PATHWAYS
Wildlife Habitat Connectivity in the Changing Climate of the Hudson Valley

Timothy G. Howard and Matthew D. Schlesinger

New York Natural Heritage Program
PATHWAYS

Wildlife Habitat Connectivity in the Changing Climate of the Hudson Valley

February 2012

Timothy G. Howard
Matthew D. Schlesinger

New York Natural Heritage Program
The Nature Conservancy
625 Broadway, 5th Floor
Albany, New York 12233-4757


Cover photos: longtail salamander (Jesse Jaycox); eastern box turtle (Steve Young); northern cricket frog (Matthew Schlesinger), gray petaltail (Kim Smith), Blanding’s turtle (Mike Losito), cerulean warbler (retrieved from sciencecastle.com), New England cottontail (retrieved from myhouserabbit.com), timber rattlesnake (Jesse Jaycox)
Executive Summary

Maintaining or restoring connectivity is a key adaptation strategy for biodiversity conservation in the face of climate change and in recent years investigators have taken advantage of the multitude of connectivity modeling options to fuel conservation planning at various spatial scales. In this study, we combined species distribution modeling with connectivity modeling using present-day and future climate regimes to identify zones of connectivity—places where management agencies might focus attention on maintaining and restoring connections among populations of rare species.

The 10-county region bordering the lower Hudson River (Figure A) contains a wide range of landforms, geologic types, land-use patterns, and biodiversity. This critical area for rare and common species alike has the opportunity to support local and regional efforts for species adaptation to climate change.

We assessed habitat connectivity under current-day and future climate in the Hudson Valley for 26 Species of Greatest Conservation Need and aggregated these results to identify the importance of land parcels for multiple species. The forest, shrubland, and wetland species in the study included salamanders and frogs (longtail salamander, blue-spotted/Jefferson salamander, four-toed salamander, marbled salamander, northern cricket frog), snakes and lizards (black rat snake, northern black racer, northern copperhead, timber rattlesnake, eastern ribbon snake, common five-lined skink), turtles (eastern box turtle, wood turtle, Blanding’s turtle, bog turtle, spotted turtle), dragonflies (arrowhead spiketail, tiger spiketail, gray petaltail), neotropical migratory birds (black-throated blue warbler, Kentucky warbler, scarlet tanager, wood thrush, worm-eating warbler, cerulean warbler), and the New England cottontail.

We modeled suitable habitat for each species by analyzing the relationship between known locations and 44 environmental variables that included climate, geology, topography, land cover, and soils. This resulted in a continuous surface depicting the probability of suitable habitat throughout the study area (Figure B, left panel). Using climate model output from the IPCC fourth assessment, we created spatial models depicting the probability of suitable habitat for each species for the decades of the 2050s and 2080s (Figure B, right panel).

Strong shifts in suitable habitat were predicted for many of our target species. Under future climate regimes, suitable habitat appeared upslope and farther north, or simply contracted from existing habitat. A common pattern was for suitable habitat to appear in the Catskills, Taconics, and Rensselaer Plateau for species where none or very little is modeled as suitable in the current-day scenario. Conversely, although
suitable habitat patches often contracted greatly at current-day habitat strongholds, small patches of suitable habitat often were predicted to remain within or nearby these habitats for current-day populations. All predictions need to be evaluated in the context of species’ presumed adaptability, dispersal abilities, and connectivity to populations outside the study area.

We modeled potential connections among habitat patches by finding the least-cost path (LCP) for every single patch-to-patch connection for each species for each time period (Figure C). Each LCP is a function of both distance and resistance—here a quantitative measure of how different a spot on the landscape is from suitable habitat. We included all potential connections in the final output, even long ones depicting paths not likely to be traveled by individuals in a single generation.

In order to encapsulate general patterns into a geographic scale that could be acted on by conservation practitioners, we aggregated all the modeled habitat patches and connections to the tax parcel. We aggregated by counting the number of species for which a certain parcel is important (Figure D), as well as by quantifying the importance of patch connectivity at the scale of the entire population (betweenness). The patterns that emerged are striking. Parcels within the Hudson Highlands, Shawangunk Ridge, Catskill Mountains, and Harlem Valley had high overlap of species, with areas upslope and northward in the valley attaining greater projected importance over time.

This modeling effort represents a novel assimilation of modern techniques in fine-scale distribution modeling, connectivity modeling, and climate change adaptation planning. In our full report, through a series of discussions and charts, we provide guidance on interpreting these current-day patterns and predicted changes. We envision that land managers and conservation planners will be able to use our results to help identify priority locations for providing for biodiversity adaptation to climate change, and our methods are easily translated to other regions and other species.
Table of Contents

Executive Summary ........................................................................................................................................ iv
Table of Tables ............................................................................................................................................... vi
Table of Figures ........................................................................................................................................ vii
Introduction ................................................................................................................................................ 1
  Project background ................................................................................................................................. 1
  Scientific underpinnings ......................................................................................................................... 1
Study area .................................................................................................................................................... 3
Methods ....................................................................................................................................................... 4
  Species selection ..................................................................................................................................... 4
  Species distribution modeling .............................................................................................................. 4
    Assembly of known locations and background .............................................................................. 4
    Environmental variables ................................................................................................................. 7
    Future climate ..................................................................................................................................... 8
    Modeling, validation, and prediction ............................................................................................... 10
Connectivity assessments ........................................................................................................................ 12
  Approach ............................................................................................................................................... 13
  Parcel-level aggregation and multispecies metrics .............................................................................. 14
Results ......................................................................................................................................................... 16
  Distribution modeling ......................................................................................................................... 16
    Validation ........................................................................................................................................... 16
    Distributional changes ..................................................................................................................... 16
    Variable importance ......................................................................................................................... 18
Connectivity ................................................................................................................................................ 19
Discussion and applications ..................................................................................................................... 28
  Species distribution models ............................................................................................................... 28
    Distribution model performance ................................................................................................. 28
    Interpreting projected distributional changes ............................................................................... 29
Connectivity models .................................................................................................................................. 31
  Least-cost path component ............................................................................................................... 31
  Centrality measures ........................................................................................................................... 32
  Interpreting connectivity .................................................................................................................... 33
Downscaling, parcels versus pixels, and the scale of conservation action ........................................ 35
Next steps .................................................................................................................................................. 36
  Accounting for concomitant changes ............................................................................................... 36
  Incorporating dispersal ....................................................................................................................... 37
Conclusion ................................................................................................................................................... 37
Acknowledgments ...................................................................................................................................... 38
Literature Cited .......................................................................................................................................... 38
Appendix 1. Modeled patches and connections for each species within each time period .................. 46
Appendix 2: Climate downscaling results for temperature, precipitation, and snowfall ................. 57
Appendix 3. Sample climate downscaling metadata and analysis results ........................................... 65
Appendix 4. Species distribution model validation metadata ............................................................... 71

Table of Tables
Table 1. The 26 species selected for this study ......................................................................................... 4
Table 2. Final number of point locations and points generated from element occurrence (EO) polygons for Species of Greatest Conservation Need used in species distribution modeling ........ 6
Table 3. Environmental layers used for suitability modeling with a short description and derivation for each. ........................................................................................................................................ 7
Table 4. Connectivity analysis steps and key settings .......................................................................... 14
Table of Figures
Figure 1. Maps of New York and southeastern New York depicting the study area for this project.............. 3
Figure 2. Average May temperature for the late 20th century (left panel), and as projected for the 2050s and the 2080s................................................................. 9
Figure 3. CRCM-modeled average mid century snow-depth for the month of March (left) and the final smoothed version of the modeled output used in this study (right) .................................................. 10
Figure 4. The probability of suitable habitat—and the inverse, resistance—for northern copperhead within the study area for current day (left), 2050s (middle), and 2080s (right)................................................. 11
Figure 5. A schematic of the process for identifying a least-cost path (LCP) and assessing patch metrics.... 15
Figure 6. Summary of the True Skill Statistic, a measure of model accuracy, for 26 species distribution models.................................................................................................................. 16
Figure 7. Modeled suitable habitat for the northern copperhead in the Hudson Valley, for current day (left), 2050s (middle), and 2080s (right)................................................................. 17
Figure 8. Modeled suitable habitat for the New England cottontail in the Hudson Valley, for current day (left panels), 2050s (middle panels), and 2080s (right panels)................................................................. 18
Figure 9. Importance of five groups of variables (Table 3) in determining the distribution of suitable habitat across 26 species in the Hudson Valley (see Table 1) .................................................. 19
Figure 10. Mean importance ranks for all 44 environmental layers, across all 32 species and guilds modeled for this study......................................................................................... 20
Figure 11. Examples of results for a small subset of the range of northern copperhead........................... 21
Figure 12. Modeled least-cost path connectivity for northern copperhead under current-day modeled habitat patches (left), 2050-modeled habitat patches (middle), and 2080-modeled habitat patches ...... 22
Figure 13. Modeled least-cost path connectivity for New England cottontail using current-day modeled habitat patches (left), 2050-modeled habitat patches (middle), and 2080-modeled habitat patches ...... 23
Figure 14. Betweenness centrality for New England cottontail current-day habitat patches .................... 24
Figure 15. Landowner parcels in the study area predicted to be important for any facet of life history (i.e., intersecting a path or patch) for one or more species. .................................................. 25
Figure 16. Average patch betweenness (increasing from blue to red; note that the same range of colors is used in each figure although the absolute numbers differ) for patches for all species, applied to landowner parcels in the Study Area, for current day (left), 2050s (middle), 2080s (right). ...................... 26
Figure 17 (details from Figure 15 and Figure 16, middle panels). Number of species per parcel (increasing from blue to yellow) and betweenness values (increasing from blue to red) for parcels intersecting a path under projected 2050s climate. The map depicts the northeast corner of the study area, with the Rensselaer Plateau to the west and the Taconic Mountains to the east. ................................................................. 26
Figure 18. Consistency in parcel importance, as measured by the number of species for which a parcel is important (increasing from blue to red; note that the same range of colors is used in each figure although the absolute numbers differ) in both current day and the 2050s (left panel) and the number of species for which a parcel is important for all three time periods (current day, 2050s, 2080s), right panel ................................................................. 27
Figure 19 (detail from Figure 18, right panel). Consistency in parcel importance in the Hudson Highlands (A), Shawangunk Ridge (B), and connecting the Shawangunk Ridge to the Catskill Mountains (C), as measured by the number of species for which a parcel is important (increasing from blue to red) for all three time periods (current day, 2050s, 2080s). ................................................................. 28
Figure 20. Model of management actions to take based on known relationships of a species with climate and other variables and four potential types of distribution responses ........................................................................... 30
Figure 21. Model of management actions for maintaining patch to patch connectivity. This model begins with the final actions from Figure 20 .............................................................................. 34
Figure 22. Model of conservation priorities based on the layout and viability of populations in patch clusters. This model follows Figure 20 and considers the broader picture of patches taken together rather than the individual path focus of Figure 21 ................................................................. 35
Introduction

Project background

Public and private conservation organizations in New York State (NYS) have identified and protected areas for the purpose of biodiversity conservation for more than 100 years. Within the Upper and Lower Hudson River watershed, many of the areas described in the Hudson River Estuary Wildlife and Habitat Conservation Framework (Penhollow et al. 2006) are the focus of NYS Department of Environmental Conservation (DEC) and multi-partner efforts to conserve these special places and the biodiversity habitats they provide. While much progress has been made in these watersheds, continued fragmentation and habitat loss resulting from development pressures, incompatible neighboring land uses, and global climate change threaten to degrade and further isolate the areas currently protected and managed for biodiversity. Further isolation of habitats would leave the animal populations they support, including those of many species of greatest conservation need (SGCN), increasingly more isolated.

This loss of habitat connectivity through habitat loss and fragmentation is a primary threat to many SGCN (New York State Department of Environmental Conservation 2005). Compounding this threat are the changes in habitat condition, composition, and location that are predicted to accompany global climate change. Thus, planners clearly need an understanding of the current and future states of connectivity among SGCN, and, even more important, an action plan that includes a prioritized strategy for maintaining habitat connectivity for focal SGCN populations. Historically, biologists and planners have focused on conservation of core habitat areas within the Hudson River Valley and NYS. We now recognize the dire need to plan for species’ movements among habitat patches, particularly in light of our rapidly changing climate.

To these ends, in 2007 we began the PATHWAYS (Planning Along The Hudson for Warming and Animal connectivitY) project. With funding from State Wildlife Grants, we set out to develop a regional action plan that identifies and prioritizes key habitat corridors that will allow for SGCN migration at multiple scales. This project will help planners target key conservation corridors that will assist animal movement and migration at multiple scales in the Upper Hudson and Lower Hudson/Long Island Bays watersheds, and in particular the 10 counties bordering the Hudson River Estuary. Our results can be used to meet relevant targets from the Hudson River Estuary Action Agenda (Hudson River Estuary Program 2010).

The project built on organizational meetings held in 2006-2007 that identified the need and enlisted stakeholder support. Supporting the project were a variety of DEC Divisions and Bureaus and non-profit and government agency partners who committed to working closely with the NY Natural Heritage Program to (1) develop statewide GIS layers of current and future predicted habitats for selected SGCN based on the development and application of element distribution models (EDMs) under three climate regimes: the present, the 2050s, and the 2080s; (2) assess connectivity for focal SGCN for these points in time within the region of interest; and (3) identify and prioritize the habitat patches and connections most important for maintaining connectivity for the focal SGCN.

Scientific underpinnings

Since Parmesan et al. (1999) documented poleward shifts in several butterfly taxa, similar patterns of association between animal movement and climate change have been documented across the globe and in many taxa. For example, an analysis of New York’s breeding birds over a 20-year span showed northward shifts in several species (Zuckerberg et al. 2010). As species distributions shift, conservation biologists have proposed many strategies to ensure that species can reach suitable
habitat, including translocation, increasing reserve size or numbers of reserves, adding buffer zones to reserves, site restoration, and most often, increasing connectivity (Heller and Zavaleta 2009).

Maintaining or restoring connectivity is a key adaptation strategy for biodiversity conservation in the face of climate change (Heller and Zavaleta 2009, Mawdsley et al. 2009, Krosby et al. 2010) and in recent years investigators have taken advantage of the multitude of connectivity modeling options to fuel conservation planning at various spatial scales (Compton et al. 2007, McRae et al. 2008, Urban et al. 2009, Beier et al. 2011, Rayfield et al. 2011). Researchers have used modeling and predictions of future climate to predict how species’ distributions might change, typically showing upslope and northward movement (e.g., Iverson et al. 2008, Paradis et al. 2008, Rodenhouse et al. 2008). Fewer researchers have examined how connectivity itself might change with changing climate (but see Williams et al. 2005, Phillips et al. 2008).

Methodologies for assessing the connectivity (also called permeability) of a landscape have received considerable attention in the ecological and modeling literature. Debates have revolved around the definition of patches (and whether patches are even appropriate), the definition of paths (and whether paths are appropriate), single paths versus multiple paths, and constriction zones versus permeability, for example. We believe that how habitat patches are defined is critical to the outcome of connectivity modeling, as patches are intended to represent the most important and limiting habitat type for the organisms in question. Many studies use expert opinion to define the suitability (and its inverse, resistance) of different landcover classes, but assignment of resistance levels to habitat types based on expert opinion often lacks repeatability and justification (Sawyer et al. 2011). Thus, we further believe that connections modeled among patches should integrate a quantitative assessment of the magnitude of difference between the matrix (the landscape where the connection is being modeled) and the suitable habitat.

Conservation rarely has the luxury of fully enacting a plan for a single species, and planning for connectivity is no exception. As integrating species distribution modeling for multiple species increases the depth of conservation planning (Carvalho et al. 2010), connectivity planning will be much stronger when performed for multiple species. Highlighting overlaps and connectivity “hotspots”—what here we’ll call “zones” of connectivity—can provide managers with greater certainty that their actions will benefit multiple species. And connectivity needs often align among species, since habitat suitability, barriers, and travel corridors have commonalities among many at-risk species (Franklin et al. 2011).

In this study, we combined species distribution modeling with connectivity modeling using present-day and future climate regimes to identify zones of connectivity—places where management agencies might focus attention on maintaining and restoring connections among populations of rare species. We modeled 26 at-risk species in the Hudson River Valley of New York State and highlighted priority areas of focus. While we are aware of studies that have integrated portions of these components (e.g., Phillips et al. 2008, Carvalho et al. 2010), we know of no other study that has similarly combined species distribution modeling, connectivity modeling, and climate change to develop conservation priorities for multiple species.
Study area

Our study area was the 10-county region bordering the Hudson River (Figure 1), the administrative boundary used by NYSDEC’s Hudson River Estuary Program (Hudson River Estuary Program 2010). This 16,180-km² area encompasses nearly the entire watershed of the Hudson River south of Stillwater and north of New York City and includes such natural features as the Hudson Highlands, Shawangunk Mountains, Taconic Mountains, and eastern Catskills Mountains, as well as the large (>75,000-person) cities of Yonkers, Albany-Troy-Rensselaer, and New Rochelle. Its proximity to the U.S.’s largest city and major shipping corridors results in a substantial human population: 2,789,254 in the 2010 census (U.S. Census Bureau 2011).

This region has a dynamic geologic history and a great diversity in geologic structure, ranging from highly metamorphic gneiss, marble, and quartzite of the Hudson Highlands; acidic to strongly calcareous shale, siltstone, sandstone, limestone, and dolostone of the Hudson Lowlands; metamorphic shale and sandstone of the Taconics; to the sandstones and conglomerates of the Shawangunk and Catskill Mountains (Isachsen et al. 2000). Continental glaciations also played a role in the region, with extensive deposition of till and glacial lake features such as well-drained sand plains and deltas (Isachsen et al. 2000). Soils range from acidic to basic and can be very thin with poor development all the way to deep, rich agricultural soils (United States Department of Agriculture Natural Resources Conservation Service 1995). Though most of the valley lies below 152 m (Edinger and Howard 2008), elevations range from over 1,200 m in the Catskills to sea level. As of 2005 the study area was composed of 60% forest; 17% agriculture, grassland, and open habitats; 9% urban development; 8% wetland, 3% water, and 2% shrubland (see Dobson et al. 1995). Habitat fragmentation has been recognized as a major threat to biodiversity in the region.
Chief agents of habitat loss and fragmentation include urbanization and roads.

Methods

To identify connectivity hotspots in the Hudson Valley, we took a three-parted approach: 1) model species distributions under present and future climate; 2) model connectivity under present and future climate, and 3) apply these results to ownership parcels, the core unit of conservation action. This approach facilitated a quantitative, transparent methodology permitting the development of information at multiple stages and levels of refinement (e.g., observed species locations, modeled species habitat, smoothed current day and future climate predictions, modeled connections, and the relative importance of patches for connectivity), in turn facilitating access and use for differing use scenarios. We outline the several steps for this process below.

Species selection

We selected 26 species for this assessment (Table 1). All were designated Species of Greatest Conservation Need (New York State Department of Environmental Conservation 2005), possessed limited dispersal capability and/or specialized habitat requirements, were sensitive to habitat fragmentation and climate change, and occurred at a minimum of five mapped locations within the study area. We chose species in several habitat groups, ensuring a spread across habitat types, and with the aim of modeling species grouped into habitat guilds.

Species distribution modeling

Species distribution modeling was achieved through five steps: 1) documentation of known locations (typically points) and establishment of a set of background points; 2) attribution of all points with environmental variables; 3) applying modeling algorithms; 4) model validation; and 5) prediction (extrapolation) to the entire landscape of interest. This method follows well-documented approaches in extensive use throughout the research community (e.g., Guisan and Zimmerman 2000, Elith et al. 2006, Lawrence et al. 2006, Prasad et al. 2006, Vincenzi et al. 2011).

Assembly of known locations and background

We assembled 5,968 known locations of the 26 species (Table 2) from a variety of sources: the New York Natural Heritage Program element occurrence (EO) database (New York Natural Heritage Program 2011), NYSDEC Herp Atlas (NYSDEC 2009), North American Breeding Bird Survey (Sauer et al. 2011), fieldwork from the Wildlife Conservation Society (2007) and others. Locations represented breeding, foraging, and movement locations. Herp Atlas, Breeding Bird Survey, and Metropolitan Conservation Alliance locations were obtained as points. After careful screening for accuracy, we removed all duplicate points and points within 10 meters of each other for each species. NY Natural Heritage EOs were available as polygons that delineated suitable habitat, following the EO methodology and specifications laid out by NatureServe (NatureServe 2002). Element Occurrences representing older observations or museum or herbarium specimens may, by necessity, be represented with very low precision, while EOs with better spatial documentation may be represented with very high precision. After applying initial rules to weed out low-precision or historical locations (locations with no observations since 1980 or where the population was considered extinct [EORANK = H or X] and records in which locations could not be pinpointed to within 4800 meters [PRECISION = M or G]), we evaluated every location of every species and determined whether the spatial representation was adequate as input for a distribution model and either modified or removed polygons that poorly represented known habitat.
Table 1. The 26 species selected for this study, listed by general habitat groups (“guilds”). Each species’ state listing status (E/T/SC is State listed as Endangered, Threatened or Special Concern, respectively; (T) is federally listed as Threatened), Natural Heritage S-Ranks, and NatureServe G-Ranks are noted in the final columns. “ORR” = open rocky ridges.

<table>
<thead>
<tr>
<th>Species</th>
<th>Habitat group</th>
<th>NYS Listing</th>
<th>S - Rank</th>
<th>G - Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black-throated blue warbler (<em>Dendroica caerulescens</em>)</td>
<td>Forest</td>
<td>S5</td>
<td>G5</td>
<td></td>
</tr>
<tr>
<td>Kentucky warbler (<em>Oporornis formosus</em>)</td>
<td>Forest</td>
<td>S2</td>
<td>G5</td>
<td></td>
</tr>
<tr>
<td>Scarlet tanager (<em>Piranga olivacea</em>)</td>
<td>Forest</td>
<td>S5</td>
<td>G5</td>
<td></td>
</tr>
<tr>
<td>Wood thrush (<em>Hylocichla mustelina</em>)</td>
<td>Forest</td>
<td>S5</td>
<td>G5</td>
<td></td>
</tr>
<tr>
<td>Worm-eating warbler (<em>Helmitheros vermivorum</em>)</td>
<td>Forest</td>
<td>S4</td>
<td>G5</td>
<td></td>
</tr>
<tr>
<td>Black rat snake (<em>Elaphe obsoleta</em>)</td>
<td>Forest</td>
<td>S4</td>
<td>G5</td>
<td></td>
</tr>
<tr>
<td>Eastern box turtle (<em>Terrapene Carolina</em>)</td>
<td>Forest</td>
<td>SC</td>
<td>S3</td>
<td>G5</td>
</tr>
<tr>
<td>Northern black racer (<em>Coluber constrictor</em>)</td>
<td>Forest</td>
<td>S4</td>
<td>G5</td>
<td></td>
</tr>
<tr>
<td>Timber rattlesnake (<em>Crotalus horridus</em>)</td>
<td>Forest (ORR)</td>
<td>T</td>
<td>S3</td>
<td>G4</td>
</tr>
<tr>
<td>Common five-lined skink (<em>Eumeces fasciatus</em>)</td>
<td>Forest (ORR)</td>
<td></td>
<td>S3</td>
<td>G5</td>
</tr>
<tr>
<td>Northern copperhead (<em>Agristodon contortrix mokasen</em>)</td>
<td>Forest (ORR)</td>
<td>S3</td>
<td>G5T5</td>
<td></td>
</tr>
<tr>
<td>Cerulean warbler (<em>Dendroica cerulea</em>)</td>
<td>Forest (Riparian)</td>
<td>SC</td>
<td>S4B</td>
<td>G4</td>
</tr>
<tr>
<td>Wood turtle (<em>Clemmys insculpta</em>)</td>
<td>Forest (Riparian)</td>
<td>SC</td>
<td>S3</td>
<td>G3</td>
</tr>
<tr>
<td>Longtail salamander (<em>Eurycea longicauda</em>)</td>
<td>Forest (Riparian)</td>
<td>SC</td>
<td>S2S3</td>
<td>G5</td>
</tr>
<tr>
<td>Arrowhead spiketail (<em>Cordulegaster obliqua</em>)</td>
<td>Forest (Seeps)</td>
<td>S2S3</td>
<td>G4</td>
<td></td>
</tr>
<tr>
<td>Gray petaltail (<em>Tachopteryx thoreyi</em>)</td>
<td>Forest (Seeps)</td>
<td>SC</td>
<td>S2</td>
<td>G4</td>
</tr>
<tr>
<td>Tiger spiketail (<em>Cordulegaster erronea</em>)</td>
<td>Forest (Seeps)</td>
<td>S1</td>
<td>G4</td>
<td></td>
</tr>
<tr>
<td>Blue-spotted/Jefferson salamander complex (<em>Ambystoma laterale/A. jeffersonianum</em>)</td>
<td>Forest (Vernal Pool)</td>
<td>SC</td>
<td>S3</td>
<td>G5/G4</td>
</tr>
<tr>
<td>Four-toed salamander (<em>Hemidactylium scutatum</em>)</td>
<td>Forest (Vernal Pool)</td>
<td>S5</td>
<td>G5</td>
<td></td>
</tr>
<tr>
<td>Marbled salamander (<em>Ambystoma opacum</em>)</td>
<td>Forest (Vernal Pool)</td>
<td>SC</td>
<td>S3</td>
<td>G5</td>
</tr>
<tr>
<td>Blanding’s turtle (<em>Emydoidea blandingii</em>)</td>
<td>Wetland</td>
<td>T</td>
<td>S2S3</td>
<td>G4</td>
</tr>
<tr>
<td>Bog turtle (<em>Clemmys muhlenbergii</em>)</td>
<td>Wetland</td>
<td>E(T)</td>
<td>S2</td>
<td>G3</td>
</tr>
<tr>
<td>Eastern ribbon snake (<em>Thamnophis sauritus sauritus</em>)</td>
<td>Wetland</td>
<td>S4</td>
<td>G5</td>
<td></td>
</tr>
<tr>
<td>Spotted turtle (<em>Clemmys guttata</em>)</td>
<td>Wetland</td>
<td>SC</td>
<td>S3</td>
<td>G5</td>
</tr>
<tr>
<td>Northern cricket frog (<em>Acris crepitans</em>)</td>
<td>Wetland</td>
<td>E</td>
<td>S1</td>
<td>G5</td>
</tr>
<tr>
<td>New England cottontail (<em>Sylvilagus transitionalis</em>)</td>
<td>Shrubland</td>
<td>SC</td>
<td>S1S2</td>
<td>G3</td>
</tr>
</tbody>
</table>

* These salamanders readily hybridize; for this effort we will consider them one “species.”

Each polygon was then converted to one or more points for attribution of environmental condition variables. Recognizing the value in acquiring a full representation of environmental conditions within each polygon we designed a sampling formula to sample small polygons on a per-area basis but large polygons with a constant number of points. The formula is a derivation of the logistic function:

$$ Y = A \left( \frac{2}{1 + e^{-kx}} \right) - 1 $$

[Equation 1]
where $Y$ is the number of points to sample, $A$ is the asymptote, $k$ is the shape of the curve, and $x$ is the baseline per-area sample size. We chose 400 points as the asymptote by evaluating a sample of 1702 species polygons, set $k$ to 0.004, and defined $x$ as $[(\text{polygon area in m}^2/900) + 1]$. Points were randomly placed using the GRTS methodology (Stevens and Olsen 2003, 2004) using the spsurvey package (Kincaid and Olsen 2007) in the statistical program R versions 2.10 – 2.12 (R Development Core Team 2011). The final number of input points per species ranged from 78 to 13,473 (Table 2). R packages used in other parts of this study include randomForest (Liaw and Wiener 2002), ROCR (Sing et al. 2005), vcd (Meyer et al. 2010), abind (Plate and Heiberger 2011), RODBC (Ripley and Lapsley 2010), and foreign (R Development Core Team et al. 2011).

All points representing known habitat or observed presence were then combined and attributed with a complete set of environmental variables, described in more detail below.

Absence data were unavailable for most species in this study and thus, for consistency across all species, we similarly attributed a set of 10,000 spatially balanced randomly placed points for comparison as a sample of the background environment. Using pseudo-absences is one of the many ways to work with presence-only data (e.g., Engler et al. 2004, Elith et al. 2006, Pearce and Boyce 2006, Tsoar et al. 2007, Buechling and Tobalske 2011). Here, we follow Elith et al. (2006) by using the same background points for each species model with the intent to sample the entire study area, even though some of these points may occur within a particular species’ suitable habitat. In comparison with presence-only methods that do not require some form of background points, Elith et al. (2006) found the highest performing methods are those that do. Evaluations of random forest have shown the method to be robust for predictive mapping, using metrics such as Kappa (Prasad et al. 2006, Lawler et al. 2006), AUC, and others (Lawler et al. 2006).

Table 2. Final number of point locations and points generated from element occurrence (EO) polygons for Species of Greatest Conservation Need used in species distribution modeling. Asterisks after the common name denote species that are tracked as element occurrences by NY Natural Heritage and that consequently had locations represented as polygons.

<table>
<thead>
<tr>
<th>Common name</th>
<th>Scientific name</th>
<th>Points</th>
<th>Points from EO polygons</th>
<th>Total input points</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Birds</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black-throated blue warbler</td>
<td><em>Dendroica caerulescens</em></td>
<td>99</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Cerulean warbler</td>
<td><em>Dendroica cerulea</em></td>
<td>243</td>
<td>0</td>
<td>243</td>
</tr>
<tr>
<td>Worm-eating warbler</td>
<td><em>Helmitheros vermivorum</em></td>
<td>151</td>
<td>0</td>
<td>151</td>
</tr>
<tr>
<td>Wood thrush</td>
<td><em>Hylocichla mustelina</em></td>
<td>661</td>
<td>0</td>
<td>661</td>
</tr>
<tr>
<td>Kentucky warbler*</td>
<td><em>Oporornis formosus</em></td>
<td>6</td>
<td>2906</td>
<td>2912</td>
</tr>
<tr>
<td>Scarlet tanager</td>
<td><em>Piranga olivacea</em></td>
<td>314</td>
<td>0</td>
<td>314</td>
</tr>
<tr>
<td><strong>Mammal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New England cottontail*</td>
<td><em>Sylvilagus transitionalis</em></td>
<td>58</td>
<td>207</td>
<td>265</td>
</tr>
<tr>
<td><strong>Odonates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tiger spiketail*</td>
<td><em>Cordulegaster erronea</em></td>
<td>2</td>
<td>141</td>
<td>143</td>
</tr>
<tr>
<td>Arrowhead spiketail*</td>
<td><em>Cordulegaster obliqua</em></td>
<td>4</td>
<td>750</td>
<td>754</td>
</tr>
<tr>
<td>Gray petaltail*</td>
<td><em>Tachopteryx iloreyi</em></td>
<td>3</td>
<td>308</td>
<td>311</td>
</tr>
<tr>
<td><strong>Amphibians</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cricket frog*</td>
<td><em>Acris crepitans</em></td>
<td>137</td>
<td>1391</td>
<td>1528</td>
</tr>
<tr>
<td>Longtail salamander*</td>
<td><em>Eurycea longicauda</em></td>
<td>14</td>
<td>462</td>
<td>476</td>
</tr>
<tr>
<td>Four-toed salamander</td>
<td><em>Hemidactylium scutatum</em></td>
<td>129</td>
<td>0</td>
<td>129</td>
</tr>
<tr>
<td>Jefferson salamander complex</td>
<td><em>Ambystoma jeffersonii × laterale</em></td>
<td>259</td>
<td>0</td>
<td>259</td>
</tr>
<tr>
<td>Marbled salamander</td>
<td><em>Ambystoma opacum</em></td>
<td>104</td>
<td>0</td>
<td>104</td>
</tr>
</tbody>
</table>
### Environmental variables

Each point was attributed with 44 environmental layers representing topography, geography, land use and land cover, soils, geology, and climate (Table 3). Each layer was represented as a 30-meter grid encompassing all of New York State. These represent both static and dynamic variables with respect to climate change, both of which are important for modeling current and future distributions (Stanton et al. 2011).

<table>
<thead>
<tr>
<th>Common name</th>
<th>Scientific name</th>
<th>Points</th>
<th>Points from EO polygons</th>
<th>Total input points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reptiles and turtles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copperhead</td>
<td>Agkistrodon contortrix</td>
<td>156</td>
<td>0</td>
<td>156</td>
</tr>
<tr>
<td>Spotted turtle*</td>
<td>Clemmys guttata</td>
<td>297</td>
<td>0</td>
<td>297</td>
</tr>
<tr>
<td>Northern black racer</td>
<td>Coluber constrictor</td>
<td>262</td>
<td>0</td>
<td>262</td>
</tr>
<tr>
<td>Rattlesnake*</td>
<td>Crotalus horridus</td>
<td>447</td>
<td>5710</td>
<td>6157</td>
</tr>
<tr>
<td>Eastern ratsnake</td>
<td>Elaphe obsoleta</td>
<td>344</td>
<td>0</td>
<td>344</td>
</tr>
<tr>
<td>Blanding’s turtle*</td>
<td>Emydoidea blandingii</td>
<td>206</td>
<td>13473</td>
<td>13679</td>
</tr>
<tr>
<td>Common five-lined skink</td>
<td>Eumeces fasciatus</td>
<td>187</td>
<td>0</td>
<td>187</td>
</tr>
<tr>
<td>Wood turtle</td>
<td>Glyptemys insculpta</td>
<td>306</td>
<td>0</td>
<td>306</td>
</tr>
<tr>
<td>Bog turtle*</td>
<td>Glyptemys muhlenbergii</td>
<td>156</td>
<td>3522</td>
<td>3678</td>
</tr>
<tr>
<td>Eastern box turtle</td>
<td>Terrapene c. carolina</td>
<td>333</td>
<td>0</td>
<td>333</td>
</tr>
<tr>
<td>Eastern ribbon snake</td>
<td>Thamnophis sauritus</td>
<td>78</td>
<td>0</td>
<td>78</td>
</tr>
</tbody>
</table>

Table 3. Environmental layers used for suitability modeling with a short description and derivation for each.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent forest at 300 m</td>
<td>CCAP neighborhood analysis, percent forest cover within 10-cell radius</td>
</tr>
<tr>
<td>Percent forest at 990 m</td>
<td>CCAP neighborhood analysis, percent forest cover within 33-cell radius</td>
</tr>
<tr>
<td>Percent open cover at 300 m</td>
<td>CCAP neighborhood analysis, percent open cover within 10-cell radius</td>
</tr>
<tr>
<td>Percent open cover at 990 m</td>
<td>CCAP neighborhood analysis, percent open cover within 33-cell radius</td>
</tr>
<tr>
<td>Percent shrub at 300 m</td>
<td>CCAP neighborhood analysis, percent shrub cover within 10-cell radius</td>
</tr>
<tr>
<td>Percent shrub at 990 m</td>
<td>CCAP neighborhood analysis, percent shrub cover within 33-cell radius</td>
</tr>
<tr>
<td>Percent water at 300 m</td>
<td>CCAP neighborhood analysis, percent water within 10-cell radius</td>
</tr>
<tr>
<td>Percent water at 990 m</td>
<td>CCAP neighborhood analysis, percent water within 33-cell radius</td>
</tr>
<tr>
<td>Percent wetland at 300 m</td>
<td>CCAP neighborhood analysis, percent wetland within 10-cell radius</td>
</tr>
<tr>
<td>Percent wetland at 990 m</td>
<td>CCAP neighborhood analysis, percent wetland within 33-cell radius</td>
</tr>
</tbody>
</table>

**Soils and geology**

- **Percent clay**
  - Percent clay in the top soil layer from STATSGO database (USDA Natural Resource Conservation Service 2004)

- **Cation exchange capacity**
  - Cation exchange capacity of surface soil layer from STATSGO

- **Percent organic matter**
  - Percent organic matter of surface soil layer from STATSGO

- **pH of top soil layer**
  - pH of surface soil layer from STATSGO

- **Soil permeability**
  - Soil permeability of surface soil layer from STATSGO

- **Water holding capacity**
  - Water holding capacity of surface soil layer from STATSGO

- **Distance to calcareous soil**
  - Distance to nearest soil polygon containing calcium carbonate (m), derived from STATSGO

- **Surficial geology class**
  - Surficial geology class as derived from NYS Geological Survey (NYS Museum / NYS Geological Survey 1999)

- **Geologic class**
  - Bedrock geology class as derived from NYS Geological Survey (New York State Museum 1999)

**Climate**

- **Annual precipitation**
  - Total annual precipitation from DAYMET (Thornton et al. 1997)

- **May precipitation**
  - Total May precipitation from DAYMET

- **June precipitation**
  - Total June precipitation from DAYMET

- **July precipitation**
  - Total July precipitation from DAYMET

- **Average Annual temp.**
  - Annual average of daily temperatures from DAYMET

- **Average May temp.**
  - May average daily temperature from DAYMET

- **Average June temp.**
  - June average daily temperature from DAYMET

- **Average July temp.**
  - July average daily temperature from DAYMET

- **March snow depth**
  - Mean snow depth for 1961-2000 as modeled by Environment Canada’s Canadian Center for Climate Modeling and Analysis and downscaled by NYNHP.

3 = (Beven and Kirby 1979), 1 = (Zimmerman 2001)


**Future climate**

The relationship between species habitat and environmental variables was modeled using current-day environmental conditions. To project this relationship into the future, we prepared GIS layers for future climatic conditions (described below) equivalent to the current-day climate variables used to build the models (temperature, rainfall, snow depth). We then followed existing approaches to project species’ future habitat suitability using this new set of 44 environmental variables (e.g., Iverson et al. 2004, Stanton et al. 2011).
For future temperature and precipitation projections, we used an ensemble average of models as provided by ClimateWizard (www.climatewizard.org). Other products were available, such as the datasets from WorldClim (Hijmans et al. 2005), and the Northeast Climate Impact Assessment (Hayhoe et al. 2008), but these no longer used the most recent global circulation models. Climate Wizard provides 16 Global Circulation Models from the IPCC fourth assessment (IPCC 2007) in a readily accessible format as well as all sixteen in an averaged surface, all downscaled to approximately 12-km cells following Maurer et al. (2007). From this dataset, we downloaded average May (Figure 2), June, and July monthly temperatures for mid-century years (2050s) and late-century years (2080s). We downloaded the equivalent time frames for total precipitation as well as mean annual temperature and mean annual precipitation. To represent these data at the finer scales required for our species modeling we smoothed each grid using an approach similar to Hijmans et al. (2005) by modeling the relationship between the climate model output and elevation, latitude (Y), and longitude (X). We added other location and interaction variables: \( X^2, Y^2, X*Y \) (Borcard et al. 1992) and sampled 1,000 spatially balanced, randomly placed points throughout the state (Stevens and Olsen 2003, 2004). We modeled the relationship between each climate variable and the independent variables using the randomForest package in R (Liaw and Wiener 2002). The mean percent variance explained for mean temperature models was 96.4 (range: 95.78 – 97.5); the mean percent variance explained for total precipitation models was 96.49 (95.88 – 96.98). We then applied these modeled relationships to New York State at 30-m resolution to create a smoothed version of each variable for application in our species models (e.g., Figure 2 and Appendix 2). Elevation was usually the most important variable for all models, with \( Y \) and \( Y^2 \) also most often in the top three; however, some of the precipitation models placed \( X*Y \) or \( X \) in the top three rankings for importance. Appendix 2 provides images for each downscaled climatic variable; Appendix 3 provides sample analysis results.

![Figure 2. Average May temperature for the late 20th century (left panel), and as projected for the 2050s and the 2080s. The projections as downloaded from Climate Wizard are in the top panels, and as downscaled by us in the bottom panels.](image-url)
We also collected modeled estimates of snow depth for the study area. The best available set that includes both estimates of historic and future snow depth came from Environment Canada’s Canadian Center for Climate Modeling and Analysis (CCCMA) ([http://www.cccma.ec.gc.ca/data/data.shtml](http://www.cccma.ec.gc.ca/data/data.shtml)). From the CCCMA Ouranos Climate Simulation Team, we used the Canadian Regional Climate Model (CRCM) which covers most of North America, including New York State (Music and Caya 2007). We used output from The Canadian Regional Climate Model (CRCM4.2, act run) to generate historical (1961-2000), 2050s, and 2080s snow depth estimates for the month of March. The CRCM data are available in gridded form with 45-km cells. To represent snow depth at a finer resolution, we followed a similar approach to Hijmans et al. (2005) by modeling the relationship between the climate model output and elevation, latitude ($Y$), and longitude ($X$). We added other location and interaction variables: $X^2$, $Y^2$, $X*Y$ (Borcard et al. 1992) and sampled every cell within 100 km of New York, for a total of 248 sample points. Random forests fit each relationship with greater than 94% of the variance explained. The three variables most important in deriving this fit were elevation, $Y^2$, and $Y$. We then applied the modeled relationship to New York State at 30-m resolution to create a smoothed version (e.g., Figure 3). Appendix 3 provides more details in the analysis results.

Modeling, validation, and prediction

We generated random forest models for all 26 species and six groups of species occupying similar habitats. These “guilds” (general forest, forest ridges, forest riparian, forest seep, forest vernal pools, and wetlands) were based on the combined known locations of all species assigned to each group (Table 1).

We used random forests (Breiman 2001, Cutler et al. 2007) to model the relationships between presence and background data with the randomForest package in R (Liaw and Wiener 2002). Random forest models are based on an integration of many classification or regression trees and often perform at the top level when compared to other approaches (Cutler et al. 2007, Prasad et al. 2006). Random forests fits among the methods dubbed as “machine learning” (Iverson et al. 2004), and is capable of modeling complex relationships among many, potentially noisy, variables without making a priori assumptions about the type of relationship (Vincenzi et al. 2011).

Model validation was based on jackknife cross-validation (a.k.a. “leave one out” [Manel et al. 1999, Jaberg and Guisan 2001, Fielding 2002]) at the polygon level. For those species in which locations were originally acquired as points, we divided the study area into a grid of 10-km × 10-km cells, attributed points by this grid, and applied cross-validation by cell. Thus, if there were four known locations for a species, represented by four polygons (or cells) with many points randomly
placed within each polygon, we modeled all points within three polygons and then evaluated the ability of that model to predict species presence for the points in the fourth polygon. Repeating this for each polygon resulted in repeated estimates of accuracy from which we derived and reported summary statistics. We generated mean, standard deviation, and standard error for the following validation metrics: Kappa (weighted and unweighted), area under the ROC curve (AUC), True Skill Statistic (TSS), Overall Accuracy, Specificity, and Sensitivity (Allouche et al. 2006, Fielding 2002).

The final model was based on all known locations and run for the entire study area. An output surface representing the probability of suitable habitat was generated for each species (e.g., Figure 4). Those species with a mean TSS > 0.51 were considered to pass validation and we moved forward with them for connectivity assessments.

![Figure 4. The probability of suitable habitat—and the inverse, resistance—for northern copperhead within the study area for current day (left), 2050s (middle), and 2080s (right). Brighter, yellow tones indicate higher habitat suitability and less resistance to travel while darker, blue tone indicate lower habitat suitability and more resistance to travel.](image)

We evaluated the representation of patches on the landscape using different cutoffs generated by varying alpha in the F-measure (Van Rijsbergen 1979, Sing et al. 2005). Our goal was to maximize the capture of true positives (known species locations) while simultaneously restricting the total area of patches and allowing for some representation in each time period (current day, 2050’s, 2080’s). We strove to maintain the same cutoff threshold for each species across time periods, but very large changes in distribution or extent of predicted suitable habitat required some changes in cutoff threshold. Maintaining within-species consistency in cutoff maintains the association between independent variables and suitable habitat so that the extent of suitable habitat can be compared between time periods.
In addition to developing validation metrics for each species distribution model, we extracted variable importance estimates from each model to better understand the factors most associated with environment-species habitat relationships in this landscape. Each tree in the random forest model is constructed from a random subset of variables and records. Construction of a large number of trees provides opportunity to evaluate model performance with and without each predictor variable. Large decreases in prediction error with the inclusion of a variable represent high variable importance, and vice-versa (Liaw and Wiener 2002).

Applying each model to the study area resulted in a continuous probability surface with integer values ranging from 0 to 10000 (scaled from 0.00-100.00%) reflecting the probability of suitable habitat. Higher probabilities signified combinations of environmental layers most similar to combinations seen where the species is known to occur.

**Connectivity assessments**

The field of connectivity study and assessment continues to be dynamic and rapidly changing. New and improved indices and methods are published and discussed in recent articles (e.g., Rayfield et al. 2011), and no one method is appropriate for answering all questions, particularly at different scales (e.g., Beier et al. 2011). Before this project began, the Hudson River Estuary Program gathered key partners in a meeting (Feb 16, 2007) to work through the options and decide on the most appropriate assessment method for this project. Given that the tools had to match the goals, we first describe the goals of this connectivity analysis, and then describe our approach.

The goal of this portion of the project is to identify potential connections between the patches developed from our species distribution modeling. Ideally, connections would be represented in a fine scale so that they may inform conservation action on the ground. Calabrese and Fagan (2004, Fagan and Calabrese 2006) point out the relationship between increased detail in a connectivity assessment and the increased data requirements to accomplish the study. The high-resolution species location information available for this project increases the data quality considerably, allowing for good potential in getting the detail desired. In recognition that patches also play an important role in connectivity at a broader scale, a secondary goal was to be able to evaluate the relative importance of individual patches to the overall connectivity of collection of patches.

A relatively new and increasing popular approach for assessing landscape connectivity uses circuit-theory modeling to assess overall landscape “electrical flow” which highlights areas of restricted connection, and conversely, high flow (McRae et al. 2008, see also www.circuitscape.org). At its core, Circuitscape uses graph theory (see below) to define the links between every cell in a GIS grid. This approach provides a landscape perspective and also allows multiple paths between patches (Rayfield et al. 2011), but generally requires a separate evaluation for each patch to patch connection and does not quantify relative patch importance (McRae and Shah 2009). Conversely, one of the most traditional ways of evaluating connectivity is to estimate the least-cost path between patches (e.g., Halpin and Bunn 2000). This approach provides specific detail about potential routes between patches but does not provide landscape perspective of overall connectivity and/or bottlenecks (but see Adriaensen et al. 2003, Pinto and Keitt 2009).

Graph theory, in which patches (“nodes”) are connected by paths (“edges”) can assess the relative importance of patches by looking at the overall patterns of connections among all patches (Cantwell and Forman 1993, Bunn et al. 2000, Urban and Keitt 2001). While providing patch-level metrics (are there key bottleneck patches or habitat or less important peripheral patches?), the level of resolution for the edges between patches can vary depending on the quality of information available (Rayfield et al. 2011). This is the ideal scenario for our application: a way to generate relatively high-resolution connections between patches while at the same time allowing for patch-
level metrics that inform us about the relative importance of patches for overall connectivity. An early effort to combine these analyses into a GIS tool was developed by Best et al. (unpublished, in Urban et al. 2009). This tool was applied to roadless areas in the Northeast to prioritize corridors and patches (Jantz and Goetz 2008). We adapted this tool (ConnMod; see http://mgel.env.duke.edu/tools) to assess connectivity in this study.

**Approach**

Each patch of suitable habitat generated from the species distribution model represented a habitat patch from which to assess connectivity. In the matrix between patches, we used the inverse of habitat suitability to represent resistance or cost. This action integrated the relationship of species habitat preference with the entire suite of environmental variables from the habitat suitability analysis and applied that relationship to assess the likelihood of use of all locations across the landscape. Because roads are not always fully represented in land-use/land-cover datasets (Thomas and Endreny 2008) and because roads constitute potentially significant barriers for many of our species, we added roads into the resistance layer in a consistent manner. We divided roads that appeared in the Accident Location Information System (ALIS; NYS Office of Cyber Security & Critical Infrastructure Coordination 2005) layer into three categories: major roads (feature class codes A10-A19), primary and secondary roads (codes A20-A39), and local roads (codes A40-A79). We buffered all roads by 6 m and assigned resistance values to these roads according to their presumed role as barriers for terrestrial animals: major roads at 100%, primary and secondary roads at 95%, and local roads at 90%. If the original resistance layer had a higher resistance value than the one about to be assigned by the road, we kept the original value. We accounted for the value of culverts and bridges as movement corridors (Clevenger et al. 2001) by punching holes in the roads at all stream and river crossings, as defined by the National Hydrography Dataset (U.S. Geological Survey and U.S. Geological Survey 2010).

We converted the resistance layer to a triangulated irregular network (TIN), incorporated the patches into the TIN (Urban et al. 2009), and then calculated the least-cost paths from each patch to each neighboring patch using the NetworkX (network.lanl.gov, version 1.5) module in Python (python.org, versions 2.6 and 2.7). Data tables not directly attributed in GIS were stored in an SQLite database (sqlite.org, version 3.7) using the APSW Python module (code.google.com/p/apsw, version 3.7.6). Final output was written to an ESRI ArcGIS file geodatabase, with a line featureclass containing the paths between patches and a polygon featureclass containing patches.

Table 4 enumerates the steps, the settings used, and the purpose for each step. Figure 5 provides a schematic showing the key steps: 1. Identifying suitable habitat patches within the variable landscape, 2. Converting the landscape to a resistance surface represented by nodes with interconnected lines, and 3. Calculating a path between the patches that minimizes a function of both distance (lines) and resistance (values at the points).

The cost to travel an edge affects travel routes (Rayfield et al. 2010). We settled on a formula for cost that best balanced the effects of the landscape and habitat suitability:

\[
\text{cost} = d \times (R_{N1} + R_{N2})^4
\]

[Equation 2]

where \( R \) is the resistance at each node (N1 and N2) and \( d \) is the distance between these nodes. Resistance (\( R \)) is standardized over the entire graph to vary between 0-100; distance is in meters.

Recognizing the high variability among our target species in their ability to move both within and among generations, we allowed paths to be created among all patches, no matter how far the distance. This allowed us to be consistent in our treatment of all species and provided more
flexibility for later in-depth assessments of paths. Because of shape of the study area and the proximity of some patches to the study area boundary, some edges were created that travelled along the study area boundary. As they did not follow true paths of least resistance, we removed these edges manually prior to any further assessment.

Graph theory provides a variety of patch-level metrics that can be calculated to determine the relative importance of each patch to landscape connectivity. We calculated several such metrics (Table 4) but focus our output on betweenness centrality, a relative measure of the number of paths passing through each patch, which has been touted as one of the best metrics for assessing patch importance (Bodin and Saura 2010).

Table 4. Connectivity analysis steps and key settings.

<table>
<thead>
<tr>
<th>Step</th>
<th>Settings</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate nearby patches</td>
<td>75-m aggregate distance.</td>
<td>Treat patches within 75 m as the same patch.</td>
</tr>
<tr>
<td>Remove small patches</td>
<td>Four cells or smaller.</td>
<td>Focus only on larger patches, small spots with high suitability may be captured in paths.</td>
</tr>
<tr>
<td>Smooth raster far from patches</td>
<td>8 × 8 cell roving mean greater than 4000 m from any path.</td>
<td>Reduce precision of path at distances where exact location of path becomes less important.</td>
</tr>
<tr>
<td>Convert to TIN</td>
<td>z-tolerance set to 100 (about 1/10 the range of the cost raster). Habitat patches are embedded after the TIN is created, with no within-patch travel costs.</td>
<td>Triangulated Irregular Network is a vector representation of the cost raster, providing a means to represent and model paths along the cost surface.</td>
</tr>
<tr>
<td>Find patch neighbors</td>
<td>Create a second TIN using patch centroids. Identify which centroids have direct triangulation connections.</td>
<td>Increases efficiency by evaluating least-cost paths using full cost surface for neighbors only.</td>
</tr>
<tr>
<td>Create undirected cost graph integrating distance and cost raster.</td>
<td>For each edge compute travel cost using Equation 2. All edges inside patches: cost = 1.</td>
<td>In contrast to a regular grid (raster) each edge in a TIN differs in length. Each node is located in a different place in the cost surface and so must also be accounted for.</td>
</tr>
<tr>
<td>Find the least-cost path between each adjacent patch</td>
<td>Use bidirectional dijkstra algorithm in NetworkX. Sum the total cost for each path between patches and write results to SQLite tables.</td>
<td>Models a connection that stays within the most suitable habitat while travelling between patches.</td>
</tr>
<tr>
<td>Calculate patch-level metrics.</td>
<td>Using stored weights between patches, calculate betweenness, closeness, degree, and load.</td>
<td>These metrics help describe how each patch contributes to the connectivity of the entire set of patches.</td>
</tr>
</tbody>
</table>

Parcel-level aggregation and multispecies metrics

Small-scale data often must be aggregated for actionable conservation decisions (Strager and Rosenberger 2007). We summarized data at the level of the ownership (tax) parcel, essentially the smallest unit of conservation action for local land trusts and other organizations (Strager and Rosenberger 2007). A further advantage of aggregation is in functionally creating a buffer for the paths currently represented as lines, a width obviously too narrow to serve as a travel corridor for most species. Finally, aggregating to tax parcels permits the merging of species in a consistent way for summation and prioritization.
Figure 5. A schematic of the process for identifying a least-cost path (LCP) and assessing patch metrics. The shaded patches represent suitable habitat, a wetland type. (A) Other wetlands exist to the northeast (cattail icon), and mountains representing the most unsuitable habitat are present to the southwest. (B) The landscape between habitat patches is converted to a continuous surface of triangles (the TIN) with nodes attributed by the level of resistance (here, mountain equals highest resistance, wetland is lowest, and the remainder is in-between). (C) This panel illustrates a path that minimizes the sum of all costs derived from a function that balances distance and resistance (Equation 2), to find the LCP. The same two patches depicted again in (D), and then as part of a cluster of patches in (E), with every patch-to-patch combination mapped (28 paths). In (F), peripheral patches through which other patch-to-patch paths do not run have low betweenness and are shaded pale red. Patches with many paths running through them have high betweenness and are shaded bright red. Maroon is intermediate betweenness.

We obtained tax parcel layers from a statewide repository and from individual county websites, stitching them together and retaining identifier attributes when available. We manually deleted large road networks grouped as single parcels, although we made no attempt to screen the entire seamless parcel layer for errors. The final parcel layer contained 921,746 parcels.

We intersected both paths and patches with parcels, retaining the presence or absence of a path or patch for each species in each parcel and calculating two statistics: the number of species intersecting a parcel and the average betweenness of patches intersecting each parcel. The former statistic highlights the richness of species for which a parcel is important and the latter measures the average importance for species whose patches intersect a parcel. Parcels may be slightly important to many species (high number, low average betweenness) or highly important to one or many species.
Results

Distribution modeling

Validation

Models for most of the species were excellent and all performed better than random, as determined by the True Skill Statistic (TSS), a balance between sensitivity (correctly classified presence points) and specificity (correctly classified background or pseudoabsence points). TSS ranged from 0.28 to 0.92, with many models showing excellent performance (Figure 6; Appendix 4). However, we determined that models for marbled salamander, eastern ribbonsnake, wood thrush, scarlet tanager, and black-throated blue warbler were insufficiently accurate (TSS < 0.515) for use in assessing connectivity, and we excluded them from the connectivity analysis. Guilds varied considerably in model accuracy and we felt were not representative of their groups. We also excluded all guilds from further analysis.

Distributional changes

Habitat for all species was predicted to shift in one way or another under future climate regimes. Generally speaking, habitat was predicted to expand northward and upslope, contract in the southern portion of the study area and downslope, and shrink overall.

Suitable habitat was predicted to become available in the higher elevations of the Catskills, where little or none previously existed, for the following species: northern cricket frog, northern copperhead, Jefferson’s/blue-spotted salamander, northern black racer, tiger spiketail, arrowhead spiketail, timber rattlesnake, black rat snake, five-lined skink, wood turtle, worm-eating warbler, Kentucky warbler, cerulean warbler, eastern box turtle, and eastern ribbon snake (Appendix 2). The Blanding’s turtle models were unique in their prediction of northward-moving suitable habitat that eventually disappears from the Hudson Valley, presumably continuing to move north out of the study area to meet existing habitat in Saratoga County and the Saint Lawrence drainage. Models for nearly all species showed suitable habitat contracting or disappearing, particularly in the southern and lower elevation portions of the study area (for examples see Figure 7 and Figure 8). How these results should be interpreted depends on the species-specific biogeographic context, questions of rapid evolution, dispersal ability, and connectivity. Each of these topics will be treated in the discussion.
Figure 7. Modeled suitable habitat for the northern copperhead in the Hudson Valley, for current day (left), 2050s (middle), and 2080s (right) time frames. The bottom panels zoom into the Schunnemunk Mountain area as indicated by the black box in the upper left panel. These bottom panels provide an example of local contractions in suitable habitat over time. Future predicted suitable habitat overlays the faded extent of previous suitable habitats.
Figure 8. Modeled suitable habitat for the New England cottontail in the Hudson Valley, for current day (left panels), 2050s (middle panels), and 2080s (right panels). The bottom panels zoom into the Hudson Highlands east of the Hudson River as indicated by the black box in the upper left panel. These bottom panels provide an example of local contractions in suitable habitat over time. Future predicted suitable habitat overlays the faded extent of previous suitable habitats.

Variable importance

Across all species modeled, climate variables were the most important in determining the distribution of suitable habitat (Figure 9). Among individual variables, precipitation variables constituted the three most important (Figure 10). Land cover classes varied in their importance, but many (% developed land within 30 m, % shrubs within 300 m, % water within 300 m, and the simplified six-category landcover classification) were consistently least important for determining a species distribution. While averaging in the middle of the pack, most of the other environmental variables fell in the top five in importance for at least one species (Figure 10). Variable-specific importance values are given for each species in Appendix 4.
**Connectivity**

We successfully modeled all combinations of patch-to-patch connections for 21 species in each of three time periods (current day, 2050s, and 2080s). All connectivity model outputs are depicted in Appendix 2. The species not modeled for connectivity had been excluded based on poor habitat suitability model performance. Computational intensity turned out to be a considerable problem for a subset of the connectivity models. Namely, habitat suitability models that predicted suitable habitat patches over much of the study area resulted in TIN surfaces with a large density of points. These points, and how they are interconnected, had to be loaded into memory in order to find the least-cost paths among habitat patches. Memory requirements quickly exceeded the maximum memory allowed for normal 32-bit Windows programs (2 GB), requiring us to move the work to a 64-bit computer with 12 GB of RAM. To increase modeling speed we divided the study area in half (down the middle of the Hudson River) and completed a connectivity assessment on each half independently. Even with this modification, we were required to reduce the number of patches from which to model for spotted turtle, timber rattlesnake, Blanding’s turtle, wood turtle, worm-eating warbler, and eastern box turtle for the 2050s and 2080s.

A zoomed-in example of the resistance layer, the TIN, and the resulting modeled patch-to-patch connections is shown in Figure 11. The total number of paths varied exponentially with the number of patches—the largest number of paths modeled for a single species in a single time period was over 7000.

A single least-cost path was a function of distance and resistance, with every single edge in the TIN getting evaluated. Paths were often relatively direct, nearly linear connections between patches, emphasizing the importance in the routine of minimizing distance. Other paths, however, made distinct turns, curves, or detours, reflecting the influence of habitat suitability and roads on path models. Examples of both can be seen in Figure 11.
Figure 10. Mean importance ranks for all 44 environmental layers, across all 32 species and guilds modeled for this study.
Figure 11. Examples of results for a small subset of the range of northern copperhead. The left panel depicts the resistance grid, which is transformed to a triangular irregular network (TIN) with inset habitat patches in the middle panel. The right panel shows the modeled least-cost paths among patches; patches with more patch-to-patch paths passing through them (=higher betweenness) are shown with darker green in the right panel.

Viewing connectivity from the perspective of an entire group of patches requires one to look at the relative importance of patches to the connectivity of the group as a whole (Bodin and Saura 2010). Betweenness centrality was calculated for each patch within each graph (full connectivity model). Betweenness values varied from near zero to about 0.5, with more central and larger patches usually having higher values. The right panels in Figure 11 and Figure 14 show examples of patch betweenness.
Figure 12. Modeled least-cost path connectivity for northern copperhead under current-day modeled habitat patches (left), 2050-modeled habitat patches (middle), and 2080-modeled habitat patches. The top panels show the entire study area while the bottom panels show a small zoomed-in zone in the Hudson Highlands showing patch contraction and more detail in path routes.
Parcel-level aggregation and multispecies metrics

Aggregation across the 21 species at the level of the tax parcel yielded clear patterns in parcel importance under the three climate regimes. Suitable habitat for most species was projected to move upward in elevation and northward (Figure 15). Parcels in the Catskill Mountains, at the west-central edge of the study area (Figure 1), were not important for the 21 species under the current-day climate, but were predicted to be increasingly important in the 2050s and the 2080s. Similarly, few parcels north of the middle of the study area, for example in and near the Taconic Mountains in the northeastern portion of the study area, were less important under current-day climate, but they were important under future climates. Parcels in the Hudson Highlands, in the southern portion of the study area on both sides of the Hudson River (Figure 1), remained important under all three climate regimes.

Parcel importance as measured by the average betweenness of patches intersecting parcels showed similar trends, with parcels with high betweenness generally appearing upslope and northward with changing climate (Figure 16). Some portions of the study area, like the northeast corner, containing the Taconic Range and Rensselaer Plateau, were not highlighted as important under current climate conditions but were projected to become important under future climate (Figure 17).

Even though the relative importance of parcels is projected to change over time with changing climate, and newly important parcels are revealed under projected future climate regimes, some parcels are projected to remain consistently important, especially those in the Hudson
Highlands, in the Shawangunk Ridge, and connecting the Shawangunk Ridge to the Catskill Mountains (Figure 18, Figure 19).

Figure 14. Betweenness centrality for New England cottontail current-day habitat patches.
Figure 15. Landowner parcels in the study area predicted to be important for any facet of life history (i.e., intersecting a path or patch) for one or more species. Parcels important for more species (up to 17; note that the full range of colors is used in each figure although the absolute numbers differ) are identified with orange and red shades, while parcels important for fewer species are identified with blue to green shades. The left panel shows current-day path and patch intersections, the middle panel is for the 2050s, and the right panel is for the 2080s.
Figure 16. Average patch betweenness (increasing from blue to red; note that the same range of colors is used in each figure although the absolute numbers differ) for patches for all species, applied to landowner parcels in the Study Area, for current day (left), 2050s (middle), 2080s (right).

Figure 17 (details from Figure 15 and Figure 16, middle panels). Number of species per parcel (increasing from blue to yellow) and betweenness values (increasing from blue to red) for parcels intersecting a patch under projected 2050s climate. The map depicts the northeast corner of the study area, with the Rensselaer Plateau to the west and the Taconic Mountains to the east.
Figure 18. Consistency in parcel importance, as measured by the number of species for which a parcel is important (increasing from blue to red; note that the same range of colors is used in each figure although the absolute numbers differ) in both current day and the 2050s (left panel) and the number of species for which a parcel is important for all three time periods (current day, 2050s, 2080s), right panel.
Discussion and applications

Species distribution models

Distribution model performance

The ability of the random forests method to discern complex patterns of association among many independent variables is well documented (Prasad et al. 2006, Cutler et al. 2007, Vincenzi et al. 2011). The success of the method is evident here as well, with all models performing better than random, and most performing exceptionally. To be conservative in our connectivity assessment, we dropped the five species with poorest performance from further analysis. Poor performing models typically were those for which we had incomplete observation data. For example, three of the species (black-throated blue warbler, scarlet tanager, and wood thrush) are interior forest birds for which we had presence data from surveys conducted only in a portion of the study area. These species are known to occur (but not documented with high enough precision) with relative high frequency in other localities in the Hudson Valley. The restricted sample set available to us did not represent a sample of habitat conditions that adequately characterized habitat preferences for
these—mostly generalist—species, particularly at the fine scale at which our models were being applied (30 m).

Alternatives to distribution modeling that could have been applied here include 1) using only known observations as patches from which to model connectivity or, 2) defining habitat more simply (e.g., forest patches, emergent wetlands), and modeling connectivity among all examples of this habitat type. Using habitat definitions (or sometimes a combination of environmental layers based on expert knowledge of a species’ requirements – a deductive model) as a basis for connectivity assessments is not uncommon (e.g., Lookingbill et al. 2010, Urban and Keitt 2001), but removes any opportunity for detecting unknown habitat relationships as well as quantifying the magnitude of habitat suitability and typically over-represents habitat availability. Conversely, using only known locations typically under-represents habitat suitability. It is common for biodiversity surveys to discover additional locations, even for rare species, due in part to the vagaries of property access and species detectability.

One beneficial outcome of statistically assessing habitat suitability (inductive modeling) is a measure of the relative importance of each environmental variable used to build each model. We found high variation in the order of variable importance among species (Figure 10), yet metrics for precipitation, temperature, geology, and snow depth were among the overall top ten. Many species were associated with the environmental variables one might expect. Spotted Turtle had wetland metrics (wetlands at 300 m and 990 m) within the top five, low elevations were important for Blanding’s Turtle, high forest percent cover and more extensive forests for Tiger Salamander, and steep slopes for Timber Rattlesnake, for example (Appendix 4).

Interpreting projected distributional changes

One consequence of employing a study area whose boundaries do not encompass the full range of the study species is that processes and conditions that could not be accounted for in our models outside the study area affect the proper interpretation of predicted future distributions. For example, the position of the species’ New York range within the entire species’ range (i.e., the southern edge, northern edge, or middle) might be important for interpretation of these results. Our models show that, for many species, based on known relationships with climate and other variables, the distribution of suitable habitat will change, typically moving upslope and northward and potentially vanishing downslope and southward. Interpreting these projected changes relies on knowledge or assumptions about each species’ pre-adaptation or adaptability to the coming climate, connectivity to populations outside the study area, and dispersal ability (Figure 20). Distributional shifts are therefore interpreted versions of the maps of suitable habitat, and require expert judgment to determine whether management action is likely necessary. Below we address these issues in more detail, with examples.

While the full range of each species modeled in this study extends far beyond the boundaries of New York State, note that these models are “trained” on New York genotypes, and only the range of environmental conditions associated with species locations in New York could be modeled. This means that the models do not take into account whether New York’s populations of these animals might be able to adapt to changing climate in a way that we could not represent here. For species known to be adapted to conditions with different temperature or precipitation regimes south of New York, therefore, it might be best to think of the 2050 and 2080 models as additive to the present-day models rather than as replacing them. Of course, simply because the distribution of suitable habitat changes does not mean that the distribution of the animal will change with it. How these results should be interpreted depends on the species-specific biogeographic context, questions of rapid evolution, dispersal ability, and connectivity. Each of these topics will be treated below.
First, we need to ask whether the habitat predicted to become unsuitable could in fact remain suitable. Two questions are relevant:

1) Could Hudson Valley populations of the species already be adapted to the climate that will arrive? It is possible that species with populations farther south (i.e., ones in the middle or at the northern edge of their range in the Hudson Valley) are adapted to the arriving climate, but it is unknown whether the particular populations (=genotypes) share the more southern populations’ tolerance for the climate that will arrive—namely, warmer and wetter conditions. Species on the southern edge of their range in the Hudson Valley may be at the limit of their physiological climatic tolerances, or they may be limited only by competition at this range edge (MacArthur 1972). Regardless, the substantial number of examples of species ranges that have contracted at their southern edges (see Parmesan 2006) does not provide encouragement for those species’ persistence in the Hudson Valley.

2) If Hudson Valley populations are not pre-adapted to the new climate (as we would assume for most southern-range-edge species), could they adapt? Parmesan (2006) notes that during Pleistocene glaciations, species tended to track the climate rather than stay in place and evolve, but it is less certain what would happen over shorter timeframes. Rapid evolution has been documented in response to recently warming climates (and to other stressors such as road salt [Brady 2012]), but not wholesale changes in climate tolerances.
Reliance on adaptation to changing climate would be a very risky conservation strategy.

Second, if local populations in the Hudson Valley are not already adapted to the new climate, or cannot adapt to the new climate, then animals will become extirpated from those locations. The next set of questions asks whether the valley might be recolonized by individuals from farther south.

3) Is there a population farther south that might serve as a source of colonists?
4) If so, are individuals capable of dispersing the necessary distance?
5) If so, is there habitat connectivity from those populations to the Hudson Valley’s?

If the answer to all three questions is “yes,” then we can infer that populations that disappear from the Hudson Valley will be recolonized. Unfortunately, however, we do not have the information, particularly on connectivity from New Jersey to New York, to answer all these questions for all species, so recolonization will not be a sure thing.

Now we examine two species in detail to illustrate these ideas: the northern copperhead and the New England cottontail. The copperhead is at its rangewide northern limit in the Hudson Valley (Gibbs et al. 2007), while the cottontail is at its southern (and western) limit (Litvaitis et al. 2006).

The copperhead model for the present day shows hotspots in the Hudson Highlands and Shawangunk Ridge (Figure 7). Over the next 70 or so years, with changes in climate, those locations are predicted to contract considerably or become outright unsuitable for the copperhead, while areas of the Catskills become suitable. We need further knowledge of the genetic makeup of the Hudson Valley populations, although rangewide the species shows low genetic diversity and is thus expected not to adapt quickly to changing climate (Douglas et al. 2009). However, populations in southern New York appear well connected to populations in northern New Jersey and eastern Pennsylvania, and copperheads are known to move several kilometers per year (Smith et al. 2009), so it is likely that Hudson Highlands populations of copperheads could be replenished from the south and west. Shawangunk populations, on the other hand, are isolated from Hudson Highlands populations and would require connectivity between the Highlands and the Shawangunks for the latter to be recolonized, and connectivity between these areas and the Catskills would be necessary for the species distribution to expand along with suitable habitat. Our connectivity models show the best places on which to focus in managing the landscape for that species.

The outlook for the New England cottontail in New York is less favorable. The model for the current day shows hotspots up and down the east side of the valley, and predicted suitable habitat all but disappears by the 2080s (Figure 7). There are no populations to the south, so if the cottontail is not able to adapt to the changing climate, there is little chance for recolonization.

**Connectivity models**

**Least-cost path component**

Least-cost path (LCP) modeling is often applied in a raster-style situation where the path segment (one grid cell to the next) follows the eight primary directions of a compass (N, NE, E, SE, S, SW, W, NW), and the distance is either a value of 1 for the cardinal directions or $\sqrt{2}$ for their intermediates (Pinto and Keitt 2009). By converting our resistance grid to a TIN, we allowed node-to-node connections not to be restricted to the grid layout and for them to occur in any direction and distance. Landscape regions that were relatively homogenous in the resistance surface resulted in a flatter surface of the TIN with more spaced-out nodes and longer edges. Landscape regions with more heterogeneity in resistance had closer nodes and shorter edges. This approach has the potential to increase analytic efficiency by reducing the number of connections (edges) a full path from habitat patch to habitat patch has to traverse (Urban et al. 2009). In this study, with the large number of connectivity assessments that needed to be done (21 species by 3 time periods $= 63$ assessments),
connectivity assessments that needed to be done (21 species by 3 time periods = 63 assessments), efficiency was an important consideration. This approach was relatively rapid unless the size of the TIN overcame the memory capacity of the computer. We discuss this issue along with other scale considerations later in this section.

Another benefit of applying a least-cost approach to connections that vary both in distance and resistance is that the model becomes more transparent to the relative importance of distance and resistance to calculating cost. Assigning more importance to distance results in straight paths between patches, while assigned higher importance to resistance results in more sinuous paths. Interestingly, we found very little discussion of this critical modeling component in the scientific literature. After much exploration of our data sets, we settled on a standardized formula that applied an exponential effect of resistance on cost (Equation 2). This had the result of forcing paths around features characterized by highly unsuitable habitat, yet at the same time maintaining relatively direct lines among patches. Additional research into this balance between distance and resistance is warranted; indeed, the potential to fine-tune this formula by species-specific life-history traits related to a species ability to traverse differing landscapes at the individual or population level is great. The toolset developed through this project facilitates this research by making it easy to modify all components of the modeling framework.

Finally, the least-cost paths as modeled here do not imply there is literally a route on the ground for individual species to follow. The intent of a modeled path is to integrate what we know about the landscape and habitat needs into an interpretation of the best area for a population to maintain connectivity. Sometimes the LCP may point directly to a barrier gap, such as a culvert, bridge, or agricultural field. Other times the LCP may be recognizing a broad swath of forest or a very coarse gradient between ridge and slope or forest and field. In all cases it is incumbent on the person using these data to consider them as supporting one tool in a collection to help understand the intricacies of landscape permeability and habitat connectivity. Two ways to “move beyond the line” of an LCP are 1) to apply them to an appropriate scale for conservation action, which is reflected in the rollup to parcels, and 2) to consider the importance of patches within patch groups, as we discuss next.

Centrality measures

Using a graph-theoretic perspective for assessing connectivity allowed us to assess the contribution of patches toward the overall connectivity of habitats in the landscape (Urban et al. 2009). Note the different perspective than evaluating a single LCP; here, we can assess individual patches with respect to clusters of patches or the entire graph as a whole. Peripheral patches only connected to one other patch scored lower than a central patch through which many other patches might be connected (see Figure 5). All else being equal (such as patch size and habitat quality), that peripheral patch contributed far less to overall population viability than the more central patch. With more connections to other patches, the central patch is more likely to receive individuals from nearby sub-populations or send individuals to nearby patches, thus contributing more to metapopulation stability. Similarly, the central patch is more likely to act as a stepping stone for distribution shifts. Betweenness, the metric we used in this study to quantify patches in this way, behaved as expected, with more central or core (and often larger) patches showing higher values. We calculated other centrality measures (degree, closeness, load), but do not report on them here.

Figure 17 is perhaps the most instructive for seeing differences between assessing habitat patches and connections and assessing betweenness. Different parts of the landscape are valued slightly differently depending on the conservation goals and perspective. We discuss this in the next section.
Interpreting connectivity

A single LCP connecting two patches represents a modeled route between those patches that strives to minimize the difference in environmental conditions along the path to those within a habitat patch and simultaneously minimize travel distance. Given a conservation goal that includes maintaining connectivity (see Figure 20) and a species-specific focus, LCPs should be included in the decision-making toolbox with a certain number of considerations.

First, consider species-specific traits and requirements. Are individuals known to easily cross habitat typically considered unsuitable? Birds, of course, might fit this criterion well, especially considering their seasonal migration to the neo-tropics. The exact path of the LCP would be much less important for these species, other flying animals such as odonates, and wide-ranging species not included in this study such as coyote or black bear. Birds, however, do have a level of site fidelity where adult birds generally return to former nesting grounds (Schlossberg 2009, Bernard et al. 2011), and adult and yearling dispersal may vary considerably in distance (Studds et al. 2008). Further, many species of birds have been shown to be reluctant to cross habitat gaps (e.g., Desrochers and Hannon 1997) and can be highly sensitive to forest fragmentation (e.g., Robinson et al. 1995). Similarly, odonates show site fidelity (Grether and Switzer 2000) and cases of extensive (Conrad et al. 1999), and restricted (Watts et al. 2004) dispersal. Thus, the distance between patches may more important for these types of species, possibly even approaching straight-line, or Euclidean, distance. What are typical (and extreme) individual dispersal distances for the species of interest?

Conversely, for species known to have more difficulty traversing unsuitable habitat we should consider the makeup of the space between patches. Discrete barriers such as roads (Bhattacharya et al. 2003, Delaney et al. 2010) and more diffuse landscape features that reduce permeability such as agriculture (Mader 1984) or housing developments ([Gibbs 1998, Forman 1999], even for birds [Hodgson et al. 2007]) influence the probability of individuals getting from one patch to the next. In this case, we need to pay more attention to the route an LCP takes through the landscape. Does the path offer suggestions for road crossings such as culverts or bridges over streams? Does the path choose one landscape over another because of land-use characteristics? Integrating travel “resistance” in this way provides a less direct path, but in effect, provides a realized distance for dispersal between patches.

Second, incorporate distance (Euclidean or realized) into a generational time-frame. What is the likelihood that an individual will make the entire distance between patches within a single lifetime? This question leads us to the decision tree in Figure 21. If the target patch falls within a distance an individual might be expected to travel, then the most important conservation strategy for maintaining connectivity would be to minimize barriers and increase landscape permeability. If the probability that a single individual might move between patches is low, however, a manager must consider the possibility for multi-generational movement between patches. Are there small patches of habitat that might provide appropriate stopover or stepping stone sites? These sites may not be adequate for long term population viability yet may be enough to support one or two generations that may create dispersers with a higher chance of making it to suitable habitat. Under this scenario, a manager must consider the availability (and conservation) of stepping stone habitats in addition to barrier reduction (Figure 21). If fieldwork determines the connection between paths is too long and intermediate patches of habitat do not exist, it may be time to consider other paths, populations, or conservation strategies.
Population viability of a species that occurs in patches often, but not always, depends on a metapopulation model of species persistence (Smith and Green 2005). Yet many of the taxonomic groups or species included in this study may depend, at least in part, on overall population persistence through metapopulation dynamics, including the New England cottontail (Litvaitis and Villafuerte 1996), marbled salamander (Gamble et al. 2007), bog turtle (Rosenbaum and Nelson 2010), and even woodland birds (Schippers et al. 2011).

From a conservation perspective, consideration of metapopulation dynamics comes to the forefront with our rapidly changing environment. Will the number of connected patches making up a metapopulation diminish over time? Similarly, how does a patch contribute to metapopulation connectivity? The decision tree in Figure 22 describes management considerations related to these questions. If species distribution models indicate the persistence (albeit diminishing size) of a network of habitat patches in a landscape, a viable conservation objective may be to maintain those patches and a network of connections among them. If, however the number or size of patches is predicted to diminish so much that population viability may be threatened, conservation attention might be better directed toward supporting a shifting population. In this scenario, those habitat patches and paths most likely to contribute positively to a shifting population should receive conservation attention. These are the patches with highest betweenness values (Figure 22).
Downscaling, parcels versus pixels, and the scale of conservation action

One of the most critical decisions to be made at the outset of a modeling effort is at what scale the modeling will be done. Factors that feed into this decision include the expected scale of conservation decisions, the scale of available data, and computer processing power (with its governing bodies, time and funding).

Distribution modeling conducted at coarser scales (e.g., Rodenhouse et al. 2008, Lawler 2009), can represent the distributions of common species primarily because presence data are available at these coarse scales, such as the breeding bird atlas blocks (McGowan and Corwin 2008), and the intent of the modelers is to understand broader patterns in distribution—or even a species’ range—rather than the fine-scale habitat preferences targeted with the approach taken here. For this project, we chose to model at the finest scale at which most data were available, 30-m grid cells.

Distribution modeling at fine scales such as the 30-m raster size we used in this study has its drawbacks, such as how to include coarse environmental variables such as climate (discussed below), but it also has its advantages. The most important reason is that this scale is necessary for applying the questions being addressed in this study. On-the-ground conservation action, whether it concerns the reduction of barriers or the protection of habitat, occurs at fine scales and thus our modeling of suitable habitat must also.

Choosing to model at such a fine scale did not come without its practical challenges. First, climate data were not available at such a fine scale, so we interpolated climate data by modeling relationships with elevation, latitude, longitude, and climate at the coarse scale and mapping these...
relationships to the finer scale. We recognize this approach did not incorporate bias-correction, stochasticity, or other downscale techniques (Maurer and Hidalgo 2008, Tryhorn and DeGaetano 2011), yet it was effective in creating a smoothed data set that could be applied to our needs. While we do not presume that the major forces shaping climate operate at such a fine scale, our approach recognizes that microclimates resulting from topographical variation can be key in shaping species’ distributions (Dobrowski 2011). Further, applying the coarse grid (12 km for temperature and precipitation, 45 km for snow depth) in which climate data were previously available (the 12-km grids themselves downscaled from much coarser native formats) would have meant that locations thousands of feet apart in elevation were assigned the same climate values.

Second, even with a reasonably powerful computer (2.40 GHz 6-core processor with 12GB RAM on a 64-bit architecture) we were limited in conducting the connectivity modeling using the TINs derived from the 30-m resistance grids for some species and needed to reduce the number of patches and/or reduce resolution away from patches. Complex resistance grids at the order of 16 million nodes maxed out the 12GB of RAM but still completed the connectivity analysis for some species in approximately four days while only slightly less complex grids of 13 million nodes would finish overnight.

There are several good reasons for modeling at the finest scale possible. Fine-scale distribution modeling allows fine-scale connectivity modeling, which is needed for applications in local areas (Hannah 2011), and allows investigators to scale up to a variety of coarser scales depending on the application. Land conservation, be it through acquisition by land trusts or government, conservation easements, or management, generally takes place at the level of the ownership parcel, especially in human-dominated landscapes like New York’s Hudson Valley. Thus, scaling up to the parcel rather than presenting results in grid cells makes the results of the study directly applicable to the missions of many agencies and organizations whose focus is biodiversity preservation.

Rolling up the 30-m data to parcels also allowed our linear paths to be reflected more accurately as broader swaths. While optimal widths of travel corridors for various target species are not always known (Andreassen et al. 1996, Haddad 2008, Gilbert-Norton et al. 2010), it is hard to imagine that a conceptual line on the landscape sufficiently represents a travel corridor for many of our target taxa. Also, while individuals may not move optimally through the landscape (Fahrig 2007), differential movement success may result in populations moving roughly along least-cost paths and offering more landscape (i.e., parcels) for this to occur may support the movement of more individuals. Further, land-use and land-cover data are variously accurate at the level of the pixel, so having paths that represent conceptual lines puts too much faith in the underlying data. Therefore, scaling up to the parcel allows the highlighting of entire parcels, which can then be scanned using other remote sensing techniques or field groundtruthing to identify the most likely areas of travel habitat within. Finally, if parcels are the primary unit of conservation, it makes sense to target large parcels, which by their largeness will intercept more paths and patches than small parcels. Large areas should always be a priority over small areas for conservation, all other things being equal.

**Next steps**

**Accounting for concomitant changes**

Clearly, climate is not the only relevant factor shaping species’ distributions that will change in the next 40 to 70 years. Predicted changes in tree species’ distributions (e.g., Schwartz et al. 2006, Iverson et al. 2008) and wholesale changes in vegetation communities (e.g., Iverson et al. 2004, Iverson and Prasad 2001, Theurillat and Guisan 2001, Tang and Beckage 2010, Hickler et al. 2012)
could have substantial impacts on animal species’ distributions. Interspecific interactions among animals, caused by, for example, the expansion of a novel predator or competitor, or the loss of a prey item, could similarly have dramatic effects on species’ distributions (Harrington et al. 1999, Araújo and Luoto 2007, Schweiger et al. 2008, Van der Putten et al. 2010). Unfortunately, predictive modeling of key interactions and vegetation communities are not as far along as for species distributions, and relevant data are not yet available for our study area.

Changes in land use are generally thought to pose an even bigger threat to biodiversity than changes in climate (Sala et al. 2000, Hannah 2011) and effects of urbanization in particular are well documented in the ecological literature (e.g., Chace and Walsh 2006). Land use is expected to change considerably this century (Haim et al. 2011). Great potential exists to incorporate build-out scenarios into our models to determine the effects of future habitat loss and fragmentation through increasing urban, suburban, and exurban development (Rhodes et al. 2006). However such models of land-use change are not yet available at the appropriate spatial scales.

**Incorporating dispersal**

While we encourage our models to be interpreted in the context of species’ dispersal abilities (Figure 21), in future efforts dispersal could be incorporated into the modeling more explicitly. Doing so would require estimates of average and maximum dispersal capabilities and rates, ideally as specific to the study area or region as possible. Least-cost paths that are longer than a threshold average or maximum dispersal distance could be removed or downweighted, with paths known to be within the dispersal range for the species being highlighted. The species- and landscape-specific data required to do so may seem like an impediment to incorporating dispersal, but recent evidence suggests that many species with seemingly different ecologies may disperse similarly in fragmented landscapes (Doerr et al. 2011). There are many recent examples of studies on which to base further work in our region (e.g., Brooker et al. 1999, Lookingbill et al. 2010, Schleicher et al. 2011).

**Conclusion**

The analyses, products, and discussion provided here represent a step forward in planning for biodiversity conservation in the Hudson River Valley. This is the first study we know of, anywhere, that incorporates the quantitative components of distribution modeling, connectivity, and climate change and then applies the findings to a scale applicable to site-based conservation practitioners: the real estate parcel.

These analyses identify some clear targets already in the sights of some conservation organizations: a Shawangunks-to-Catskills linkage, the Hudson Highlands, the Harlem Valley wetlands and linkages farther north, the Taconic Mountains, and the Rensselaer Plateau. Within the general domain of each targeted linkage or habitat complex, the data and assessments provided here offer a way to zoom in to a specific landscape and assess the conservation options on the ground.

These analyses also provide other situations and potential linkages that warrant additional scrutiny: is there a way to link up the Hudson Highlands to the Shawangunk Ridge? How about north of the Catskills? How might we maintain metapopulation dynamics for multiple species with discrete conservation actions?

Overall, these analyses provide a place to begin the discussion, with quantitative, transparent data to support any decisions being made by conservation practitioners. Nonetheless, they should always be used as one tool in the toolset and supported by on-the-ground assessments of both population and linkage viability.
Acknowledgments

Thanks to T. Kerpez, L. Zucker, K. Strong, A. DeWan, P. Riexinger, D. Rosenblatt, D. Evans, F. McKinney, T. Tear, J. Corser, J. Jaycox, E. Spencer, G. Kenney, D. VanLuven, A. Mahar, and R. Shirer, for conceptual discussions, feedback, and administrative help. We obtained species locations with the help of A. Chaloux, K. Perkins, H. Shaw, E. White, S. Barker, A. DeWan, and J. Ozard. Assorted help was coaxed from H. Krahling and A. Feldmann. Funding for this project was through New York State Department of Environmental Conservation New York State Wildlife Grant T-9, Project 1, Job 4 in cooperation with the U.S. Fish and Wildlife Service Division of Wildlife and Sport Fish Restoration.

Literature Cited


New York State Museum. 1999. New York State bedrock geology GIS layer.


R Development Core Team, S. DebRoy, R. Bivand, and others: see copyrights file in the sources. 2011. foreign: Read Data Stored by Minitab, S, SAS, SPSS, Statat, Systat, dBase,... R package version 0.8-43.


Van der Putten, W. H., M. Macel, and M. E. Visser. 2010. Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2025–2034.
Appendix 1. Modeled patches and connections for each species within each time period

Cerulean warbler (*Dendroica cerulea*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

Worm-eating warbler (*Helmitheros vermivorum*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Kentucky warbler (*Oporornis formosus*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

New England cottontail (*Sylvilagus transitionalis*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Tiger spiketail (*Cordulegaster erronea*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

Arrowhead spiketail (*Cordulegaster obliqua*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Gray petaltail (*Tachopteryx thoreyi*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

Northern cricket frog (*Acris crepitans*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Blue-spotted/Jefferon salamander complex (*Ambystoma laterale/A. jeffersonianum*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

Longtail salamander (*Eurycea longicauda*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Four-toed salamander (*Hemidactylium scutatum*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

Northern copperhead (*Agkistrodon contortrix mokasen*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Spotted turtle (*Clemmys guttata*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

Northern black racer (*Coluber constrictor*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Timber rattlesnake (Crotalus horridus) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

Black rat snake (Elaphe obsoleta) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Blanding’s turtle (*Emydoidea blandingii*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

Common five-lined skink (*Eumeces fasciatus*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Wood turtle (*Glyptemys insculpta*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).

Bog turtle (*Glyptemys muhlenbergii*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Eastern box turtle (*Terrapene carolina*) modeled suitable habitat patches with least-cost paths connecting them for current day (left), 2050s (middle), and 2080s (right).
Appendix 2: Climate downscaling results for temperature, precipitation, and snowfall.

Figure 1. New York State maps depicting the average May temperature for the late 20th century (left panel) and as projected for the 2050s and the 2080s. The projections as downloaded from Climate Wizard are in the top panels, and as downscaled by us in the bottom panels. Historical climatological data are from the DayMet climate data center (www.daymet.org).
Figure 2. New York State maps depicting the average June temperature for the late 20th century (left panel) and as projected for the 2050s and the 2080s. The projections as downloaded from Climate Wizard are in the top panels, and as downscaled by us in the bottom panels. Historical climatological data are from the DayMet climate data center (www.daymet.org).

Figure 3. New York State maps depicting the average July temperature for the late 20th century (left panel) and as projected for the 2050s and the 2080s. The projections as downloaded from Climate Wizard are in the top panels, and as downscaled by us in the bottom panels. Historical climatological data are from the DayMet climate data center (www.daymet.org).
Figure 4. New York State maps depicting the average annual temperature for the late 20th century (left panel) and as projected for the 2050s and the 2080s. The projections as downloaded from Climate Wizard are in the top panels, and as downscaled by us in the bottom panels. Historical climatological data are from the DayMet climate data center (www.daymet.org).

Figure 5. New York State maps depicting the average May precipitation for the late 20th century (left panel) and as projected for the 2050s and the 2080s. The projections as downloaded from Climate Wizard are in the top panels, and as downscaled by us in the bottom panels. Historical precipitation data are from the PRISM climate group (www.prism.oregonstate.edu).
Figure 6. New York State maps depicting the average June precipitation for the late 20th century (left panel) and as projected for the 2050s and the 2080s. The projections as downloaded from Climate Wizard are in the top panels, and as downscaled by us in the bottom panels. Historical precipitation data are from the PRISM climate group (www.prism.oregonstate.edu).

Figure 7. New York State maps depicting the average July precipitation for the late 20th century (left panel) and as projected for the 2050s and the 2080s. The projections as downloaded from Climate Wizard are in the top panels, and as downscaled by us in the bottom panels. Historical precipitation data are from the PRISM climate group (www.prism.oregonstate.edu).
Figure 8. New York State maps depicting the average annual precipitation for the late 20th century (left panel) and as projected for the 2050s and the 2080s. The projections as downloaded from Climate Wizard are in the top panels, and as downscaled by us in the bottom panels. Historical precipitation data are from the PRISM climate group (www.prism.oregonstate.edu).
Downscaled representations

Figure 9. New York State maps depicting the average January snow depth for the late 20th century (left panels) and as projected for the 2050s and the 2080s. The original data are modeled by The Canadian Centre for Climate Modeling in their Canadian Regional Climate Model (CRCM v4.2). The bottom panel depicts the results of downscaling these data by modeling the relationship of the original model output to latitude, longitude, and elevation.
Downscaled representations

Figure 10. New York State maps depicting the average February snow depth for the late 20th century (left panels) and as projected for the 2050s and the 2080s. The original data are modeled by The Canadian Centre for Climate Modeling in their Canadian Regional Climate Model (CRCM v4.2). The bottom panel depicts the results of downscaling these data by modeling the relationship of the original model output to latitude, longitude, and elevation.
Figure 11. New York State maps depicting the average March snow depth for the late 20th century (left panels) and as projected for the 2050s and the 2080s. The original data are modeled by The Canadian Centre for Climate Modeling in their Canadian Regional Climate Model (CRCM v4.2). The bottom panel depicts the results of downscaling these data by modeling the relationship of the original model output to latitude, longitude, and elevation.
Appendix 3. Sample climate downscaling metadata and analysis results

**Tme0550wz50**

Average temperature for May, modeled for years 2040-2069, and averaged from an ensemble of sixteen General Circulation Models

**ArcGIS GRID:** tme0550wz50

("temperature mean – month 5 – 2050 – ClimateWizard ensemble 50th percentile")

**Developed by:** Tim Howard,
New York Natural Heritage Program
January, 2011

![Figure 1. Average temperature (degrees F) in May for 2040-2069 as predicted by an ensemble of climate models, provided by Climate Wizard.](image)

This document describes the methods and sources used to create and validate the model depicted above and used by The New York Natural Heritage Program in its modeling efforts.

**Step 1. Base Grid Pre-processing:**
The source of these data is Climate Wizard, an online provider of accessible climate change model data ([www.climatewizard.org](http://www.climatewizard.org)). The source from Climate Wizard is the World Climate Research Program's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset. These CMIP3 data are downscaled by Lawrence Livermore National Laboratory, Reclamation, and Santa Clara University following Maurer *et al.* (2007). Additional information about the data sources and methods are available here: [http://www.climatewizard.org/documentation.html](http://www.climatewizard.org/documentation.html).

This dataset was downloaded in the fall of 2010, when IPCC Fourth Assessment climate models were being provided. These are the options chosen for download:

- Analysis Area: United States
- Time Period: Mid Century (2050s)
- Measurement: Average Temperature
- Emission Scenario: High A2
- General Circulation Model: Ensemble Average

The file name as downloaded from Climate Wizard:
map_mean_ensemble_50_AR4_US_12k_a2_tmean_5_2040_2069.asc

This raster was processed with the following steps:
1. Convert downloaded ascii file to ArcGIS GRID format, assign the correct projection (WGS 84).
2. Clip the US GRID to a buffered version of NY and reproject to NAD83 UTM zone 18.

The base grid ready for evaluation and downscaling has a cell size of 12km, with 57 columns and 44 rows to cover New York (Figure 1).
Step 2. Model relationship and apply model at finer scale
To represent this variable at a finer resolution, we followed a similar approach to Hijmans (2005) by modeling the relationship between climate model output and elevation, latitude, and longitude. We created this relationship using the regression-trees portion of the random forests statistical procedure as follows.

1. Generate 1,000 spatially-balanced randomly-placed points throughout the state. In this procedure, we used the Generalized Random Tessellation Stratified (GRTS) spatially explicit sampling design developed by the EPA (Stevens & Olsen 2003, 2004). We used the spsample package (Kincaid & Olsen 2007) within the R statistics program (R Development Core Team 2007) to generate these points and the R statistics program for all other statistics computations.

2. Attribute these points at the same 12km cell size with the targeted climate variable and the following variables: elevation, x-coordinate, y-coordinate. We then derived (x-coord)², (y-coord)², (x-coord)*(y-coord). See Borcord et al. 1992 for more information on the spatial components.

3. Model the relationship between the climate variable and the static variables using the randomForest package in R (Liaw & Wiener 2002).

4. Apply the relationship to the entire state, using an elevation grid at 30-meters cell-size resolution.

Random forests was successful at fitting the relationship (Figure 2); the percent variance explained and mean R² are indicated below. The relative importance of each independent variable is indicated in Figure 3, the final GIS output is depicted in Figure 4.

Mean of squared residuals (=MSE, an estimate of the error, smaller is better): 0.2220705
% Var explained (=pseudo R², an estimate of the fit, closer to 100 is better): 95.78
Figure 4. The final downscaled product.

Literature Cited

Kincaid, T. and T. Olsen. 2007. spsurvey: Spatial Survey Design and Analysis. [R package version 1.6.2].
Zn2050_03: Mean mid century snow depth for the month of March in New York State

ArcGIS GRID: zn2050_03
("snow depth– 2050 -- month 3")

Developed by: Tim Howard
New York Natural Heritage Program
8 December 2010

Figure 1. Mean snow depth in March for 2040-2069 as produced by the CRCM V4.2 and averaged and downscaled as described in this document. Although represented as inches in this figure, the layer provides snow depth in meters.

This document describes the methods and sources used to create and validate the model of mid-century March snow depth used by The New York Natural Heritage Program in its modeling efforts.

Step 1. Base Grid Pre-processing:
The source of these data is Environment Canada’s Canadian Center for Climate Modeling and Analysis (CCCMA). In addition to developing Global Climate Models for the IPCC, the CCCMA also develops The Canadian Regional Climate Model that covers most of North America, including New York State. See the following links for more information on the CCCMA and their regional modeling.
http://www.cccma.ec.gc.ca/data/data.shtml

We used output from The Canadian Regional Climate Model CRCM V4.2 (CRCM4.2) to generate this data layer. More specifically the data are labeled by CCCMA as ‘aet 1961-2100’ and represent: CRCM4.2.3 time-slice simulation for years 1961 to 2100 driven by CCM3.1/T47 member #4 following the IPCC observed 20th century 20C3M scenario for years 1961-2000 and the SRES A2 scenario for years 2001-2100, over the North-American domain (referred to as AMNO) with a 45-km horizontal grid-size mesh (true at 60°N) and 29 vertical levels. Spectral nudging of large-scale winds was applied within the regional domain.


CCCMA provides output from the entire simulation as a single text file with information provided for each month representing daily averages. To extract and summarize the data we:
1. Created a python script to extract each monthly average from the original file and write it to a new text file.
2. Created an R script to read in the appropriate years for the appropriate month and export a text file representing averages for the interval.
   a. To make the output comparable, we followed ClimateWizard time brackets as closely as possible:
b. “Historical” = “observed 20th century” = 1961-2000
c. “Mid Century” = 2050’s = 2040-2069
d. “End Century” = 2080’s = 2070-2099

3. Created a python script to extract the following invariant variables from CCCMA in order to use consistent inputs at the same scale with which to model relationships:
   a. Latitude
   b. Longitude
   c. Elevation

The base grid averaged and ready for downscaling is shown in Figure 2. Cell size is about 45 km to a side with the X and Y axes following the projection as provided by CCCMA. Values in each cell represent the modeled snow depth for the time period, in meters.

![Figure 2. CRCM-modeled average mid century snow depth for March at the CCCMA output of 45km grid cells.](image1)

![Figure 3. Sampling frame for modeling the relationship between the climate model and environmental variables. The elevation grid is provided in the background.](image2)

**Step 2. Model relationship with elevation and location**

To represent snow depth at a finer resolution, we followed a similar approach to Hijmans (2005) by modeling the relationship between climate model (CRCM) output and elevation, latitude, and longitude. We created this relationship using the regression-trees portion of the random forests statistical procedure as follows.

5. Create a sampling frame representing the center of every grid cell in and within 125 km of the state (Figure 2). The total number of points in this sample frame is 248.
6. Attribute these points at the same 45km cell size with the following variables: elevation, temperature, x-coordinate, y-coordinate. We then derived \((x\text{-coord})^2\), \((y\text{-coord})^2\), \((x\text{-coord})(y\text{-coord})\). See Borcord *et al.* 1992 for more information on the spatial components.
7. Model the relationship using the randomForest package in R (Liaw & Wiener 2002).
8. Apply the relationship to the entire state, using an elevation grid at 30-meters cell-size resolution.

Random forests was successful at fitting to the relationship with 95.59% of the variance explained. The most important independent variable for modeling this variable was elevation (Figure 4).
Figure 4. The relative importance of variables in fitting snow depth using the random forests statistical routine.

Literature Cited


Appendix 4. Species distribution model validation metadata
**Acris crepitans**  
Element Distribution Model (EDM) assessment metrics and metadata  
Common name: Northern Cricket Frog  
Date: 13 Apr 2011  
Code: acricrep4

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 9 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Grid cells = number of input cells with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid cells</td>
<td>30</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>1528</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.96</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.98</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>TSS</td>
<td>0.93</td>
<td>0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.93</td>
<td>0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>AUC</td>
<td>0.99</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 500 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

**Figure 1.** ROC plot for all 9 validation runs, averaged along cutoffs. The first cutoff indicated (0.015) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.127) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

**Figure 2.** Relative importance of each environmental variable based on the full model using all sites as input.

**Figure 3.** Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.

The New York Natural Heritage Program, 625 Broadway, Albany NY 12233-4757. www.nynhp.org
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.015) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.127) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine \[1, 2\] in the R statistical environment \[3, 4\]. We validated the model by jackknifing (also called leave-one-out, see \[5, 6, 7\]) by 10 km grid for a total of 31 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>31</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>156</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see \[8, 9, 6\].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.84</td>
<td>0.23</td>
<td>0.04</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.97</td>
<td>0.10</td>
<td>0.02</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.70</td>
<td>0.43</td>
<td>0.08</td>
</tr>
<tr>
<td>TSS</td>
<td>0.67</td>
<td>0.46</td>
<td>0.08</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.67</td>
<td>0.46</td>
<td>0.08</td>
</tr>
<tr>
<td>AUC</td>
<td>0.96</td>
<td>0.12</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 31 validation runs, averaged along cutoffs. The first cutoff indicated (0.046) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.054) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 \[10\].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.046) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.054) may be more appropriate.

References

Please cite this document and its associated EDM as:
**Ambystoma jeffersonianum x laterale**
Element Distribution Model (EDM) assessment metrics and metadata
Common name: Jefferson Salamander Complex
Date: 09 Jun 2011
Code: ambjefla1

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 77 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>77</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>259</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.77</td>
<td>0.23</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.17</td>
<td>0.02</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.58</td>
<td>0.45</td>
<td>0.05</td>
</tr>
<tr>
<td>TSS</td>
<td>0.54</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.54</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>AUC</td>
<td>0.93</td>
<td>0.19</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 77 validation runs, averaged along cutoffs. The first cutoff indicated (0.086) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.018) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.086) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.018) may be more appropriate.

References

Please cite this document and its associated EDM as:
**Ambystoma opacum**

Element Distribution Model (EDM) assessment metrics and metadata

Common name: Marbled Salamander
Date: 09 Jun 2011
Code: ambyopac2

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 36 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>36</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>104</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.75</td>
<td>0.24</td>
<td>0.04</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.98</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.53</td>
<td>0.46</td>
<td>0.08</td>
</tr>
<tr>
<td>TSS</td>
<td>0.51</td>
<td>0.48</td>
<td>0.08</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.51</td>
<td>0.48</td>
<td>0.08</td>
</tr>
<tr>
<td>AUC</td>
<td>0.98</td>
<td>0.08</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 36 validation runs, averaged along cutoffs. The first cutoff indicated (0.1) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.023) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.1) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.023) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 71 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

**Table 1. Input statistics.** Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>71</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>297</td>
</tr>
</tbody>
</table>

**Table 2. Validation statistics for jackknife trials.**

Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.83</td>
<td>0.24</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.18</td>
<td>0.02</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.71</td>
<td>0.41</td>
<td>0.05</td>
</tr>
<tr>
<td>TSS</td>
<td>0.65</td>
<td>0.49</td>
<td>0.06</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.65</td>
<td>0.49</td>
<td>0.06</td>
</tr>
<tr>
<td>AUC</td>
<td>0.95</td>
<td>0.19</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 71 validation runs, averaged along cutoffs. The first cutoff indicated (0.106) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.027) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.106) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.027) may be more appropriate.

References

Please cite this document and its associated EDM as:
*Coluber c. constrictor*
Element Distribution Model (EDM) assessment metrics and metadata
Common name: Racer
Date: 09 Jun 2011
Code: colucons2

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 52 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>52</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>262</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.81</td>
<td>0.21</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.99</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.63</td>
<td>0.43</td>
<td>0.06</td>
</tr>
<tr>
<td>TSS</td>
<td>0.62</td>
<td>0.43</td>
<td>0.06</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.62</td>
<td>0.43</td>
<td>0.06</td>
</tr>
<tr>
<td>AUC</td>
<td>0.97</td>
<td>0.10</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 52 validation runs, averaged along cutoffs. The first cutoff indicated (0.104) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.019) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with bargraphs.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.104) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.019) may be more appropriate.

References


Please cite this document and its associated EDM as:
Cordulegaster erronea
Element Distribution Model (EDM) assessment metrics and metadata
Common name: Tiger Spiketail
Date: 09 Jun 2011
Code: corderro3

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 8 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>8</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>143</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.77</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.58</td>
<td>0.45</td>
<td>0.16</td>
</tr>
<tr>
<td>TSS</td>
<td>0.53</td>
<td>0.41</td>
<td>0.15</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.53</td>
<td>0.41</td>
<td>0.15</td>
</tr>
<tr>
<td>AUC</td>
<td>0.83</td>
<td>0.20</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 500 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 8 validation runs, averaged along cutoffs. The first cutoff indicated (0.021) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.176) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.021) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.176) may be more appropriate.

References


Please cite this document and its associated EDM as:
Cordulegaster obliqua
Element Distribution Model (EDM) assessment metrics and metadata

Common name: Arrowhead Spiketail
Date: 09 Jun 2011
Code: cordobli3

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 17 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>17</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>754</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.87</td>
<td>0.18</td>
<td>0.04</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.97</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.77</td>
<td>0.37</td>
<td>0.09</td>
</tr>
<tr>
<td>TSS</td>
<td>0.74</td>
<td>0.36</td>
<td>0.09</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.74</td>
<td>0.36</td>
<td>0.09</td>
</tr>
<tr>
<td>AUC</td>
<td>0.92</td>
<td>0.17</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 17 validation runs, averaged along cutoffs. The first cutoff indicated (0.039) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.136) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.039) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.136) may be more appropriate.

References

Please cite this document and its associated EDM as:
**Crotalus horridus**

Element Distribution Model (EDM) assessment metrics and metadata

Common name: Timber Rattlesnake

Date: 09 Jun 2011

Code: crothorr3

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 73 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

### Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>73</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>6157</td>
</tr>
</tbody>
</table>

### Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.95</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.96</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.94</td>
<td>0.18</td>
<td>0.02</td>
</tr>
<tr>
<td>TSS</td>
<td>0.90</td>
<td>0.18</td>
<td>0.02</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.90</td>
<td>0.18</td>
<td>0.02</td>
</tr>
<tr>
<td>AUC</td>
<td>0.99</td>
<td>0.02</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 73 validation runs, averaged along cutoffs. The first cutoff indicated (0.21) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.088) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.21) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.088) may be more appropriate.

References

Please cite this document and its associated EDM as:

The New York Natural Heritage Program, 625 Broadway, Albany NY 12233-4757. www.nynhp.org
**Dendroica caerulescens**

Element Distribution Model (EDM) assessment metrics and metadata

Common name: Black-throated Blue Warbler

Date: 09 Jun 2011

Code: dendcaer1

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 40 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>40</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>99</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.71</td>
<td>0.26</td>
<td>0.04</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.22</td>
<td>0.03</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.47</td>
<td>0.47</td>
<td>0.07</td>
</tr>
<tr>
<td>TSS</td>
<td>0.42</td>
<td>0.52</td>
<td>0.08</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.42</td>
<td>0.52</td>
<td>0.08</td>
</tr>
<tr>
<td>AUC</td>
<td>0.85</td>
<td>0.32</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 40 validation runs, averaged along cutoffs. The first cutoff indicated (0.063) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.023) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.063) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.023) may be more appropriate.

References


Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 68 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>68</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>243</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.76</td>
<td>0.25</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.92</td>
<td>0.24</td>
<td>0.03</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.59</td>
<td>0.46</td>
<td>0.06</td>
</tr>
<tr>
<td>TSS</td>
<td>0.52</td>
<td>0.50</td>
<td>0.06</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.52</td>
<td>0.50</td>
<td>0.06</td>
</tr>
<tr>
<td>AUC</td>
<td>0.88</td>
<td>0.30</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.064) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.033) may be more appropriate.

References

Please cite this document and its associated EDM as:
**Elaphe obsoleta**

Element Distribution Model (EDM) assessment metrics and metadata

Common name: Eastern Ratsnake

Date: 09 Jun 2011

Code: elapobso1

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 64 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>64</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>344</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.79</td>
<td>0.22</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>1.00</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.58</td>
<td>0.44</td>
<td>0.05</td>
</tr>
<tr>
<td>TSS</td>
<td>0.57</td>
<td>0.44</td>
<td>0.05</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.57</td>
<td>0.44</td>
<td>0.05</td>
</tr>
<tr>
<td>AUC</td>
<td>0.93</td>
<td>0.24</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 64 validation runs, averaged along cutoffs. The first cutoff indicated (0.203) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.022) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.203) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.022) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 58 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>58</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>13418</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.91</td>
<td>0.17</td>
<td>0.02</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.96</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.85</td>
<td>0.33</td>
<td>0.04</td>
</tr>
<tr>
<td>TSS</td>
<td>0.81</td>
<td>0.34</td>
<td>0.04</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.81</td>
<td>0.34</td>
<td>0.04</td>
</tr>
<tr>
<td>AUC</td>
<td>0.96</td>
<td>0.13</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 300 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 58 validation runs, averaged along cutoffs. The first cutoff indicated (0.102) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.353) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.102) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.353) may be more appropriate.

References

Please cite this document and its associated EDM as:
**Eumeces fasciatus**

Element Distribution Model (EDM) assessment metrics and metadata

Common name: Five-lined Skink

Date: 09 Jun 2011

Code: eumefasc1

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 18 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>18</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>187</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.88</td>
<td>0.19</td>
<td>0.04</td>
</tr>
<tr>
<td>Specificity</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.76</td>
<td>0.37</td>
<td>0.09</td>
</tr>
<tr>
<td>TSS</td>
<td>0.76</td>
<td>0.37</td>
<td>0.09</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.76</td>
<td>0.37</td>
<td>0.09</td>
</tr>
<tr>
<td>AUC</td>
<td>1.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 18 validation runs, averaged along cutoffs. The first cutoff indicated (0.296) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.038) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.296) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.038) may be more appropriate.

References


Please cite this document and its associated EDM as:
**Eurycea longicauda**
Element Distribution Model (EDM) assessment metrics and metadata

Common name: Longtail Salamander
Date: 27 Apr 2011
Code: eurylong3

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 20 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

### Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>20</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>490</td>
</tr>
</tbody>
</table>

### Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.91</td>
<td>0.14</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.93</td>
<td>0.11</td>
<td>0.03</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.89</td>
<td>0.29</td>
<td>0.07</td>
</tr>
<tr>
<td>TSS</td>
<td>0.81</td>
<td>0.29</td>
<td>0.06</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.81</td>
<td>0.29</td>
<td>0.06</td>
</tr>
<tr>
<td>AUC</td>
<td>0.95</td>
<td>0.13</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 20 validation runs, averaged along cutoffs. The first cutoff indicated (0.02) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.147) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.02) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.147) may be more appropriate.

References

Please cite this document and its associated EDM as:
Glyptemys insculpta
Element Distribution Model (EDM) assessment metrics and metadata
Common name: Wood Turtle
Date: 09 Jun 2011
Code: glypinsc1

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid, grouped to 100 levels for a total of 100 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>100</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>306</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.79</td>
<td>0.23</td>
<td>0.02</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.19</td>
<td>0.02</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.64</td>
<td>0.44</td>
<td>0.04</td>
</tr>
<tr>
<td>TSS</td>
<td>0.59</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.59</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>AUC</td>
<td>0.95</td>
<td>0.19</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 100 validation runs, averaged along cutoffs. The first cutoff indicated (0.13) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.02) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.13) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.02) may be more appropriate.

References

Please cite this document and its associated EDM as:
Glyptemys muhlenbergii
Element Distribution Model (EDM) assessment metrics and metadata
Common name: Bog Turtle
Date: 09 Jun 2011
Code: glypmuhl3

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 31 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>31</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>3678</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.91</td>
<td>0.16</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.94</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.87</td>
<td>0.32</td>
<td>0.06</td>
</tr>
<tr>
<td>TSS</td>
<td>0.81</td>
<td>0.32</td>
<td>0.06</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.81</td>
<td>0.32</td>
<td>0.06</td>
</tr>
<tr>
<td>AUC</td>
<td>0.97</td>
<td>0.07</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 31 validation runs, averaged along cutoffs. The first cutoff indicated (0.038) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.151) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.038) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.151) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 84 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

### Table 1. Input statistics.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>84</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>6473</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 84 validation runs, averaged along cutoffs. The first cutoff indicated (0.176) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.073) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.176) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.073) may be more appropriate.

References


Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid, grouped to 100 levels for a total of 100 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>100</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>1039</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.77</td>
<td>0.23</td>
<td>0.02</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.93</td>
<td>0.21</td>
<td>0.02</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.61</td>
<td>0.44</td>
<td>0.04</td>
</tr>
<tr>
<td>TSS</td>
<td>0.54</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.54</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>AUC</td>
<td>0.89</td>
<td>0.26</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 100 validation runs, averaged along cutoffs. The first cutoff indicated (0.146) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.022) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.146) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.022) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 31 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>31</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>1202</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.86</td>
<td>0.19</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.93</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.79</td>
<td>0.37</td>
<td>0.07</td>
</tr>
<tr>
<td>TSS</td>
<td>0.72</td>
<td>0.37</td>
<td>0.07</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.72</td>
<td>0.37</td>
<td>0.07</td>
</tr>
<tr>
<td>AUC</td>
<td>0.92</td>
<td>0.19</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.038) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.199) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine \([1, 2]\) in the R statistical environment \([3, 4]\). We validated the model by jackknifing (also called leave-one-out, see \([5, 6, 7]\)) by 10 km grid, grouped to 100 levels for a total of 100 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>100</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>487</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see \([8, 9, 6]\).

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.79</td>
<td>0.22</td>
<td>0.02</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.16</td>
<td>0.02</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.62</td>
<td>0.44</td>
<td>0.04</td>
</tr>
<tr>
<td>TSS</td>
<td>0.58</td>
<td>0.44</td>
<td>0.04</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.58</td>
<td>0.44</td>
<td>0.04</td>
</tr>
<tr>
<td>AUC</td>
<td>0.94</td>
<td>0.19</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 100 validation runs, averaged along cutoffs. The first cutoff indicated (0.118) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.019) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 \([10]\).

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.118) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.019) may be more appropriate.

References


Please cite this document and its associated EDM as:

The New York Natural Heritage Program, 625 Broadway, Albany NY 12233-4757. www.nynhp.org
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid, grouped to 100 levels for a total of 100 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>100</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>4818</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.69</td>
<td>0.26</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.93</td>
<td>0.20</td>
<td>0.02</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.45</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>TSS</td>
<td>0.38</td>
<td>0.51</td>
<td>0.05</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.38</td>
<td>0.51</td>
<td>0.05</td>
</tr>
<tr>
<td>AUC</td>
<td>0.78</td>
<td>0.35</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 100 validation runs, averaged along cutoffs. The first cutoff indicated (0.192) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.03) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with bargraphs.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.192) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.03) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid, grouped to 100 levels for a total of 100 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>100</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>18940</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 300 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 100 validation runs, averaged along cutoffs. The first cutoff indicated (0.188) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.049) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.188) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.049) may be more appropriate.

References

Please cite this document and its associated EDM as:
**Helmitheros vermivorum**
Element Distribution Model (EDM) assessment metrics and metadata
Common name: Worm-eating Warbler
Date: 25 Apr 2011
Code: helmverm1

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 27 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>27</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>151</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.79</td>
<td>0.26</td>
<td>0.05</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.96</td>
<td>0.19</td>
<td>0.04</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.61</td>
<td>0.44</td>
<td>0.08</td>
</tr>
<tr>
<td>TSS</td>
<td>0.57</td>
<td>0.53</td>
<td>0.10</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.57</td>
<td>0.53</td>
<td>0.10</td>
</tr>
<tr>
<td>AUC</td>
<td>0.94</td>
<td>0.21</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

![Figure 1](image1.png)

**Figure 1.** ROC plot for all 27 validation runs, averaged along cutoffs. The first cutoff indicated (0.133) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.088) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

![Figure 2](image2.png)

**Figure 2.** Relative importance of each environmental variable based on the full model using all sites as input.

![Figure 3](image3.png)

**Figure 3.** Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.133) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.088) may be more appropriate.

References

Please cite this document and its associated EDM as:
**Hemidactylium scutatum**

Element Distribution Model (EDM) assessment metrics and metadata

Common name: Four-toed Salamander

Date: 25 Apr 2011

Code: hemiscut1

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 30 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>30</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>129</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.80</td>
<td>0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>Specificity</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.59</td>
<td>0.49</td>
<td>0.09</td>
</tr>
<tr>
<td>TSS</td>
<td>0.59</td>
<td>0.49</td>
<td>0.09</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.59</td>
<td>0.49</td>
<td>0.09</td>
</tr>
<tr>
<td>AUC</td>
<td>0.87</td>
<td>0.28</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 30 validation runs, averaged along cutoffs. The first cutoff indicated (0.058) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.032) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.058) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.032) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid, grouped to 100 levels for a total of 100 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>100</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>660</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

![Figure 1](image1.png)

Figure 1. ROC plot for all 100 validation runs, averaged along cutoffs. The first cutoff indicated (0.11) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.013) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

![Figure 2](image2.png)

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

![Figure 3](image3.png)

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.11) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.013) may be more appropriate.

References

Please cite this document and its associated EDM as:
Oporornis formosus
Element Distribution Model (EDM) assessment metrics and metadata
Common name: Kentucky Warbler
Date: 25 Apr 2011
Code: oporform2

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 13 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>13</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>2912</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.81</td>
<td>0.22</td>
<td>0.06</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.90</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.72</td>
<td>0.43</td>
<td>0.12</td>
</tr>
<tr>
<td>TSS</td>
<td>0.62</td>
<td>0.43</td>
<td>0.12</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.62</td>
<td>0.43</td>
<td>0.12</td>
</tr>
<tr>
<td>AUC</td>
<td>0.87</td>
<td>0.21</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 500 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 13 validation runs, averaged along cutoffs. The first cutoff indicated (0.013) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.197) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.

The New York Natural Heritage Program, 625 Broadway, Albany NY 12233-4757. www.nynhp.org
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.013) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.197) may be more appropriate.

References


Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 98 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>98</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>314</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.75</td>
<td>0.27</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.82</td>
<td>0.32</td>
<td>0.03</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.67</td>
<td>0.42</td>
<td>0.04</td>
</tr>
<tr>
<td>TSS</td>
<td>0.49</td>
<td>0.55</td>
<td>0.06</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.49</td>
<td>0.55</td>
<td>0.06</td>
</tr>
<tr>
<td>AUC</td>
<td>0.81</td>
<td>0.35</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 98 validation runs, averaged along cutoffs. The first cutoff indicated (0.046) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.025) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.046) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.025) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 17 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>17</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>265</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.96</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.96</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.95</td>
<td>0.13</td>
<td>0.03</td>
</tr>
<tr>
<td>TSS</td>
<td>0.91</td>
<td>0.16</td>
<td>0.04</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.91</td>
<td>0.16</td>
<td>0.04</td>
</tr>
<tr>
<td>AUC</td>
<td>0.98</td>
<td>0.05</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 17 validation runs, averaged along cutoffs. The first cutoff indicated (0.023) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.377) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.023) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.377) may be more appropriate.

References

Please cite this document and its associated EDM as:
This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 10 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>10</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>311</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.90</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>Specificity</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.80</td>
<td>0.42</td>
<td>0.13</td>
</tr>
<tr>
<td>TSS</td>
<td>0.80</td>
<td>0.42</td>
<td>0.13</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.80</td>
<td>0.42</td>
<td>0.13</td>
</tr>
<tr>
<td>AUC</td>
<td>0.90</td>
<td>0.22</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 500 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 10 validation runs, averaged along cutoffs. The first cutoff indicated (0.073) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.2) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.073) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.2) may be more appropriate.

References

Please cite this document and its associated EDM as:


**Terrapene c. carolina**

Element Distribution Model (EDM) assessment metrics and metadata

Common name: Eastern Box Turtle

Date: 25 Apr 2011

Code: terrcaro2

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 82 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

### Table 1. Input statistics.

Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>82</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>333</td>
</tr>
</tbody>
</table>

### Table 2. Validation statistics for jackknife trials.

Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.80</td>
<td>0.24</td>
<td>0.03</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.15</td>
<td>0.02</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.65</td>
<td>0.46</td>
<td>0.05</td>
</tr>
<tr>
<td>TSS</td>
<td>0.60</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.60</td>
<td>0.47</td>
<td>0.05</td>
</tr>
<tr>
<td>AUC</td>
<td>0.93</td>
<td>0.23</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

![Figure 1](image1.png)

**Figure 1.** ROC plot for all 82 validation runs, averaged along cutoffs. The first cutoff indicated (0.093) is generated by finding the point along this curve closest to the upper left-most corner. Validation statistics requiring a cutoff use this value. The second (0.034) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

![Figure 2](image2.png)

**Figure 2.** Relative importance of each environmental variable based on the full model using all sites as input.

![Figure 3](image3.png)

**Figure 3.** Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.

The New York Natural Heritage Program, 625 Broadway, Albany NY 12233-4757. www.nynhp.org
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.093) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.034) may be more appropriate.

References

Please cite this document and its associated EDM as:
**Thamnophis sauritus**

Element Distribution Model (EDM) assessment metrics and metadata

Common name: Eastern Ribbonsnake

Date: 25 Apr 2011

Code: thamsaur1

This EDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by 10 km grid for a total of 37 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Groups = number of input grid cells or cell groupings with PR points; BG points = background points; PR points = presence points placed throughout all polygons.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>37</td>
</tr>
<tr>
<td>BG points</td>
<td>10210</td>
</tr>
<tr>
<td>PR points</td>
<td>78</td>
</tr>
</tbody>
</table>

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [8, 9, 6].

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.76</td>
<td>0.24</td>
<td>0.04</td>
</tr>
<tr>
<td>Specificity</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.51</td>
<td>0.48</td>
<td>0.08</td>
</tr>
<tr>
<td>TSS</td>
<td>0.51</td>
<td>0.48</td>
<td>0.08</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.51</td>
<td>0.48</td>
<td>0.08</td>
</tr>
<tr>
<td>AUC</td>
<td>0.95</td>
<td>0.15</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Validation runs used 44 environmental variables, with 5 variables tried at each split (mtry) and 200 trees built. The final model was built using 600 trees, all presence and background points, with an mtry of 5, and the same number of environmental variables.

Figure 1. ROC plot for all 37 validation runs, averaged along cutoffs. The first cutoff indicated (0.079) is generated by finding the point along this curve closest to the upperleft-most corner. Validation statistics requiring a cutoff use this value. The second (0.028) uses the full model and maximizes the precision-recall F-measure using alpha=0.01 [10].

Figure 2. Relative importance of each environmental variable based on the full model using all sites as input.

Figure 3. Separation between presence and background points. Red and blue lines show densities of absence and presence points, respectively. One of each is drawn for each validation run. Arrows indicate cutoff locations as in Fig. 1.
Figure 4. Partial dependence plots for the eight environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin. Categorical variables are depicted with barplots.

Important! Element distribution models map places of similar environmental conditions to the submitted locations (PR points). No model will ever depict sites where a targeted element will occur with certainty, it can only depict locations it interprets as appropriate habitat for the targeted element. EDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an EDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower cutoff depicting more land area such as that derived from the validation ROC plots (0.079) may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher cutoff such as that derived from the final model (0.028) may be more appropriate.

References

Please cite this document and its associated EDM as: