through effects on population density. Population density can also directly affect language diversity, or influence diversity by contributing to the carrying capacity with group size limits mechanism. Therefore, large groups of people can occupy small areas if population density is high, which affects the total number of groups in a given region. We designed two types of path analysis models, one assuming that the relationship between predictors is constant over space (i.e. Stationary Path Analysis), and another assuming that the relationship between predictors may vary over space (i.e. GWPath). Our analysis examines the strength of associations between the hypothesized predictors and language diversity, and how these effects vary over space. The only variable that explicitly captures a causal relationship is carrying capacity, which is produced by a mechanistic simulation model (see Methods and [49]).

(b) Geographical domain

We applied our models to understand the spatial pattern of language diversity in North America. We obtained the distribution of languages in North America from Goddard [53], which provides information about the approximate spatial distribution, around the time of colonial contact, of languages north of Mexico, and the Survey of California and Other Indian Languages, which provides additional detail in a particularly diverse region. Using these data, we calculated the number of languages occupying geographical cells on a gridded map at the resolution of 300 × 300 km (figure 1; See Sensitivity analysis in the electronic supplementary material).

North America provides an ideal setting to examine how the relative effects of explanatory factors vary over space, as the continent contains a wide range of environmental and sociocultural conditions and a wide spectrum of language diversity. Prior to European contact, the continent supported hundreds of languages [53,54], unevenly distributed over the continent, with greater richness along the west coast and at lower latitudes [53,55]. Prior research has proposed many factors to explain the empirical pattern of North American language diversity (e.g. [55]), but no empirical study has tested them. Here, we explore the direct and indirect effects of river density, topographic complexity, ecoregion diversity, climate, population density, and carrying capacity with group size limits on the spatial pattern of North American language diversity. These factors encompass proposed drivers of language richness in North America and are also expected to drive global patterns of language diversity [29].

(c) Results and discussion

To explore both indirect and direct effects of each factor, we first conducted a stationary path analysis that assumes the effects of environmental and sociocultural variables are constant over space. The variables included in our model vary in the direction of effect (i.e. negative and positive; figure 2). Population density, carrying capacity with group size limits, and ecoregion richness had the strongest direct effects, suggesting a role for multiple mechanisms in shaping language richness patterns (figure 2).

Population density had the strongest direct effect on language diversity ($\beta = 0.44$; figure 2), supporting the proposed mechanism that a larger number of individuals should lead to a greater accumulation of languages. The simple mechanistic model, simulating the effects of varying carrying capacity with group size limits was also one of the strongest predictors of language diversity ($\beta = 0.25$,

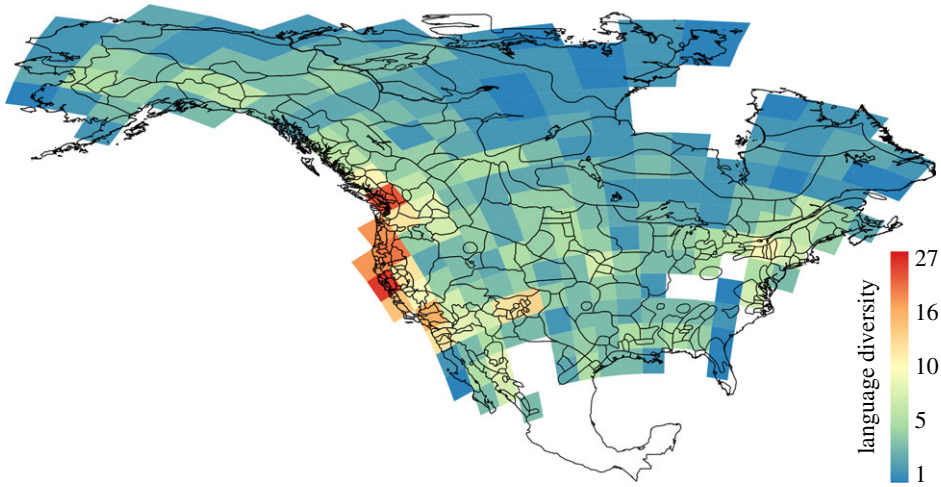

Figure 1. Observed language diversity. Language ranges are shown in the gridded map. Blank spaces on the map indicate regions in which no information about language distribution is available and thus were not compiled in the grid map. (Online version in colour.)

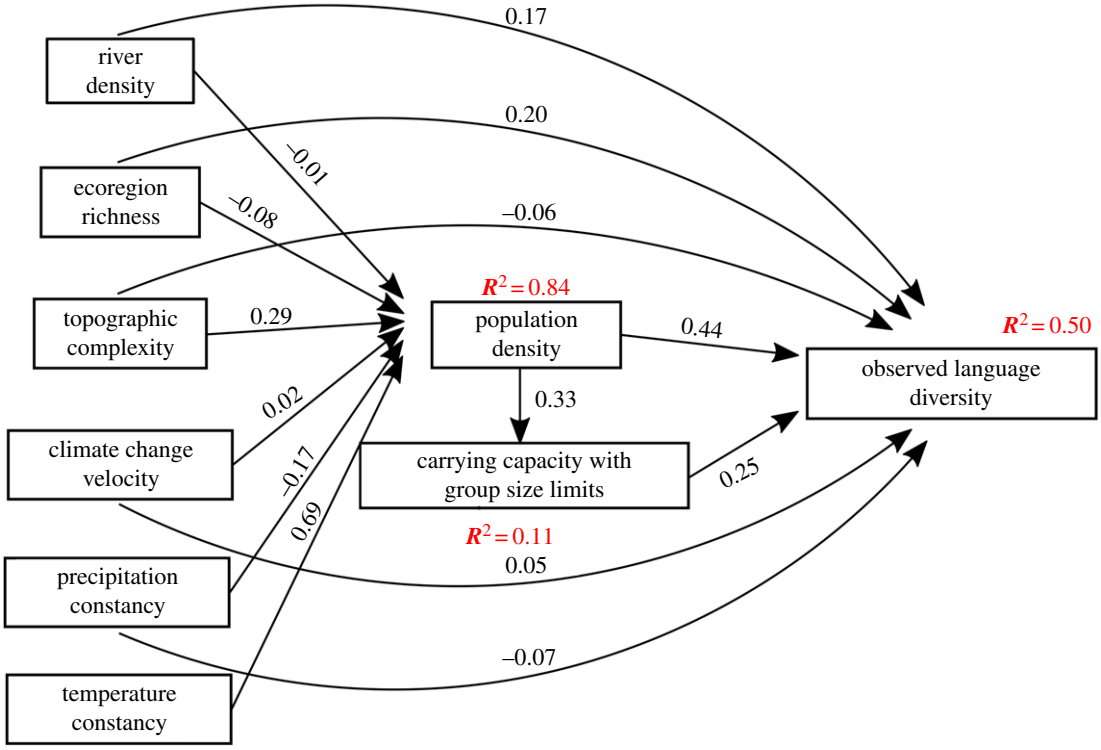

Figure 2. Global path model quantifying direct and indirect effects of environmental and sociocultural factors on North American language richness. The numbers marking each arrow represent the standardized β coefficients (i.e. path coefficients) for language diversity. Model fits (R^2) are shown for variables directly affected by other factors. (Online version in colour.)

figure 2). Therefore, in regions with higher potential carrying capacity, limits on the size of human groups tended to lead to greater language richness [49]. Finally, the strength of the direct effect of ecoregion richness ($\beta = 0.20$, figure 2)[1] implies that resource partitioning may contribute to language diversification [11], as unique subsistence strategies and technologies could allow different human groups to thrive within different ecoregions.

We emphasize here that carrying capacity with group size limits is the only component of our path analysis that is modelled in a mechanistic, explicitly causal manner. The correlations used to explore all the other components indicate an association with language diversity, but future simulation modelling will be needed to verify the causal mechanisms that link these components with language diversification.

The stationary path analysis approach also demonstrates the indirect roles played by several variables. For example, if we evaluated only the direct effects of variables, as was commonly done in prior language diversity studies [11], we would conclude that topographic complexity has little influence on language diversity. However, each of these variables does have a substantial indirect effect by shaping population density (figure 2). Topographic complexity may indirectly affect population density through its positive association with resource availability [56–58], which, in turn, may influence the number of people that can live in a given location (i.e. population density; [41]).

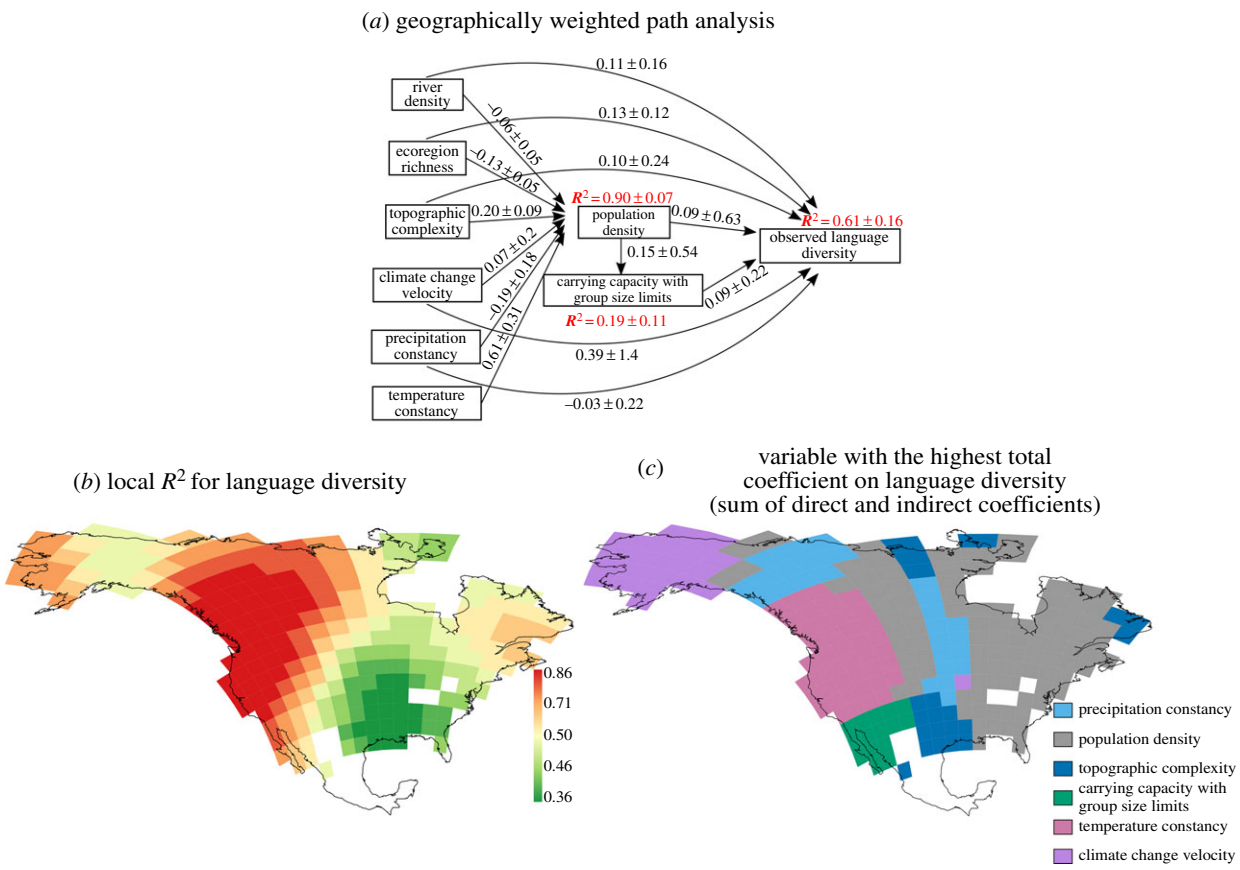

Figure 3. GWPath applied to North American linguistic diversity. (*a*) In the GWPath model, the standardized β coefficients of variables, as well as the R^2 for the direct relationships are represented by the average value over the continent, followed by its standard deviation. (*b*) Model fit varies over the geographical domain of North America. (*c*) Variables with the highest total coefficient (sum of direct and indirect effects) also vary across the continent. (Online version in colour.)

2. Geographically weighted path analysis

The combination of environmental and demographic variables in our stationary path analysis explains 50% of the variation in the spatial pattern of language richness in North America (figure 2). The stationary path analysis has a large statistical effect (*effect-error* ratio = 28.430) relative to the magnitude of error given the null expectation (see Comparison to a Null Model in the electronic supplementary material). However, this analysis does not allow us to explore how drivers of linguistic richness vary over space. To overcome this limitation, we conducted a GWPath, which assumes that the effects of hypothesized factors may vary over geographical space. To the best of our knowledge, this is the first study to apply a GWPath to examine human diversity patterns.

The effects of the predictors we tested vary widely over space (figure 3*a*). The overall model performs well in some regions of North America (e.g. the northwest region where $R^2 \sim 0.80$, figure 3*b*), but the model fit varies over space (36–86%), with an average R^2 of 0.61. Our model also has a large statistical effect over space relative to the magnitude of errors given the null expectation (minimum *effect-error* ratio = 3.7, see Comparison to a Null Model in the electronic supplementary material). In addition, we find no universal predictor of language richness. Instead, the variables that most strongly affect language richness change from one region to another across the continent (figure 3*c*), implying that the mechanisms of language diversification also vary over space. This result helps to explain why the variables tested in previous global-scale studies tend to explain only a limited portion of the variability in language richness, and why different regional analyses point to the importance of distinct sets of variables [1]. Spatial variation in explanatory variables is also found in macroecological analyses of species diversity patterns (e.g. [15,59,60]). For example, although species diversity is strongly limited by water availability in southern regions, in northern regions energy availability is more important [59]. Our results show not only that the most important predictor varies over space, but also that predictors can vary in the direction of their effects in different regions (figure 4). Climate change velocity presents different directions of effect in two different regions of North America: the northern region and eastern region (figure 4*d*). In the northern region, climate change velocity has a positive direct effect on language richness, while the effect is negative in the eastern region (figure 4*d*). The high rate of climate change in the northern region reflects rapid warming following the Last Glacial Maximum (LGM) (e.g. ice sheet melting, [61]), which likely opened ecological opportunities for human populations to obtain more resources given the positive effect of past climate change on many aspects of biodiversity in these northern regions [62]. Conversely, in the eastern region (figure 3*c*), the effect of climate change velocity is negative (figure 4), suggesting that climatic instability since the LGM prevented or reduced language diversity. The effect of climate change velocity across both regions is consistent with a long-term version of the ecological risk hypothesis [9,13]. Nettle [13] proposed that in areas with high seasonal variation in food availability, humans will experience high levels of ecological risk. An increased probability of food deficiencies may force people to form social bonds across wider areas, to ensure access to

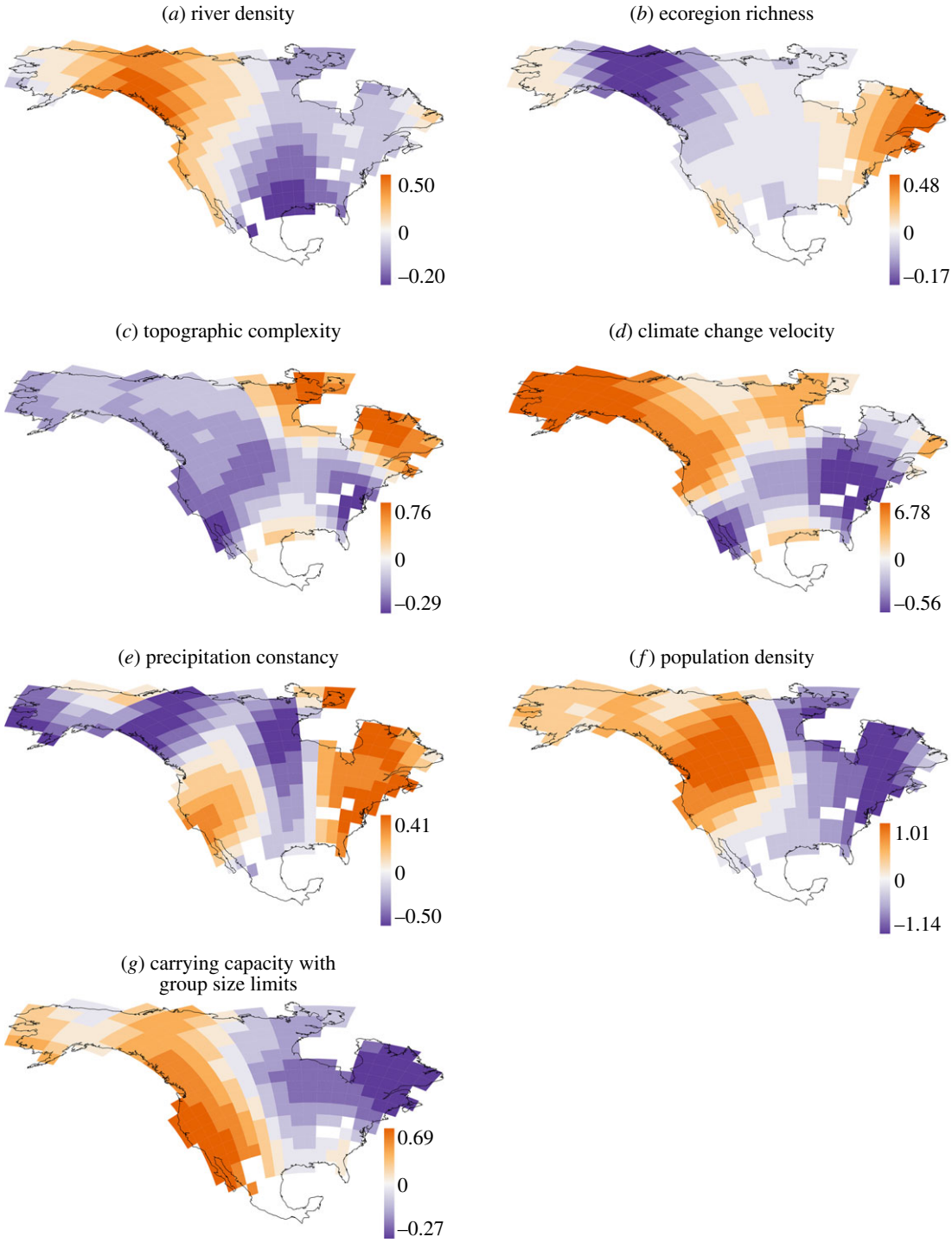

Figure 4. Direct effect of predictors mapped over the North American domain. The standardized β coefficient is mapped for (*a*) river density, (*b*) ecoregion richness, (*c*) topographic complexity, (*d*) climate change velocity, (*e*) precipitation constancy, (*f*) population density, and (*g*) carrying capacity with group size limits. (Online version in colour.)

sufficient resources. Wider social networks may increase the geographical range of a language and reduce language diversity in areas that pose greater ecological risk. Over thousands of years of human spread in North America, higher climate change velocity likely decreased ecological risk in northern regions, while climatic change may have increased ecological risk farther south. The strong indirect effect of temperature constancy (figure 2; electronic supplementary material, figure S5b) on language diversity is another indication of the importance of ecological risk for shaping population density and language diversity.

Our GWPath also reveals that river density is not the primary predictor of language diversity in any region of North America (figure 3*c*). River density has been proposed as a global universal predictor of language diversity [7], but it does not show substantial effects in any region of North America when compared to other variables (figure 3*c*).

Where our model performs best ($R^2 > 0.5$; red areas in figure 3*b*), population density and climate (i.e. temperature or precipitation constancy) are the variables most strongly affecting language diversity (figure 3*c*). The strong association of these variables in the areas of highest model fit provides support for several of the proposed pathways of language diversification (See *factors contributing to language diversity patterns*). Therefore, in those regions we can identify the best predictors of language diversity and better understand what is driving the performance of our model. However, in other regions (green in figure 3*b*), the model

explains less than 50% of the variation in language diversity ($R^2 < 0.5$). One possible reason for the poorer model performance in these regions is that pre-colonial human groups may have used rivers differently in different regions. The observed effect of river density on language diversity in the areas of lower model performance is the opposite (negative effect) to what has been hypothesized in the literature (figure 4*a*). One potential mechanism that may explain this negative correlation involves the impact of rivers on transportation. Compared to the west, many of the rivers in the central part of the continent flow through plains with fewer rapids, making them more navigable. Therefore, these rivers may have served to connect human groups and reduce language diversity, as opposed to acting as a barrier and means of group boundary formation. Finally, there are multiple sociocultural and historical factors that cannot be summarized in gridded map cells, and thus are absent from our model, including subsistence strategies, agricultural development, trade, and political complexity [12,29,63] that may be part of the unexplained percentage of variation. For example, the spread of politically complex agricultural societies may be a dominant factor in the reduction of language diversity [12].

To the best of our knowledge, this is the first study to investigate the complex web of predictors underlying geographical patterns of language diversity. We show that the strongest effects on North American language diversity involve variables associated with previously developed hypotheses that assume the effect of resource availability, resource diversity, and climate affecting population density, and thus language diversification. The many factors are connected in a complex web of causality, consisting of both direct and indirect effects. Moreover, no single predictor explains the pattern of language diversity in North America, and the best predictors of language diversity vary over space. Thus, our study sheds light on important points that should be taken into consideration in future studies of language diversity, namely that the ecological drivers of language diversity are neither perfectly universal nor entirely direct. The combination of path analysis techniques with the exploration of non-stationarity in predictors' effects can help us to examine these complexities, and better understand a more complete picture of human biogeography. The methodological approach outlined here may serve as a template for exploring the potential interaction between multiple factors that have shaped geographical patterns of human diversity across the planet.

3. Methods

(a) Data

We obtained the approximate distribution of languages in North America immediately prior to European contact from two sources. We used the Survey of California and Other Indian Languages map (http://linguistics.berkeley.edu/~survey/resources/language-map.php) for the approximate spatial extents of California language ranges, and we digitized language ranges for other regions from Goddard [53]. The final map consisted of 344 language ranges. The geographical domain of North America was represented by an equal-area, gridded map at the resolution of 300 × 300 km. Our choice of this grid resolution ensured that grid cells were small enough to capture the variation in language diversity across space. We tested the sensitivity of our results to different grid resolutions; and we concluded that the results remained qualitatively insensitive to grid resolution (see Sensitivity Analysis in the electronic supplementary material). We computed the number of languages (i.e. language diversity) and extracted each predictor variable for each grid map cell (electronic supplementary material, figure S6).

High-resolution river maps for North America were obtained from the Global Self-Consistent Hierarchical High-resolution Shoreline dataset ([64], www.soest.hawaii.edu/wessel/gshhg/). Following Axelsen & Manrubia [7], we defined river density as the number of river branches within a geographical cell. We obtained data on ecoregions from the Terrestrials Ecoregions of the World dataset ([36]; www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world), and we used the number of terrestrial ecoregions within each geographical cell as a measure of ecoregion richness. We measured topographic complexity as the standard deviation of elevation above the sea level (m) within a cell ([65]; www.worldclim.org/). We used climate change velocity since the LGM [62] as a measure of long-term ecological risk. Climate change velocity measures the rate of displacement of climate over the geographical space by dividing the climatic difference between two periods by climate change over space. We calculated the inter-annual variability (i.e. constancy) of temperature and precipitation following the Colwell index of constancy [66]. Constancy is used to describe the time-independent magnitude of variability of temperature and precipitation. We calculated precipitation and temperature constancy using data from ecoClimate [67] for 1900–1949 from the CCSM4 model. We extracted the estimated population density (people per km^2) for foraging societies [42] in each grid cell (see Population Density in the electronic supplementary material).

The effect of carrying capacity with group size limits on language diversity was simulated through a recently proposed mechanistic simulation model of language diversity (see Simulation Model section in the electronic supplementary material for additional details) [49]. The model's basic assumption is that the carrying capacity of a region is a function of the environment. Thus, locations that support more humans per unit area can also support more languages. The model accurately predicted the diversity of Australian languages [49], and here we apply it to North America. After running the model, replicated 120 times, we used the simulated geographical distribution of language ranges to summarize the model's prediction in the 300 × 300 km grid of North America. The prediction extracted from the model and used in our path analysis was a ratio between the number of languages predicted in each cell and total number of languages predicted for the geographical domain. We used the average among 120 model replicates as our carrying capacity with group size limits estimation in the path model.

(b) Statistics

Based on the hypothesized roles of the predictors used in our study on language and cultural diversity, we designed a path analysis model including the direct and indirect effects of our predictors on language diversity (figure 1). We evaluated the proposed direct and indirect effect of each variable on language diversity while controlling for the effects of the remaining predictor variables. We used the standardized partial slope coefficient of a multiple regression (i.e. path coefficient) to represent the strength of the effect of each variable on language diversity. This modelling technique allows us to explore direct, indirect (i.e. multiplication of direct coefficients), and total effects (i.e. sum of direct and indirect coefficients) of each predictor.

Path analysis assumes stationarity in the relationship among variables, but no theory would suggest that mechanisms of language diversification must be the same in all locations. In

order to explore the potential for non-stationarity in our results, we also employed a GWPath, in which we estimated the coefficients for the predictor variables for each geographical cell following a Geographically Weighted Regression (GWR) [14] with a Gaussian distance function. We estimated a bandwidth for the GWR by visual inspection [14] and Akaike criteria model selection, which considers the likelihood of the model as well as its complexity. The best bandwidth obtained was 8° (approx. 880 km), which avoids overfitting and has a good fit to empirical data. Statistical analysis was conducted in R. GWPath used the 'gwr' function of the 'spgwr' package ([68]; also see electronic supplementary material for data and code). We also 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 (see Contrast Against a Null Model in the electronic supplementary material).

Data accessibility. The data used in this study are available as electronic supplementary material.

Authors' contributions. M.T.P.C. and M.G. jointly conceived the study. M.T.P.C. led the writing and created the figures with input from all authors. E.B.P. performed statistical analysis. H.H., produced the language distribution map. M.T.P.C., E.B.P., K.K., H.H. and P.K. processed the spatial data. T.F.R. programed the mechanistic simulation and M.T.P.C. applied it for North America. All authors contributed conceptually to the design of the study and interpretation of results.

Competing interests. We declare we have no competing interests.

Funding. Research was supported by the National Science Foundation (award no. 1660465). M.T.P.C. and E.B.P. are supported by PhD scholarships provided by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES - Finance Code 001). T.F.R. is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant nos. PQ309550/2015-7), and by INCT in Ecology, Evolution and Biodiversity Conservation (grant nos. MCTIC/CNPq 465610/2014-5 and FAPEG 201810267000).

Endnote

[1]In Australia, there are language-origin stories explicitly linking language regions of clans to ecological differentiation through staple foods, such as the tradition of the founding ancestress Warramurrungunji [25], who placed different plant foods (lily roots, yams, etc.) in different parts of the landscape at the same time as she placed people there and instructed them in what their clans would be, what their languages would be, and what they would eat.

References

1. Gavin MC *et al.* 2013 Toward a mechanistic understanding of linguistic diversity. *Bioscience* **63**, 524–535. (doi:10.1525/bio.2013.63.7.6)
2. Hammarström H, Bank S, Forkel R, Haspelmath M. 2018 Glottolog 3.2. See http://glottolog.org, (accessed on 14 May 2018).
3. Collard IF, Foley RA. 2002 Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? *Evol. Ecol. Res.* **4**, 371–383.
4. Sutherland WJ. 2003 Parallel extinction risk and global distribution of languages and species. *Nature* **423**, 276–279. (doi:10.1038/nature01607)
5. Maffi L. 2005 Linguistic, cultural, and biological diversity. *Annu. Rev. Anthropol.* **34**, 599–617. (doi:10.1146/annurev.anthro.34.081804.120437)
6. Burnside WR, Brown JH, Burger O, Hamilton MJ, Moses M, Bettencourt LM. 2012 Human macroecology: linking pattern and process in big-picture human ecology. *Biol. Rev.* **87**, 194–208. (doi:10.1111/j.1469-185X.2011.00192.x)
7. Axelsen JB, Manrubia S. 2014 River density and landscape roughness are universal determinants of linguistic diversity. *Proc. R. Soc. B* **281**, 20141179. (doi:10.1098/rspb.2014.1179)
8. Gavin MC, Sibanda N. 2012 The island biogeography of languages. *Glob. Ecol. Biogeogr.* **21**, 958–967. (doi:10.1111/j.1466-8238.2011.00744.x)
9. Nettle D. 1996 Language diversity in West Africa: an ecological approach. *J. Anthropol. Archaeol.* **15**, 403–438. (doi:10.1006/jaar.1996.0015)
10. Nettle D. 1999 Linguistic diversity of the Americas can be reconciled with a recent colonization. *Proc. Natl Acad. Sci.* **96**, 3325–3329. (doi:10.1073/pnas.96.6.3325)
11. Moore JL, Manne L, Brooks T, Burgess ND, Davies R, Rahbek C, Williams P, Balmford A. 2002 The distribution of cultural and biological diversity in Africa. *Proc. R. Soc. B* **269**, 1645–1653. (doi:10.1098/rspb.2002.2075)
12. Currie TE, Mace R. 2009 Political complexity predicts the spread of ethnolinguistic groups. *Proc. Natl Acad. Sci. USA* **106**, 7339–7344. (doi:10.1073/pnas.0804698106)
13. Nettle D. 1998 Explaining global patterns of language diversity. *J. Anthropol. Archaeol.* **17**, 354–374. (doi:10.1006/jaar.1998.0328)
14. Fotheringham AS, Brunsdon C, Charlton M. 2002 *Geographically weighted regression: the analysis of spatially varying relationships*. New York, NY: Wiley.
15. Cassemiro FAS, Barreto BS, Rangel TFLVB, Diniz-Filho JAF. 2007 Non-stationarity, diversity gradients and the metabolic theory of ecology. *Glob. Ecol. Biogeogr.* **16**, 820–822. (doi:10.1111/j.1466-8238.2007.00332.x)
16. Gouveia SF, Hortal J, Cassemiro FAS, Rangel TF, Diniz-Filho JAF. 2013 Nonstationary effects of productivity, seasonality, and historical climate changes on global amphibian diversity. *Ecography (Cop.)* **36**, 104–113. (doi:10.1111/j.1600-0587.2012.07553.x)
17. Derungs C, Köhl M, Weibel R, Bickel B. 2018 Environmental factors drive language density more in food-producing than in hunter–gatherer populations. *Proc. R. Soc. B* **285**, 20172851. (doi:10.1098/rspb.2017.2851)
18. Labov W. 1963 The social motivation of a sound change. *Word* **19**, 273–309. (doi:10.1080/00437956.1963.11659799)
19. Milroy L. 1982 Language and group identity. *J. Multiling. Multicult. Dev.* **3**, 207–216. (doi:10.1080/01434632.1982.9994085)
20. Dorian NC. 1994 Choices and values in language drift and its study. *Int. J. Soc. Lang.* **110**, 113–124.
21. Lehmann P, Malkiel Y. 1968 *Directions for historical linguistics*. Austin, TX: University of Texas Press.
22. Luraghi S. 2010 Causes of language change. In *Continuum companion to historical linguistics* (eds Luraghi, Bubenik), pp. 354–366. London/New York: Continuum International Publishing Group.
23. Levinson SC, Gray RD. 2012 Tools from evolutionary biology shed new light on the diversification of languages. *Trends Cogn. Sci.* **16**, 167–173. (doi:10.1016/j.tics.2012.01.007)
24. Bowern C. 2013 Relatedness as a factor in language contact. *J. Lang. Contact* **6**, 411–432. (doi:10.1163/19552629-00602010)
25. Evans N. 2010 *Dying words: endangered languages and what they have to tell us*. Maldon, UK: Wiley-Blackwell.
26. Hock HH. 1991 *Principles of historical linguistic*, 2nd edn. Berlin, Germany: Moutoon de Gruyter.
27. Pawley A, Ross MD. 1994 *Austronesian terminologies: continuity and change*. Canberra, Australia: Pacific Linguistics.
28. Stepp JR, Castaneda H, Cervone S. 2005 Mountains and biocultural diversity. *Mt Res. Dev.* **25**, 223–227. (doi:10.1659/0276-4741(2005)025[0223:MABD]2.0.CO;2)
29. Greenhil SJ. 2014 Demographic correlates of language diversity. In *Historical linguistics*, (eds C Bowern, B Evans), pp. 555–578. London/New York: Routledge Taylor & Francis Group.
30. Diller A. 2008 Mountains, rivers or seas? Ecology and language history in Southeast Asia. In *SEALSXIV: Papers from the 14th meeting of the Southeast Asian Linguistics Society* (eds W Khanittanan, P Sidwell). Canberra, Australia: Pacific Linguistics.
31. Drake NA, Blench RM, Armitage SJ, Bristow CS, White KH. 2011 Ancient watercourses and

biogeography of the Sahara explain the peopling of the desert. *Proc. Natl Acad. Sci. USA* **108**, 458–462. (doi:10.1073/pnas.1012231108)

32. Evans N. 2012 Even more diverse than we thought: the multiplicity of Trans-Fly languages. In *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century* (eds N Evans. M Klamer), pp. 109–149. Language Documentation and Conservation Special Publication.
33. Evans N *et al.* 2017 The languages of Southern New Guinea. In *The languages and linguistics of New Guinea: A comprehensive guide* (ed. B Palmer), pp. 641–774. Berlin, Germany: Walter de Gruyter.
34. Stepp JR, Cervone S, Castaneda H, Lasseter A, Stocks G, Gichon Y. 2004 Development of a GIS for global biocultural diversity. *Policy Matters* **13**, 6.
35. Fincher CL, Thornhill R. 2008 A parasite-driven wedge: infectious diseases may explain language and other biodiversity. *Oikos* **117**, 1289–1297. (doi:10.1111/j.0030-1299.2008.16684.x)
36. Olson DM *et al.* 2001 Terrestrial ecoregions of the world: a new map of life on earth. *Bioscience* **51**, 933–938. (doi:10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2)
37. Shaul D. 1986 Linguistic adaptation and the great basin. *Am. Antiq.* **51**, 415–416. (doi:10.2307/279958)
38. Harcourt A. 2012 *Human biogeography*. Berkeley, CA: University of California Press.
39. Marlowe FW. 2005 Hunter-gatherers and human evolution. *Evol. Anthropol.* **14**, 54–67. (doi:10.1002/evan.20046)
40. Belovsky GE. 1988 An optimal foraging-based model of hunter-gatherer population dynamics. *J. Anthropol. Archaeol.* **7**, 329–372. (doi:10.1016/0278-4165(88)90002-5)
41. Kavanagh PH, Vilela B, Haynie HJ, Tuff T, Lima-Ribeiro M, Gray RD, Botero CA, Gavin MC. 2018 Hindcasting global population densities reveals forces enabling the origin of agriculture. *Nat. Hum. Behav.* **2**, 478–484. (doi:10.1038/s41562-018-0358-8)
42. Hassan FA. 1975 *Determination of the size, density, and growth rate of hunting-gathering populations*. In *Population, ecology, and social evolution* (ed. S Polgar), pp. 27–52. The Hague, The Netherlands: Mouton.
43. Binford LR. 2001 *Constructing frames of reference: an analytical method for archaeological theory building using hunter–gatherer and environmental data sets*. Berkeley, CA: University of California Press.
44. Brown JH. 1981 Two decades of homage to santa rosalia: toward a general theory of diversity. *Integr. Comp. Biol.* **21**, 877–888.
45. Coelho MTP, Dambros C, Rosauer DF, Pereira EB, Rangel TF. 2018 Effects of neutrality and productivity on mammal richness and evolutionary history in Australia. *Ecography* **42**, 478–487. (doi:10.1111/ecog.03784)
46. Bromham L, Hua X, Fitzpatrick TG, Greenhill SJ. 2015 Rate of language evolution is affected by population size. *Proc. Natl Acad. Sci. USA* **112**, 201419704. (doi:10.1073/pnas.1419704112)
47. Kosse K. 1990 Group size and societal complexity: thresholds in the long-term memory. *J. Anthropol. Archaeol.* **9**, 275–303. (doi:10.1016/0278-4165(90)90009-3)
48. Dunbar RIM. 2008 Cognitive constraints on the structure and dynamics of social networks. *Res. Pract.* **12**, 7–16. (doi:10.1037/1089-2699.12.1.7)
49. Gavin MC *et al.* 2017 Process-based modelling shows how climate and demography shape language diversity. *Glob. Ecol. Biogeogr.* **26**, 584–591. (doi:10.1111/geb.12563)
50. Nichols J. 1990 Linguistic diversity and the first settlement of the new world. *Language (Baltim)* **66**, 475–521.
51. Nichols J. 1997 Modeling ancient population structures and movement in linguistics. *Annu. Rev. Anthropol.* **26**, 359–384. (doi:10.1146/annurev.anthro.26.1.359)
52. Peck SL. 2004 Simulation as experiment: a philosophical reassessment for biological modeling. *Trends Ecol. Evol.* **19**, 530–534. (doi:10.1016/j.tree.2004.07.019)
53. Goddard I. 1996 Native languages and language families of North America. In *Handbook of North American Indians volume 17: languages*. Washington, DC: Smithsonian Institution.
54. Mithun M. 2001 *The languages of native North America*. Cambridge, UK: Cambridge University Press.
55. Mace R, Pagel M. 1995 A latitudinal gradient in the density of human languages in North America. *Proc. R. Soc. B* **261**, 117–121. (doi:10.1098/rspb.1995.0125)
56. Kerr JT, Packer L. 1997 Habitat heterogeneity as a determinant of mammal species richness in high-energy regions. *Nature* **385**, 252–254. (doi:10.1038/385252a0)
57. Jetz W, Rahbek C. 2002 Geographic range size and determinants of avian species richness. *Science* **297**, 1548–1551. (doi:10.1126/science.1072779)
58. Kreft H, Jetz W. 2007 Global patterns and determinants of vascular plant diversity. *Proc. Natl Acad. Sci. USA* **104**, 5925–5930. (doi:10.1073/pnas.0608361104)
59. Hawkins BA *et al.* 2003 Energy, water, and broad-scale geographic patterns of species richness. *Ecology* **84**, 3105–3117. (doi:10.1890/03-8006)
60. Hillebrand H. 2004 On the generality of the latitudinal diversity gradient. *Am. Nat.* **163**, 192–211. (doi:10.1086/381004)
61. Clark PU, Mix AC. 2002 Ice sheets and sea level of the Last Glacial Maximum. *Quat. Sci. Rev.* **21**, 1–7. (doi:10.1016/S0277-3791(01)00118-4)
62. Sandel B, Arge L, Dalsgaard B, Davies RG, Gaston KJ, Sutherland WJ, Svenning J-C. 2011 The influence of Late Quaternary climate-change velocity on species endemism. *Science* **334**, 660–664. (doi:10.1126/science.1210173)
63. Bowern C. 2010 Correlates of language change in hunter-gatherer and other 'small' languages. *Lang. Lang. Compass* **4**, 665–679. (doi:10.1111/j.1749-818X.2010.00220.x)
64. Wessel P, Smith WHF. 1996 A global, self-consistent, hierarchical, high-resolution shoreline database. *J. Geophys. Res. Solid Earth* **101**(B4), 8741–8743. (doi:10.1029/96JB00104)
65. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005 Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978. (doi:10.1002/joc.1276)
66. Colwell RK. 1974 Predictability, constancy, and contingency of periodic phenomena. *Ecology* **55**, 1148–1153. (doi:10.2307/1940366)
67. Lima-Ribeiro MS, Varela S, González-Hernández J, Oliveira G, Diniz-Filho JAF, Terribile LC. 2015 ecoClimate: a database of climate data from multiple models for past, present, and future for Macroecologists and Biogeographers. *Biodivers. Inform.* **10**, 1–21.
68. Bivand R, Yu D. 2017 spgwr: geographically weighted regression (R software package (2014).