Supporting Information. Healy Hamilton, Regan L. Smyth, Bruce E. Young, Timothy G. Howard, Christopher Tracey, Sean Breyer, Richard Cameron, Anne Chazal, Amy K. Conley, Charles Frye, and Carrie Schloss. Increasing taxonomic diversity and spatial resolution clarifies opportunities for protecting US imperiled species. Ecological Applications.

Appendix S1: Modeling Methods

Inductive Models

All model scripts were developed in R (Microsoft and R Core Team 2018 with repository snapshot 2019-02-01, R Core Team 2018) and are available on GitHub (https://github.com/HeritageNetwork/Regional_SDM) as well as in Data S1 (terrestrial models) and Data S2 (riverine models). R packages used include abind (Plate and Heiberger 2011), DBI (R Special Interest Group on Databases (R-SIG-DB) et al. 2018), doParallel (Corporation and Weston n.d.), dplyr (Wickham et al. 2018), fasterize (Ross 2018), foreign (R Core Team 2017), here (Muller 2017), iterators (Analytics and Weston n.d.), knitr (Xie 2018), maptools (Bivand and Lewin-Koh 2018), odbc (Hester and Wickham 2018), OpenStreetMap (Fellows and Stotz 2016), randomForest (Liaw and Wiener 2002), raster (Hijmans 2018), rasterVis (Perpiñán and Hijmans 2018), RColorBrewer (Neuwirth 2014), ROCR (Sing et al. 2005), RSQLite (Muller et al. 2018), sf (Pebesma 2018), smoothr (Strimas-Mackey 2018), snow (Tierney et al. 2018), snowfall (Knaus and Knaus 2015), stringi (Gagolewski 2018), tables (Murdoch 2018), tmap (Tennekes 2018a), tmaptools (Tennekes 2018b), vcd (Meyer et al. 2010), and xtable (Dahl et al.
All model scripts were developed in R (Microsoft and R Core Team 2018 with repository snapshot 2019-02-01, R Core Team 2018a) and are available on GitHub (https://github.com/HeritageNetwork/Regional_SDM). R packages used include abind (Plate and Heiberger 2016), DBI ( R-SIG-DB et al. 2018), doParallel (Calaway et. al. 2017), dplyr (Wickham et al. 2018), fasterize (Ross 2018), foreign (R Core Team 2018b), here (Müller 2017), iterators(Revolution Analytics and Weston 2018), knitr (Xie 2018), maptools (Bivand and Lewin-Koh 2018), odbc (Hester and Wickham 2018), OpenStreetMap (Fellows and Stotz 2016), randomForest (Liaw and Wiener 2002), raster (Hijmans 2018), rasterVis (Perpiñán and Hijmans 2018), RColorBrewer (Neuwirth 2014), ROCR (Sing et al. 2005), RSQLite (Müller et al. 2018), sf (Pebesma 2018), smoothr (Strimas-Mackey 2018), snow (Tierney et al. 2018), snowfall (Knaus2015), stringi (Gagolewski 2018), tables (Murdoch 2018), tmap (Tennekes 2018a), tmaptools (Tennekes 2018b), vcd (Meyer et al. 2010), and xtable (Dahl et al. 2018).

The riverine and terrestrial/palustrine modeling frameworks took a similar approach, with slight differences based on the structure of the training data and the modeling unit. Our overall approach followed standard modeling methods: attribute presence and background records, variable removal, model validation, final model development, predict model to modeling area, model review, and output final spatial representation of predicted suitable habitat based on review.

Variable removal, model validation, and the final models (below) were all built using the randomForest package (Liaw and Wiener 2002) in R. The number of variables randomly drawn to test at each split (mtry) was chosen to minimize out-of-bag (OOB) error. To minimize pseudoreplication we defined a sample size for groups of points (defined below) based on representation accuracy (RA, see
http://help.natureserve.org/biotics/Content/Record_Management/Element_Occurrence/EO_Representation_Accuracy_Value.htm), with very high, high, medium, and low RA groups getting 5, 4, 3, and 2 points drawn from them, respectively.

Location records for terrestrial species were grouped spatially such that known locations that were close together were assigned to the same group. These groupings were already defined in NatureServe’s Biodiversity Location Data (NatureServe, unpublished data) following the element occurrence data standard (https://www.natureserve.org/conservation-tools/standardsmethods/element-occurrence-data-standard) and represented as one or more polygons. For species where we used additional point-based observation data, points within 1000 meters of each other were assigned to the same group. We converted these points to polygon by buffering each point 90 meters and merging the resulting polygons. These polygons were then merged with any element occurrence polygons from the BLD with the result being a polygon-based representation of species occurrence with each polygon assigned to a group. To sample the environmental variability within these occurrence polygons, we placed random points within each polygon at approximately one point per 30 m cell for very small polygons but the total number of points being capped at 400 such that even large polygons never received over 400 points (Howard and Schlesinger 2013). The points maintained their grouping attributions, allowing us to sub-select a very small number of points by group while modeling, minimizing pseudo-replication while maximizing our ability to capture the environmental heterogeneity within each presence polygon.

Riverine species were prepped into groups for the models by assigning species occurrence locations to flowlines and including all occurrence flowlines within a specified upstream/downstream separation distance (typically two kilometers) to the same group. We
manually removed extraneous stream reaches (e.g. side tributaries) that may have been selected in the previous step. We attributed all presence records with the full set of terrestrial or freshwater environmental predictors.

We also attributed 20 million randomly placed points throughout the CONUS to use as background records available to our terrestrial modeling algorithms. We used the full set of 2.6 million flowlines as the background records available to our riverine methods. While modeling, only those records occurring within the modeling area for the species were used as the pool of records to randomly draw from during model iterations. Also, background samples coincident or near presence data (30 meters on land, or one reach up or downstream) were removed from the background pool of records before modeling.

From the compiled set of attributed presence and background records, we removed environmental variables in a three-step process. First, distance-to variables (e.g. distance to wetland, distance to calcareous bedrock geology) were removed if all presence records had distances greater than 5,000 meters. This step only applied to terrestrial models, as there were no distance-to variables in the riverine framework. Second, we assessed correlations among variables, *a priori*, and used this information to remove less-important correlated variables by building a preliminary random forest model and retaining only the top-importance variable within each set of correlated variables. Finally, using the same preliminary model, we removed the variables with the lowest importance, based on mean decrease in accuracy (Liaw and Wiener 2002) of the overall model. In most cases, we removed the bottom 25th percentile of variables. This preliminary model was built using 1,000 classification trees.

Using the groups assigned earlier, we used a jackknifing procedure called leave-one-out validation (Fielding 2002) as external validation of the modeling process. This procedure
consisted of removing all records of one group from the data set, building a model using the remainder of the data, and then testing the model (running a prediction) on those data left out. With each group held out, an equal-sized set of background points was randomly drawn and also held out and used in the validation set. We held out each group in turn, cycling through them all for up to 50 groups. We compiled the mean and error (SD, SEM) for overall accuracy, specificity, sensitivity, true skill statistic, kappa, and area under the curve (Fielding and Bell 1997) for all the jackknifing runs, combined.

We built a final model using all input groups and 2,000 trees in the forest. Each tree was built using all groups but only one to five randomly drawn points from each group such that the number drawn was weighted by the representation accuracy of the polygons defining the group. The number of background records drawn from the full pool of background points was set equal to the number of presence points. Thus, for each tree in the forest, a different set of presence and background points were drawn from the full pool of available presence and background points (using the 'sampsize' argument in the randomForest call).

Once a model of the relationship between presence and background points was built, we then predicted the probability that the environment at each raster cell or stream reach throughout the modeling area was similar to the environment at known locations. The result is a continuous raster surface or stream network of probability values within the modeling area. As the riverine modeling unit was the vector flowlines, but the NHD also includes polygon features that represent wider rivers and lakes that are linked to those flowlines, we applied the average probability value of the linked ‘artificial path’ to each polygon feature. Both vector and polygon features were included in all subsequent analyses.
We converted the continuous representation to a binary representation of predicted habitat/non-habitat by selecting a probability threshold above which all values were classified as habitat. We calculated a suite of thresholds, including equal sensitivity and specificity, maximum of sensitivity plus specificity, minimum training presence, minimum training presence by group, and tenth percentile of training presences. For riverine models, our default threshold for model review was defined to include 90% of input training reaches (called 10th percentile threshold). For terrestrial and palustrine models, we used as our default the highest threshold possible while still including at least one presence record from each group (minimum training presence by group; MTPG). In cases where the number of presence records classified as present was very low (proportion of presence records classified as present less than 0.4), we used the same tenth percentile threshold as the riverine methods.

Model run information was stored in an SQLite tracking database. A detailed metadata document was also generated for each species, including information on training data, variable importance, thresholds, validation statistics, and evaluation of the modeling components using a rubric specifically designed for communication of model attributes and appropriate use (Sofaer et al. 2019). These species-specific metadata are available at the ArcGIS Living Atlas of the World (https://livingatlas.arcgis.com).

**Model Review & Development of Alternative Distribution Maps**

Field biologists from state Natural Heritage Programs familiar with the species modeled were asked to review the thresholded output of the habitat suitability modeling process and register comments via an online model review tool. From the comments we decided whether to (1) use
the model as-is, (2) revise the model via threshold adjustments, clipping of extent, or other means, or (3) pursue development of an alternative map for the species.

Models that were not reviewed via this process were either species with existing vetted models, species for which alternative maps were developed based on biologist input (though some alternative maps did receive review), or species for which no reviewers were identified. For the latter, we took several steps to ensure that final habitat models were as accurate as possible. For unreviewed fish, mussels, and freshwater crayfish (148 species) we used HUC-8 range maps recently completed by NatureServe to clip the modeling extent, retaining modeled habitat only within mapped watershed boundaries. For other unreviewed vertebrates and pollinators, we manually reviewed each model, visually examining the model results in light of the element occurrence records and habitat description information maintained in NatureServe Explorer (http://explorer.natureserve.org), and, when available, comparing our outputs to USGS Gap Analysis Project (GAP) habitat maps (USGS 2019). For plants and terrestrial crayfish not reviewed by field biologists, we reviewed the representational accuracy of training data, the number of training records, the threshold selected (review of other species models suggested models using the Tenth Percentile threshold were more likely to need review), the total amount of habitat predicted, the difference between modeled habitat area and area of occupancy as reported on NatureServe Explorer, and model validation statistics to make adjustments as necessary. Models for 197 species without any of these indicators of possible modeling inaccuracies did not receive review.

Following review, we clipped all models to remove predicted habitat from any HUC-10 watersheds that reviewers indicated were unlikely to support the species. Guided by reviewer
feedback, we adjusted threshold values for 522 species to expand or constrict areas of predicted suitable habitat.

We developed 243 alternative models for species identified as problematic during the model review stage or for those identified early in the project as poor candidates for modeling because of a paucity of locality data or life history traits not well matched to available predictor data (e.g., cave species). Alternative models fell in the categories listed below. The specific approach taken for each individual species is captured in the species-specific metadata.

- **Freshwater species with too few locality records (e.g. <3 reach groups).** We developed habitat maps for these species by selecting all stream reaches that intersected locality data as well as one additional upstream/downstream reach.

- **Bat generalist species.** We intersected the species’ IUCN range map with HUC-10 polygons (NHD Plus V2). The selected HUC-10 polygons were dissolved and used as a mask for the alternate map. Within this mask, we reclassified landcover from the National Land Cover Database (Homer et al. 2015) as either suitable habitat or not using information about the species’ habitat in NatureServe Explorer.

- **Karst species.** We created habitat maps for karst species by applying a 250-m buffer to the locality data. Due to varying precision of locality data, some species contained large polygons, and in these cases, we applied a smaller buffer.

- **Terrestrial species with only a single locality record.** For species known from only a single locality, we converted the polygon record from the NatureServe Biodiversity Location Data directly to a raster.
• *Species with no reliable training data.* In the absence of any reliable training data, alternative maps were delineated based on maps in the published scientific literature (see *Additional sources for range maps*, below).

• *Other terrestrial species with poorly performing models.* For species in which model parameters could not be modified to generate an acceptable habitat model, an alternative map was developed through one the following three approaches.

  1. For species that were well surveyed and represented in the NatureServe Biodiversity Location Data, polygons were converted directly to raster format as for species with only a single locality record.

  2. If a reasonable USGS GAP model existed (i.e., it overlapped most of the locality records without markedly overextending the range), the GAP habitat raster was used as the habitat model. If the GAP model extent was inconsistent with the current accepted species range (as was sometimes the case due to changing taxonomic concepts), we clipped the GAP model to known range boundaries using IUCN range maps.

  3. For more widespread and/or less well surveyed species, we developed a deductive model using information about species habitat requirements to define environmental parameters from land cover, elevation, and terrain ruggedness layers.

**References cited**

Bivand, R., and M. Lewin-Koh. 2018. maptools: Tools for handling spatial objects. R Package
Calaway, R., Microsoft Corporation, S. Weston, and D. Tenenbaum. 2017. “doParallel”: foreach parallel adaptor for the parallel package. R Package Version 1.0, 11. https://CRAN.R-project.org/package=doParallel

Dahl, D. B., D. Scott, C. Roosen, A. Magnusson, and J. Swinton. 2018. xtable: Export Tables to LaTeX or HTML. https://CRAN.R-project.org/package=xtable

Fellows, I., and J. P. Stotz. 2016. OpenStreetMap: Access to open street map raster images. R package version 0.3.3. https://CRAN.R-project.org/package=OpenStreetMap

Fielding, A. H. 2002. What are the appropriate characteristics of an accuracy measure. Pages 271–280 in J. M. Scott, P.J. Hegland, and M. L. Morrison, editors. Predicting species occurrences: issues of accuracy and scale. Island Press, Washington, D.C., USA.

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

Gagolewski, M. 2018. stringi: Character String Processing Facilities, R package version 1.2.4. http://www.gagolewski.com/software/stringi/

Hester, J., and H. Wickham. 2018. odbc: Connect to ODBC Compatible Databases (using the DBI Interface). R package version, 1.1.6. https://CRAN.R-project.org/package=odbc

Hijmans, R. J. 2018. raster: Geographic Data Analysis and Modeling. R package version 2.8-4. https://CRAN.R-project.org/package=raster

Homer, C. G., J. A. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N. D. Herold, et al. 2015. Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. Photogrammetric Engineering and Remote Sensing 81:345–354.
Howard, T.G., and M. D. Schlesinger. 2013. Wildlife habitat connectivity in the changing climate of New York's Hudson Valley. Annals of the New York Academy of Sciences 1298:103–119.

Knaus, J., 2015. snowfall: Easier cluster computing (based on snow). R package version 1.84-6.1. https://CRAN.R-project.org/package=snowfall

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R news 2:18–22.

Meyer, D., A. Seileis, and K. Hornik. 2010. vcd: Visualizing Categorical Data. R package version 1.2-8.

Microsoft and R Core Team. 2018. The enhanced R distribution from Microsoft. R: A language and environment for statistical computing, version 3.5.1. R Foundation for Statistical Computing, Vienna, Austria and Microsoft, Redmond, Washington, USA.

Müller, K., 2017. here: A simpler way to find your files. R package version 0.1. https://CRAN.Rproject.org/package=here

Müller, K., H. Wickham, D. A. James, and S. Falcon. 2018. RSQLite: ‘SQLite’interface for R. R package version 2.1.1. https://CRAN.R-project.org/package=RSSQLite

Murdoch, D. 2018. tables: Formula-driven table generation. R package version 0.8.7. https://CRAN.R-project.org/package=tables

Neuwirth, E. 2014. RColorBrewer: ColorBrewer palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer

Pebesma, E. 2018. Simple features for R: standardized support for spatial vector data. The R Journal, 10:439-446. https://journal.r-project.org/archive/2018/RJ-2018-009/

Perpiñán, O., and R. Hijmans. 2018. rasterVis. R package version 0.45. https://CRAN.Rproject.org/package=rasterVis
Plate, T., and R. Heiberger. 2016. abind: Combine multi-dimensional arrays. R package version 1.4-5. https://CRAN.R-project.org/package=abind

R Core Team 2018a. R: A language and environment for statistical computing, version 3.5.1. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team 2018b. foreign: Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', .... R package version 0.8-71. https://CRAN.R-project.org/package=foreign

R Special Interest Group on Databases (R-SIG-DB), H. Wickham, and K. Müller. 2018. DBI: R database interface. R package version 1.0.0. https://CRAN.R-project.org/package=DBI

Revolution Analytics and S. Weston. 2018. iterators: Provides Iterator Construct. R package version 1.0.10. https://CRAN.R-project.org/package=iterators

Ross, N. 2018. fasterize: Fast Polygon to Raster Conversion. R package version 1.0.0 https://CRAN.R-project.org/package=fasterize

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21:3940–3941.

Sofaer, H. R., C. S. Jarnevich, I. S. Pearse, R. L. Smyth, S. Auer, G. L. Cook, T. C. Edwards, Jr., G. F. Guala, T. G. Howard, J. T. Morisette, and H. H. Hamilton. 2019. Development and delivery of species distribution models to inform decision-making. BioScience 69:544–557.

Strimas-Mackey, M. 2018. smoothr: Smooth and Tidy Spatial Features. R package version 0.0.1 https://CRAN.R-project.org/package=smoothr

Tennekes, M. 2018a. tmap: Thematic Maps in R. Journal of Statistical Software 84:1–39. doi: 10.18637/jss.v084.i06

Tennekes, M. 2018b. tmaptools: Thematic Map Tools. R package version 2.0. https://CRAN.Rproject.org/package=tmaptools
Tierney, L., A. J. Rossini, N. Li, and H. Sevcikova. 2018. snow: Simple Network of Workstations. R package version 0.4-3. https://CRAN.R-project.org/package=snow

USGS. 2019. Gap Analysis Project. https://www.usgs.gov/core-science-systems/scienceanalytics-and-synthesis/gap

Wickham, H., R. Francois, L. Henry, and K. Müller. 2018. dplyr: A grammar of data manipulation, R package version 0.7. 6. https://CRAN.R-project.org/package=dplyr

Xie, Y. 2018. knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.21. https://CRAN.R-project.org/package=knitr

Additional sources for range maps

Benkman, C. W., and C. K. Porter. 2018. Cassia Crossbill (*Loxia sinesciuris*), version 1.0. In P. G. Rodewald, editor, The Birds of North America. Cornell Lab of Ornithology, Ithaca, New York USA.

Hammerson, G. A. 2019. *Anniella alexanderae*. The IUCN Red List of Threatened Species 2019: e.T89929911A89929920.

Hammerson, G. A. 2019. *Anniella grinnelli*. The IUCN Red List of Threatened Species 2019: e.T89930126A89930151.

Hammerson, G. A. 2019. *Charina umbratica*. The IUCN Red List of Threatened Species 2019: e.T4567A18979867.

Krysko, K. L., L. P. Nuñez, C. E. Newman, and B. W. Bowen. 2017. Phylogenetics of kingsnakes, *Lampropeltis getula* complex (Serpentes: Colubridae), in eastern North America. Journal of Heredity 108:226–238.

NatureServe. 2017. Mojave desert tortoise (*Gopherus agassizii*). NatureServe: Arlington, VA.
Papenfuss, T. J., J. F. Parham. 2013. Four new species of California legless lizards (*Anniella*).

Breviora 536:1–17.