Using Bayesian non-linear models to uncover broad macroecological patterns

Bernat Bramon Mora[1,*], Antoine Guisan[2,3] and Jake M. Alexander[1]

[1]Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland; [2]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland; [3]Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland; [*]bernat.bramon@gmail.com

Abstract

Species' realized niches are classically pictured as bell-shaped probability distributions. These distributions, however, can actually take many different forms, including fat-tailed or skewed responses. While one does not need to know the shape of species' distributions to effectively model them, studying their basic form can teach us a lot about the ways climatic processes and historical contingencies shape ecological communities. Unfortunately, we still lack a general understanding of the properties describing the shape of species' distributions, and how these compare to each other across gradients. Here, we use a set of Bayesian non-linear models to quantify such properties in empirical plant distributions. With this approach, we are able to distil the shape of plant distributions, tackling long-standing hypotheses regarding the way ecological communities are assembled across space. Studying the relationship between distribution properties, we revealed the existence of broad macroecological patterns along environmental gradients. We also shed light on the extent to which some aspects of the shape of observed realized niches—such as kurtosis and skewness of the distributions—could be intrinsic properties or the result of species' historical contexts. Overall, our results provide a novel perspective on the way systems of many species are distributed along environmental gradients.

Introduction

One of the central goals of ecology is to understand the ways species are distributed across space and time. Many ecological textbooks assume the shape of species' realized niches to be unimodal and symmetric along environmental gradients (Krebs, 1972). In practice, there is a strong argument to be made in favour of assuming these to be bell shaped. Namely, if all that we are willing to assume about species' distributions is that these occupy finite geographic ranges, the most conservative statistical approach is to model their distribution as Gaussian (i.e. the corresponding maximum entropy distribution; Frank 2009). However, some have warned that empirical distributions can take many different forms (Austin, 1987, 2002), and there is currently no general agreement on the basic shape of species' realized niches (Sagarin & Gaines, 2002; **?**; Sagarin *et al.*, 2006). Indeed, many factors can play a role in defining their shape, and several natural processes can lead to non-normal distributions.

Fat-tailed and skewed distributions are very common across scientific fields. The former naturally emerges as a result of processes involving intermittent (e.g. in email communications patterns; Malmgren *et al.* 2008) or stochastic events (e.g. in the spread of infectious diseases; Wong & Collins 2020). Indeed, species' dispersal patterns have been shown to have fat tails due to the natural variability among individuals (Petrovskii *et al.*, 2009). This is important because one might expect environmental and individual variation to also be crucial factors determining the presence and absence of species along gradients, and fat tails are therefore a plausible property of species' realized niches. Similarly, several processes can lead to skewed distributions. For example, species might have asymmetric environmental tolerances along altitudinal gradients, allowing them to withstand different temperature extremes (Sunday *et al.*, 2011). Species might also experience abiotic and biotic pressures that increase or decrease along a temperature gradient, which could result in species' distributions presenting steeper declines towards warmer or colder environments (Normand *et al.*, 2009). Overall, many different properties could characterize species' realized niches, and every new shape entails different and potentially competing underlying hypotheses regarding the way communities are assembled over time (D'Amen *et al.*, 2017).

Comparing these properties across species allows us to study broad macroecological patterns that could be critical from a conservation and management perspective (Stevens, 1992;

Channell & Lomolino, 2000a). For example, Rapoport's rule, a classic biogeographical hypothesis, predicts species' range size to increase with latitude or elevation (Stevens, 1992), hinting at the existence of general biogeographical constraints that shape species' distributions along gradients. These sorts of macroecological patterns are interesting because they provide insights into the way different species assemble and establish in different environments (Linder *et al.*, 2000). That is, the differences in species' responses to the environment can shed light on how climatic processes and historical contingencies have differently shaped their distributions (Rohde, 1992; Helmuth *et al.*, 2004; Siefert *et al.*, 2015). Uncovering the shape of species' realized niches and the extent to which these vary across species is nevertheless a challenging statistical problem. Indeed, to this date, we do not have an effective way to parsimoniously compare the shape of the realized niches of many species along environmental gradients.

Over the last two decades, ecologists have developed a plethora of distribution models to try to untangle the factors that play a role in defining species' realized niches (Guisan & Zimmermann, 2000; Zurell *et al.*, 2019). These models are fundamental to the scientific community for predicting changes in species' geographic distributions and the effects of environmental disturbances. Such frameworks, however, commonly assume an underlying linear relationship between covariates (but see 'semiparametric models'; Norberg *et al.* 2019). This is useful because it simplifies the optimization process, but it might not be ideal when studying and comparing the shape of species' distributions along environmental gradients. First and foremost, a linear relationship between covariates often comes with a set of implicit mathematical constraints. These might not hinder the predictive performance of the models (Norberg *et al.*, 2019), but a direct biological interpretation of parameter estimates in linear models becomes increasingly difficult as one moves from unimodal and symmetric distributions (ter Braak & Looman, 1986; Jamil & ter Braak, 2013) to fat-tailed or skewed responses (Huisman *et al.*, 1993). Second, these mathematical constraints also limit our ability to include any prior information to our parameter estimates. Observations of species' geographic variation and optimal climatic conditions have long been documented, with extensive databases compiled by botanists and field ecologists documenting basic knowledge of species' realized niches (e.g. Landolt *et al.* 2010). That said, this information is rarely accounted for in most modelling approaches, likely because there is not a straightforward way to feed this information into the parameters of a linear model (Scherrer & Guisan 2019; but see ter Braak & Looman 1986; Ovaskainen *et al.* 2017). Finally, some have proposed several non-linear structures to characterize several features of individual species' response curves (Huisman *et al.*, 1993). Setting aside the fact that the interpretation and comparison of parameter estimates becomes challenging following most of these model structures, these are generally not designed to jointly study different species, and therefore taking full advantage of modern statistical approaches (e.g. sharing information among species or accounting for parameter uncertainty; Evans *et al.* 2016).

In this work, we develop a set of Bayesian hierarchical distribution models to study the shape of empirical plant distributions. We start by considering species' response curves as Gaussian distributed, and then we adapt our model to allow non-linear responses characterizing skewed and long-tailed distributions. This allows us to measure different properties of species' realized niches while accounting for all prior information regarding these distributions, including expert knowledge of species' environmental indicator values, range sizes, and plant ecological strategies. Using our statistical framework, we study the distribution of plant species along an elevation gradient in the Swiss Alps, revisiting some classic hypotheses in ecology and biogeoraphy. Specifically, we are able to compare the basic properties of the realized niche of multiple species, testing for the existence of broad macroecological patterns. Comparing the posterior distribution of those parameters that control for the shape of distributions, we are also able to showcase variation in the way different types of species, such as native or non-native, might respond to the environment. More generally, we are able to uncover the approximate shape of empirical plant distributions and answer fundamental questions regarding the way systems of many species are distributed along environmental gradients.

Methods

Empirical data

We studied the distribution of plant communities along an elevation gradient. To do so, we combined two different datasets: i) one describing the co-occurrence of species across multiple open grasslands in the Swiss Alps (Randin *et al.*, 2009), and ii) an extensive plant-attribute database containing environmental and life-history traits for all plant species across Switzerland (Landolt *et al.*, 2010).

Distribution data

We used data describing the distribution of 798 species across 912 sites covering most of the mountain region of the Western Alps in the Canton de Vaud (Switzerland; Scherrer & Guisan 2019). Each of these sites is a 8×8 m plot placed somewhere along an elevation range from 375 m to 3210 m. In all sites, presence/absence data as well as Braun-Blanquet abundance-dominance classes were recorded for all species. Additionally, we used meteorological data provided by Scherrer & Guisan (2019), containing multiple variables characterizing the climate in each site at high spatial resolution (25 m). This dataset was compiled based on 30 years (1961–1990) of records from national weather stations. Since most of the data are highly correlated, we used a Principal Component Analysis to calculate the main axes of variation of the following scaled variables: daily minimum, maximum and average temperature; sum of growing degree-days above 5°C; mean temperature of wettest quarter; annual precipitation, precipitation seasonality, and precipitation of driest quarter (Supplementary Fig. 1). Notice that the first axis of variation is positively correlated with elevation and negatively correlated to temperature, while the second axis is positively correlation to precipitation seasonality (Supplementary Fig. 1).

Plant attributes

To complement the aforementioned distribution data, we used a attribute database of around 5500 vascular plants across Switzerland. Notice that some of the information in this database has been previously shown to account for unexplained variation when used as explanatory variables in species' distribution models (Scherrer & Guisan, 2019). It was built based on expert knowledge and phytosociological field experience of botanists and ecologists, and contains information regarding plants' environmental preferences and ecological strategies.

Species' environmental preferences in this database can be used to inform distribution models—e.g. as an informative prior in a Bayesian framework. These are characterized following the ecological indicator values developed by Landolt *et al.* (2010), providing both an estimate of the average conditions in which a species can be found as well as a broad description of their range of variation. These values are provided for a range of 8 environmental variables, including temperature, continentality, light conditions, as well as moisture, acidity and nutrient content of the soil (see a full list and description of the ecological indicators in the Supplementary Table 1; Landolt *et al.* 2010). In addition to species' environmental preferences, the attribute data also contain information on species' introduction status (e.g. identifying those species that are recent and historical range expanders) and change tendency (e.g. indicating species that have shown decline or increase in their populations over the recent decades). We describe this information in more detail in Supplementary Table 1.

Baseline model

There is a long list of model structures well suited to characterizing species' distributions (see Norberg *et al.* 2019). As a baseline model, however, we were interested in a hierarchical model that does not make any assumptions regarding the shape of the distributions, and yet explicitly incorporates all information that we have regarding plant's environmental preferences. More specifically, we wanted to account for the climatic indicator values and range of variation registered in the attribute database for all plants in our dataset. These two values provide basic information regarding plants' optimal environmental conditions and the width of their distributions.

Response curve

To choose an appropriate response curve, we first need to agree on what we truly know about the system. Given the prior information that we have about the system, we know that species occupy specific geographic ranges; therefore, we know that their distributions have finite variance. While we could also assume that many other factors might influence species' presence in a given site—e.g. the biotic interactions among species in the site—we do not necessarily have an *a priori* expectation of how exactly these factors will influence the shape of species' distributions. Therefore, for this baseline model, we choose the maximum entropy distribution with finite variance: a Gaussian distribution (Fig 1a). That is, given the presence/absence or abundance y_{ij} of any species i in any given site j, and an environmental variable x_j, we can define species' responses to the environment as

$$
\begin{aligned}
y_{ij} &\sim F\left(p_{ij}\right) \\
\log\left(p_{ij}\right) &= -\alpha_i - \gamma_i\left(x_j - \beta_i\right)^2,
\end{aligned}
\tag{1}
$$

where F is the likelihood function, and α_i, β_i, and γ_i^{-1} describe the amplitude of the probability p_{ij}, species' average climatic suitability and range of variation along the environmental gradient, respectively. Notice that F characterizes a Binomial distribution when considering binary data, and it characterizes an ordered categorical likelihood function when we consider Braun-Blanquet abundance-dominance classes as response variables (see the full description of both models in the Supplementary Methods). For the sake of simplicity, we use only one environmental variable to characterize the species' probability distribution. That said, this model can easily be generalized to account for multiple predictors (see Supplementary Methods).

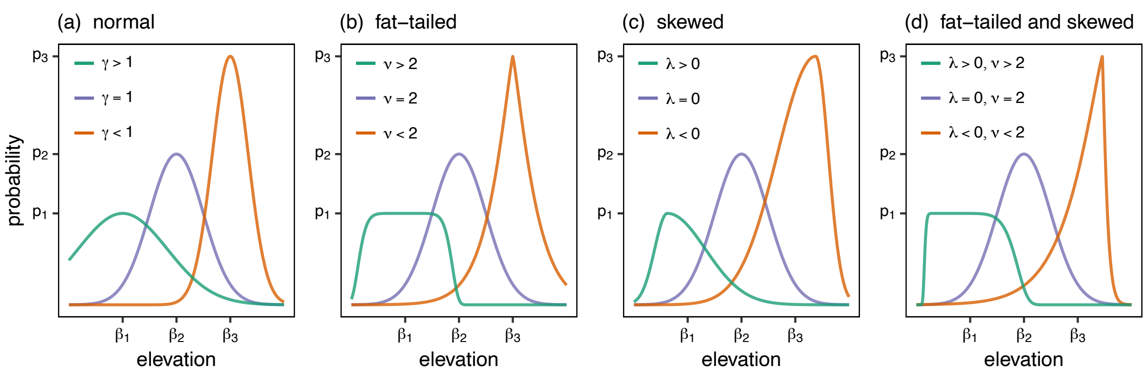

Figure 1: Different response curves. Panel (a) shows the distribution shapes characterized by Eq. (1) for different values of β, α and γ. Panel (b) shows the distribution shapes characterized by Eq. (4) for different values of β, α and ν, when $\gamma = 1$. Panel (c) shows the distribution shapes characterized by Eq. (5) for different values of β, α and λ, when $\gamma = 1$. Panel (d) shows the distribution shapes characterized by Eq. (6) for different values of β, α, λ and ν, when $\gamma = 1$. Notice that for all panels, we chose α values such that $\mathrm{p}_i = \exp(-\alpha_i)$.

Model priors

The model structure described above allows us to explicitly incorporate all prior knowledge that we have regarding species' distributions contained in the attribute database. To do so, we define the prior distributions for the parameters in model (1) as:

$$
\begin{aligned}
\log(\alpha_i) &\sim \text{Normal}\left(\hat{\alpha}, \sigma_\alpha\right) \\
\beta_i &\sim \text{MVNormal}\left(\hat{\beta}, \Sigma^\beta\right) \\
\log(\gamma_i) &\sim \text{MVNormal}\left(\hat{\gamma}, \Sigma^\gamma\right) \\
\hat{\alpha}, \hat{\beta}, \hat{\gamma} &\sim \text{Normal}\left(0, 1\right) \\
\sigma_\alpha &\sim \text{Exponential}\left(1\right)
\end{aligned}
\tag{2}
$$

where parameters γ_i and β_i are expressed as multivariate normal distributions—i.e. Gaussian processes—such that Σ^β and Σ^γ are variance-covariance matrices describing species' similarity in terms of their average climatic suitability and range of variation along the different environmental gradients, respectively. We define these variance-covariance matrices as follows:

$$\Sigma_{ij} = \eta \exp\left(-\rho {D_{ij}}^2\right) + \delta_{ij}\sigma, \tag{3}$$

where Σ_{ij} characterizes the covariance between any pair of species i and j, and δ_{ij} is the Kronecker delta. Such a covariance structure declines exponentially with the square of a distance matrix D_{ij}, which characterize differences between species computed using our prior information. In the attribute database, this information is represented by the set of ordinal traits specified for the different species. While there are many different ways to turn ordinal data into distance matrices, we choose to use a mixed-membership stochastic block model because it allows us to deal with cases of missing data (see Supplementary Methods for extended details; Godoy-Lorite *et al.* 2016). In each covariance matrix, the hyperparameter ρ determines the rate of decline of the covariance between any two species, and η defines its maximum value. The hyperparameter σ describes the additional covariance between the different observations for any given species. For all these hyperparameters, we choose weakly informative priors such that $\sigma, \eta \sim \text{Exponential}\left(1\right)$ and $\rho \sim \text{Exponential}\left(0.5\right)$. Notice that other structures can be used to define the covariance matrices of the different Gaussian processes (McElreath, 2020), including structures that account for multiple distance matrix D_{ij} for any given parameter (Supplementary Methods).

Sampling the posterior

We generated the posterior samples for the Bayesian models with the Hamiltonian Monte Carlo algorithm implementation provided by the R packages 'rstan' and 'cmdstanr' (Stan Developent Team, 2021). Sampling models like the ones described above can be computationally very intensive. This is especially true when using ordered categorical likelihood functions (see Stan Development Team 2021). Therefore, we focus on those species for which we have at least 20 occurrences when modelling both binary data and ordinal data (251 species in total).

To test the performance of the model as well as our choice of prior distributions, we modelled simulated data and compared the sampled posterior distributions to the data-generating parameters (e.g. Supplementary Fig. 2; see Code Availability section). Notice that using the link function in Eq. (1) could cause problems when sampling the model, and some adjustments need to be made when specifying the model (see Code Availability section). To perform the data analysis and generate the figures, we used some of the functions available with the R package 'rethinking' (McElreath, 2020).

Modifying the baseline model

We proposed a baseline model that is naive regarding how the data is distributed, and yet accounts for all prior information that we have about the system. Now, we want to modify this model to test the extent to which empirical species' distributions showcase different shapes. We focused on two properties: fat-tailed and skewed responses. While there are several model structures that could account for these properties, we propose new species' response curves following three criteria. First, the probability distribution of a species along an environmental gradient must have a defined mean and variance. This is important because we know that species naturally have different environmental preferences as well as finite geographic ranges. Second, the Gaussian shape must be a special case of the probability distribution, allowing species to showcase variation regarding the presence (or lack thereof) of any given pattern. Finally, there must be a re-parametrization of the model that allows us to keep the same prior information and interpretable parameters.

Fat-tailed response curve

Fat-tailed distributions represent distributions with relatively high representation of extreme events. While many different distributions exhibit this property, we decided to accommodate this feature into our baseline model by considering a response curve that follows a generalized error distribution. Such a distribution is useful because the Gaussian shape is a special case of it, and it contains a parameter that regulates the level of kurtosis—ranging from longer to shorter tails than the Gaussian case (Fig **1**b). In particular, we can adapt Eq. (1) to present this non-linear form as follows:

$$\log\left(p_{ij}\right) = -\alpha_i - \gamma_i' \left|x_j - \beta_i\right|^{\nu_i}, \tag{4}$$

where $\gamma_i' = g(\gamma_i, \nu_i)$, and ν_i is a parameter that describes the kurtosis of the distribution, which we define as $\nu_i \in (1, \infty)$. Following this, we choose an adaptive prior for this set of new parameters such that $\log(\nu_i - 1) \sim \text{Normal}(\hat{\nu}, \sigma_\nu)$, where $\hat{\nu} \sim \text{Normal}(0, 1)$ and $\sigma_\nu \sim \text{Exponential}(2)$. Given the relationship between γ_i' and γ_i, we can re-parametrize the model and follow Eq. (2) to define the prior distributions (see Supplementary Table 2; Nadarajah 2005). Notice that the Gaussian distribution will naturally emerge when $\nu_i = 2$.

Alternatively, we could have used other distributions that present fat tails and fulfil the selection criteria described above. For example, the non-standardized Student's t-distributions is an interesting distribution because, as opposed to the generalized error distribution, it allows for fat tails without generating a cusp at the center (see Fig **1**b). However, we avoided using the non-standardized Student's t-distributions because it does not allow for tails that are lighter than normal (e.g. $\nu_i > 2$ in Eq. 4; Fig **1**b), and the sampling of the model can be somewhat more challenging.

Skewed response curve

Skewed responses present steeper declines towards either side of the distribution. One way to accommodate this feature in our models is by considering a skewed normal distribution. As for the case described above, the Gaussian is a special case of this distribution, and it contains a parameter that controls for the level and direction of 'skewness' (Fig **1**c). Importantly, this distribution presents normal-like tails; therefore, the added skewness does not make additional assumptions regarding how species are distributed along the gradient. To test for the existence of this feature, we modified Eq. (1) as

$$\log\left(p_{ij}\right) = -\alpha_i - \gamma_i' \left(\frac{x_j - \beta_i'}{1 + \lambda_i \operatorname{sgn}\left(x_j - \beta_i'\right)} \right)^2, \tag{5}$$

where $\gamma_i' = q_1\left(\gamma_i, \lambda_i\right)$, $\beta_i' = q_2\left(\gamma_i, \beta_i, \lambda_i\right)$, and λ_i is a parameter that describes the skewness of the distribution such that $\lambda_i \in (-1, 1)$. The function $\operatorname{sgn}(x)$ characterizes the sign function. We chose λ_i to have an adaptive prior such that $\operatorname{logit}\left(\frac{\lambda_i+1}{2}\right) \sim \text{Normal}\left(\hat{\lambda}, \sigma_\lambda\right)$, where $\hat{\lambda} \sim \text{Normal}\,(0, 1)$ and $\sigma_\lambda \sim \text{Exponential}\,(1)$. Notice that this model can be re-parametrized following q_1 and q_2, allowing us to set the rest of the prior distributions as described for the baseline model (see Supplementary Table 2; Code Availability section). In this case, the Gaussian distribution is a special case of Eq. (5) when $\lambda_i = 0$ (Ashour & Abdel-hameed, 2010).

Fat-tailed and skewed response curve

Finally, one could consider a response curve with both kurtosis and skewness. A convenient way to achieve this is by using a response curve that follows a skewed generalized error distribution. This is a combination of the two distributions described above, containing two parameters that control for both the level and direction of kurtosis and skewness (Fig **1**d). The skewed generalized error distribution can be considered by modifying the species' response curve in Eq. (1) as

$$\log\left(p_{ij}\right) = -\alpha_i - \left(\frac{\gamma_i' \left|x_j - \beta_i'\right|}{1 + \lambda_i \operatorname{sgn}\left(x_j - \beta_i'\right)} \right)^{\nu_i}, \tag{6}$$

where $\gamma_i' = f_1\left(\gamma_i, \nu_i, \lambda_i\right)$, $\beta_i' = f_2\left(\gamma_i, \beta_i, \nu_i, \lambda_i\right)$, and ν_i and λ_i are parameters that control the kurtosis and skewness of the distribution, respectively. We define ν_i, λ_i and their prior distributions as in Eq. 4 and 5, respectively. Again, we can re-parametrize the model following f_1 and f_2, and set the rest of the prior distributions as in the baseline model (see Supplementary Table 2; Code Availability section). Notice that the generalized error distribution (Eq. 4) and the skew normal distribution (Eq. 5) are special cases of Eq. (6) when $\lambda_i = 0$

and $\nu_i = 2$, respectively.

Evaluating the log-likelihood

The log-likelihood values measure the goodness of fit of a statistical model to any data point, for a given sample of the posterior distributions. These values can be used to understand where the models fail to capture the variation of our empirical data. For example, high log-likelihood values indicate those data points that are well captured by a given distribution model, while low values signal those points that are instead unexpected by the model. Therefore, one could use these values to understand what aspects of the shape of distributions that are missing. To do so, for every sample of the model, we computed the log-likelihood values and the normalized probability distribution. This normalized probability is defined such that its maximum is set to 1 for all species in our dataset. In particular, for a heavy-tailed and skewed response, the normalized probability distribution was calculated for every sample of the Bayesian model using Eq. (6), where α_i was set to 0 for any value of x_j. Notice that the normalized probability distribution is interesting when comparing the log-likelihood values across species because it can be used to understand whether the model errors are at the tails of the distributions or their center.

Results

We studied the distribution data to characterize species' realized niches along the main axis of variation of all environmental variables. Using the presence and absence of species across sites as the response variable, we sampled the posterior distributions of the baseline model, accounting for the information in the attribute database regarding species' indicator values and range of variation. This allowed us to map the center and variance of species' distributions along the environmental gradient (Fig. 2). Studying the relationship between these properties, we found these to be negatively correlated (i.e. β_i and γ_i in the baseline model were positively correlated; Fig. 2). This means that species found at the lower end of the gradient (i.e. higher temperature and lower elevation) have generally wider distributions than those found at the higher end (i.e. lower temperature and higher elevation). The same relationship was found when independently using elevation or mean temperature as explanatory variables (Supplementary Fig. 3) as well as when using ordinal data (Supplementary Fig. 4); however, the pattern was not present along the second main axis of variation of our environmental variables (i.e. correlated to precipitation seasonality; Supplementary Fig. 1 and 5). The comparison between the other parameter estimates revealed additional, somewhat more expected, relationships. In particular, we found the amplitude of distributions (i.e. height of the distributions) to be positively and negatively correlated with their mean and variance, respectively (i.e. α_i is positively correlated with β_i and γ_i; Supplementary Fig. 6). This implies that, at higher elevations, species' distributions generally have lower amplitudes (i.e. lower overall probability of occurrence).

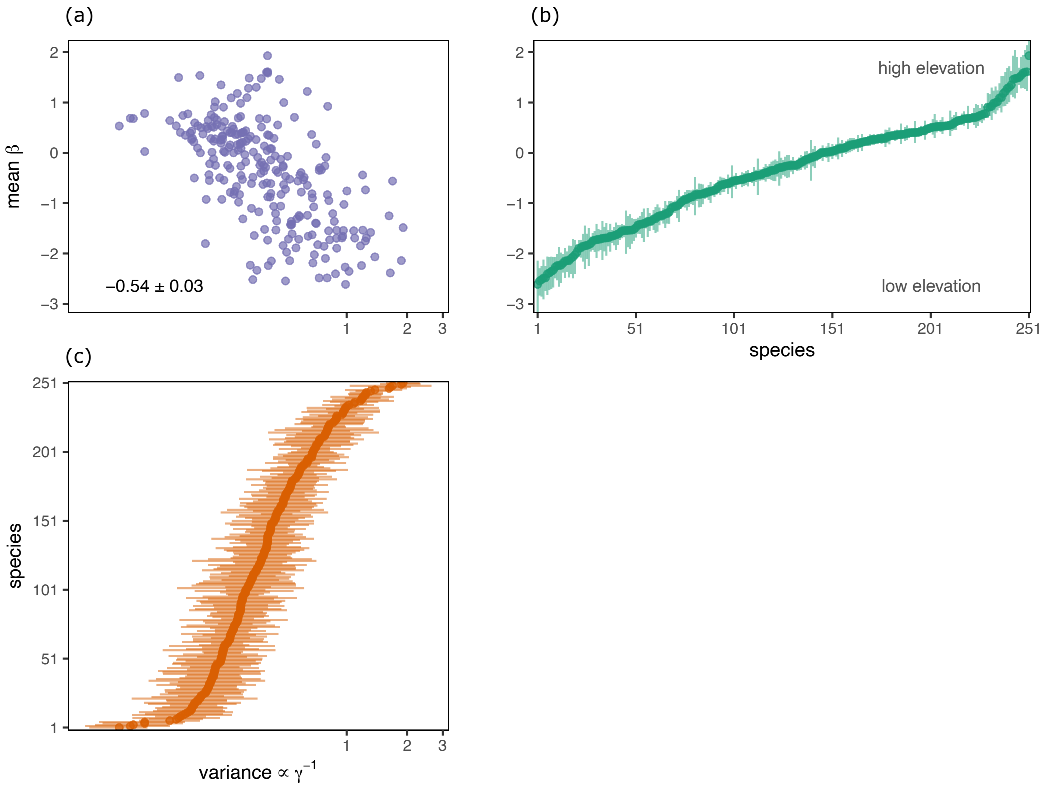

Figure 2: Relationship between the posterior distributions for parameters β_i and γ_i from Eq. (1) across species. Panel (a) describes the relationship between the mean (β_i) and variance ($\propto \gamma_i^{-1}$) of distributions. Each point represents the average value of the corresponding posterior distributions for any given species. The value in the bottom-left corner of the plot displays the Pearson's correlation coefficient between the parameters calculated across all samples. Panel (b) displays the estimates for the center of species' distributions along the environmental gradient. Panel (c) displays the estimates for the variance of species' distributions along the environmental gradient. In (b) and (c), the points represent the mean of the posterior distributions, the corresponding lines characterize the 89% confidence intervals, and species are sorted according to the mean of the posterior distributions along the x and y axes, respectively.

Maintaining the symmetry of species' distributions, we then allowed the kurtosis—or shape of the tails—of these to vary in different ways. To do so, we changed the response curve of our Bayesian model to follow a generalized error distribution (Eq. 4). A comparison of the WAIC values showed this non-linear regression to outperform the baseline model (Supplementary Fig. 7). Studying the resulting posterior distributions, we found the average kurtosis of the distributions to be slightly greater than zero, which corresponds to distributions with longer tails than the Gaussian case (Fig. **3**). However, the parameter controlling for the kurtosis ν_i displayed a lot of variation across species (Supplementary Fig. 8), which might indicate that the shape of the tails is species-specific and potentially explained by species' ecological traits.

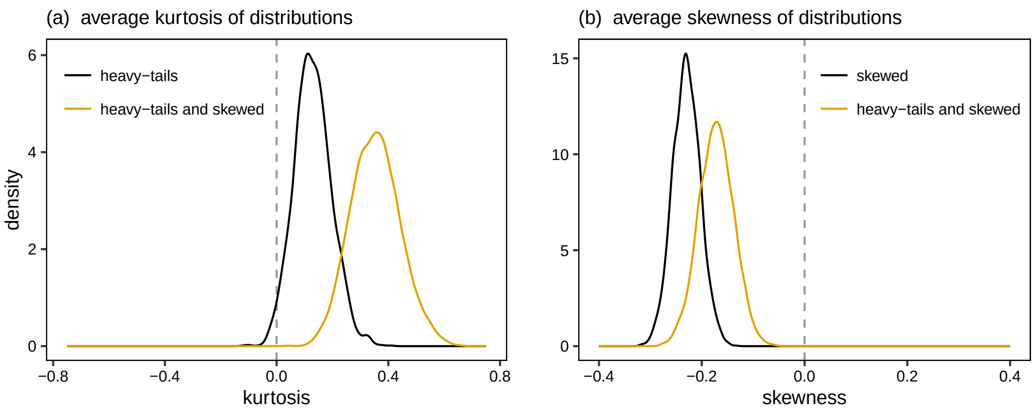

Figure 3: Average kurtosis and skewness of species' distributions. Calculated using the posterior distributions of parameters $\hat{\nu}$ and $\hat{\lambda}$ from the models (see Supplementary Table 2 and Kerman & McDonald 2013), the two panels describe the average (a) kurtosis and (b) skewness of distributions. Panel (a) displays the results obtained using response curves that follow a generalized error distribution (black line) and a skewed generalized error distribution (yellow line). Panel (b) displays the results obtained using response curves that follow a skewed normal distribution (black line) and a skewed generalized error distribution (yellow line). In both cases, the gray dotted line indicates the conditions by which species are normally distributed along the environmental axis.

Using Eq. (5), we next studied the skewness of species' distributions. Based on the estimates for the WAIC values, this model outperformed the first two (Supplementary Fig. 7), which sheds light on the naturally skewed nature of species' distributions. Perhaps most importantly, studying the mean value of the skewness across species, we found this to be consistently below zero (Fig. **3**). This indicates that species' distributions generally present steeper declines towards higher elevations (i.e. $\hat{\lambda} < 0$; Fig. **1**). The same was true when using a model that allowed for both fat-tailed and skewed response curves (Eq. 6). This model outperformed the rest, presenting Akaike weights close to 1 (Supplementary Fig. 7), suggesting that both the kurtosis and skewness are useful properties to describe empirical distributions (Fig. 3). Notice that these properties are also not fully independent, as the consideration of both fat-tails and skewed responses substantially change the results obtained using the previous two models (Eq. 4 and 5; Fig. 3). A study of how the prior knowledge we had regarding species' environmental preferences and range of variation informed the different parameters of the model is presented in the Supplementary Note 1 and Supplementary Fig. 9.

The model characterizing fat-tailed and skewed distributions allowed us to study the posterior distributions for the parameters describing the mean, variance, amplitude, kurtosis and skewness of species' realized niches simultaneously. We observed that different types of species seem to present characteristically different distributions (Fig. 4). Focussing on the negative correlation between the mean and variance of species' distributions, we found some species to escape such macroecological constraints (Supplementary Fig. 10). Moreover, recent and historical range expanders are often found at the lower end of the environmental axis (i.e. higher temperature and lower elevation), presenting higher amplitudes, and distributions that appear to showcase steeper declines towards lower elevations (Fig. 4)—potentially showing these species to spread towards the higher elevations. Notice the nature of these results does not depend on the presence or absence of a species at the edge of the sampling area, as the same model produced comparable results when using simulated and bootstrapped data (Supplementary Note 2 and Supplementary Fig. 11). Moreover, these results did not substantially change when using ordinal data (Supplementary Fig. 12).

Finally, we wanted to identify what aspects of the shape of distributions were still missing. We used the computed log-likelihood values and normalized probabilities to understand where our best performing model fails to capture the variation in empirical plant distributions. We found most data points to be located at the tails of distributions (normalized probability ≈ 0) and to present high log-likelihood values (Supplementary Fig. 13). This is not surprising as the study area spans an extensive elevation gradient, and species' distributions are generally narrow relative to it; therefore, the model accurately predicts that species are usually absent in those sampling sites that fall relatively far from the center of their distributions. However, studying instead only those points for which the model did not perform well (with

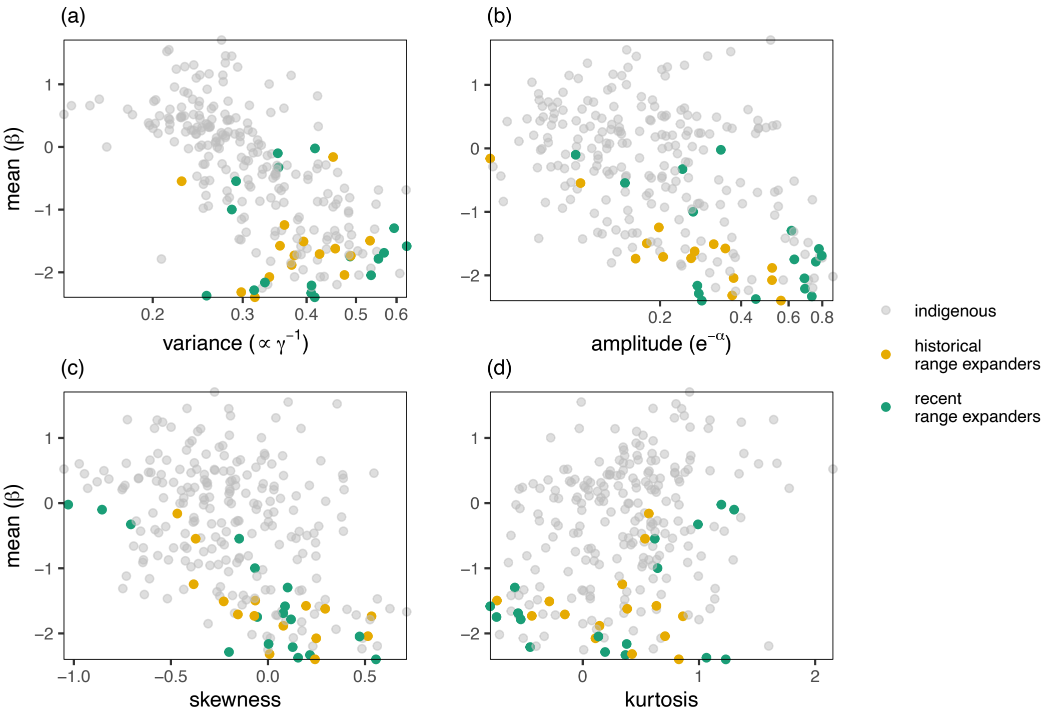

Figure 4: Comparing the distributions of different types of species. Focussing on the differences between indigenous, historical range expanding and recent range expanding species, the panels describe the relationship between the basic properties of their distributions. Panels (a-d) characterize the relationship between the mean, and the variance, amplitude, skewness and kurtosis of the species' distributions (Supplementary Table 2 and Kerman & McDonald 2013). The points in every panel are calculated as the average value across all samples of the model.

a likelihood ≤ 0.5), we found these to generally be associated with high normalized probabilities (Fig. 5). This indicates that the unexplained variation is often found at the center of species' distributions. Similar results were found when using ordinal data (Supplementary Fig. 14).

Discussion

In this work, we used non-linear response curves to model the distribution of species across an environmental gradient. First, we used a baseline model that considered these as bell-shaped, and we studied the relationship between the basic parameters characterizing them. We found both the amplitude and variance of distributions to be negatively correlated with elevation. Considering more complex response curves, we then found species' distributions to also present non-normal tails and skewed shapes. Specifically, we found species' distribu-

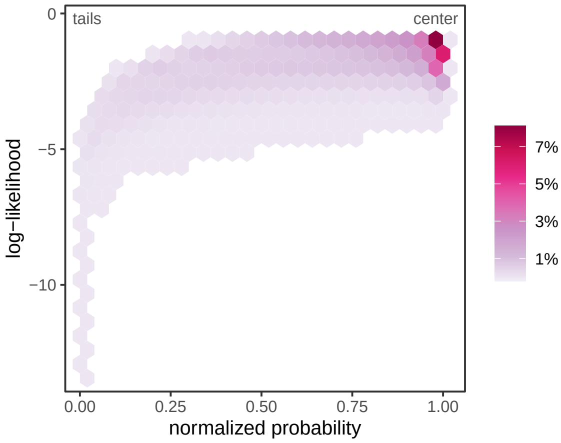

Figure 5: Studying the distribution of log-likelihood values. The graph maps the log-likelihood and normalized probability values for all species across all samples. The colour characterizes the overall percentage of points falling within a given hexagon. Notice that in this figure there are only displayed those points that present a likelihood smaller than 0.5. The mapping of all log-likelihood values is presented in Supplementary Fig. 13.

tions to generally be characterized by fat tails and steeper declines towards higher elevations. That said, the nature of these distributions was not homogeneous across species, as some of them presented singularly different properties. This is the case of rapid and historical range expanders, often found in warmer environments, with distributions presenting higher amplitudes and skewed responses towards high altitudes. Finally, we studied the variation that remained unexplained by the best performing model. We found this unexplained variation to be generally located at the center of distributions, which identified potential general properties of empirical distributions that were missed by our model. Putting this all together, our results uncovered several aspects of the shape of empirical plant distributions and revealed crucial differences between the way species are assembled along environmental gradients.

Our approach allowed us to parsimoniously compare the shape of the species' realized niches along an altitude gradient, testing for the existence of several macroecological patterns. For example, the Rapoport's rule predicts wider ranges of species at higher latitudes and altitudes (Stevens, 1992); and therefore, one might expect a positive correlation between the mean and variance of species distributions. A common explanation for the Rapoport's rule is that climatic variability selects for species with greater climatic tolerances. But while this pattern has been largely studied for multiple systems and across gradients (McCain & Knight, 2013), contrasting evidence suggests that this rule is not pervasive across species (Ribas & Schoereder, 2006; Bhattarai & Vetaas, 2006; McCain & Knight, 2013).

Our results seem to contradict the predictions of the Rapoport's rule, as we observed a negative correlation between species' range width and elevation. Moreover, other properties of species' distributions—such as their amplitude—were also significantly correlated with species' ranges. This is interesting because it hints at the existence of some general macroecological constraints that dictate the way different species assemble across environments. That said, our results also suggest that species such as neophytes and archeophytes might not obey this same constraints, as these were singularly positioned along the gradient (Supplementary Fig. 10).

The level of skewness of species' distribution as well as the variability in the shape of theirs tails diverged from traditionally assumed bell-shaped curves. This allowed us to focus on other interesting macroecological hypotheses. For instance, the so-called abiotic stress limitation hypothesis predicts species' distributions to present steeper declines towards stressful conditions (Austin, 1990). Normand *et al.* (2009) tested this for vegetation data using Huisman *et al.*'s statistical models for several independent species, finding no clear support for such a hypothesis (but see Ziffer-Berger *et al.* 2014). Our results, however, showcased species' distributions to generally present steeper declines towards higher elevations, providing clear evidence of this geographical pattern. Moreover, we were able to highlight the degree to which different species might present different levels of decline towards stressful conditions, as plants found at low elevations—such as recent and historical range expanding species—displayed contrasting levels of skewness. This is important because it could provide glimpses of the different stages of species' assembly processes, with range expanders' distributions trending towards higher elevations.

There are many other properties characterizing empirical distributions that might not have been captured by the different models. One possible way to untangle these properties is by studying the unexplained variation in the empirical data. We observed that this variation is often located at the center of distributions, which suggests that the aspects of their shape not picked up by the models involve those points at the peak of the distributions. This observation is directly linked to another macroecological pattern: the so called abundant-center hypothesis (Sagarin & Gaines, 2002). This hypothesis predicts species to be most abundant at the center of their distributions, and it is an implicit assumption at the core of most modelling approaches. Namely, if one is only willing to assume that species have finite geographic ranges, the abundant-center hypothesis is a consequence of our state of ignorance (i.e. the maximum entropy distribution). That said, several studies have pointed out that the abundant-center hypothesis is not pervasive in empirical distributions (Wagner *et al.*, 2011; Pironon *et al.*, 2017; Dallas *et al.*, 2017), suggesting that population abundance could often be more strongly driven by interactions and community structure than the environment (Dallas *et al.*, 2017). Our results, for both binary and ordinal data, support these observations, suggesting that the species' probability of appearance—as well as likelihood of presenting high abundance at a given site—might not ubiquitously be highest at the center of their distributions. Allowing species to showcase other distribution shapes, such as those including multimodal or plateau peaks, could potentially resolve some of the unexplained variation. Indeed, studying the tails of species' distributions, we observed several species presenting low kurtosis levels. While this implied that these distributions had shorter tails than normal, it also reflected plateau-shaped response curves (e.g. $\nu > 2$ in Fig. **1**b).

The different hypotheses regarding the shape of species' distributions address central topics in ecology and evolution (Sagarin & Gaines, 2002). These distributions are the result of environmental variability (Helmuth *et al.*, 2002; Butterfield, 2015), biotic interactions (Hastings *et al.*, 1997) and historical contingencies (Frick *et al.*, 2010), and their shape determines gene flow (Haldane & Ford, 1956; Lesica & Allendorf, 1995; Pironon *et al.*, 2017) and energy balances along gradients (Hall *et al.*, 1992). Perhaps most importantly, the shape of species' distributions will influence their responses to environmental changes (Channell & Lomolino, 2000a), and it could therefore be used as an ecological compass to inform conservation and management decisions (Channell & Lomolino, 2000b). In this context, we identify two areas we feel represent key steps from which to move forward. First, trait data could crucially inform the different parameters controlling the shape of distributions. For example, if the skewness of species' distributions is the result of uneven environmental tolerances along the gradient (Sunday *et al.*, 2011), this information should be accounted for analogously to the way we used the expert knowledge on plants' environmental preferences. The same is true for species' ecological strategies, with aspects regarding their competitive ability potentially informing the shape of distributions. Second, from a performance standpoint, the models presented here will likely do a worse job at predicting species' occurrences than some of the distribution models developed over the recent years (Norberg *et al.*, 2019), including those accounting for spatial autocorrelation (Ovaskainen *et al.*, 2016), associations between species (Tikhonov *et al.*, 2020), and some non-parametric approximations (Harris, 2015). However, our models have clear interpretable parameters, and can be used to directly compare the shape of species' realized niches. These comparisons could be used to generate hypotheses regarding where and when different species might strongly interact with one another along an environmental gradient (Louthan *et al.*, 2015), making ecologically-informed predictions regarding the presence and absence of these relationships (Callaway *et al.*, 2002; He *et al.*, 2013).

References

Ashour, S. K. & Abdel-hameed, M. A. (2010). Approximate skew normal distribution. *Journal of Advanced Research*, 1, 341–350.

Austin, M. P. (1987). Models for the analysis of species' response to environmental gradients. *Vegetatio*, 69, 35–45.

Austin, M. P. (1990). Community theory and competition in vegetation. *Community theory and competition in vegetation.*, 215–238.

Austin, M. P. (2002). Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling*, 157, 101–118.

Bhattarai, K. R. & Vetaas, O. R. (2006). Can Rapoport's rule explain tree species richness along the Himalayan elevation gradient, Nepal? *Diversity and Distributions*, 12, 373–378.

Butterfield, B. J. (2015). Environmental filtering increases in intensity at both ends of climatic gradients, though driven by different factors, across woody vegetation types of the southwest USA. *Oikos*, 124, 1374–1382.

Callaway, R. M., Brooker, R. W., Choler, P., Kikvidze, Z., Lortie, C. J., Michalet, R., Paolini, L., Pugnaire, F. I., Newingham, B., Aschehoug, E. T., Armas, C., Kikodze, D. & Cook, B. J. (2002). Positive interactions among alpine plants increase with stress. *Nature*, 417, 844–848.

Channell, R. & Lomolino, M. V. (2000a). Dynamic biogeography and conservation of endangered species. *Nature*, 403, 84–86.

Channell, R. & Lomolino, M. V. (2000b). Trajectories to extinction: Spatial dynamics of the contraction of geographical ranges. *Journal of Biogeography*, 27, 169–179.

Dallas, T., Decker, R. R. & Hastings, A. (2017). Species are not most abundant in the centre of their geographic range or climatic niche. *Ecology Letters*, 20, 1526–1533.

D'Amen, M., Rahbek, C., Zimmermann, N. E. & Guisan, A. (2017). Spatial predictions at the community level: From current approaches to future frameworks. *Biological Reviews*, 92, 169–187.

Evans, M. E. K., Merow, C., Record, S., McMahon, S. M. & Enquist, B. J. (2016). Towards Process-based Range Modeling of Many Species. *Trends in Ecology & Evolution*, 31, 860–871.

Frank, S. A. (2009). The Common Patterns of Nature. *Journal of evolutionary biology*, 22, 1563–1585.

Frick, W. F., Pollock, J. F., Hicks, A. C., Langwig, K. E., Reynolds, D. S., Turner, G. G., Butchkoski, C. M. & Kunz, T. H. (2010). An Emerging Disease Causes Regional Population Collapse of a Common North American Bat Species. *Science*, 329, 679–682.

Godoy-Lorite, A., Guimerà, R., Moore, C. & Sales-Pardo, M. (2016). Accurate and scalable social recommendation using mixed-membership stochastic block models. *Proceedings of the National Academy of Sciences*, 113, 14207–14212.

Guisan, A. & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135, 147–186.

Haldane, J. B. S. & Ford, E. B. (1956). The relation between density regulation and natural selection. *Proceedings of the Royal Society of London. Series B - Biological Sciences*, 145, 306–308.

Hall, C. A. S., Stanford, J. A. & Hauer, F. R. (1992). The Distribution and Abundance of Organisms as a Consequence of Energy Balances along Multiple Environmental Gradients. *Oikos*, 65, 377–390.

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, 6, 465–473.

Hastings, A., Harrison, S. & McCann, K. (1997). Unexpected spatial patterns in an insect outbreak match a predator diffusion model. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264, 1837–1840.

He, Q., Bertness, M. D. & Altieri, A. H. (2013). Global shifts towards positive species interactions with increasing environmental stress. *Ecology Letters*, 16, 695–706.

Helmuth, B., Harley, C. D. G., Halpin, P. M., O'Donnell, M., Hofmann, G. E. & Blanchette, C. A. (2002). Climate Change and Latitudinal Patterns of Intertidal Thermal Stress. *Science*, 298, 1015–1017.

Helmuth, B., Kingsolver, J. G. & Carrington, E. (2004). Biophysics, Physiological Ecology, and Climate Change: Does Mechanism Matter? *Annual Review of Physiology*, 67, 177–201.

Huisman, J., Olff, H. & Fresco, L. F. M. (1993). A hierarchical set of models for species response analysis. *Journal of Vegetation Science*, 4, 37–46.

Jamil, T. & ter Braak, C. J. F. (2013). Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, 1, e95.

Kerman, S. C. & McDonald, J. B. (2013). Skewness–kurtosis bounds for the skewed generalized T and related distributions. *Statistics & Probability Letters*, 83, 2129–2134.

Krebs, C. J. (1972). *Ecology: The Experimental Analysis of Distribution and Abundance/by Charles J. Krebs.* 4th edn. Harper & Row, New York.

Landolt, E., Bäumler, B., Ehrhardt, A., Hegg, O., Klötzli, F., Lämmler, W., Nobis, M., Rudmann-Maurer, K., Schweingruber, F. H., Theurillat, J.-P., Urmi, E., Vust, M. & Wohlgemuth, T. (2010). *Flora indicativa: Okologische Zeigerwerte und biologische Kennzeichen zur Flora der Schweiz und der Alpen.* Haupt, Bern. ISBN 978-3-258-07461-0.

Lesica, P. & Allendorf, F. W. (1995). When Are Peripheral Populations Valuable for Conservation? *Conservation Biology*, 9, 753–760.

Linder, E. T., Villard, M.-A., Maurer, B. A. & Schmidt, E. V. (2000). Geographic range structure in North American landbrids: Variation with migratory strategy, trophic level, and breeding habitat. *Ecography*, 23, 678–686.

Louthan, A. M., Doak, D. F. & Angert, A. L. (2015). Where and When do Species Interactions Set Range Limits? *Trends in Ecology & Evolution*, 30, 780–792.

Malmgren, R. D., Stouffer, D. B., Motter, A. E. & Amaral, L. A. N. (2008). A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105, 18153–18158.

McCain, C. M. & Knight, K. B. (2013). Elevational Rapoport's rule is not pervasive on mountains. *Global Ecology and Biogeography*, 22, 750–759.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC press.

Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*, 32, 685–694.

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., Husby, M., Kålås, J. A., Lehikoinen, A., Luoto, M., Mod, H. K., Newell, G., Renner, I., Roslin, T., Soininen, J., Thuiller, W., Vanhatalo, J., Warton, D., White, M., Zimmermann, N. E., Gravel, D. & Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89, e01370.

Normand, S., Treier, U. A., Randin, C., Vittoz, P., Guisan, A. & Svenning, J.-C. (2009). Importance of abiotic stress as a range-limit determinant for European plants: Insights from species responses to climatic gradients. *Global Ecology and Biogeography*, 18, 437–449.

Ovaskainen, O., Roy, D. B., Fox, R. & Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, 7, 428–436.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T. & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20, 561–576.

Petrovskii, S., Morozov, A., Taylor, A. E. P. D. & DeAngelis, E. D. L. (2009). Dispersal in a Statistically Structured Population: Fat Tails Revisited. *The American Naturalist*, 173, 278–289.

Pironon, S., Papuga, G., Villellas, J., Angert, A. L., García, M. B. & Thompson, J. D. (2017). Geographic variation in genetic and demographic performance: New insights from an old biogeographical paradigm. *Biological Reviews*, 92, 1877–1909.

Randin, C. F., Engler, R., Normand, S., Zappa, M., Zimmermann, N. E., Pearman, P. B., Vittoz, P., Thuiller, W. & Guisan, A. (2009). Climate change and plant distribution: Local models predict high-elevation persistence. *Global Change Biology*, 15, 1557–1569.

Ribas, C. R. & Schoereder, J. H. (2006). Is the Rapoport effect widespread? Null models revisited. *Global Ecology and Biogeography*, 15, 614–624.

Rohde, K. (1992). Latitudinal Gradients in Species Diversity: The Search for the Primary Cause. *Oikos*, 65, 514–527.

Sagarin, R. D. & Gaines, S. D. (2002). The 'abundant centre' distribution: To what extent is it a biogeographical rule? *Ecology Letters*, 5, 137–147.

Sagarin, R. D., Gaines, S. D. & Gaylord, B. (2006). Moving beyond assumptions to understand abundance distributions across the ranges of species. *Trends in Ecology & Evolution*, 21, 524–530.

Scherrer, D. & Guisan, A. (2019). Ecological indicator values reveal missing predictors of species distributions. *Scientific Reports*, 9, 1–8.

Siefert, A., Lesser, M. R. & Fridley, J. D. (2015). How do climate and dispersal traits limit ranges of tree species along latitudinal and elevational gradients? *Global Ecology and Biogeography*, 24, 581–593.

Stan Developent Team (2021). RStan: The R interface to Stan.

Stan Development Team (2021). Stan Modeling Language Users Guide and Reference Manual.

Stevens, G. C. (1992). The Elevational Gradient in Altitudinal Range: An Extension of Rapoport's Latitudinal Rule to Altitude. *The American Naturalist*, 140, 893–911.

Sunday, J. M., Bates, A. E. & Dulvy, N. K. (2011). Global analysis of thermal tolerance and latitude in ectotherms. *Proceedings of the Royal Society B: Biological Sciences*, 278, 1823–1830.

ter Braak, C. J. F. & Looman, C. W. N. (1986). Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio*, 65, 3–11.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M. J., Oksanen, J. & Ovaskainen, O. (2020). Joint species distribution modelling with the r-package Hmsc. *Methods in Ecology and Evolution*, 11, 442–447.

Wagner, V., von Wehrden, H., Wesche, K., Fedulin, A., Sidorova, T. & Hensen, I. (2011). Similar performance in central and range-edge populations of a Eurasian steppe grass under different climate and soil pH regimes. *Ecography*, 34, 498–506.

Wong, F. & Collins, J. J. (2020). Evidence that coronavirus superspreading is fat-tailed. *Proceedings of the National Academy of Sciences*, 117, 29416–29418.

Ziffer-Berger, J., Weisberg, P. J., Cablk, M. E. & Osem, Y. (2014). Spatial patterns provide support for the stress-gradient hypothesis over a range-wide aridity gradient. *Journal of Arid Environments*, 102, 27–33.

Zurell, D., Engler, J. O., Dunn, P. O. & Møller, A. P. (2019). Ecological niche modeling. In: *Effects of Climate Change on Birds*. Oxford University Press, pp. 60–73.