

RESEARCH  
PAPER



# Process-based modelling shows how climate and demography shape language diversity

Michael C. Gavin<sup>1,11\*†</sup>, Thiago F. Rangel<sup>2†</sup>, Claire Bown<sup>3</sup>, Robert K. Colwell<sup>4</sup>, Kathryn R. Kirby<sup>5</sup>, Carlos A. Botero<sup>6</sup>, Michael Dunn<sup>7</sup>, Robert R. Dunn<sup>8</sup>, Joe McCarter<sup>9</sup>, Marco Túlio Pacheco Coelho<sup>2</sup> and Russell D. Gray<sup>10,11</sup>

<sup>1</sup>Department of Human Dimensions of Natural Resources, Colorado State University, Fort Collins, CO, USA, <sup>2</sup>Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, GO, Brazil, <sup>3</sup>Department of Linguistics, Yale University, New Haven, CT, USA, <sup>4</sup>Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA and Museum of Natural History, University of Colorado, Boulder, CO, USA, <sup>5</sup>Department of Ecology & Evolutionary Biology and Department of Geography and Planning, University of Toronto, ON, Canada, <sup>6</sup>Department of Biology, Washington University in St Louis, St Louis, MO, USA, <sup>7</sup>Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden, <sup>8</sup>Department of Applied Ecology, North Carolina State University, Raleigh, NC, USA, <sup>9</sup>Center for Biodiversity and Conservation, American Museum of Natural History, New York, NY, USA, <sup>10</sup>School of Psychology, University of Auckland, Auckland, New Zealand and Research School of the Social Sciences, Australian National University, Canberra, Australian Capital Territory, Australia, <sup>11</sup>Max Planck Institute for the Science of Human History, Jena, Germany

\*Correspondence: Michael Gavin, Department of Human Dimensions of Natural Resources, Colorado State University, Fort Collins, CO 80523-1480, USA.

E-mail: michael.gavin@colostate.edu

<sup>†</sup>These authors contributed equally.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

## ABSTRACT

**Aim** Two fundamental questions about human language demand answers: why are so many languages spoken today and why is their geographical distribution so uneven? Although hypotheses have been proposed for centuries, the processes that determine patterns of linguistic and cultural diversity remain poorly understood. Previous studies, which relied on correlative, curve-fitting approaches, have produced contradictory results. Here we present the first application of process-based simulation modelling, derived from macroecology, to examine the distribution of human groups and their languages.

**Location** The Australian continent is used as a case study to demonstrate the power of simulation modelling for identifying processes shaping the diversity and distribution of human languages.

**Methods** Process-based simulation models allow investigators to hold certain factors constant in order to isolate and assess the impact of modelled processes. We tested the extent to which a minimal set of processes determines the number and spatial distribution of languages on the Australian continent. Our model made three basic assumptions based on previously proposed, but untested, hypotheses: groups fill unoccupied spaces, rainfall limits population density and groups divide after reaching a maximum population.

**Results** Remarkably, this simple model accurately predicted the total number of languages (average estimate 406, observed 407), and explained 56% of spatial variation in language richness on the Australian continent.

**Main conclusions** Our results present strong evidence that current climatic conditions and limits to group size are important processes shaping language diversity patterns in Australia. Our study also demonstrates how simulation models from macroecology can be used to understand the processes that have shaped human cultural diversity across the globe.

## Keywords

Culture, language diversity, macroecology, simulation modelling.

## INTRODUCTION

The approximately 7000 human languages are unevenly distributed across the globe (Lewis *et al.*, 2014). Although geographical patterns of language diversity are well documented, little is known about the processes that have shaped them. For centuries, scholars from multiple disciplines have proposed hypotheses to explain the origin of language diversity (Gavin *et al.*, 2013). Despite the existence of these long-standing and diverse hypotheses, only a limited number of studies have empirically assessed the relationship between geographical patterns in language richness and socio-cultural, environmental and geographical factors (Gavin *et al.*, 2013). Because results from these empirical studies contradict one another, no consensus exists regarding the relative importance of different causal factors (Gavin *et al.*, 2013). All previous empirical studies have relied on correlative, curve-fitting methods. Correlative analyses have the potential to identify factors associated with language diversity patterns. However, because correlation does not equate to causation, seeking the specific processes that drive language diversity patterns requires a process-based approach (Gotelli *et al.*, 2009; Gavin *et al.*, 2013).

A shift from correlative studies and narrative explanations to the use of simulation models (Rangel *et al.*, 2007; Gotelli *et al.*, 2009; Colwell & Rangel, 2010) has recently revolutionized the field of macroecology. These simulation models allow investigators to hold certain factors constant in order to isolate and assess the impact of others (Rangel *et al.*, 2007). Geographical simulation models can also be predictive, producing maps with taxon ranges that can be compared with observed range maps to provide a more statistically robust means of inferring the roles that specific variables and underlying processes play (or fail to play) in determining spatial diversity patterns (Rangel *et al.*, 2007). Here we present, to our knowledge, the first application of these tools derived from macroecology to examine the processes underlying the distribution of human groups and their languages.

To demonstrate the power of this new approach for understanding patterns in human diversity, we examine the spatial distribution of languages in Australia. The continent is an ideal domain for investigating questions of language diversity. Just as at the global scale, the number of languages spoken per unit area varies widely within Australia (Fig. 2a, c). To a greater extent than most regions, language ranges (i.e. the geographical area occupied by speakers of a language) prior to European settlement are known for the entire continent of Australia with a fair degree of confidence (Bowern, 2016). Moreover, by focusing on Australia we can examine continent-wide expansions of hunter-gatherer groups without any confounding effects from major differences in subsistence technology that characterize other continents.

The degree to which different processes influence language diversity patterns in Australia is contested. Some argue that human groups spread across the continent soon after arrival (40,000–50,000 years ago) (Birdsell, 1957). If regular contact between speakers is required to prevent linguistic divergence (Labov, 2001; Hock & Joseph, 2009), then environmental barriers, such as rivers and mountains, may have divided the

early, rapidly spreading populations into linguistically distinct groups (Gavin *et al.*, 2013; Axelsen & Manrubia, 2014). Others argue that changing environmental conditions during the Last Glacial Maximum (c. 23,000–15,000 years ago) and the Antarctic Cold Reversal (c. 14,500–12,500 years ago) would have driven populations into refugia that offered more resources (Evans & McConvell, 1997; Nettle, 1998; Sutton & Koch, 2008; Williams *et al.*, 2013). Isolation in refugia for substantial periods, along with neutral changes in languages, could have led to diversification (Gavin *et al.*, 2013; Williams *et al.*, 2015). In contrast to these views based on demic diffusion (i.e. languages spreading with populations) other researchers have emphasized the importance of cultural diffusion, in which languages spread and converge without mass migration (Dixon, 1997). Finally, limits on group size may facilitate the division of social groups (Birdsell, 1979; Kosse, 1990; Dunbar, 2008). Researchers have hypothesized that limits on group size represent a balance between, on the one hand, the benefits that larger groups provide for increased mate choice and cooperation in resource procurement and, on the other hand, the costs of maintaining social ties, including remembering and processing the information needed to recognize unrelated individuals (Kosse, 1990; Nettle, 1999; Dunbar, 2008). If limits on group size divide growing populations, and environmental carrying capacity varies in a region, spatial variation in language diversity patterns may arise. Despite the diversity of hypotheses offered in the literature, none of these ideas have been rigorously tested.

Our aim here is not to examine all possible hypotheses, but rather to test the extent to which a minimal set of processes may determine the number and spatial distribution of languages in Australia. Our model makes only three assumptions. First, human groups move to fill unoccupied spaces. Second, environmental carrying capacity drives local population density (an assumption supported by empirical analyses of hunter-gatherer groups; Birdsell, 1953). Third, linguistically defined social groups have a maximum population size. We assume that when the population speaking one language reaches a maximum size it will divide into two populations, each speaking a distinct language. Overall, our model is not driven by pure environmental determinism, but rather is based on a process by which environmental carrying capacity shapes population density and social limits on group size divide populations. We expect this intentionally simplistic model to perform poorly in areas where processes not explicitly modelled here (e.g. topographical barriers, historical movement in and out of refugia, cultural diffusion of language) have been major drivers of language diversity patterns.

## METHODS

The objective of the simulation is to evaluate the extent to which we can reproduce both the total number of languages and the spatial pattern of language richness (i.e. languages per unit area) of the pre-colonial distribution of Australian languages. Our model is stochastic, spatially explicit and based on hypothetical rather than actual languages.

## Environmental data

Data for analyses are presented in the figures in the main text and Supporting Information, or are available from the references cited. A hexagonal grid was laid out on a map of Australia (17,136 cells), and the area of each hexagonal cell was computed and recorded (average area 450 km<sup>2</sup>). We designed the grid to ensure that cell sizes would be large enough to encompass a group of individuals, but smaller than the range of 99% of languages in the empirical Australian language map. The mean annual precipitation for each map cell was extracted from WorldClim, which is based on data from 1950–2000 (Hijmans *et al.*, 2005) (Fig. S1 in the Supporting Information).

## Environmental carrying capacity

On the basis of previous correlation-based studies that found a relationship between precipitation and range size of hunter-gatherer-fisher (HGF) groups (Birdsell, 1953), we assumed that the carrying capacity for HGF groups, in spatial density of individuals, should also be a function of precipitation. The only free parameters in the model define the relationship between mean annual precipitation (using modern climate data) and a cell's carrying capacity (i.e. the number of people that will occupy the cell), with each model run exploring a different set of parameter combinations. Although there is evidence that carrying capacity increases with precipitation in multiple species, including humans, the exact relationship is unknown (Birdsell, 1953; Pascual *et al.*, 1997; Georgiadis *et al.*, 2003). For this reason, we evaluated some of the most widely used functions: power, exponential and logistic. The power function took the following form:

$$K = \alpha P^\beta$$

where  $K$  is the carrying capacity, measured in the number of individuals per square kilometre,  $P$  is mean annual precipitation, measured in millimetres of rain per year, and  $\alpha$  and  $\beta$  are unknown parameters to be estimated. While  $\alpha$  serves as a simple scaling factor,  $\beta$  governs the rate of growth of carrying capacity with the increase in precipitation. We also evaluated a two-parameter exponential function

$$K = \alpha + e^{\beta P}$$

and a three-parameter logistic function

$$K = \frac{\eta}{1 + e^{-\beta(P-\alpha)}}$$

in which  $\eta$  is the curve's maximum value,  $\beta$  governs the steepness of the curve and  $\alpha$  is the level of precipitation at the midpoint of the sigmoid.

## Group size

We used published information to establish typical numbers of members for HGF groups. These data implicitly

incorporate the social, cultural, political and cognitive factors that constrain group size. We used the dataset compiled by Binford (2001) for HGF group sizes from around the world. From the 339 groups included in the list, we excluded estimates of HGF groups located in arctic and subarctic biomes because these climate zones do not occur in Australia (Olson *et al.*, 2001), and we merged estimates for contiguous groups sharing a language. We also excluded estimates for HGF groups from Australia to avoid circularity in our arguments. This procedure left 178 group-size estimates, from which we built a statistical distribution (Fig. S2).

## Model algorithm

Given trial values for the parameters of the carrying capacity functions and the estimated precipitation values for the hexagonal map cells in Australia, for each cell ( $i$ ) we calculated the carrying capacity of HGFs ( $K_i$ ). Each simulation started with 10 individuals of a single language group occupying a single, randomly chosen map cell ( $i$ ). For each new language, a maximum population size was sampled from the empirical distribution of HGF group size (Binford, 2001). The model advanced in discrete time steps (see Movie S1). At any given time step, each language range covered a specific region of the map, constrained to be a contiguous set of cells. Time was modelled strictly as algorithmic steps, with no intended relationship to historic time. At each time step ( $t$ ), a regional carrying capacity ( $K_{i,j}$ ) of each occupied cell ( $i$ ) was the carrying capacity of the cell  $i$  ( $K_i$ ) plus the carrying capacity of all its  $p$  adjacent cells (cells that share an edge with a cell already occupied by the focal language or an unoccupied cell).

The increase in population size ( $N$ ) that was attributable to individuals of cell  $i$  between time step  $t$  and the next time step ( $t + 1$ ) was given by

$$N_{i,t+1} = rN_{i,t} \times \left[ 1 - \frac{N_{i,t} + \sum_{j=1}^p N_{i,j,t}}{K_i + \sum_{j=1}^p K_{i,j}} \right],$$

where the per capita intrinsic rate in population growth  $r = 1.01$ ,  $\sum N_{i,j,t}$  is the regional population size (the number of individuals at time  $t$  in all  $p$  cells, indexed by  $j$ , that are adjacent to cell  $i$ ), and  $\sum K_{i,j}$  is the regional carrying capacity (for the same  $p$  cells adjacent to cell  $i$ ). Thus, the equation above takes into account the potential population growth of individuals present in the cell  $i$ , but also the opportunity for colonization of the adjacent cells. The new individuals ( $N_{i,t+1} - N_{i,t}$ ) were distributed across all regional cells as a function of availability of resources in each cell ( $N/K$ ). The size of a population speaking a language increased and expanded its geographical range until it reached its maximum population size, which had been sampled, a priori, from the

empirical distribution of HGF group sizes. Notice that the population growth rate ( $r$ ) does not affect the outcome of the model, but only the rate of expansion of the simulated language family.

When a population reached its maximum size, an empty cell at the edge of the previous growing population was randomly chosen. The new cell was assumed to be colonized by 10 individuals representing a new language, which then experienced population growth and range expansion, just as for the first language. The new population was allowed to colonize only adjacent cells, but not cells already occupied by other language groups. Just as for each previously simulated language, the expansion of the new population was halted when the population reached its pre-determined maximum population size, or sooner if no adjacent empty cells remained available. For each run, the simulation stopped when all cells of the map had been colonized (see Movie S1).

### Observed patterns

To evaluate the capacity of the model to replicate observed patterns, we overlaid an equal-area grid on the map of Australia, with each cell measuring 300 km  $\times$  300 km. Our choice of grid resolution ensured that each cell encompassed more than one language, but also that cells were small enough to capture the variation in the richness of languages across space. Any grid cell with less than 20% of its area occupying land was removed from the grid. The final grid was composed of 98 cells.

Data for the Australian language map (Fig. 2a) came from all available sources on language locations (Bown, 2016). Numerous compilations exist in the literature (O'Grady *et al.*, 1966; Wurm, 1972; Tindale & Jones, 1974; Wurm & Hattori, 1981). Regional surveys were also consulted, including unpublished or locally published maps and pamphlets from regional Aboriginal language centres. For the most part, sources were in broad agreement, as might be expected since they were not compiled independently. Where sources did not agree on boundary locations, we favoured sources that were explicit about the basis of their evidence, local sources over national surveys, and primary sources (e.g. statements of locations in reference grammars) over secondary compilations. Where no single source was obviously more reliable than another, we used the majority placement.

A further issue concerned how to treat the boundaries between languages and dialects. Australian linguists have tended to 'lump' rather than 'split' (treating, for example, Yolŋu and Western Desert as single languages rather than as subgroups with several mutually unintelligible languages). The question of mutual intelligibility was challenging, since in areas of extensive multilingualism, speakers can understand quite different varieties through long acquaintance. Our choices for inclusion of 'dialects' were based on a combination of intelligibility, the need for even sampling and consistency across the continent.

We quantified the map of language richness by counting the number of observed language polygons that mapped into each 300 km  $\times$  300 km grid cell – intentionally much coarser (about 200 times) than the grid used for the simulations themselves. Because of the complex shape of Australia's coastline and the removal of coastal cells with a small land area, the occurrence of four small-ranged coastal languages was not recorded in our final map of language richness. We excluded island languages from our analysis. Our final list comprised 407 languages.

### Model test and validation

We tested the ability of the simulation model to replicate two observed patterns: (1) the map of language richness and (2) total number of languages in Australia (407).

To compare observed and predicted maps of language richness, the 300 km  $\times$  300 km grid was also overlaid on the simulated language map. We then produced the predicted map of language richness by counting the total number of simulated language ranges that intersected or were contained by each grid cell. The coefficient of determination ( $r^2$ ) of a linear regression was used to measure the fit between observed (response variable) and simulated (explanatory variable) maps of language richness. The difference between the observed and predicted number of languages for each cell yielded the map of residuals (Fig. 2e), which indicates regions where our model underestimated (positive values) or overestimated (negative values) the observed number of languages.

In addition to the map of language richness, we also evaluated the ability of the model to predict the number (407) of observed languages in the grid. We calculated a standardized measure of similarity ( $s$ ) between observed ( $O$ ) and predicted ( $P$ ) number of languages as

$$s = 1 - \frac{|O - P|}{P}.$$

To evaluate the combined ability of the model to simultaneously predict the map of language richness and the number of languages, we created an ad hoc goodness of fit index ( $f$ ) that took into account both measures of fit. The goodness of fit index was calculated as the average between (1) the coefficient of determination of the regression ( $r^2$ ) between the observed and predicted maps of language richness and (2) the standardized similarity index ( $s$ ) between the observed and predicted number of languages

$$f = \frac{r^2 + s}{2}.$$

The maximum possible value of  $f$  is 1, which would occur if the model predicted exactly 407 languages, and the pattern of language richness precisely matched the observed map ( $r^2 = 1$ ). However, because standardized similarity ( $s$ ) is not constrained by a minimum value,  $f$  may take negative values.

## Stochasticity

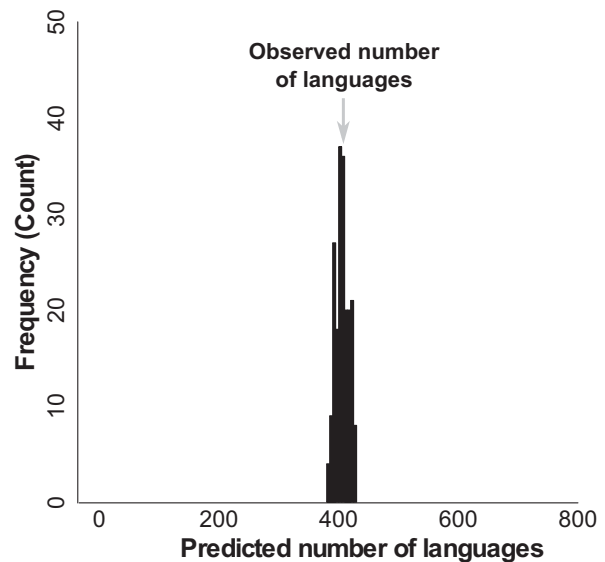
Two features of our simulation model were defined stochastically: (1) the seed cell for each language and (2) the maximum population size for each language. The seed cell for the initial language was chosen randomly from among all hexagonal cells on the map, whereas seed cells for subsequent languages were chosen from among the empty cells adjacent to the range of any previously simulated language. The maximum population size for each language was randomly chosen from the documented global distribution of HGF population sizes (Binford, 2001), before simulating the corresponding range, regardless of the location where the new language arose.

To assess the effect of stochasticity in the model prediction, we ran 120 replicates of the simulation for each parameter combination. Subsequently, to calculate the predicted map of language richness for each given parameter combination, we averaged predicted language richness, among the 120 replicates, for each grid cell. Similarly, the predicted number of languages for each parameter combination was calculated as the number of predicted languages averaged over all 120 replicates.

## Parameter estimation

Because the values of the parameters of the carrying capacity functions were not known, we constructed a Markov chain Monte Carlo Gibbs sampler (Gelman *et al.*, 2013) to stochastically explore the parameter space and evaluate how multiple parameter values and alternative carrying capacity functions affect the predictive power of the model. Using the goodness of fit index  $f$ , the Gibbs sampler algorithm estimates how likely it is that each of our models, as defined by different parameter combinations of the carrying capacity functions, would produce the observed language map data. We ran the sampler multiple times during preliminary evaluation of the model and used previous runs as burn-ins for subsequent runs by feeding the best parameter combinations of previous runs as seed parameters of the new run. For the final analysis, we ran  $8.4 \times 10^5$  replicates of the simulation to explore 7000 different parameter combinations (total chain length of the last analysis) for each of the three alternative carrying capacity functions. To compare the parameter combinations with regard to the model's explanatory power, we ranked the 7000 parameter combinations according to their goodness of fit index ( $f$ ).

The best fitting parameter combination for the power function provided substantially better predictive power for carrying capacity than the exponential and logistic functions. Thus, here we only provide results for the power function. The power function is a monotonic increasing function. The lack of an asymptote in the best-fitting function suggests that, based on current climate data, nowhere in Australia is wet enough to allow human population density to reach its potential maximum (see Fig. S3).



**Figure 1** Distribution of total number of languages predicted by the 200 best models. Average predicted total languages = 406. Observed total languages = 407. The lower limit of the x-axis was defined based on one being the theoretical minimum number of languages on the continent. The maximum extent of the x-axis (*c.* 800 languages) was set assuming all languages on the continent had the same range size, which was defined as the median language range size from the distribution of observed language ranges.

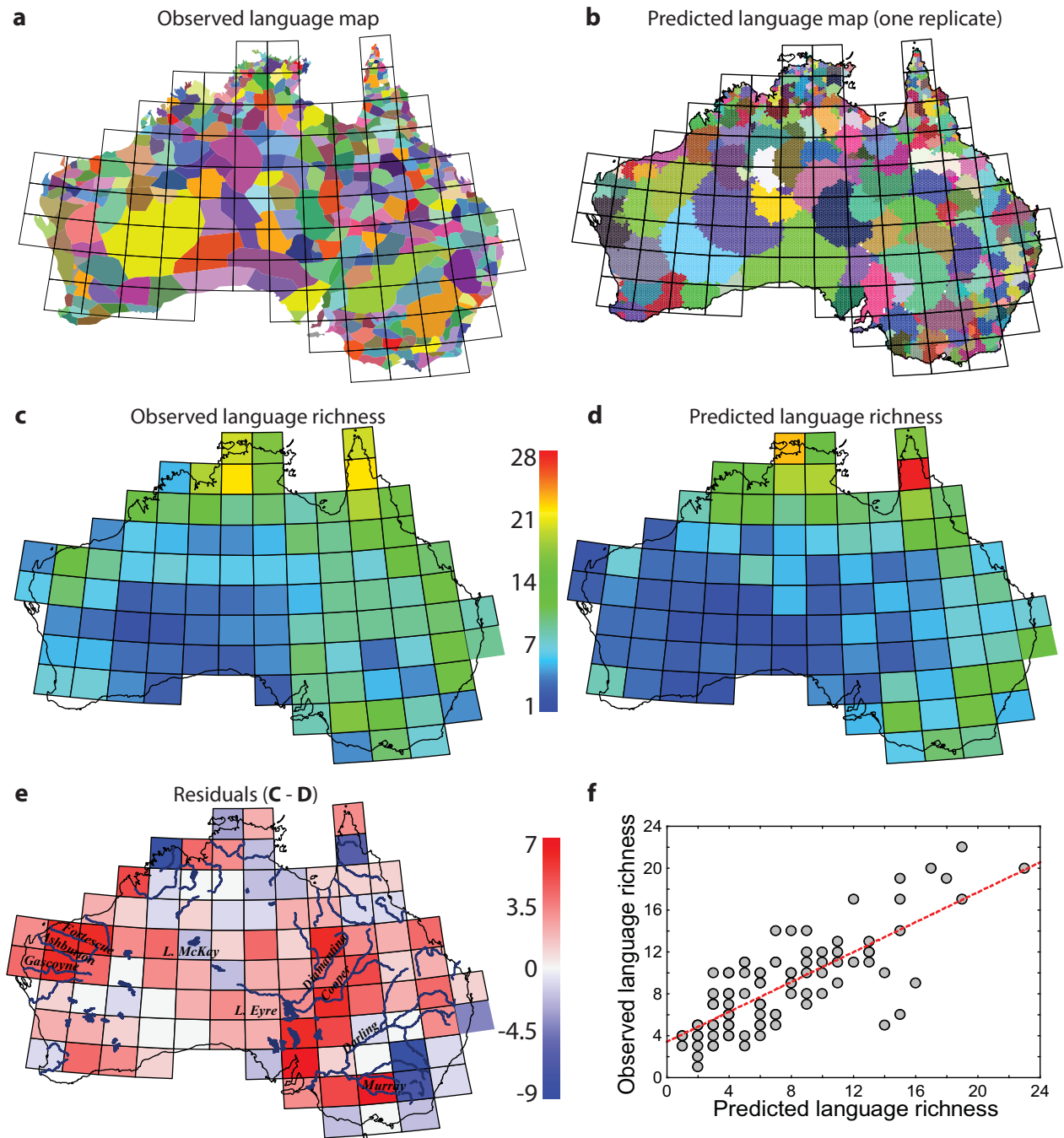
## RESULTS

Our model had strong predictive power across all criteria we evaluated. The estimates of the total languages in Australia (383–427 languages, average 406) produced by replicates of the 200 best parameter combinations (see Methods and Fig. S4) bracket the observed number (407 languages; Fig. 1). To be able to predict an aspect of human social behaviour with such precision and so few parameters is remarkable.

Visually, the maps for observed and estimated language richness per cell are strikingly similar (Fig. 2c, d), and the estimated language richness per cell explains 56% of the variation in observed richness (Fig. 2f; for the best 200 models  $0.53 < r^2 < 0.58$ , Fig. S5). The location of initial colonization had no effect on model results (see Methods and Fig. S6). We also used a traditional correlational approach to examine the relationship between observed language richness and mean precipitation value in each  $300 \text{ km} \times 300 \text{ km}$  grid cell ( $r^2 = 0.43$ ). Not only did our simulation model explicitly incorporate processes shaping language diversity, but it also had greater predictive power (56% *vs* 43%) than the simple correlation approach.

## DISCUSSION

Our results provide evidence that current climatic conditions and limits to group size have a causal role in shaping spatial patterns of indigenous language richness on the Australian



**Figure 2** Observed and predicted language diversity patterns for the continent of Australia. The observed language map (a) shows ranges for a total of 407 languages (based on data from Bovern (2016)). The predicted language map (b) depicts results from one model replicate (final analysis includes  $8.4 \times 10^5$  replicates). The observed (c), predicted (d), and residual (e) maps of language richness shown were based on the best parameter combination. The scale in (c) and (d) depicts the number of languages with ranges intersected or contained by a given grid cell. The scale in (e), which illustrates fit, represents (observed – predicted) languages in each grid cell. The model predicts 56% of the variation in observed language richness patterns (f).

continent. The causal link between climate and carrying capacity suggests that climatic conditions similar to those currently prevailing in Australia drove processes responsible for the majority of the pattern observed.

Because we deliberately built a minimal model for a complex process, it is not surprising that some variation in the spatial patterns of diversity remains unexplained (Fig. 2e, f). Future work could incrementally incorporate a wide variety

of historical, social and environmental factors to investigate more fully the determinants of geographical patterns in language diversity, both in Australia and on a global scale. The map of residuals (Fig. 2e) offers clues regarding additional factors that may be at play. For example, the model underestimated language richness (red cells in Fig. 2e) in areas with a greater density of large rivers and lakes, such as the confluence of the Murray and Darling Rivers in the south, the region around Lake Eyre and adjoining rivers, and the Fortescue, Ashburton and Gascoyne Rivers in the west. Additional water resources may increase carrying capacity and potential linguistic diversity beyond limits shaped by precipitation alone. In other regions, overestimation of language diversity may relate to inaccuracies in the observed language map. For example, an early and sustained colonial impact in the south-east may have caused extinction of languages prior to documentation. Also, the degree to which certain languages are distinct or represent dialectal clusters is contentious. Under-splitting of these languages in the creation of the observed map could lead to model overestimation (e.g. the Ngarinyin language cluster in the north-west; dark blue cell in Fig. 2e).

Recent work in biogeography has demonstrated the value of process-based simulation models for uncovering mechanisms driving species diversity patterns (Rangel *et al.*, 2007; Gotelli *et al.*, 2009; Colwell & Rangel, 2010). These previous studies reveal that mechanisms tend to vary across different taxa and regions (Gotelli *et al.*, 2009). We believe that the mechanisms shaping the patterns of diversity of human languages will also be variable in their impact and context dependent. For example, patterns of species richness are strongly influenced by the overlap of ranges, but the ranges of human languages tend not to overlap in space, as we see in the Australian case presented here (Gavin & Stepp, 2014). Instead, the forces that determine language range sizes shape spatial patterns in language richness, with smaller ranges leading to greater richness per unit area. In Australia, we conclude that environmental carrying capacity and human group sizes are major influences on language range sizes and diversity patterns. However, we expect that either different degrees of the same processes or different processes will drive language diversity patterns in other regions, where environmental, social and historical conditions differ substantially from Australia. The methodological approach we outline here can be used to compare the degree to which different processes shape patterns of language diversity across different regions of the world.

Generations of anthropologists and linguists have documented the impressive diversity of human cultures and languages. What remains largely unknown is how this diversity evolved, and why some regions support far more diversity than others. The factors causing these patterns of diversity undoubtedly vary across space and time. The Australian case we present here exemplifies a new and promising approach using simulation models derived from macroecology to

determine the processes that have shaped human diversity across the globe.

## ACKNOWLEDGEMENTS

Research was supported in part by the National Evolutionary Synthesis Center in Durham, NC, USA, the Max Planck Institute for the Science of Human History and the National Science Foundation (award no. 1519987).

## REFERENCES

- Axelsen, J.B. & Manrubia, S. (2014) River density and landscape roughness are universal determinants of linguistic diversity. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20133029.
- Binford, L.R. (2001) *Constructing frames of reference: an analytical method for archaeological theory building using ethnographic and environmental data sets*. University of California Press, Berkeley, CA.
- Birdsell, J.B. (1979) Ecological influences on Australian aboriginal social organization. In *Primate ecology and human origins: ecological influences on social organization* (ed. by I.S. Bernstein and E.O. Smith), pp. 117–151. Garland STPM Press, New York.
- Birdsell, J.B. (1953) Some environmental and cultural factors influencing the structuring of Australian Aboriginal populations. *The American Naturalist*, **87**, 171–207.
- Birdsell, J.B. (1957) Some population problems involving Pleistocene man. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 47–69.
- Bowern, C. (2016) The Australian comparative lexical database. *Language Documentation and Conservation*, **9**, 1–45.
- Colwell, R.K. & Rangel, T.F. (2010) A stochastic, evolutionary model for range shifts and richness on tropical elevational gradients under Quaternary glacial cycles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 3695–3707.
- Dixon, R.M. (1997) *The rise and fall of languages*. Cambridge University Press, Cambridge.
- Dunbar, R.I.M. (2008) Cognitive constraints on the structure and dynamics of social networks. *Group Dynamics: Theory, Research, and Practice*, **12**, 7–16.
- Evans, N. & McConvell, P. (1997) The enigma of Pama-Nyungan expansion in Australia. *Archaeology and Language*, **II**, 174–191.
- Gavin, M.C. & Stepp, J.R. (2014) Rapoport's rule revisited: geographical distributions of human languages. *PLoS One*, **9**, e107623.
- Gavin, M.C., Botero, C.A., Bowern, C., Colwell, R.K., Dunn, M., Dunn, R.R., Gray, R.D., Kirby, K.R., McCarter, J., Powell, A., Rangel, T., Stepp, J., Trautwein, M., Verdolin, J.L. & Yanega, G. (2013) Toward a mechanistic understanding of linguistic diversity. *BioScience*, **63**, 524–535.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2013) *Bayesian data analysis*. CRC Press, Boca Raton, FL.

- Georgiadis, N., Hack, M. & Turpin, K. (2003) The influence of rainfall on zebra population dynamics: implications for management. *Journal of Applied Ecology*, **40**, 125–136.
- Gotelli, N.J., Anderson, M.J., Arita, H.T. *et al.* (2009) Patterns and causes of species richness: a general simulation model for macroecology. *Ecology Letters*, **12**, 873–886.
- Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.
- Hock, H.H. & Joseph, B.D. (2009) *Language history, language change, and language relationship: an introduction to historical and comparative linguistics*. Walter de Gruyter, Berlin.
- Kosse, K. (1990) Group size and societal complexity: thresholds in the long-term memory. *Journal of Anthropological Archaeology*, **9**, 275–303.
- Labov, W. (2001) *Principles of linguistic change, Vol. 2 Social factors*. Blackwell Publishers, Malden, MA.
- Lewis, M.P., Simons, G.F. & Fennig, C.D. (2014) *Ethnologue: languages of the world*, 17th edn. SIL International, Dallas, TX. Online version: <http://www.ethnologue.com>.
- Nettle, D. (1998) Explaining global patterns of language diversity. *Journal of Anthropological Archaeology*, **17**, 354–374.
- Nettle, D. (1999) *Linguistic diversity*. Oxford University Press, Oxford.
- O'Grady, G.N., Voegelin, C.F. & Voegelin, F.M. (1966) Languages of the world: Indo-Pacific fascicle six. *Anthropological Linguistics*, **8**, 1–197.
- Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V., Underwood, E.C., D'Amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P. & Kassem, K.R. (2001) Terrestrial ecoregions of the world. *BioScience*, **51**, 933–938.
- Pascual, M.A., Kareiva, P. & Hilborn, R. (1997) The influence of model structure on conclusions about the viability and harvesting of Serengeti wildebeest. *Conservation Biology*, **11**, 966–976.
- Rangel, T.F.L.V.B., Diniz-Filho, J.A.F. & Colwell, R.K. (2007) Species richness and evolutionary niche dynamics: a spatial pattern-oriented simulation experiment. *The American Naturalist*, **170**, 602–616.
- Sutton, P. & Koch, H. (2008) Australian languages: a singular vision. *Journal of Linguistics*, **44**, 471–504.
- Tindale, N.B. & Jones, R. (1974) *Aboriginal tribes of Australia: their terrain, environmental controls, distribution, limits, and proper names*. University of California Press, Berkeley, CA.
- Williams, A.N., Ulm, S., Cook, A.R., Langley, M.C. & Collard, M. (2013) Human refugia in Australia during the Last Glacial Maximum and Terminal Pleistocene: a geospatial analysis of the 25–12 ka Australian archaeological record. *Journal of Archaeological Science*, **40**, 4612–4625.
- Williams, A.N., Ulm, S., Turney, C.S.M., Rohde, D. & White, G. (2015) Holocene demographic changes and the emergence of complex societies in prehistoric Australia. *PLoS One*, **10**, e0128661.
- Wurm, S.A. (1972) *Languages of Australia and Tasmania*. Mouton, The Hague.
- Wurm, S.A. & Hattori, S. (1981) *Language Atlas of the Pacific Areas: New Guinea Area, Oceania, Australia*. Australian Academy of the Humanities in collaboration with the Japan Academy, Canberra.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site:

**Movie S1** Example simulation of one replicate of the model. The final analysis was based on  $8.4 \times 10^5$  replicates. The geographical location of model initiation does not affect the results (see Fig. S6).

**Figure S1** Map of mean annual precipitation on the Australian mainland.

**Figure S2** Empirical distribution of group size for hunter-gatherer-fishers.

**Figure S3** Power functions used to predict carrying capacity (individuals/km<sup>2</sup>) from mean annual precipitation (mm), as estimated by the 10 best models.

**Figure S4** Relationship between carrying capacity parameters.

**Figure S5** Distribution of the coefficient of determination ( $r^2$ ) between observed and predicted maps of language richness for the 200 best models.

**Figure S6** Predictive power of the model when each cell is used as seed cell.

## DATA ACCESSIBILITY

As detailed in the Methods, all data are from published sources. Precipitation data were extracted from WorldClim (Hijmans *et al.*, 2005). Data on group size are from Binford (2001), and the observed Australian language map is available from Bowern (2016).

## BIOSKETCH

The team for this project included researchers with expertise in: biogeography (Thiago Rangel, Robert Colwell, Robert Dunn, Marco T. Coelho), linguistics (Claire Bowern, Michael Dunn), evolutionary studies (Carlos Botero, Russell Gray), geography and ethnobiology (Michael Gavin, Kathryn Kirby and Joe McCarter). The team's joint research has focused on the synthesis of theory, methods, and data across multiple disciplines to study spatial patterns of human cultural and linguistic diversity.

Editor: Allen Hurlbert