ECOGRAPHY

Research

An evaluation of transferability of ecological niche models

Huijie Qiao, Xiao Feng, Luis E. Escobar, A. Townsend Peterson, Jorge Soberón, Gengping Zhu and Monica Papeş

H. Qiao (https://orcid.org/0000-0002-5345-6234), Key Laboratory of Animal Ecology and Conservation Biology, Inst. of Zoology, Chinese Academy of Sciences, Beijing, P. R. China. – X. Feng (http://orcid.org/0000-0003-4638-3927) (fengxiao@email.arizona.edu), Inst. of the Environment, Univ. of Arizona, Tucson, AZ, USA. – L. E. Escobar, Dept of Fish and Wildlife Conservation, Virginia Tech, Blacksburg, VA, USA. – A. T. Peterson (http://orcid.org/0000-0003-0243-2379), J. Soberón and G. Zhu, Biodiversity Inst., Univ. of Kansas, Lawrence, KS, USA. – GZ also at: College of Life Sciences, Tianjin Normal Univ., Tianjin, P. R. China. – M. Papeş, Dept of Ecology and Evolutionary Biology, Univ. of Tennessee, Knoxville, TN, USA.

Ecography
42: 521–534, 2019
doi: 10.1111/ecog.03986

Subject Editor: Christine Meynard
Editor-in-Chief: Miguel Araújo
Accepted 6 August 2018

Ecological niche modeling (ENM) is used widely to study species' geographic distributions. ENM applications frequently involve transferring models calibrated with environmental data from one region to other regions or times that may include novel environmental conditions. When novel conditions are present, transferability implies extrapolation, whereas, in absence of such conditions, transferability is an interpolation step only. We evaluated transferability of models produced using 11 ENM algorithms from the perspective of interpolation and extrapolation in a virtual species framework. We defined fundamental niches and potential distributions of 16 virtual species distributed across Eurasia. To simulate real situations of incomplete understanding of species' distribution or existing fundamental niche (environmental conditions suitable for the species contained in the study area; N^*_F), we divided Eurasia into six regions and used 1–5 regions for model calibration and the rest for model evaluation. The models produced with the 11 ENM algorithms were evaluated in environmental space, to complement the traditional geographic evaluation of models. None of the algorithms accurately estimated the existing fundamental niche (N^*_F) given one region in calibration, and model evaluation scores decreased as the novelty of the environments in the evaluation regions increased. Thus, we recommend quantifying environmental similarity between calibration and transfer regions prior to model transfer, providing an avenue for assessing uncertainty of model transferability. Different algorithms had different sensitivity to completeness of knowledge of N^*_F, with implications for algorithm selection. If the goal is to reconstruct fundamental niches, users should choose algorithms with limited extrapolation when N^*_F is well known, or choose algorithms with increased extrapolation when N^*_F is poorly known. Our assessment can inform applications of ecological niche modeling transference to anticipate species invasions into novel areas, disease emergence in new regions, and forecasts of species distributions under future climate conditions.

Keywords: extrapolation, interpolation, non-analog environment

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

Ecological niche modeling (ENM) refers to the analysis of relationships between species' distribution and environments; estimating the fundamental niche (N_F) is a typical goal, to permit model transfers to other places and times (Soberón 2007, Peterson and Soberón 2012, Warren 2012). Transferring models involves two processes in environmental space: interpolation and extrapolation (Peterson et al. 2011, Heikkinen et al. 2012). Interpolation involves transfers to environmental conditions among those used to calibrate the model, whereas extrapolation is applying the model to environmental conditions beyond the values used to calibrate the model or to novel combinations of environments (Fitzpatrick and Hargrove 2009, Peterson et al. 2011, Zurell et al. 2012, Owens et al. 2013).

Assessing model performance is fundamental in ENM; researchers have focused on optimizing ENM algorithms, but not necessarily distinguishing between interpolation and extrapolation (but see Escobar et al. 2018). For instance, Elith et al. (2006) evaluated 16 algorithms applied to species from five regions; because calibration and evaluation localities were from the same area, that study investigated interpolative abilities. Peterson et al. (2007) compared two algorithms based on spatial subsets of known distributions, thus potentially including effects of both interpolation and extrapolation. Muscarella et al. (2014) proposed strategies to address model selection in Maxent and presented evaluations of model spatial transferability, potentially including both interpolation and extrapolation.

Although transferring a model may involve both interpolation and extrapolation, extrapolation may be more important in studies striving to make predictions of distributional shifts in the face of global change (Elith and Leathwick 2009), as non-analog conditions are common (Williams and Jackson 2007). However, model extrapolation is statistically challenging, as it forces the algorithm to make predictions for novel environmental conditions (Gelman and Hill 2007), which may often be erroneous (Williams and Jackson 2007, Elith and Leathwick 2009, Fitzpatrick and Hargrove 2009, Owens et al. 2013).

Many algorithms have been developed for ENM. Envelope algorithms (e.g. BIOCLIM; Busby 1991) and ellipsoids (Farber and Kadmon 2003) assume a regular shape of N_F in environmental space, and determine the parameters of that shape based on environmental conditions associated with known presences. Statistical algorithms, such as generalized linear models (GLM; McCullagh and Nelder 1989, Guisan et al. 2002) and generalized additive models (GAM; Hastie and Tibshirani 1990, Guisan et al. 2002), use logistic regression to estimate species' responses (presence/absence) to environmental conditions. Cluster algorithms, such as kernel density estimation (KDE; Blonder et al. 2014) and Marble (MA; Qiao et al. 2015b), estimate niches based on the density or clustering of presences in environmental space. Finally, machine-learning algorithms, such as boosted regression trees (BRT; Elith et al. 2008), maximum entropy (Maxent; Phillips et al. 2004, Elith et al. 2006), and genetic algorithms (GARP; Stockwell 1999), make less restrictive assumptions about niches, and maximize model fit to calibration data. Machine-learning algorithms generally show limited extrapolation abilities; however, extrapolation may be improved by controlling model complexity via internal cross-validation (e.g. BRT; De'ath 2007, Elith et al. 2008) or regularization (e.g. Maxent; Merow et al. 2013).

ENM algorithms also differ in extrapolation strategies, which can be classified broadly into truncation, clamping, and actual extrapolation (Owens et al. 2013). Truncation simply designates all conditions outside of the calibration data range as unsuitable; clamping uses the marginal values in the calibration area as the prediction for more extreme conditions in transfer areas; and actual extrapolation extends the response curve based on trends obtained from calibration conditions or assumptions about the niche.

Previous studies have compared transferability of different ENM algorithms across space (Randin et al. 2006, Duque-Lazo et al. 2016) and time (Roberts and Hamann 2012, Veloz et al. 2012) based on occurrence data of real species; however, real occurrence and/or absence data may not be optimal for assessing model performance given sampling bias (Hortal et al. 2008), limited sample size (Jiménez-Valverde et al. 2009), limited dispersal ability, and complex species interactions (Soberón and Peterson 2005), thus affecting the generality of algorithm comparisons. In contrast, virtual species with known niche properties could provide abundant, controlled occurrence (Moudrý 2015, Leroy et al. 2016, Qiao et al. 2016) and absence data (Feng and Papeş 2017a, Hattab et al. 2017). Therefore, ecological niches from virtual species are more appropriate for algorithm assessment.

We used virtual species to develop detailed evaluations of transferability performance of 11 ENM algorithms. We distinguished interpolation from extrapolation based on overlap between calibration and evaluation conditions in environmental space (Supplementary material Appendix 1 Fig. A1), and refined our evaluations to consider degrees of environmental similarity between evaluation and calibration conditions. We termed evaluation conditions in environmental space as 1) 'overlapping' if evaluation conditions were inside a concave hull estimated from the calibration conditions (Lafarge and Pateiro-Lopez 2016); 2) 'novel' if conditions exceeded the range of calibration conditions (Owens et al. 2013); and 3) 'novel-combination' if evaluation conditions were within the range of calibration conditions, but represented combinations of variables that were absent from the calibration set (Zurell et al. 2012; Supplementary material Appendix 1 Fig. A1). Because ecological niches exist in environmental space and are manifested in geography (Hutchinson 1957), we employed both classic ENM evaluation indices in geographic space and novel, shape-based indices in multivariate environmental space.

Glossary

N_F: fundamental niche (sometimes termed the theoretical fundamental niche), the set of all environmental conditions, whether available or not in geographic space, that would permit a species to maintain populations without immigrational subsidy; the fundamental niche is difficult to estimate in practice (Peterson et al. 2016); in this study, for simplicity, the fundamental niche of a virtual species is defined as a multidimensional sphere.

N^*_F: existing fundamental niche (sometimes termed the potential niche or empirical fundamental niche), the environmental conditions that fall within the fundamental niche (N_F) and that are actually present across the area relevant and accessible to a species; the projection of N^*_F in geography is the species' potential distribution. Compared with the fundamental niche, the existing fundamental niche can be studied empirically much more readily.

N^*_{Fx}: a sample from the existing fundamental niche (N^*_F) from geographic region x used in model calibration.

$\widehat{N}$: the niche estimated from N^*_{Fx} using ecological niche modeling algorithms.

Material and methods

Virtual species in a real landscape

We created 16 virtual species distributed across mainland Eurasia (Figs. 1a, Supplementary material Appendix 1 Fig. A2). We chose Eurasia because it is the largest continuous landmass on Earth and possesses diverse environments. We used an Eckert IV equal-area map projection at 10 km spatial resolution for 19 bioclimatic variables used widely in ENM studies (Hijmans et al. 2005). We used principal components analysis to reduce dimensionality and collinearity among environmental variables and facilitate quantification of environmental overlap. We retained the first three principal components (PC1, PC2, PC3), which together accounted for 82.6% of overall variation.

In the environmental space delineated by these three variables across Eurasia, we defined fundamental niches (N_F) for 16 virtual species as spheres using NicheA ver. 3.0 (Qiao et al. 2016), with eight distinct niche centers and two radii (1.5 and 2.0, corresponding to narrow and wide niches, respectively; Supplementary material Appendix 1 Fig. A2). We assumed that N_F is convex, as suggested by empirical evidence, and used spherical shapes for convenience (Birch 1953, Maguire Jr 1967, Hooper et al. 2008, Angilletta 2009, Soberón and Nakamura 2009). Because not all environmental conditions within N_F may be available across the study area, we denote environmental conditions falling inside N_F as the species' existing fundamental niche (N^*_F), and its projection onto geography as its potential distribution (Fig. 1; Peterson et al. 2011).

Spatial segregation of potential distribution

In real-world situations, only a portion of the species' potential distribution is known and used in model calibration, with consequently incomplete knowledge of the existing fundamental niche, N^*_F. To simulate these scenarios, for each species, we divided the study area into six regions (Fig. 1a), each covering an equal portion of the virtual species' potential geographic distribution (Fig. 1b; similar to the 'block' partitioning method in Muscarella et al. 2014). We termed data characterizing N^*_F for any region x as N^*_{Fx}, which was the information subsequently used in model calibration. We did not aim to make the six regions represent equal portions of N^*_F in environmental space (Fig. 1b), which is only possible in an artificial landscape that would lack generality.

ENM algorithms

We estimated ecological niches using 11 ENM algorithms (Supplementary material Appendix 1 Table A1). We included a) four envelope algorithms, BIOCLIM (Busby 1991), ecological niche factor analysis (ENFA; Hirzel et al. 2002), CONVEXHULL (Guisan and Zimmermann 2000), and minimum-volume ellipsoids (MVE; Van Aelst and Rousseeuw 2009, Qiao et al. 2016); b) two cluster algorithms, kernel density estimation (KDE; Blonder et al. 2014) and Marble (MA; Qiao et al. 2015b); c) two statistical algorithms, generalized linear models (GLM; McCullagh and Nelder 1989, Guisan et al. 2002) and generalized additive models (GAM; Hastie and Tibshirani 1990, Guisan et al. 2002); and d) three machine-learning algorithms, boosted regression trees (BRT; Elith et al. 2008), Maxent (Phillips et al. 2004), and GARP (Stockwell 1999). Modeling algorithms can also be distinguished by data input needs (Peterson et al. 2011), i.e., presence-only (BIOCLIM, ENFA, CONVEXHULL, MVE, KDE and MA), presence and background (GARP and Maxent), and presence and absence (GLM, GAM, and BRT; Supplementary material Appendix 1 Table A1). We followed default or commonly used settings for each algorithm (see detailed parameterizations in Supplementary material Appendix 1 Table A1). We did not aim to tune each algorithm to an optimal performance, because our goal was to evaluate model performance under commonly adopted parameters, thus replicating practical use by the community. Full details of algorithms, parameter settings, and acronyms are in Supplementary material Appendix 1 Table A1.

Experiment 1 – Modeling existing fundamental niches (N^*_F) from incomplete knowledge

Experimental design

Within each of the six calibration regions (Fig. 1a) we randomly selected 10% of suitable pixels as calibration presences (i.e. N^*_{Fx}) and 10 000 pixels (from across the whole study area) as background data for algorithms that need background input and as pseudo-absences for algorithms that need absence input; this source of absence information is thus comparable to real-world ENM applications, as true absence data are usually not available (Mackenzie 2005, Elith and Leathwick 2007, Barbet-Massin et al. 2012). To eliminate influence of inadequate sampling, we excluded models with < 10 000 background pixels available. In this experiment, we used each region once to calibrate models and the remaining five regions to evaluate predictions.

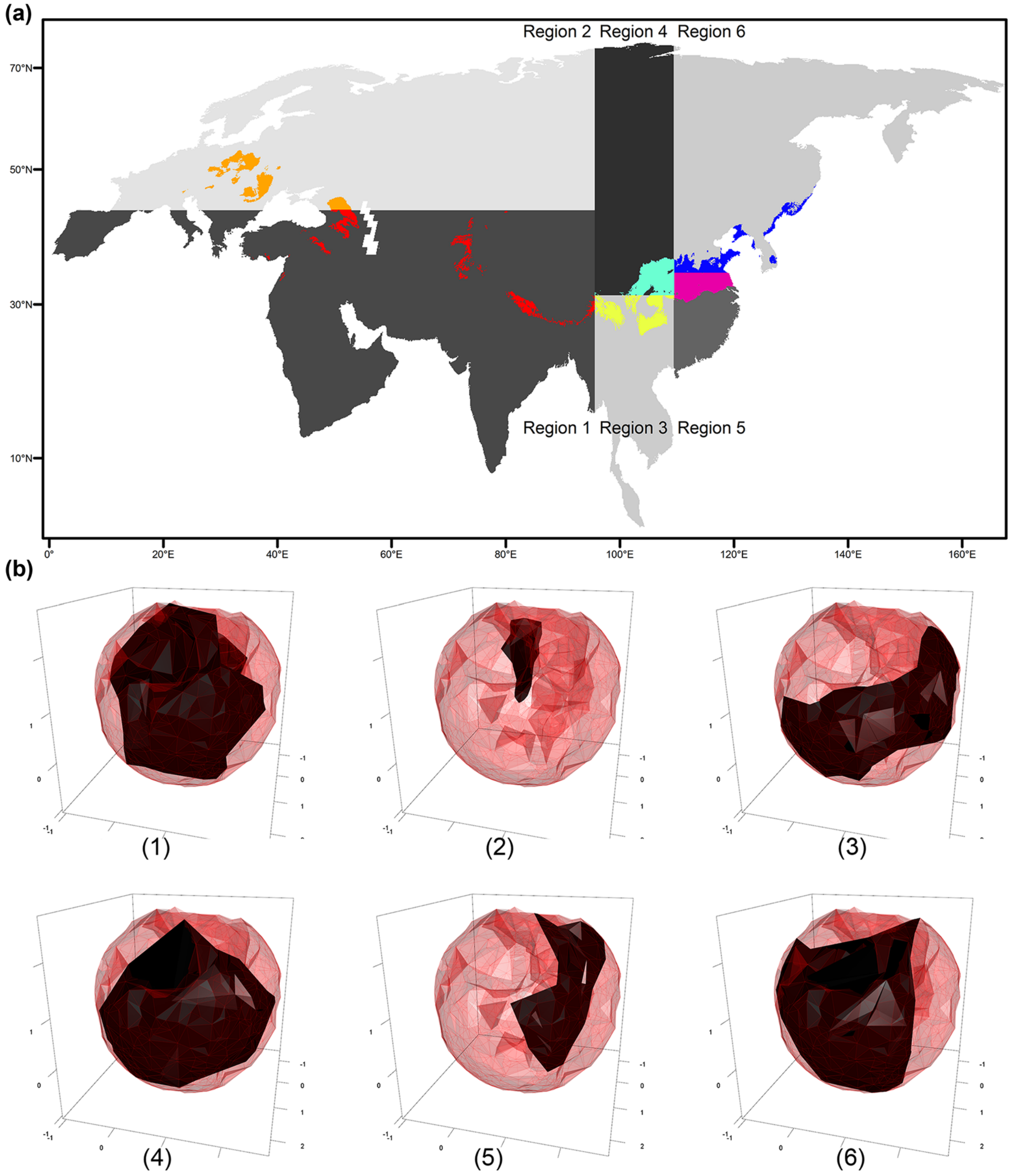

Figure 1. An example of one virtual species used in this study. The virtual species is displayed in (a) geographic space and (b) environmental space (axes are the first three principal components from the climate dataset). In panel (a), the various colors represent a virtual species' potential distribution and the shades of gray represent unsuitable areas. We divided the study area into six regions (1–6) containing equal portions of the virtual species' potential distribution. In panel (b), the species' potential distribution within each region corresponds to different portions of the existing fundamental niche (N^*_F; red 3D hull), with the sampled existing fundamental niche (N^*_{Fx}) represented by the black 3D hull.

Models were transferred to all of Eurasia, and raw model outputs were transformed to binary (suitable and unsuitable) maps. Given that all calibration presences were true presences of our virtual species, we used the minimum predicted value for calibration presences (least presence threshold; Pearson et al. 2007) as a threshold to generate binary predictions (Peterson et al. 2011). We then explored the predicted distribution in environmental space to obtain the estimated niche ($\hat{N}$). The binary models were used to evaluate algorithms in environmental and geographic spaces (see below).

Model evaluation in environmental space

To characterize performance of algorithms in estimating niches in environmental space, we developed an evaluation metric in a three-dimensional principal component space. All pixels of Eurasia were represented as points in this space. We delineated the 3D space of sampled existing fundamental niche (N^*_{Fx}) (Fig. 1b) and estimated niche ($\widehat{N}$; Supplementary material Appendix 1 Fig. A3), using the three-dimensional concave hull based on the 3D alpha-hull method (Lafarge and Pateiro-Lopez 2016), which fits a hull around occurrences based on a parameter α. We simulated a range of α (0.1 to 100, with 0.1 increments), and selected the smallest α that achieved a continuous volume for a target cloud (similar to Capinha et al. 2014), therefore representing the most conservative estimate of the hull.

To calculate the volume of each 3D hull, we split each axis (i.e. PC1, PC2 and PC3) into 100 segments, generating 10^6 cubes in environmental space. The number of cubes inside a 3D hull was used to estimate hull volume and volume overlap between any pairs of 3D hulls. We calculated volume ratio and Jaccard similarity coefficient (Jaccard 1912) using either the N^*_F or N^*_{Fx} as a reference and using N^*_F, N^*_{Fx}, and $\widehat{N}$ as targets, as follows. Volume ratio was calculated by dividing the volume of the reference hull by the volume of the target hull. A Jaccard similarity coefficient was used to measure similarity of pairs of 3D hulls, calculated as:

$$\frac{Volume\left(X\bigcap Y\right)}{Volume\left(X\bigcup Y\right)} \qquad (1)$$

where X and Y are reference and target hulls, respectively and $\cap$ and $\cup$ denote the intercept and union of the 3D hulls. Using either the existing fundamental niche (N^*_F) or sample of the existing fundamental niche (N^*_{Fx}) as a reference helped set the context of model evaluation. We calculated these indices for each species (16), calibration region (6) and algorithm (11), and calculated means across all species and calibration regions for each algorithm. We also evaluated means across species with wide or narrow niches.

Model evaluation in geographic space

The model performance evaluation in geographic space was based on presences and absences from the evaluation regions (regions not used in model calibration). We used three metrics: sensitivity, specificity (Fielding and Bell 1997), and true skill statistic (TSS; Allouche et al. 2006), at a minimum training presence threshold (Pearson et al. 2007). Sensitivity (true positive rate) measures proportion of presences correctly identified. Specificity (true negative rate) measures proportion of absences correctly identified. TSS accounts for both ($TSS = sensitivity + specificity - 1$), and ranges from –1 to 1, with values >0 indicating models better than random (Allouche et al. 2006).

Based on similarity between environments in calibration and evaluation regions (see below), we classified evaluation points into three categories: overlapping, novel-combination, and novel (Supplementary material Appendix 1 Fig. A1). To obtain the three categories in each model calibration, we built a 3D hull (with the same methods described above) and a cube in the environmental space. We built the cube with length, width, and height corresponding to the range of environmental variables (PC1, PC2, or PC3) of each calibration dataset (presences and background pixels). The category of similarity between environments in the evaluation region and calibration region was determined using the 3D hull and the cube from the calibration dataset: overlapping (conditions inside the 3D hull), novel-combination (outside the 3D hull but inside the cube), and novel (extending outside the cube and the 3D hull; Supplementary material Appendix 1 Fig. A1). Here, novel has the same meaning as 'novel' in two other published methods accounting for environmental dissimilarity (Elith et al. 2011, Owens et al. 2013). The novel-combination group is theoretically similar to ideas from Zurell et al. (2012), although our 3D hull approach will result in a finer boundary of the calibration environments. Compared with other studies that quantify environmental novelty in a continuous manner (e.g. Euclidean distance or Mahalanobis distance; Fitzpatrick et al. 2018), our approach to defining environmental novelty is categorical; we note that extrapolation is defined in statistics as making predictions outside training data, so environmental distance or similarity may not directly help distinguish extrapolation from interpolation; further, even in the case of extrapolation, we may not discriminate between less vs more novel conditions, because both scenarios are extrapolation by definition. This separation of evaluation datasets allowed us to measure algorithm performance along a gradient of novelty of environmental conditions. Conditions classified as overlapping corresponded to interpolation, whereas conditions classified as novel-combination and novel offered two levels of extrapolation. Including all evaluation data corresponded to an overall transferability situation, combining model interpolation and extrapolation. We calculated evaluation metrics for each species (16), calibration region (6), algorithm (11), and category of evaluation data (3), and calculated means across species (all species, or species with wide niches, or species with narrow niches) and calibrating regions for each algorithm and category of evaluation data. Models with < 20 evaluation presences or with extremely unbalanced presences and absences (ratio < 0.001) were excluded from the analysis.

Model evaluation via response curves

Essentially, a calibrated model portrays species' response to environmental gradients, so response curves offer a powerful way to visualize models (Elith and Graham 2009, Owens et al. 2013). Here, our virtual species were based on threshold responses (Meynard and Kaplan 2013), which lay a simple baseline for our algorithm comparisons. Further, we used the environmental range of training data as a reference system, which allowed us to compare interpolation and extrapolation among ENM algorithms. To compare shapes of response curves of algorithms across environmental conditions, we projected models onto a series of simulated conditions (similar as evaluation strip in Elith et al. 2005). Briefly, we examined shapes of model response curves according to PC1 with values ranging from –10 to 10, keeping PC2 and PC3 constant (i.e. $PC2 = 0$ and $PC3 = 0$). To facilitate visualization of results, we normalized raw predictions of all models to a uniform scale (0–1) with a simple rescaling method:

$$p' = \frac{p - \min(p)}{\max(p) - \min(p)} \tag{2}$$

where p is the original predicted value from one model and $p´$ is the normalized value. We applied this assessment with respect to PC2 and PC3 using similar methods.

Experiment 2: sensitivity of algorithms to environmental representations of existing fundamental niche (N^*_F)

Experimental design and model evaluation

To extend experiment 1, which consisted of calibrating models with a single region, we calibrated models across 1–5 regions to simulate gradients of completeness of knowledge of the existing fundamental niche (N^*_F). We included all possible combinations (62) of 1–5 regions for model calibration. To ensure equal presence contribution from each region, models were calibrated using calibration datasets comprising 10% random presence samples from the species' distribution in each region. We also randomly selected 10 000 background points within the calibration region(s) for algorithms needing background or absences for calibration. To reduce influence of inadequate absence input, we excluded models with < 10 000 background pixels available. We evaluated model performance by calculating TSS, sensitivity, and specificity, based on all presences and absences, as a representation of the whole existing fundamental niche, N^*_F. Data preparation, model training and evaluation, and data analyses were conducted in R (ver. 3.1.2; Supplementary material Appendix 1 Table A1).

Data deposition

Bioclimatic data: available from the WorldClim database at <www.worldclim.org/current>. Other relevant data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.kg3d57r> (Qiao et al. 2018).

Results

Experiment 1

Model evaluation in environmental space

The mean volumes of estimated niches obtained with all algorithms were larger than the sampled existing fundamental niche (Fig. 2a–b, 3, Supplementary material Appendix 1 Fig. A3), reflecting some degree of inference and infilling of niche estimates. The mean volumes of estimated niches from BRT, ENFA, GAM, and Maxent models were also larger than mean volumes of the existing fundamental niche, thus overestimating the existing fundamental niche. Among overestimating algorithms, estimated niches from Maxent (Fig. 3), followed by ENFA (Supplementary material Appendix 1 Fig. A3), were most dissimilar from the existing fundamental niche in terms of Jaccard similarity. Estimated niches from GLM were most similar to the existing fundamental niche in terms of volume and Jaccard similarity coefficients (Fig. 2a). In contrast, mean volumes of estimated niches from BIOCLIM, CONVEXHULL, KDE, MA, and MVE were smaller than the existing fundamental niche, thus underestimating the existing fundamental niche. Estimated niches from CONVEXHULL, KDE, and MA were most similar to the sampled existing fundamental niche (Fig. 2b). The pattern

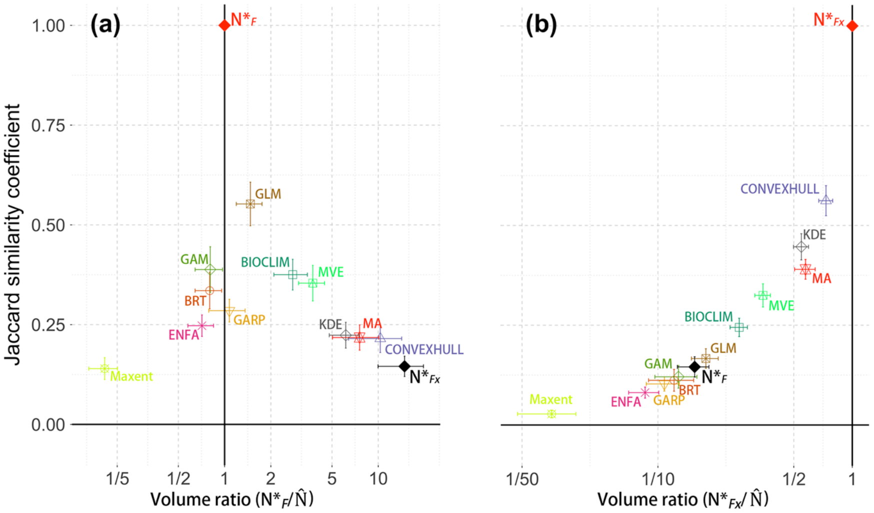

Figure 2. Volume ratio and Jaccard similarity between 3D hulls of the existing fundamental niche (N^*_F), sampled existing fundamental niche (N^*_{Fx}), and estimated niche ($\widehat{N}$) in experiment 1. In panel (a), the existing fundamental niche (N^*_F) is used as a reference for the sampled existing fundamental niche (N^*_{Fx}) and estimated niche ($\widehat{N}$) when calculating the volume ratio and Jaccard similarity. In panel (b), the sampled existing fundamental niche (N^*_{Fx}) is used as a reference. Volume ratio represents the geometric volume ratio between the reference 3D hull and target 3D hull, and the Jaccard similarity coefficient measures the geometric similarity between the target 3D hull and reference 3D hull. The error bars represent the 95% confidence intervals.

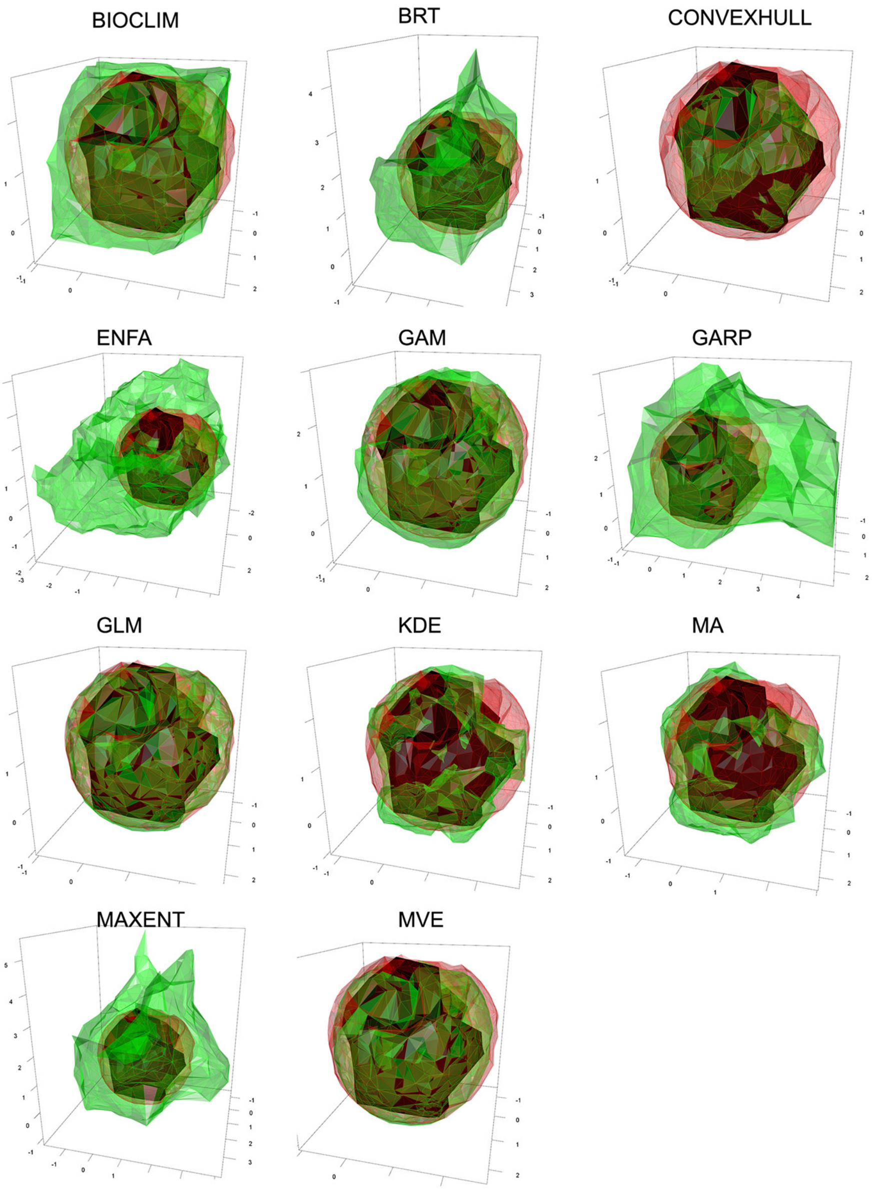

Figure 3. Example of model extrapolation in environmental space. Representations in environmental space of a broad existing fundamental niche (N^*_F; red hull), the portion of N^*_F used in model calibration (N^*_{Fx}; black hull), and the estimated niche ($\widehat{N}$; green hull). The study area (Eurasia) is divided into six geographic regions, each containing different portions of N^*_F, represented by different N^*_{Fx} in panels 1 to 6.

was consistent for different niche breadths (Supplementary material Appendix 1 Fig. A4).

Model evaluation in geographic space

All algorithms performed well when calibration data were overlapping (Fig. 4a); however, as environmental novelty increased, TSS values decreased for all algorithms (Fig. 4a). Different algorithms had different sensitivity to inclusion of novel environments. For example, TSS of models obtained with BIOCLIM, CONVEXHULL, GARP, KDE, MA, and MVE decreased rapidly as environmental novelty changed from overlapping to combinational-novel, approaching zero predictive ability under novel environmental conditions. The remaining algorithms produced models with decreased TSS, but predictive ability remained better than random even under novel environmental conditions (Fig. 4a).

As evaluation conditions became increasingly novel, BRT, GAM, and Maxent models had steady decreased sensitivity and specificity (Fig. 4), whereas BIOCLIM, CONVEXHULL, GARP, KDE, MA, and MVE models had sharp decreases in sensitivity but little reduction of specificity (Fig. 4b, c). ENFA and GLM showed reduced sensitivity when evaluation conditions became increasingly novel; however, specificity of ENFA had an increasing trend while that of GLM was generally invariant. Under novel environmental conditions, Maxent models generally had highest sensitivity

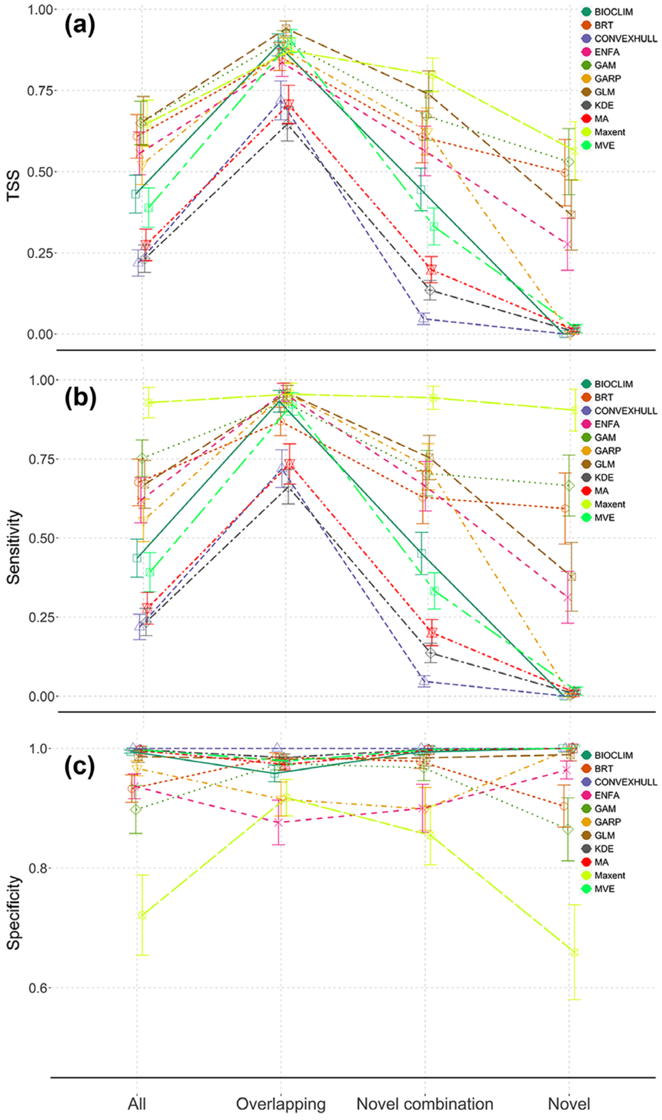

Figure 4. Model evaluation results for Experiment 1. The values of (a) true skills statistic (TSS), (b) sensitivity, and (c) specificity are differentiated by the category of data used for model evaluation. 'All' represents all evaluation data, 'overlapping' represents evaluation data within the 3D hull of the calibration data, 'novel-combinations' represents evaluation data outside the 3D hull but within the range of the calibration data, and 'novel' represents evaluation data outside the range of the calibration data. The error bars represent the 95% confidence intervals.

but lowest specificity, showing a general tendency to overestimate the species range (Fig. 4b, c). The pattern was consistent for different niche breadths (Supplementary material Appendix 1 Fig. A5).

Model evaluation via response curves

Patterns were similar along all three axes, so we only show PC1 (Fig. 5, Supplementary material Appendix 1 Fig. A6). Three envelope algorithms (BIOCLIM, CONVEXHULL, MVE) showed clear truncation of predictions around the limits of the sampled fundamental niche (N^*_{Fx}) (Fig. 5), but differed in environmental values at which truncation started: exactly at the environmental limits of N^*_{Fx} for BIOCLIM, slightly inside the limits for CONVEXHULL, and slightly beyond the limits of N^*_{Fx} for MVE (Fig. 5). ENFA showed extrapolation immediately outside the range of conditions used for calibration, or N^*_{Fx}, though only to a limited degree (Fig. 5). Both GAM and GLM showed extrapolation (Fig. 5), but GLM used a simpler curve whereas GAM used a more complex curve within N^*_{Fx}, and near-linear decay outside N^*_{Fx}. BRT showed clamping starting at the limits of N^*_{Fx} (Fig. 5). GARP showed extrapolation outside the limits of N^*_{Fx} with a non-smooth curve. Maxent showed extrapolation outside N^*_{Fx} and clamping at a distance outside the limits of N^*_{Fx} (Fig. 5, Supplementary material Appendix 1 Fig. A6). KDE and MA produced distinct models from other algorithms, showing truncation with little relation to the limits of N^*_{Fx} and gaps in the middle of the response curve. Response curves of GARP, Maxent, GLM, and BRT were very similar as those found in Elith and Graham (2009).

Experiment 2

Mean TSS increased as more regions were used in calibration, but rate of increase varied among algorithms (Fig. 6a). ENFA and Maxent models had relatively low increases of mean TSS with number of calibration regions, and reached a plateau earlier than other algorithms. ENFA and Maxent models had relatively high TSS values when one region was used in calibration, but some of the lowest TSS values among algorithms considered when five regions were used. When five regions were used in calibration, BIOCLIM, BRT, GAM, GLM, and MVE were the better-performing algorithms.

Mean sensitivity of all algorithms increased as more regions were used in model calibration (Fig. 6b, c). Maxent models had the highest mean sensitivity when only one region was used, but showed little improvement as more regions were used in calibration. Unlike sensitivity, specificity was more stable among algorithms (Fig. 6b, c). Except for ENFA, GARP, and Maxent, all other algorithms had high mean specificity (> 0.9), regardless of the number of regions used in model calibration. This results were consistent for different niche breadths (Supplementary material Appendix 1 Fig. A7).

Discussion

Model transferability depends on novelty of evaluation data

Model transferability involves situations of both interpolation and extrapolation, so accuracy depends on the novelty of the environments between calibration and evaluation regions. In statistical modeling techniques, interpolation should be less challenging than extrapolation (Gelman and Hill 2007); in the same way, transferring ecological niche models to

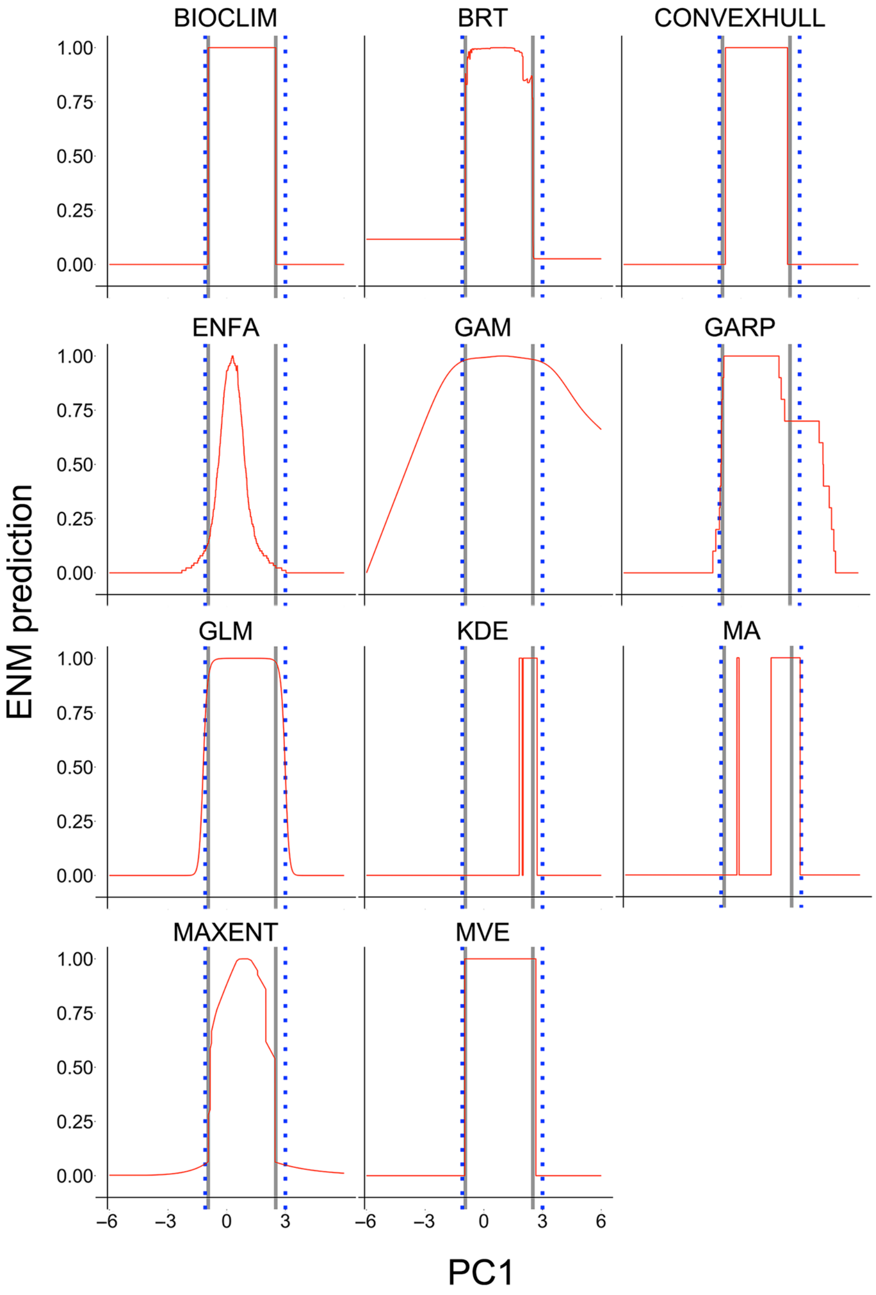

Figure 5. Illustration of extrapolation strategies used by different algorithms in experiment 1 (using region 1 of species with a wide niche, whose centroid is closest to the origin of the environmental space). The blue dotted lines represent the limits of the existing fundamental niche (N^*_F), along one environmental variable based on the first principal component (PC1) and the gray lines represent limits of a sample from the existing fundamental niche (N^*_{Fx}). The red line represents the prediction by one algorithm, which is rescaled to 0–1 for easier comparison.

similar environments (overlapping or novel-combination in this study) should be easier to achieve than transferring to novel environments. Our results confirmed these expectations for all algorithms.

Our findings are in broad agreement with previous transferability studies. Murray et al. (2011) found that algorithms performed well when models were projected to regions adjacent to the calibration area. From an environmental perspective, projection regions are similar to the calibration area in environmental dimensions, thus analogous to overlapping or novel-combination environments scenarios in our study. Sequeira et al. (2016) found better transferability of models when calibration data and evaluation data had similar spatio-temporal scales, matching again our findings of increased model transferability in environmentally matching conditions and poorer performance in non-matching conditions. Additionally, Fitzpatrick et al. (2018) found declining performance in

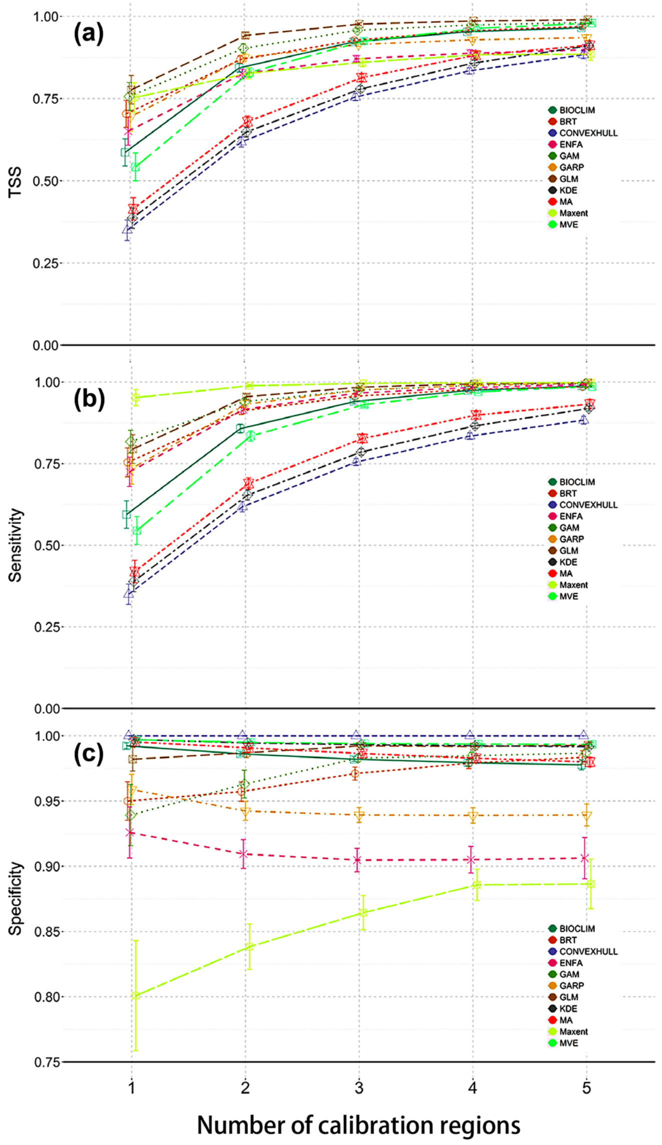

Figure 6. Trends in model evaluation metrics, true skill statistic (TSS), sensitivity, and specificity, aligned with the number of regions used in calibrating models in experiment 2. The error bars represent the 95% confidence intervals.

response to increasing climate novelty in a hindcasting experiment.

Algorithms often estimate the existing fundamental niche (N^*_F) inadequately

No algorithm included in this study accurately estimated the existing fundamental niche when a small portion of the existing fundamental niche was available for model calibration (Fig. 2a), though GLM provided the best approximation of N^*_F. Estimated niches deviated from both a sample from the existing fundamental niche and the existing fundamental niche, which resonates with Soberón and Nakamura (2009), who stated that ENM estimates something in between the realized niche (which in some cases approximates N^*_{Fx}) and fundamental niche. On the other hand, presence-only algorithms CONVEXHULL, KDE, and MA provided better estimates of sampled existing fundamental niche.

Interestingly, given incomplete information of existing fundamental niche in model calibration, different algorithms performed differently: some algorithms (presence–absence (BRT) and presence-background (Maxent) algorithms) were more liberal and tended to make broader predictions, whereas others (presence-only, e.g. BIOCLIM) were more conservative and tended to make narrower predictions (Fig. 2). The different 'behavior' of algorithms, liberal or conservative, led to distinct performance characteristics under novel environments. Liberal algorithms achieved higher sensitivity via broader predictions at the cost of lower specificity; TSS values were still above zero, indicating better-than-random predictions. Conservative algorithms maintained high specificity at the cost of low sensitivity under novel environments. GLM showed a rather intermediate behavior, which yielded better performance in terms of volume ratio and Jaccard similarity. Curiously, algorithms of the same type (e.g. machine learning algorithms) did not always perform the same: GARP was conservative compared with the other machine-learning presence-absence/background algorithms (Maxent and BRT). ENFA generated a broader estimate compared with other presence-only envelope algorithms (e.g. BIOCLIM). Additionally, the contrasting values of sensitivity and specificity highlight the importance of dissecting TSS, otherwise TSS can mask or mix the accuracy in predicting presences and absences (Qiao et al. 2015a).

Should we expect algorithms to estimate the existing fundamental niche?

GLM showed a balanced trade-off between under- and over-prediction. Estimated niches from GLM had a higher geometric similarity to the existing fundamental niche in environmental space than those from other algorithms. In a real species case study, Duque-Lazo et al. (2016) found that presence-absence/background regression algorithms (GLM and GAM) had better transferability indices than machine-learning algorithms (BRT). However, we caution about any optimistic interpretation of GLM transferability, because in principle we should not expect algorithms to predict the existing fundamental niche or fundamental niche, based on data from N^*_{Fx}, unless some supplementary information on the existing fundamental niche or fundamental niche is contained in the calibration data. That is to say, a sample from N^*_{Fx} would be a good proxy of the existing fundamental niche or fundamental niche if and only if the range of environmental tolerances are contained in N^*_{Fx}. Otherwise, the existing fundamental niche and fundamental niche will be always underestimated, although their dimensions are generally unknown (but see Brady et al. 2013).

In general, model extrapolation is not recommended in statistical modeling, because inferences beyond the range of the calibration conditions must rely on assumptions without support from data (Gelman and Hill 2007). Recovering the existing fundamental niche or fundamental niche based on the sampled existing fundamental niche clearly involves extrapolation, and decreased model performance inevitably accompanies this process, since correlative ENM algorithms are no different from other statistical modeling processes. Without additional knowledge of species' responses to novel conditions, a single N^{*}_{Fx} may be compatible with an infinite number of existing fundamental niches or fundamental niches, so correlative algorithms are unlikely to recover the existing fundamental niche or fundamental niche accurately.

However, when sufficient supplementary information is contained in the sampled existing fundamental niche, this expectation may change. The sampled existing fundamental niche was frequently situated at the periphery of the existing fundamental niche or fundamental niche in environmental space (Supplementary material Appendix 1 Fig. A2), so the information in the sampled existing fundamental niche may have been related to the environmental boundary of fundamental niche. Given such boundary information, some algorithms can partly recover the existing fundamental niche or fundamental niche, and often perform better than random. The algorithm that best recovers the existing fundamental niche or fundamental niche will be the one that best uses the information in the sampled existing fundamental niche. In our study, it seemed that more than one algorithm detected such boundary information because the algorithms made more predictions of absence outside these boundaries and more presence predictions within the boundaries (Supplementary material Appendix 1 Fig. A3). However, without knowledge from novel environments, no algorithm could detect boundary limits at the opposite end. These general points echo the results of Saupe et al. (2012) and Owens et al. (2013), who showed that correlative niche models can approximate fundamental niches only when boundaries of the niche are represented in the calibration region.

How well do algorithms estimate niches?

Different algorithms use different strategies for inference under novel environmental conditions. These strategies are based on different assumptions about what is estimated; thus, user's assumptions about the extent, shape, and position of the species' niche are crucial in selecting the algorithm that can better reconstruct the fundamental niche. Strikingly, the characteristics of fundamental niche are generally unknown, making algorithm selection challenging in real situations. It is worth noting that Maxent made considerable over-predictions of the existing fundamental niche when minimum training presence thresholding was used (Fig. 2, 4, Supplementary material Appendix 1 Fig. A3) (see also Peterson et al. 2007), even though the occurrence data included no error. We caution that performance of Maxent clamping will depend on the point where clamping begins: if clamping begins at or near the peak of a response curve, the prediction will be overly broad (Fig. 5).

Different algorithms had different sensitivity to completeness of information on the existing fundamental niche, and therefore suggested a possible strategy of selecting an algorithm that depends on confidence in the completeness of information of the existing fundamental niche. If little is known about the species' existing fundamental niche, we may need a liberal algorithm, such as Maxent, to extend beyond calibration data to approximate existing fundamental niches more closely (i.e. a broad prediction that has a larger chance of covering the full existing fundamental niche, thus higher sensitivity values). On the other hand, when knowledge of the existing fundamental niche is extensive, the advantage of choosing a liberal algorithm is minimal, and selecting a conservative algorithm, such as cluster algorithms and most envelope algorithms, becomes favorable as it fits calibration data closely and better approximate existing fundamental niche, i.e., a narrow prediction that avoids unsuitable conditions, thus generating higher specificity values. In reality, the fundamental niche is rarely known, so quantifying the completeness of information of existing fundamental niches may require external knowledge, such as expert opinion or physiological information (Feng and Papeş 2017b).

Practical considerations

Since the goal of our study was to assess model performance under a framework mirroring common practices of ENM users, we did not optimize or tune each algorithm separately. However, ENM algorithms can be intensively explored or tuned to find a parameterization that provides the best fit to the data available (e.g. Maxent; Moreno-Amat et al. 2015). In practice, an improved setting for model transfer may be achieved by calibrating models with data spatially or temporally segregated (Veloz et al. 2012, Muscarella et al. 2014, Roberts et al. 2017), though the availability of data or the improvement of model transfer are not always guaranteed.

Our evaluation of model estimations of niches was based on binary predictions, mimicking real applications of ENM and laying a baseline for virtualization of niche in environmental space, though scholars have advocated to use a probabilistic approach to simulate the distribution of virtual species for the purpose of better evaluating ENM methodology (Meynard and Quinn 2007, Meynard and Kaplan 2012, 2013). However, if niches are assumed to be continuous (e.g. probability of suitability; Hirzel et al. 2001, Elith and Graham 2009), other evaluation metrics, such as AUC (area under the receiver operating characteristic curve) and partial AUC (Peterson et al. 2008), that function across a range of thresholds, or simply the Pearson correlation between model prediction and the true suitability function, could be more appropriate. However, some evaluation metrics have received criticism (Lobo et al. 2008) and the true niche suitability function is usually unknown in reality.

In our experiment, for generalization propose, we intentionally made the six regions geographically equal. Admittedly, the six regions are therefore not equally represented in environmental space in terms of extent and density, but this scenario should be common in actual ENM applications. Simulation experiments in an artificial landscape could better manipulate environmental representation, and thus better test some relevant questions. These important aspects need to be explored in future studies.

Final remarks

Transferring models across space or time is one of the most frequent applications of ENM (Heikkinen et al. 2012, Wenger and Olden 2012, Duque-Lazo et al. 2016, Sequeira et al. 2016). However, under different spatial or temporal regimes, one must consider the changing characteristics of environmental conditions. A different spatial or temporal regime means that not only ranges of conditions change, but also possible combinations of conditions change. Our study showed that both scenarios led to decreased model performance, which suggests that one should not expect transferred models to have the same performance as in the calibration region. We also found that, given a portion of the existing fundamental niche and prior knowledge of fundamental niches, several algorithms could estimate existing fundamental niche better than random. With increasing information about species' existing fundamental niches, all algorithms investigated here had improved ability to estimate existing fundamental niches. Therefore, we do not discourage model transfers, but we do caution that investigators should quantify environmental difference or similarity between the calibration and projected areas, information that could be used as a proxy to infer or quantify the underlying uncertainty. In addition, as pointed out by Saupe et al. (2012), avoiding 'Wallacean' situations, in which representation of environments in existing fundamental niches is limited severely with respect to fundamental niches, is by far the best strategy for avoiding these problems of inference, estimation, and extrapolation, as no boundary information is available.

Previous authors have cautioned about the risks of under- and over-estimation of niches based on incorrect assumptions and data (Sinclair et al. 2010). Many ENM algorithms have been developed and many comparative analyses of algorithms have been carried out (Elith et al. 2006, Qiao et al. 2015a); in contrast with previous studies, however, we conducted a comprehensive, structured evaluation of 11 ENM algorithms in geographic and environmental spaces, from the perspective of interpolation and extrapolation in a virtual species framework. Our study provided novel insights into performance of ENM algorithms under different transferability conditions. Given incomplete knowledge of existing fundamental niches, algorithms had distinct capacities to estimate existing fundamental niche in novel environments: some gave conservative estimates, and some gave liberal, broad estimates, whereas others gave intermediate estimates. We also illustrated the different strategies of extrapolation implemented by different algorithms. We emphasize that we do not have a recommendation of a best-performing algorithm, nor do we think that one is likely to exist (Qiao et al. 2015a); instead, we suggest that users choose the appropriate method based on the situation and outcomes of preliminary tests. If input data have good representations of existing fundamental niche, one should choose a conservative algorithm; if the species is poorly known, one should choose a liberal algorithm.

Ecological niche models transferred to different areas or times are used to better understand and forecast invasiveness of non-native species and epidemic potential of infectious diseases. The usefulness of model transfer, however, lays on the biological realism of forecasts. In view of the excessive extrapolation of some algorithms, the scientific community should remain skeptical about predictions of dramatic changes of species' ranges from ecological niche models that are transferred without model evaluation and visualization in environmental dimensions.

Funding – This study was supported by the National Key Research and Development Project of China (2017YFC1200603), National Science Foundation – Advances in Biological Informatics (1458640), and Natural Science Foundation of China (31772432). Part of the computing for this project was performed at the Oklahoma State Univ. High Performance Computing Center.
Conflicts of interest – None of the authors have any conflict of interest to declare.

References

Allouche, O. et al. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

Angilletta, M. J. 2009. Thermal adaptation: a theoretical and empirical synthesis. – Oxford Univ. Press.

Barbet-Massin, M. et al. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods Ecol. Evol. 3: 327–338.

Birch, L. C. 1953. Experimental background to the study of the distribution and abundance of insects: III. The relation between innate capacity for increase and survival of different species of beetles living together on the same food. – Evolution 7: 136–144.

Blonder, B. et al. 2014. The n-dimensional hypervolume. – Global Ecol. Biogeogr. 23: 595–609.

Brady, O. J. et al. 2013. Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. – Parasites Vectors 6: 351–351.

Busby, J. 1991. BIOCLIM – a bioclimate analysis and prediction system. – Plant Prot. Q. 6: 8–9.

Capinha, C. et al. 2014. Macroclimate determines the global range limit of *Aedes aegypti*. – EcoHealth 11: 420–428.

De'ath, G. 2007. Boosted trees for ecological modeling and prediction. – Ecology 88: 243–251.

Duque-Lazo, J. et al. 2016. Transferability of species distribution models: the case of *Phytophthora cinnamomi* in southwest Spain and southwest Australia. – Ecol. Model. 320: 62–70.

Elith, J. and Graham, C.H. 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. – Ecography 32: 66–77.

Elith, J. and Leathwick, J. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. – Divers. Distrib. 13: 265–275.

Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.

Elith, J. et al. 2005. The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. – Ecol. Model. 186: 280–289.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Elith, J. et al. 2008. A working guide to boosted regression trees. – J. Anim. Ecol. 77: 802–813.

Elith, J. et al. 2011. A statistical explanation of MaxEnt for ecologists. – Divers. Distrib. 17: 43–57.

Escobar, L. E. et al. 2018. Ecological niche modeling re-examined: a case study with the Darwin's fox. – Ecol. Evol. 8: 4757–4770.

Farber, O. and Kadmon, R. 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. – Ecol. Model. 160: 115–130.

Feng, X. and Papeş, M. 2017a. Can incomplete knowledge of species' physiology facilitate ecological niche modelling? A case study with virtual species. – Divers. Distrib. 23: 1157–1168.

Feng, X. and Papeş, M. 2017b. Physiological limits in an ecological niche modeling framework: a case study of water temperature and salinity constraints of freshwater bivalves invasive in USA. – Ecol. Model. 346: 48–57.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.

Fitzpatrick, M. C. and Hargrove, W. W. 2009. The projection of species distribution models and the problem of non-analog climate. – Biodivers. Conserv. 18: 2255–2261.

Fitzpatrick, M. C. et al. 2018. How will climate novelty influence ecological forecasts? Using the Quaternary to assess future reliability. – Global Change Biol. 24: 3575–3586.

Gelman, A. and Hill, J. 2007. Data analysis using regression and multilevel/hierarchical models. – Cambridge Univ. Press.

Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.

Guisan, A. et al. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. – Ecol. Model. 157: 89–100.

Hastie, T. and Tibshirani, R. 1990. Generalized additive models. Chapman and Hall.

Hattab, T. et al. 2017. A unified framework to model the potential and realized distributions of invasive species within the invaded range. – Divers. Distrib. 23: 806–819.

Heikkinen, R. K. et al. 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? – Ecography 35: 276–288.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.

Hirzel, A. H. et al. 2001. Assessing habitat-suitability models with a virtual species. – Ecol. Model. 145: 111–121.

Hirzel, A. H. et al. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? – Ecology 83: 2027–2036.

Hooper, H. L. et al. 2008. The ecological niche of *Daphnia magna* characterized using population growth rate. – Ecology 89: 1015–1022.

Hortal, J. et al. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. – Oikos 117: 847–858.

Hutchinson, G. E. 1957. Concluding remarks. – Cold Spring Harbor Symp. Quant. Biol. 22: 415–427.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. – New Phytol. 11: 37–50.

Jiménez-Valverde, A. et al. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Community Ecol. 10: 196–205.

Lafarge, T. and Pateiro-Lopez, B. 2016. alphashape3d: implementation of the 3D alpha-shape for the reconstruction of 3D sets from a point cloud. – In: R package ver. 1.2, <http://CRAN.R-project.org/package=alphashape3d>.

Leroy, B. et al. 2016. virtualspecies, an R package to generate virtual species distributions. – Ecography 39: 599–607.

Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.

Mackenzie, D.I. 2005. Was it there? Dealing with imperfect detection for species presence/absence data. – Aust. N. Z. J. Stat. 47: 65–74.

Maguire Jr, B. 1967. A partial analysis of the niche. – Am. Nat. 101: 515–526.

McCullagh, P. and Nelder, J. A. 1989. Generalized linear models. – Chapman and Hall.

Merow, C. et al. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. – Ecography 36: 1058–1069.

Meynard, C. N. and Kaplan, D. M. 2012. The effect of a gradual response to the environment on species distribution modeling performance. – Ecography 35: 499–509.

Meynard, C. N. and Kaplan, D. M. 2013. Using virtual species to study species distributions and model performance. – J. Biogeogr. 40: 1–8.

Meynard, C. N. and Quinn, J. F. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. – J. Biogeogr. 34: 1455–1469.

Moreno-Amat, E. et al. 2015. Impact of model complexity on cross-temporal transferability in MaxEnt species distribution models: an assessment using paleobotanical data. – Ecol. Model. 312: 308–317.

Moudrý, V. 2015. Modelling species distributions with simulated virtual species. – J. Biogeogr. 42: 1365–1366.

Murray, J. V. et al. 2011. Evaluating model transferability for a threatened species to adjacent areas: implications for rock-wallaby conservation. – Austral Ecol. 36: 76–89.

Muscarella, R. et al. 2014. ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. – Methods Ecol. Evol. 5: 1198–1205.

Owens, H. L. et al. 2013. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. – Ecol. Model. 263: 10–18.

Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.

Peterson, A. T. and Soberón, J. 2012. Species distribution modeling and ecological niche modeling: getting the concepts right. – Nat. Conservacao 10: 1–6.

Peterson, A. T. et al. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. – Ecography 30: 550–560.

Peterson, A. T. et al. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. – Ecol. Model. 213: 63–72.

Peterson, A. T. et al. 2011. Ecological niches and geographic distributions. – Princeton Univ. Press.

Peterson, A. T. et al. 2016. Mechanistic and correlative models of ecological niches. – Eur. J. Ecol. 1: 28–38.

Phillips, S. J. et al. 2004. A maximum entropy approach to species distribution modeling. – Proc. 21st Int. Conf. Machine Learn., pp. 655–662. ACM press, Banff, AB, Canada.

Qiao, H. et al. 2015a. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. – Methods Ecol. Evol. 6: 1126–1136.

Qiao, H. et al. 2015b. Marble algorithm: a solution to estimating ecological niches from presence-only records. – Sci. Rep. 5: 14232–14232.

Qiao, H. et al. 2016. NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. – Ecography 39: 805–813.

Qiao, H. et al. 2018. Data from: An evaluation of transferability of ecological niche models. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.kg3d57r>.

Randin, C. F. et al. 2006. Are niche-based species distribution models transferable in space? – J. Biogeogr. 33: 1689–1703.

Roberts, D. R. and Hamann, A. 2012. Predicting potential climate change impacts with bioclimate envelope models: a palaeoecological perspective. – Global Ecol. Biogeogr. 21: 121–133.

Roberts, D. R. et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. – Ecography 40: 913–929.

Saupe, E. E. et al. 2012. Variation in niche and distribution model performance: the need for a priori assessment of key causal factors. – Ecol. Model. 237–238:11–22.

Sequeira, A. M. M. et al. 2016. Transferability of predictive models of coral reef fish species richness. – J. Appl. Ecol. 53: 64–72.

Sinclair, S. J. et al. 2010. How useful are species distribution models for managing biodiversity under future climates? – Ecol. Soc. 15: 8.

Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. – Ecol. Lett. 10: 1115–1123.

Soberón, J. and Nakamura, M. 2009. Niches and distributional areas: concepts, methods, and assumptions. – Proc. Natl Acad. Sci. USA 106: 19644–19650.

Soberón, J. and Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. – Biodivers. Inf. 2: 1–10.

Stockwell, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. – Int. J. Geogr. Inf. Sci. 13: 143–158.

Van Aelst, S. and Rousseeuw, P. 2009. Minimum volume ellipsoid. – Wiley Interdiscip. Rev. Comp. Stat. 1: 71–82.

Veloz, S. D. et al. 2012. No–analog climates and shifting realized niches during the late quaternary: implications for 21st–century predictions by species distribution models. – Global Change Biol. 18: 1698–1713.

Warren, D. L. 2012. In defense of 'niche modeling'. – Trends Ecol. Evol. 27: 497–500.

Wenger, S. J. and Olden, J. D. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. – Methods Ecol. Evol. 3: 260–267.

Williams, J. W. and Jackson, S. T. 2007. Novel climates, no-analog communities, and ecological surprises. – Front. Ecol. Environ. 5: 475–482.

Zurell, D. et al. 2012. Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions. – Divers. Distrib. 18: 628–634.

Supplementary material (Appendix ECOG-03986 at <www.ecography.org/appendix/ecog-03986>). Appendix 1.