Ecography

ECOG-00107
Smith, A. B., Santos, M. J., Koo, M. S., Rowe, K. M. C., Patton, J. L., Perrine, J. D., Beissinger, S. R. and Moritz, C. 2013. Evaluation of species distribution models by resampling of sites surveyed a century ago by Joseph Grinnell. – Ecography 36: xxx–xxx.

Supplementary material

Appendix 1: Change in climatic conditions between the historic Grinnell surveys and modern resurveys and the PRISM climate data set

Table A1. Descriptive statistics for climatic variables in the training region (western US) and at test sites (Grinnell sites) for both eras. Values are means across the given time interval and across all sites in the respective region, and values in parentheses are the range of the factor. Isothermality is the ratio of the mean range of monthly temperature to the mean range of annual temperature, and precipitation seasonality is the coefficient of variation of precipitation across months. In general Grinnell sites were higher, wetter, cooler, and had more variation in precipitation compared to the western US as a whole. Overall Grinnell sites generally became wetter and warmer and had more annual variation in temperature and precipitation than the western US at large.

	Western US		Grinnell Sites	
Factor	1900-1939	1970-2009	1900-1939	1970-2009
		General		
Mean annual temperature (°C)	9.0 (-11.0 to 24.0)	9.4 (-11.0 to 25.9)	7.8 (-1.3 to 19.3)	8.1 (-1.3 to 19.6)
Mean annual precipitation (mm)	497 (0 to 6829)	531 (0 to 7147)	882 (167 to 2705)	892 (165 to 2832)
Elevation (m)	1440 (-121 to 4275)		1921 (50 to 3940)	
		Predictors Used in Modeling		
BIO02: Mean diurnal temperature range (°C)	15.0 (4.7 to 22.2)	14.5 (4.9 to 22.0)	14.8 (10.3 to 18.4)	14.4 (9.6 to 17.4)
BIO03: Isothermality	40 (23 to 71)	39 (23 to 70)	45 (38 to 52)	44 (37 to 51)
BIO05: Temperature of the warmest month (°C)	29.5 (2.6 to 46.9)	29.6 (2.4 to 47.0)	26.0 (12.6 to 38.6)	25.8 (12.8 to 37.5)
BIO06: Temperature of the coldest month (°C)	-7.8 (-22.3 to 10.5)	-6.9 (-23.0 to 11.1)	-6.5 (-15.0 to 4.7)	-5.7 (-14.0 to 5.0)
BIO07: Temperature annual range (°C)	37.3 (10.2 to 52.2)	36.5 (9.7 to 50.7)	32.5 (25.8 to 38.0)	31.5 (24.7 to 37.4)
BIO13: Precipitation of the wettest month (mm)	80 (0 to 1106)	84 (0 to 1215)	184 (32 to 524)	172 (32 to 530)
BIO14: Precipitation of the driest month (mm)	12 (0 to 94)	14 (0 to 141)	4 (1 to 15)	6 (0 to 21)
BIO18: Precipitation of the warmest quarter (mm)	91 (1 to 533)	94 (1 to 575)	26 (2 to 79)	30 (3 to 94)
BIO15: Precipitation seasonality	50 (0 to 296)	49 (0 to 264)	86 (63 to 103)	82 (60 to 102)

The PRISM climate data set

Details on the development of the parameter-elevation regression on independent slopes model (PRISM) data set are presented in Daly et al. (2000, 2002, and 2004; see also http://www.prism.oregonstate.edu), so here we provide only a cursory description of the methods used to develop the PRISM climate layers with a focus on aspects relevant to climate of the training region (the western conterminous US) and the test region (the Sierra Nevada).

PRISM is a hybrid expert knowledge-based/statistical system in which the climatic factor of interest (i.e., precipitation or temperature) are predicted with simple linear regression equations describing the response as a function of elevation. Elevation is obtained from a digital elevation model. The regression parameters are local in the sense that they can vary for each grid cell, and are estimated from a moving window of a user-defined radius (usually ≥50 km) such that at least a minimum number of weather stations are within the windows (usually ≥15). Stations are weighted by their horizontal and vertical distance to the target cell, as well as their similarity in aspect (calculated at multiple spatial scales), atmospheric layer in which they reside (boundary layer or free atmosphere), coastal proximity, terrain, and clustering among stations. PRISM assumes a two-layer atmosphere and imposes an inversion layer in months and regions in which inversions have been recorded (esp. the montane western US, including the Sierra Nevada; Daly et al. 2002). The regression parameters are required to fall within a user-defined range established from observed relationships between temperature or precipitation and elevation. If they fall outside this range the lowest-weighted station is removed from the regression in an iterative fashion until the parameters are satisfactory or until the minimum number of stations remain in the search window. If the latter occurs, then the parameters are set to their default values (usually the mean across the landscape). By weighting stations according to their

similarity in coastal effects and aspect, PRISM accounts for the maritime influences on climate and rain shadow effects. These are especially relevant to the climate of our study region as it is affected along the western side by the Pacific Ocean and many of the training and test sites are at high altitudes.

Our climate data was derived by averaging estimates of minimum temperature, maximum temperature, and precipitation across each month of the years of our study eras (1900-1939 and 1970 to 2009, inclusive). An exact tally of the number of stations is not possible since different stations were online in different years and some had missing data. Nevertheless, the overall network of stations was roughly stable in each time period, providing approximately 2300 stations for the historic era and 8900 for the modern era (Wayne Gibson, PRISM Group, Oregon State University, *personal communication*; Matthias Falk, University of California, Davis, *personal communication*). As a result, we expect historical climate records to be less accurate than modern records. As a result, we expected projections across eras to be less accurate than within the same time period since inaccuracies in the historical climate records could be expected to cause a mischaracterization of the climatic niche of species.

PRISM has been reviewed by scores of independent scientists and cross-validated using a wide range of methods, but the only comparison between PRISM and a similar set of climate layers of which we are aware was performed by Parra and Monahan (2008) who compared it to the ANUSPLIN system (which has been used to generate the commonly-used WORLDCLIM layers; Hijmans et al. 2005). In comparison to ANUSPLIN PRISM was more precise but less accurate. They also modeled the historic and current distributions of 57 Californian mammals, with time periods roughly equivalent to our historic and modern eras. Compared to ANUSPLIN, MAXENT models based on 4×4-km resolution PRISM data tended to predict more range

stability (vs. contraction or expansion). We note that these results are not necessarily indicative of performance by our models since we used the updated, 30-arcsec (~1×1-km) resolution PRISM data set.

Literature cited

Daly, C., G.H. Taylor, W. P. Gibson, T.W. Parzybok, G. L. Johnson, P. Pasteris. 2001. High-quality spatial climate data sets for the United States and beyond. Transactions of the American Society of Agricultural Engineers, 43:1957-1962.

Daly, C., W. P. Gibson, G.H. Taylor, G. L. Johnson, P. Pasteris. 2002. A knowledge-based approach to the statistical mapping of climate. Climate Research 22: 99-113.

Daly, C., Gibson, W.P., M. Doggett, J. Smith, and G. Taylor. 2004. Up-to-date monthly climate maps for the conterminous United States. Proc., 14th AMS Conf. on Applied Climatology, 84th AMS Annual Meeting Combined Preprints, Amer. Meteorological Soc., Seattle, WA, January 13-16, 2004, Paper P5.1.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., and Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965-1978.

Parra, J.L. and W.B. Monahan. 2008. Variability in 20th century climate change reconstructions and its consequences for predicting geographic responses of California mammals. Global Change Biology 14:2215-2231.

Appendix 2: Model implementation, data cleaning, and occupancy modeling

Contents

Tables

Data cleaning

Museum records for all mammals in the Western United States were downloaded from MaNIS (www.manisnet.org) and Arctos (http://arctos.database.museum/) in July of 2010. Records with obvious problems (e.g., only genus listed or no coordinate uncertainty or records georeferenced to locations where the animals were kept in captivity) were removed. Initially all records with coordinate uncertainty ≤5000 m were retained. Hence, many of our training sites had associated spatial error larger than the resolution of the environmental data, but we do not consider this a substantial problem since others have shown that similar uncertainty does not degrade model performance (Graham et al. 2008). To further filter incorrectly georeferenced specimens, for each species we removed records that were below the 0.25th and above the 99.75th percentiles in either mean annual temperature or precipitation relative to the other sites where the species was located in both eras (Chapman 2005). We also removed records collected before year 2000 with coordinate uncertainties <3 m since these had unlikely accuracies and records from 2000 onward

with coordinate uncertainties >200 m since the widespread use of GPS and cessation of selective availability improved spatial accuracy. For each species, records were further thinned so that no presence points were within 1 km of one another.

Details on model implementation

Included here are details on implementation of the SDMs used in the main paper. All models were trained and tested using custom code and the dismo (for BIOCLIM, GLMs, and MAXENT; Hijmans et al. 2012), raster (Hijmans & van Etten 2012), mgcv (for GAMs; Wood 2012), gbm (for BRTs; Ridgeway 2007, Elith et al. 2008), and e1071 (for SVMs; Dimitriadou et al. 2011) packages in R Ver. 2.15.2 (R Core Development Team 2011).

The original BIOCLIM model gave presence/absence predictions based on whether a site was within a user-defined percentile range of the distribution of the training data (Busby 1991). In the current exercise we used a modification that yields continuous predictions across the range [0, 1] (Hijmans et al. 2011). The environmental suitability of a site is predicted to be the minimum of the percentile distribution across all environmental predictors measured at that site, where percentiles >50 are subtracted from 100 so that, for example, a site at the 80th percentile of a factor has the same value as one at the 20th percentile. This value is then divided by 50 so that values range from 0 (least suitable) to 1 (most suitable). Predictors do not interact, and the predictor with the minimum score determines the suitability of the site.

BRTs use a series of linked classification and regression trees, with each tree trained on a randomly drawn subset of the residuals from the previous one (Hastie et al 2001). BRTs were implemented with a modified version of the gbm.step function from Elith et al. (2008). We used a learning rate of 0.0001, tree complexity of 12, bag fraction of 0.7, and 2000 trees per model, all of which are within the range suggested by Elith et al. (2008).

GLMs are a regression technique that assumes a linear relation between predictors (or their higher-order terms) and species' occurrence. Initial models for GLMs were constructed from a set of linear, quadratic,

quartic, and two-way interaction terms. Initial model form was determined by entering each term alone in a GLM provided there were ≥10 presence sites per term (forcing inclusion of linear and quadratic terms when testing higher-order terms, so that, for example, there needed to be ≥30 presence sites to test the term $X + X^2 + X^3$ but between 20 and 29 to test $X + X^2$ and fewer than 19 to test just X). We then constructed an initial multivariate model using the eight terms with the lowest AIC, provided there were ≥20 presences per term. Terms in this initial model were then dropped using forward and backward AIC-based model selection until the most parsimonious model was found. All predictors were centered and standardized before training.

GAMs are a more flexible implementation of GLMs that apply non-linear smoothers to predictor variables before relating them to the dependent variable (Wood 2006). GAMs were implemented with cubic splines and with shrinkage, which allows a term's influence to go to zero if it is unimportant relative to the other terms in the model. Model form was determined in the same manner as for GLMs using single-term and joint-term (interaction term) smoothers. Joint smoothing terms used tensor products because they allow for non-stationary responses and differences in scale of the predictors (Wood 2006).

MAXENT first estimates the probability of each predictor across presence sites given the distribution of the predictor in the study area, and then inverts the relation with Bayes' Theorem, yielding the probability of presence given the environment (Phillips et al. 2006, Phillips and Dudík 2008). The distribution is derived using maximization of information entropy, which produces the mathematically smoothest distribution possible given constraints (e.g., predictor mean, variance, covariance with other predictors). We implemented MAXENT using Ver. 3.3.3e (Phillips et al. 2009) called from R using the dismo package (Hijmans et al. 2011). We used AIC-based tuning of the master regularization parameter to find its optimal value within the range 0.25 to 20 with a custom R function following procedures described in Warren and Siefert (2011). As with all other SDMs, we evaluated output using dismo's "evaluate" function to calculate AUC and other test metrics.

SVMs find the gradient in environmental space that declines most steeply perpendicular to the distribution of presences and absences given a user-set tolerance (Guo et al. 2005). We implemented two-class SVMs with a radial (Gaussian) kernel (Tax et al. 2004). An iterative grid search in log space was used to find the best γ and cost combination across the intervals 2^{-15} to 2^{3} for γ and 2^{-5} to 2^{15} for cost (Hsu et al. 2010).

The EMEAN and EMED models calculate the environmental suitability of a site as the mean or median score, respectively, across all of the other six models. Since different model types produced very different ranges of output (e.g., most GAMs produced prediction values between 0 and ~0.3, while GLMs produced values between ~0 and ~1), we standardized prediction maps to the range [0, 1] before calculating the mean or median by subtracting from each map its minimum value across all cells, then dividing by its (new) maximum value across all cells.

Evaluation data: The original Grinnell Surveys and the Grinnell Resurvey Project

Apart from the museum data used to train the models we used data from historical and contemporary surveys of three elevational transects in the Sierra Nevada to test model projections. The historical surveys were led by Joseph Grinnell, the founding director of the University of California Berkeley's Museum of Vertebrate Zoology. Grinnell and his colleagues conducted surveys of mammals, birds, reptiles, and amphibians at these transects and many more sites across California and the greater American West between 1900 and 1939 (Grinnell and Storer 1924; Grinnell et al. 1930, Sumner and Dixon 1953). Grinnell's meticulous note-taking style was unique for the era (Perrine and Patton 2011); altogether he and his students amassed ~50,000 pages of field notes during this time period. The notes consist of records from traplines (using snap traps, Macabee gopher traps, mole traps, and steel traps), sightings, and shot animals, though to maintain consistency between historic and modern methods we utilized only the trapline data since it allowed us to best quantify survey effort for occupancy modeling (below). Sites were generally surveyed for 1 to 16 nights (median 5) with 6 to 335 traps (median 96)

during seasons when the animals were expected to be present (i.e., not in hibernation). Eight thousand six hundred eighty eight of the 15,277 historic mammal records used in this study were deposited in the Museum collection, allowing us to verify Grinnell's identifications, especially of cryptic taxa (e.g., *Tamias*). Following Moritz et al. (2008) and Tingley et al. (2012) we defined a "site" as an area with a 2-km radius around an aggregate of traplines (usually centered on a campsite) that was within 100 m of the elevation of the traplines' centroid since this generally encompassed the area in which aggregates of traplines were set and described in the historic field notes. Coordinates and extent of each trapline were georeferenced using field notes, historic photographs taken by Grinnell and colleagues, areal photographs and maps. Coordinate uncertainty was obtained using the point-radius method (Wieczorek et al. 2004). Elevation was obtained from a 1-arcsec digital elevation model (DEM). Scanned copies of the original field notes used for this project are available from the Museum of Vertebrate Zoology's web site (http://bscit.berkeley.edu/mvz/volumes.html) and data for historic and modern voucher specimens are available from the Arctos database (http://arctos.database.museum/home.cfm).

Since 2003, as part of the Grinnell Resurvey Project, we have been conducting surveys along three elevational transects that straddle the Sierra Nevada centered on Lassen National Park and National Forest, Yosemite National Park, and Sequoia and Kings Canyon National Parks and Sequoia, Sierra, and Inyo National Forests (Fig. 1; Moritz et al. 2008; Rubidge et al. 2011; Perrine and Patton 2011; Morelli et al. 2012; Tingley et al. 2012). We chose these regions specifically because of the high historical sampling density and because they have experienced less development than other regions of the Sierra. Detailed descriptions of each transect can be found in Tingley et al. (2012). We used historical field notes, maps, and historical photographs to locate census sites near historical survey sites, though we also opportunistically added sites not originally censused in the historic surveys. Modern sites were georeferenced using GPS. When resurveying matching sites the modern traplines followed historical traplines as closely as possible. Surveys were conducted from 1-11 nights (median 6) with 3-339 (median 65) Sherman and Tomahawk traps plus pitfall traps made from 32-oz plastic cups sunk into the ground.

We also surveyed sites during the seasons we expected the mammals to be active and detectable. Of the mammals caught, 6,144 of 14,316 were deposited in the collection of the Museum of Vertebrate Zoology. Records are available for these specimens at the aforementioned web sites.

We restricted our analysis to historic and modern sites on the west slope of the Sierra and down to the lower limits of the yellow pine belt on the eastern slope, excluding sites more characteristic of deserts east of the Sierra. Records of the Great Basin subspecies of *Peromyscus truei* were excluded following Moritz et al. (2008) and Yang et al. (2011). Our analysis focused on species that could be regularly caught given historical and modern detection methods (i.e., excluding carnivores) and those which are characteristic of the Californian mammalian fauna (i.e., excluding desert species such as *Neotoma lepida*).

We refer to these historic and modern sites as "Grinnell sites" and use them to test output from the SDMs.

Generating the high-quality absence (HCA) test data set with occupancy modeling

To determine if a Grinnell site in which a species was not observed was truly absent of the species, we used the single-season occupancy framework for each era to estimate the probability of a false absence at each site. Models were implemented in R and MARK using the RMark package to link the two programs (White and Burnham 1999, White et al. 2001). Models of the probability of detection given that the species was present were parameterized using measures of trapping effort and era. A set of 34 candidate models for detectability with era, number of nights a site was trapped, trap effort (number of traps and log(number of traps)), and the interactions between era and trap night and between era and trap effort were used as covariates. The best 16 models from this set were then incorporated into 25 models of the probability of occupancy, which used era, elevation and its square, region, and all two- and three-way combinations of these variables, plus the constant model (the "dot" model). Finally, the probability of detection each night at a site, p_n, was calculated from the AIC-weighted average across all possible combinations of the 16 detectability and 25 occupancy models. The probability of a false absence at a site was calculated from $1 - \prod_n^N (1 - p_n)$ where N is the total number of nights a site was censused

(Moritz et al. 2008, Tingley and Beissinger 2009). Sites where the target species was not detected were assumed to be true absences if the probability of false absence was ≤0.10 (Rubidge et al. 2011). Test sites where a species was not detected and with probabilities of false absence >0.10 were discarded. Table A2 lists the number of presence test sites and LCA and HCA absence sites per species in each era.

Table A2. Number of training presences and test presences and absences for each species in each era and absence set. A species had to have >32 training presences, 5 test presences, low-confidence absences, and high-confidence absences in each era to be included in this list. The number of training presences in each era was equalized by subsampling from the more numerous era. For the PSA set 1000 sets of randomly chosen sites equal to the number of test sites in the test era were used. HCA absences are a subset of LCA absences and created by removing from the LCA set sites at which the probability of false absence was >0.10. Since prevalence (proportion of presences to presences and absences) differed between species and across eras within a species, it was included as a nuisance term in statistical evaluations of model performance of the LCA and HCA sets.

		Low Confidence Absence Test Set				High Confidence Absence Test Set			
		Presences		Absences		Presences		Absences	
Species	No. training presences	Hist.	Mod.	Hist.	Mod.	Hist.	Mod.	Hist.	Mod.
Callospermophilus lateralis	208	34	45	56	91	34	45	9	25
Chaetodipus californicus	112	13	16	77	120	13	16	8	50
Microtus californicus	157	22	18	68	118	22	18	30	33
Microtus longicaudus	176	46	65	44	71	46	65	32	19
Microtus montanus	119	23	35	67	101	23	35	39	76
Neotoma fuscipes	83	5	5	85	131	5	5	23	52
Neotoma macrotis	121	15	27	75	109	15	27	7	48
Peromyscus boylii	139	37	39	53	97	37	39	32	66
Peromyscus maniculatus	1003	82	106	8	30	82	106	6	22
Peromyscus truei	232	25	28	65	108	25	28	31	21
Reithrodontomys megalotis	202	23	27	67	109	23	27	47	47
Sorex monticolus	70	19	49	71	87	19	49	30	13
Sorex vagrans	141	8	13	82	123	8	13	52	94
Tamias amoenus	168	9	13	81	123	9	13	63	104
Tamias senex	50	13	14	77	122	13	14	21	48
Tamias speciosus	65	34	49	56	87	34	49	32	44
Urocitellus beldingi	59	5	16	85	120	5	16	37	47
Zapus princeps	84	23	32	67	104	23	32	45	18
Mean	177.2	24.2	33.2	65.8	102.8	24.2	33.2	30.2	45.9
Minimum	50	5	5	8	30	5	5	6	13
Maximum	1003	82	106	8	30	82	106	63	104

Supplemental Literature Cited

Busby, J. R. 1991. BIOCLIM – a bioclimate analysis and prediction system. In: Margules, C. R. and Austin, M. P. (eds.), Nature conservation: Cost effective biological surveys and data analysis. CSIRO, pp. 64–68.

Chapman, A. D. 2005. Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel. 2011. "e1071", ver. 1.5-25, package for R. http://cran.r-project.org/.

Elith, J., J.R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77:802-813.

Graham, C.H., J. Elith, R.J. Hijmans, A. Guisan, A.T. Peterson, B.A. Loiselle, and the NCEAS Predicting Species Distributions Working Group. 2008. The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology 45:239-247.

Guo, Q., Kelly, M., and Graham, C.H. 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling 182:75-90.

Grinnell, J. and Storer, T. 1924. Animal Life in the Yosemite. University of California Press, Berkeley.

Grinnell, J., Dixon, J., and Lindsdale, J.M. 1930. Vertebrate natural history of a section of northern California through Lassen Peak. University of California Publications in Zoology 35:1-584.

Hastie, T., Tibshirani, R. and Friedman, J.H. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.

Hijmans, R.J. and van Etten, J. 2012. "raster", ver. 1.8-9, package for R. http://cran.r-project.org/.

Hijmans, R.J., Phillips, S., Leathwick, J., and Elith, J. 2011. "dismo", ver. 0.6-3, package for R. http://cran.r-project.org/.

Hsu, C-w., C-c. Chang, and C-j. Lin. 2010. A practical guide to support vector classification. Bioinformatics 1:1-16.

Morelli, T.L., A.B. Smith, C. Kastely, I. Mastroserio, C. Moritz, and S. Beissinger. 2012. Anthropogenic refugia ameliorate the severe climate-related decline of a montane mammal along its trailing edge. Proceedings of the Royal Society of London B 279:4279-4286.

Moritz, C., Patton, J.L., Conroy, C.J., Parra, J.L., White, G.C., and Beissinger, S.R. 2008. Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. Science 322:261-264.

Perrine, J.D. and Patton, J.L. 2011. Letters to the future. Pp. 211-250 in Canfield, M.R. (ed.) Field Notes on Science and Nature. Harvard University Press, Cambridge, MA.

Phillips, S.J. and Dudík, M. 2008. Modeling species distributions with Maxent: New extensions and a comprehensive evaluation. Ecography 31:161-175.

Phillips, S.J., Anderson, R.P., and Schapire, R.E. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231-259.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., and Ferrier, S. 2009. Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. Ecological Applications 19:181-197.

R Core Development Team. 2011. R, The Project for Statistical Computing. http://cran.r-project.org/.

Ridgeway, G. 2007. "gbm," ver. 1.6-3.1, package for R. http://cran.r-project.org/.

Rubidge, E., Monahan, W., Parra, J.L,. Cameron, S.E., and Brashares, J.S. 2011. The role of climate, habitat, and species co-occurrence as drivers of change in small mammal distributions over the past century. Global Change Biology 17:696-708.

Sumner, L., and Dixon, J.S. 1953. Birds and Mammals of the Sierra Nevada. University of California Press, Berkeley, CA.

Tax, D.M.J. and Müller, K.-R. A consistency-based model selection for one-class classification. Proceedings 17th International Conference on Pattern Recognition (22–26 August 2004, Cambridge UK) (eds J. Kittler, M. Petrou and M. Nixon), pp. 363-366. IEEE Computer Society, Los Alamitos, CA.

Tingley, M.W. and S.R. Beissinger. 2009. Detecting range shifts from historical species occurrences: New perspectives on old data. Trends in Ecology and Evolution 24:625-633.

Tingley, M., Koo, M.S., Moritz, C., Rush, A.C., and Beissinger, S.R. 2012. The push and pull of climate change causes heterogeneous shifts in avian elevational ranges. Global Change Biology 18:3279-3290.

Warren, D.L. and S.N. Siefert. 2011. Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. Ecological Applications 21:335-342.

White, G. C., and K. P. Burnham. 1999. Program MARK: survival estimation from populations of marked animals. Bird Study 46 Supplement:120-138.

White, G. C., K. P. Burnham, and D. R. Anderson. 2001. Advanced features of Program Mark. Pages 368-377 in R. Field, R. J. Warren, H. Okarma, and P. R. Sievert, editors. Wildlife, land, and people: priorities for the 21st century. Proceedings of the Second International Wildlife Management Congress. The Wildlife Society, Bethesda, Maryland, USA.

Wieczorek, J., Guo, Q., and Hijmans, R.J. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Information Science 18:745-767.

Wood, S.N. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall, Boca Raton.

Wood, S.N. 2012. "mgcv", ver. 1.7-12, package for R. http://cran.r-project.org/.

Yang, D-S., Conroy, C.J., and Moritz, C. 2011. Contrasting responses of *Peromyscus* mice of Yosemite National Park to recent climate change. Global Change Biology 17:2559-2566.

Appendix 3: Ancillary results: Analysis of COR; variance partitioning; within- vs. cross-era performance correlations; turnover at Grinnell sites; analyses of omission and commission rates; and transferability of thresholds

Contents

Tables and Figures

Analysis of COR

In addition to the analysis of AUC presented in the text of the paper, we also analyzed performance of the SDMs using COR, the point biserial correlation, which is a measure of the correlation between predicted environmental suitability and the probability of presence and absence (Elith et al. 2006). COR is the Pearson correlation coefficient between model output at test sites, which ranges from [0, 1], and a set of 1's and 0's indicating whether the species of interest was present or absent at the test sites. Positive values indicate the model output correlates positively with the probability of presence and absence. Since COR has the range [-1, 1], we first transformed it to the interval [0, 1] using $0.5\times(x+1)$, then applied the modified logit transformation of Warton and Hui (2011).

We repeated the same analyses described in the main text for AUC for COR. Briefly, we examined two models for each absence type, one of the form logit(COR) ~ prevalence + prevalence^2 + number of test sites + SDM + projection + SDM × projection + species, and the other logit(COR) ~ prevalence + prevalence² + number of test sites + activity cycle + annual rhythm + diet + log(adult mass) + litter size + litters per year + young per year + log(range area) + ENFA marginality + ENFA breadth. Prevalence and its square was not included when analyzing PSA data since the number of test presences and absences was kept equal in each case, and number of test sites was not use for the analysis of LCA since they did not differ within an era across species or SDMs (90 historic sites and 136 modern sites). Variance partitioning was performed on both models after forwards/backwards variable selection with $P\leq0.05$ for inclusion of a variable.

Results of regression on COR using "species" were qualitatively very similar to the analysis of AUC in that the "species" term was always significant, absence type was significant in the analysis including all absence sets, and SDM was significant in the LCA and HCA tests (and its interaction with projection was significant in the PSA analysis; Tables 1 and A3).

Results from variance partitioning of the COR models also suggested projection had negligible influence, SDM had a consistent but somewhat minor effect, and "species" had a consistently large effect on variation (Tables A5). However, all autecological traits combined did not explain as much variation as did the "species" term when the former replaced the latter, suggesting that unmeasured traits of characteristics of the data are captured by the "species" term but not our list of traits. Of the traits, ENFA marginality and diet usually contributed most to total variance in COR.

Table A3. Results from analysis of variance of COR for each absence type and all three absence types together (compare to Table 1). Sums of squares are calculated for each term when it is entered last into the model. Species is significant in each case, and absence type is significant in the full analysis. Bold values highlight significant results.

Term	*df*	Sum of Squares	*F*	*P*
Performance against PSA				
Projection	3	0.911	8.388	**10^{-5}**
SDM	7	1.735	6.848	**10^{-8}**
Projection × SDM	21	1.765	2.322	**0.001**
Species	17	22.246	36.145	**10^{-16}**
Number of test sites	1	0.221	6.117	**0.014**
Error	526	19.043		
Performance against LCA				
Projection	3	0.041	0.982	0.401
SDM	7	0.472	4.744	**10^{-5}**
Projection × SDM	21	0.208	0.697	0.837
Species	17	5.251	21.716	**10^{-16}**
Test prevalence	1	0.046	3.248	0.072
(Test prevalence)^2	1	0.192	13.547	**10^{-4}**
Error	525	7.468		
HCA				
Projection	3	0.152	1.525	0.206
SDM	7	0.596	2.559	**0.013**
Projection × SDM	21	0.246	0.352	0.997
Species	17	12.286	21.716	**10^{-16}**
Test prevalence	1	0.276	8.318	**0.004**
(Test prevalence)^2	1	0.030	0.908	0.340
Number of test sites	1	0.160	4.806	**0.028**
Error	524	17.439		
Performance against all absence types together				
Absence Type	2	4.993	70.535	**10^{-16}**
Projection	3	0.309	2.905	**0.034**
SDM	7	8.116	32.756	**10^{-16}**
Projection × SDM	21	1.144	1.539	0.056
Species	17	29.811	49.543	**10^{-16}**
Test prevalence	1	0.044	1.243	0.264
(Test prevalence)^2	1	9.413	265.928	**10^{-16}**
Number of test sites	1	0.146	4.117	**0.042**
Error	1674	59.251		

Table A4. Results from variance partitioning of AUC. Results from Table 2 in the main text are replicated here for ease of comparison. Values represent each term's contribution to total R^2. For each absence type a regression with projection, SDM, projection × SDM, and species as factors was analyzed. The species term was then replaced with species-level traits that were expected to influence SDM. Terms were only included in the final partitioning if they were significant ($P \leq 0.05$) in a forwards/backwards model selection procedure. Pluses and minuses in parentheses indicate the direction of the relationship for non-categorical variables in the final model. † Absence type was only included in the analysis using all absence types. †† Prevalence of the test data was only included in analyses with LCA and HCA. ††† Number of test sites was only included in analyses of PSA and HCA. * autecological trait; ns not significant.

	R^2			
Term	PSA	LCA	HCA	All together
Regressions with "species" as a term				
Absence Type †	(not included)	(not included)	(not included)	0.03
Projection	ns	0.01	0.01	0.00
SDM	0.06	0.06	0.05	0.05
Species	0.50	0.42	0.36	0.34
Prevalence + (Prevalence)^2 ††	(not included)	0.06	0.04	ns
Number of test sites †††	0.01	(not included)	0.00 (–)	ns
Total	0.58	0.55	0.47	0.42
Regressions replacing "species" with autecological traits				
Absence Type †	(not included)	(not included)	(not included)	0.03
Projection	ns	0.01	ns	0.00
SDM	0.06	0.06	0.05	0.05
Prevalence + (Prevalence)^2 ††	(not included)	0.06	0.05	0.01
Number of Test Sites †††	0.01 (+)	(not included)	ns	ns
Detectability in Test Era	ns	0.01 (+)	0.01 (–)	ns
Activity Cycle*	0.06	0.10	0.04	0.05
Annual Rhythm*	0.06	0.04	0.03	0.04
Diet*	0.06	0.05	0.03	0.03
Adult Mass*	ns	0.02 (–)	ns	0.01 (–)
Litter Size*	0.04 (+)	0.02 (+)	0.02 (+)	0.02 (+)
Litters per Year*	0.06 (+)	ns	0.02 (–)	ns
Young per Year*	0.06 (–)	0.04 (–)	ns	0.04 (+)
Range Area*	0.02 (+)	0.02 (+)	0.01 (+)	0.01 (+)
Niche (ENFA) Breadth*	0.01 (+)	ns	0.01 (–)	ns
Niche (ENFA) Marginality*	0.10 (+)	0.08 (+)	0.12 (+)	0.10 (+)
Total	0.55	0.52	0.39	0.40
Total of autecological traits	0.48	0.36	0.28	0.30

Table A5. Results from variance partitioning of COR. For each absence type a regression with projection, SDM, projection × SDM, and species as factors was analyzed. The species term was then replaced with species-level traits that were expected to influence SDM performance. Terms were only included in the final partitioning if they were significant ($P \leq 0.05$) in a forwards/backwards model selection procedure. Values indicate each term's contribution to R^2. Pluses and minuses in parentheses indicate the direction of the relationship for non-categorical variables. † Absence type was only included in the analysis using all absence types. †† Prevalence of the test data was only included in analyses with LCA and HCA. ††† Number of test sites was only included in analyses of PSA and HCA. * autecological trait; ns not significant.

	R^2			
Term	PSA	LCA	HCA	All together
Regressions with "species" as a term				
Absence Type †	(not included)	(not included)	(not included)	0.03
Projection	ns	ns	ns	0.00
SDM	0.13	0.05	0.05	0.07
Species	0.46	0.42	0.45	0.29
Prevalence + (Prevalence)^2 ††	(not included)	0.12	ns	0.04
Number of Test Sites †††	ns	(not included)	ns	0.03
Total	0.59	0.60	0.51	0.47
Regressions replacing "species" with autecological traits				
Absence Type †	(not included)	(not included)	(not included)	0.03
Projection	ns	ns	ns	ns
SDM	0.13	0.05	0.05	0.07
Prevalence + (Prevalence)^2 ††	(not included)	0.14	ns	0.06
Number of Test Sites †††	0.01 (+)	(not included)	ns	ns
Detectability in Test Era	0.00 (+)	ns	0.02 (–)	ns
Activity Cycle*	0.07	0.05	ns	0.03
Annual Rhythm*	0.06	0.03	0.02	0.02
Diet*	0.06	0.07	0.10	0.04
Adult Mass*	ns	0.02 (–)	0.01 (+)	ns
Litter Size*	0.03 (+)	0.01 (+)	0.01 (+)	0.01 (+)
Litters per Year*	ns	0.02 (–)	0.01 (–)	0.02 (–)
Young per Year*	0.07 (–)	0.02 (–)	0.01 (+)	0.02 (–)
Range Area*	0.02 (+)	0.02 (+)	0.01 (+)	0.01 (+)
Niche (ENFA) Breadth*	ns	0.03 (–)	0.03 (–)	0.01 (–)
Niche (ENFA) Marginality*	0.11 (+)	0.09 (+)	0.14 (+)	0.10 (+)
Total	0.56	0.55	0.42	0.43
Total of autecological traits	0.41	0.35	0.34	0.24

Correlations between within- and cross-era performance

In the main text we present analyses using within-era AUC to predict cross-era AUC where the latter is measured against HCA data. Here we expand those results to comparisons of other types of absences (Table A6). The correlation analysis indicates that no SDM but BIOCLIM consistently predicts cross-era performance using within-era performance across all absence type combinations (Table A6). Unfortunately, BIOCLIM is one of the poorest-performing models, regardless of the absence type of the test set (Fig. 3b). The results suggest that LCA within-era performance is a better predictor of cross-era performance against HCA than is within-era HCA performance (Table A6, top two middle columns vs. top two right columns). The ability to predict performance of a projection in one direction (e.g., from historic to modern) did not necessarily imply equivalent ability to predict in the other direction (modern to historic; e.g., compare GAM PSA vs. GAM HCA).

Table A6. Pearson correlation coefficients between AUC of within-era projections vs. AUC of cross-era projections. This table contains the data from Table 2 in the main text and expands it for more absence types. Stronger correlations indicate performance of a cross-era projection can be predicted from performance of a within-era projection. When absence types are different (e.g., PSA vs. HCA), the within-era performance is against the lesser-quality data set (PSA in this example) and cross-era performance is against the higher-quality data set (HCA). Bolded values are significant at $P \leq 0.05$ ($n = 18$ in each case).

	PSA AUC (within-era) vs. PSA AUC (cross-era)		LCA AUC (within-era) vs. LCA AUC (cross-era)		HCA AUC (within-era) vs. HCA AUC (cross-era)	
SDM	HH vs. HM	MM vs. MH	HH vs. HM	MM vs. MH	HH vs. HM	MM vs. MH
BIOCLIM	**0.73**	**0.73**	**0.75**	**0.82**	**0.67**	**0.80**
BRT	**0.87**	0.36	**0.69**	0.46	**0.52**	**0.47**
GAM	**0.88**	**0.68**	**0.78**	**0.69**	0.46	0.36
GLM	**0.85**	**0.83**	**0.65**	**0.92**	0.42	**0.67**
MAXENT	**0.85**	**0.75**	**0.88**	**0.86**	**0.47**	0.42
SVM	**0.81**	**0.62**	**0.73**	**0.80**	**0.63**	**0.58**
EMEAN	**0.89**	**0.67**	**0.74**	**0.75**	0.44	0.40
EMED	**0.87**	**0.70**	**0.78**	**0.70**	0.42	0.28

	PSA AUC (within-era) vs. LCA AUC (cross-era)		PSA AUC (within-era) vs. HCA AUC (cross-era)		LCA AUC (within-era) vs. HCA AUC (cross-era)	
SDM	HH vs. HM	MM vs. MH	HH vs. HM	MM vs. MH	HH vs. HM	MM vs. MH
BIOCLIM	**0.58**	**0.82**	**0.53**	**0.82**	**0.70**	**0.82**
BRT	**0.50**	**0.56**	0.33	0.40	**0.65**	0.44
GAM	**0.57**	**0.57**	0.22	**0.47**	**0.63**	**0.57**
GLM	0.13	**0.61**	-0.09	0.46	**0.64**	**0.83**
MAXENT	**0.63**	0.41	0.42	0.25	**0.68**	**0.69**
SVM	**0.71**	0.36	**0.68**	0.24	**0.73**	**0.75**
EMEAN	**0.50**	**0.49**	0.28	0.30	**0.62**	**0.64**
EMED	**0.52**	**0.49**	0.30	0.26	**0.66**	**0.56**

Transferability of models across eras

To compare values of AUC calculated against the HCA, we applied a modified version of the transferability index (*TI*) from Randin et al. (2006). The original *TI* was developed for comparison of AUC between regions or eras, and was intended to have the range [0, 1], where 1 indicates that the model has perfect transferability (the models' performance when trained and tested in era A and is equal to its performance in when tested in era B and vice versa), and 0 indicates that the model transfers to the opposing era very poorly. The formula for *TI* in Randin et al. (2006) is based on the assumption that AUC has the range [0.5, 1], but AUC actually ranges from 0 to 1 (Mason & Graham, 2002*)*, making the original *TI* score range from -1 to 1. Here we present a modified *TI* index, *TI′*, which is based on the value being compared (AUC, threshold, etc.) having the range [0, 1]:

Eq. A1
$$TI' = \frac{\frac{1}{2}[(1 - |AA - AB|) + (1 - |BB - BA|)]}{1 + \big||AA - AB| - |BB - BA|\big|}$$

where *AA*, *BB*, *AB*, and *BA* are the values of AUC (or threshold, etc.) corresponding to the HH, MM, HM, and MH projections, respectively. *TI'* has the range [0, 1] and is appropriate for comparing transferability of model performance within an absence type but not across absence types since performance varied by absence type. Note that high transferability does not necessarily mean the model performs well, just that its performance within an era is nearly equal to its performance across eras.

Average transferability of AUC was 0.87±0.01 using the PSA test data, 0.90±0.00 against LCA, and 0.86±0.01 against HCA. Though the modified transferability index varied between SDMs within an absence type, none of the differences were significant (one-way ANOVA for each absence type, $P>0.19$ in each case, $df=7$), suggesting that performance within an absence type can be predicted equally well across SDMs (Fig. A1).

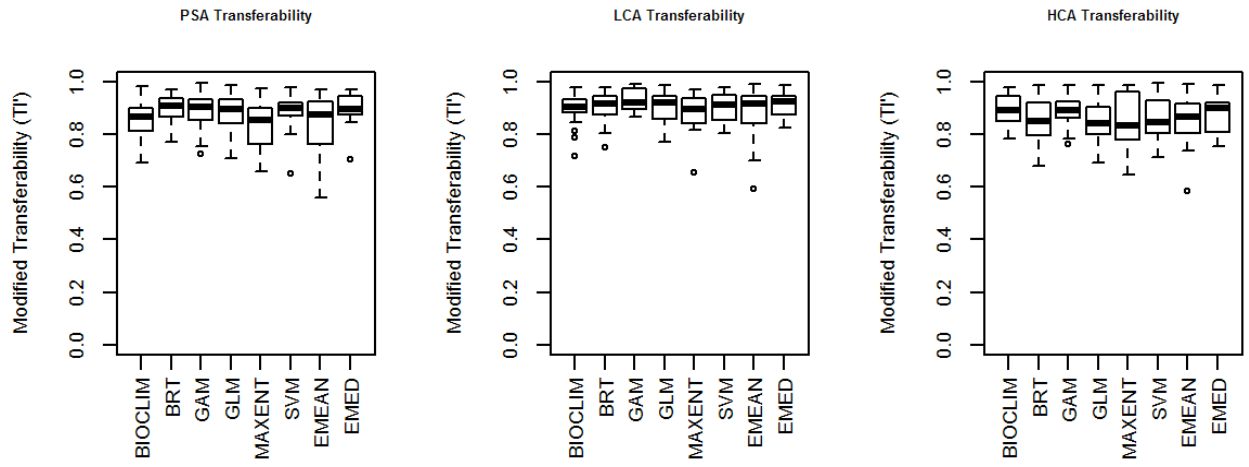

Figure A1. Transferability of SDMs within absence types. Each panel shows the modified transferability of AUC (Eq. A1) for each SDM across the 18 species. Within an absence type none of the differences between SDMs was significant. Tops of boxes, horizontal lines within boxes, and bottoms of boxes represent the upper 75%, median, and lower 25% quartiles, respectively. Dashed vertical lines extend to the lesser/greater of the maximum/minimum value and 2 standard derivations from the mean.

Turnover at Grinnell sites

Overall mean turnover (number of sites where a species changed status from present to absent or absent to present as a proportion of all sites where a species' status was certain—detected or a high-confidence absence) was 0.17±0.03 (Table A7). Turnover was unrelated to average cross-era HCA AUC (the mean of HM and MH AUC) except for SVMs, for which the relationship was negative ($r=-0.62$, $P=0.01$, $n=18$).

Table A7. Turnover across the 61 matching Grinnell sites surveyed in both historic and modern times. Values are number of sites experiencing no turnover (present in both eras or absent in both eras), extirpation (present-absent), or colonization (absent-present). Turnover is calculated as the number of sites at which there was a change in status divided by the total number of sites at which the status of the species could be confidently assigned in both eras. The total number of sites for each species is <61 because some sites were excluded for each species because they belonged to the low-confidence absence (LCA) set (the species was undetected at the site and the probability of false absence was ≤0.10).

	Status of species at sites in each era				
Species	Present-Present	Absent-Absent	Present-Absent	Absent-Present	Turnover
Callospermophilus lateralis	20	3	4	0	0.15
Chaetodipus californicus	4	6	1	1	0.17
Microtus californicus	10	18	1	1	0.07
Microtus longicaudus	28	5	5	5	0.23
Microtus montanus	11	17	5	4	0.24
Neotoma fuscipes	2	16	2	1	0.14
Neotoma macrotis	5	2	3	1	0.36
Peromyscus boylii	17	19	8	2	0.22
Peromyscus truei	8	7	1	3	0.21
Peromyscus maniculatus	48	1	7	3	0.17
Reithrodontomys megalotis	8	24	1	1	0.06
Sorex monticolus	15	5	1	5	0.23
Sorex vagrans	5	28	0	2	0.06
Tamias amoenus	6	39	2	0	0.04
Tamias senex	5	1	0	0	0.00
Tamias speciosus	21	12	7	4	0.25
Urocitellus beldingi	3	19	0	2	0.08
Zapus princeps	10	4	2	8	0.42
Mean					0.17
SE					0.03

Analysis of omission and commission error rates using the threshold that minimizes the difference between sensitivity and specificity (MDSS)

In the main text of the paper we present analyses of omission and commission rates using the threshold that maximizes the sum of sensitivity and specificity (MSSS; Fig. 4). Here we present the same analysis for the threshold that minimizes the difference between sensitivity and specificity

(MDSS). Table A8 presents the same results for analyses of variance of omission and commission rates using the MSSS threshold as in Table 5 but also includes the analysis of variance for the MDSS threshold. As with the MSSS threshold, projection (and/or interactions between projection and other factors) was always a significant factor in the models, unlike in the analysis of threshold-independent performance (Tables 1 and Fig. 3). Fig. A2 shows omission and commission error rates using the MDSS threshold. Although the MDSS threshold makes omission and commission rates as equal as possible, they are not strictly equal in each case.

Table A8. Analyses of variance on omission and commission error rates for the MSSS and MDSS thresholds (some results repeated from Table 5 for ease of comparison). Bold values highlight significant factors. See also Figs. 4 and A3.

Source	*df*	Sum of Squares	*F*	*P*
MSSS Threshold: Omission Error Rate				
Absence type	2	2.985	5.022	**0.007**
Projection	3	11.281	12.655	**10^{-8}**
SDM	7	3.774	1.814	0.080
Species	17	167.272	33.112	**10^{-16}**
SDM × Projection	21	19.733	3.162	**10^{-6}**
SDM × Absence type	14	1.730	0.416	0.970
Projection × Absence type	6	14.072	7.892	**10^{-8}**
Error	1657	492.4		
MSSS Threshold: Commission Error Rate				
Absence type	2	4.356	9.178	**10^{-4}**
Projection	3	9.057	12.722	**10^{-8}**
SDM	7	1.679	1.011	**0.421**
Species	17	85.485	21.191	**10^{-16}**
SDM × Projection	21	13.318	2.672	**10^{-5}**
SDM × Absence type	14	1.758	0.529	0.917
Projection × Absence type	6	2.614	1.836	0.088
Error	1657	393.2		
MDSS Threshold: Omission Error Rate				
Absence type	2	1.774	7.513	**0.001**
Projection	3	4.899	13.833	**10^{-9}**
SDM	7	1.971	2.385	**0.019**
Species	17	58.787	29.295	**10^{-16}**
SDM × Projection	21	9.506	3.835	**10^{-8}**
SDM × Absence type	14	1.348	0.815	0.652
Projection × Absence type	6	0.650	0.917	0.481
Error	1657	195.59		
MDSS Threshold: Commission Error Rate				
Absence type	2	0.839	3.197	**0.041**
Projection	3	4.466	11.339	**10^{-7}**
SDM	7	1.492	1.623	0.124
Species	17	77.594	34.764	**10^{-16}**
SDM × Projection	21	10.382	3.765	**10^{-8}**
SDM × Absence type	14	0.775	0.421	0.969
Projection × Absence type	6	0.131	0.166	0.985
Error	1657	217.55		

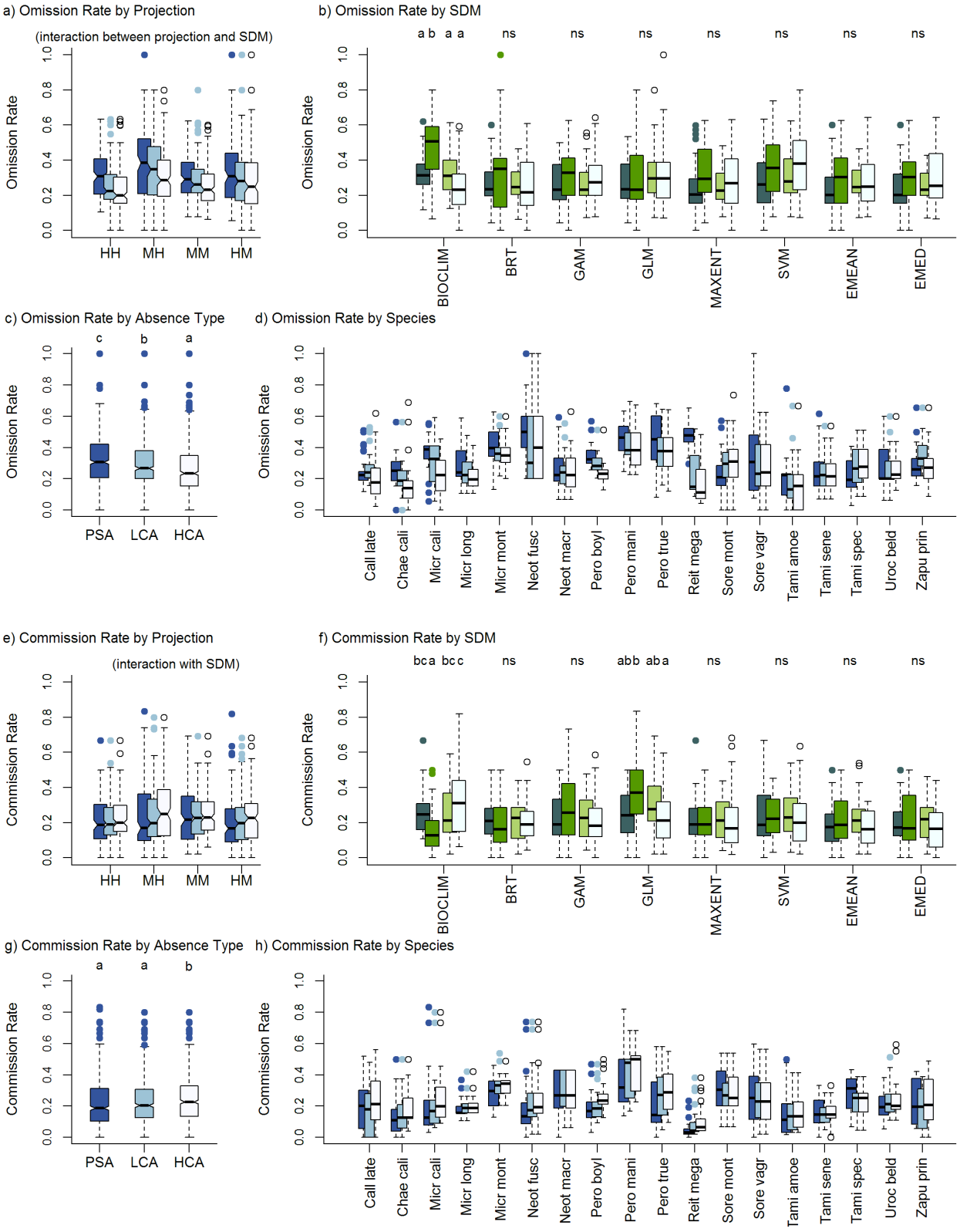
a) Omission Rate by Projection
(interaction between projection and SDM)
Omission Rate
HH
MH
MM
HM
b) Omission Rate by SDM
a b a a
ns
BIOCLIM
BRT
GAM
GLM
MAXENT
SVM
EMEAN
EMED
c) Omission Rate by Absence Type
c
b
a
PSA
LCA
HCA
d) Omission Rate by Species
Call late
Chae cali
Micr cali
Micr long
Micr mont
Neot fusc
Neot macr
Pero boyl
Pero mani
Pero true
Reit mega
Sore mont
Sore vagr
Tami amoe
Tami sene
Tami spec
Uroc beld
Zapu prin
e) Commission Rate by Projection
(interaction with SDM)
Commission Rate
f) Commission Rate by SDM
bca bcc
abb aba
g) Commission Rate by Absence Type
h) Commission Rate by Species

Figure A2. Analysis of omission (a-d) and commission (e-h) error rates using the threshold that minimizes the difference between sensitivity and specificity (MDSS; see Fig. 4 for results from the threshold that maximizes the sum of sensitivity and specificity). In panels (b) and (f) the darkest bars represent the HH projection, the second darkest the MH projection, the second lightest the MM projection, and the lightest the HM projection. In all other panels the darkest bars represent the PSA set, the gray bars the LCA set, and the lightest bars the HCA set. Projection had a significant interaction with SDM (a), so significance groupings are not noted. In (b) only BIOCLIM had significantly different omission rates when projecting within- vs. across eras. Absence type significantly affected omission rates (c). Species was a significant factor in omissions and commissions (d) and (h), but significance groupings are not shown to aid visual clarity. Commission rates varied by a combination of projection and SDM (e and f) and by absence type (g). Omission or commission error rate increases with the order of the significance group code (e.g., group "a" has the lowest error rate within a comparison, group "b" the second-lowest, etc.). Abbreviations: Call late: *Callospermophilus lateralis*, Chae cali: *Chaetodipus californicus*, Micr cali: *Microtus californicus*, Micr long: *M. longicaudus*, Micr mont: *M. monticolus*, Neot fusc: *Neotoma fuscipes*, Neot macr: *N. macrotis*, Pero boyl: *Peromyscus boylii*, Pero mani: *P. maniculatus*, Pero true: *P. truei*, Reit mega: *Reithrodontomys megalotis*, Sore mont: *Sorex monticolus*, Sore vagr: *S. vagrans*, Tami amoe: *Tamias amoenus*, Tami sene: *T. senex*, Tami spec: *T. speciosus*, Uroc beld: *Urocitellus beldingi*, Zapu prin: *Zapus princeps*. ns = not significant.

Transferability of thresholds across absence types and time

We found that some models which performed well in threshold-independent assessments (e.g., BRTs) performed poorly in threshold-dependent assessments of omission and commission rates (Figs. 4b and A2). We hypothesized that this occurred because models differed in the transferability of their thresholds across eras, which would cause presences and absences to be differentiated less than optimally given a certain rule (e.g., maximization of the sum of sensitivity and specificity or MSSS, or minimization of the difference between sensitivity and specificity or MDSS). As a post hoc test of this hypothesis, we calculated modified transferability scores (Eq. A1) for thresholds within and across absence types and eras. We use within-era PSA, LCA, or HCA thresholds as values of *AA* and *BB* in Eq. A1 and cross-era values of HCA thresholds for *AB* and *BA*. High values of the index indicate near-equality of the threshold between the sets being compared.

Fig. A3 displays the transferability of the MSSS and MDSS threshold using within-era PSA, LCA, or HCA test data to calculate the threshold, then comparing it to the cross-era HCA threshold.

Regardless of the type of absence or threshold uses, BRTs consistently had the highest transferability, meaning that the value of a threshold determined using PSA, LCA, or HCA data within an era was nearly equal to the value had it been calculated using HCA data from the opposing era. In contrast most other models displayed poorer transferability between eras and absence types.

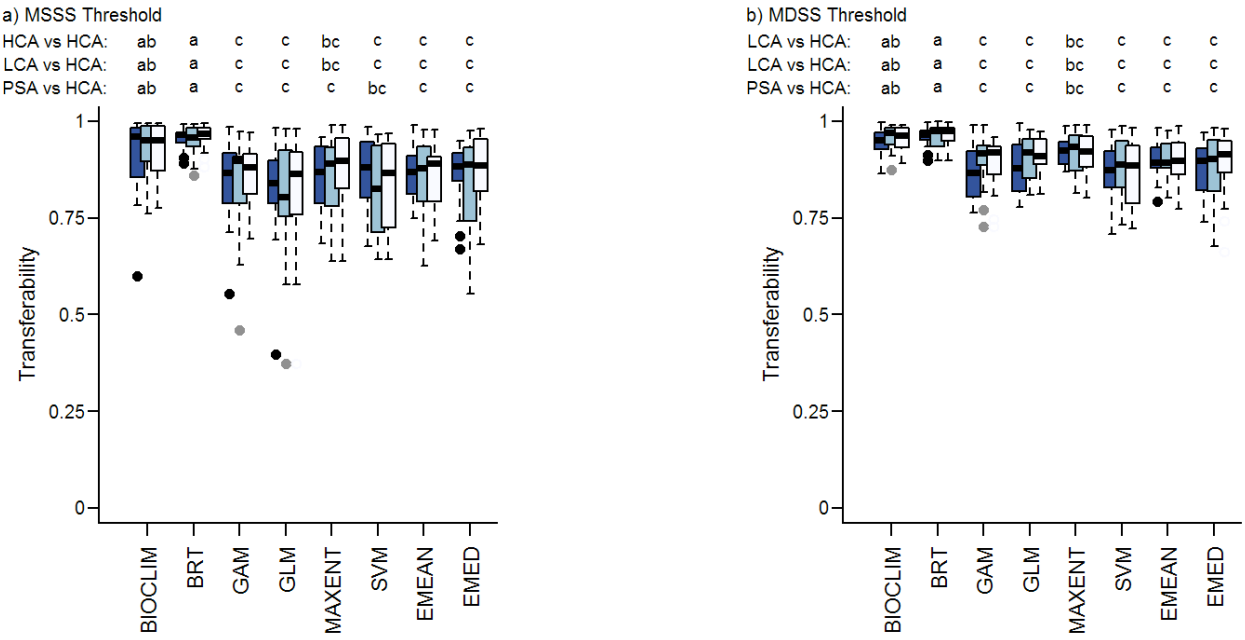

Figure A3. Transferability of thresholds *across* eras and between absence types for the MSSS (a) and MDSS (b) thresholds. Thresholds calculated for three data types (PSA, LCA, and HCA) using within-era test data were compared to thresholds calculated for the cross-era test HCA data. Significance groups within each absence-type pair are shown, with the order of the letters indicating rank of mean transferability (e.g., category "a" has the highest transferability, "b" the second-highest, etc.). There is a remarkable consistency in significance groups between comparisons within and between threshold types (these are not typos!). Darkest bars represent transferability of PSA within-era thresholds to HCA cross-era thresholds, lighter bars represent transferability of LCA to HCA thresholds, and the lightest bars represent transferability of within-era HCA thresholds to cross-era HCA thresholds. Overall BRTs had the consistently highest transferability between absence types and eras.

Finally, we were interested in determining the transferability of thresholds *within* an era when lesser-quality test data is used to calculate the threshold. Modelers may not always have access to HCA data

and so must assume that thresholds calculated against lower-quality data adequately delineate presences from absences. In this case we compared thresholds from PSA and LCA data for HH and MM projections to thresholds calculated using HCA data for the same projections, meaning that *AA* and *BB* in Eq. A1 was either the PSA or LCA threshold for the HH or MM projection and *AB* and *BA* were the HCA threshold in the HH or MM projections.

We found that BIOCLIM and BRTs had the highest within-era transferability of thresholds across absence types, except for using the LCA MSSS threshold to estimate the HCA MSSS threshold, for which all models performed equally well (Fig. A4). Using the LCA threshold to estimate the HCA threshold was better than using the PSA threshold to estimate the HCA threshold.

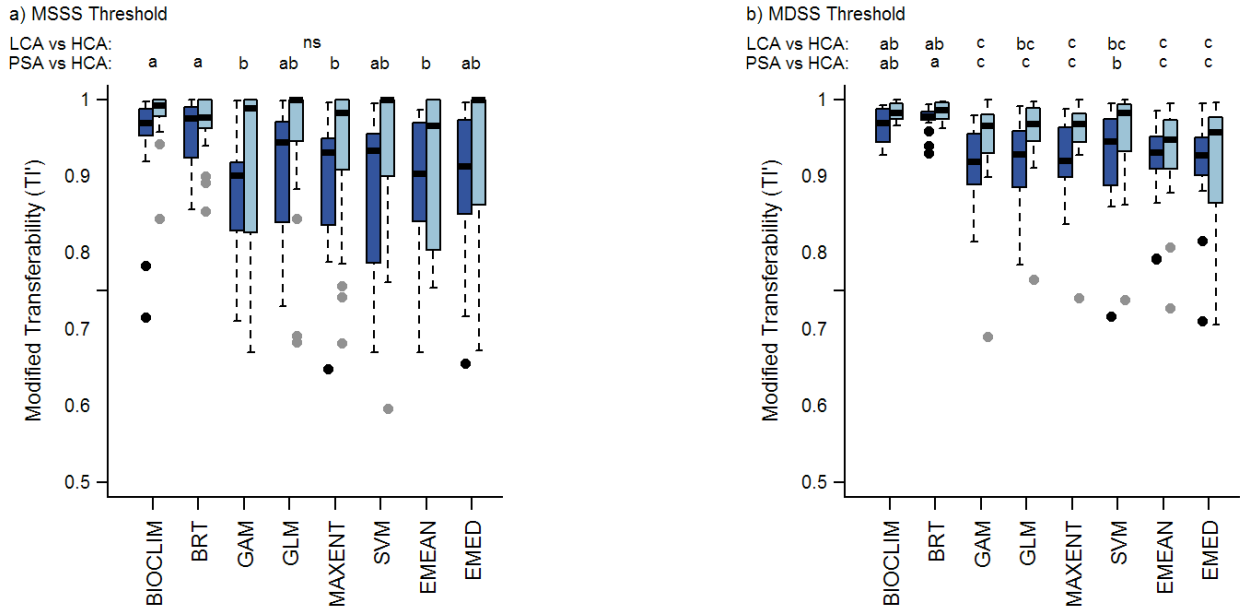

Figure A4. Transferability of thresholds between different absence types *within* the same era. (a) Transferability of the MSSS threshold. (b) Transferability of the MDSS threshold. Thresholds calculated using lower-quality absence types (PSA or LCA) were used to estimate thresholds calculated for the same within-era projections using HCA data. Dark bars represent PSA thresholds used to estimate HCA thresholds. Lighter bars represent LCA thresholds used as proxies for HCA thresholds. Significance groups are shown within each pair of absence types and order denotes rank of the average transferability of a group (e.g., category "a" has the highest transferability, "b" the second-highest, etc.).

Supplemental literature cited

Elith, J., C.H. Graham, R.P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J.McC. Overton, A.T. Peterson, S.J. Phillips, K. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soberón, S. Williams, M.S. Wisz, and N.E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

Mason, S.J. and Graham, N.E. 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society 128, 2145-2166.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M., and Guisan, A. 2006. Are niche-based species distribution models transferable in space? Journal of Biogeography 33:1689-1703.