Ecological Responses to Climate Change: Species Distribution Modelling using Maxent

Drew Kerkhoff

November 3, 2016

Species Distribution Models

One of the fundamental challenges for current ecologists is to anticipate the responses of complex ecological systems to anthropogenic climate change. The distribution and abundance of species, the diversity and structure of communities, and the functioning of whole ecosystems is strongly influenced by climatic patterns of temperature, precipitation, and their seasonal changes. And due fossil fuel combustion, deforestation, and other anthropogenic impacts, Earth's climate is warming at an unprecedented rate, changing not just local mean temperatures, but also atmospheric and oceanic circulation patterns, which determine the seasonality and extremes of precipitation and temperature.

One way to estimate how ecological systems will change in response to climate is to model how the favorable habitats available to a particular species will shift under different climate change scenarios. In a sense, this approach is based on considering changing climate in terms of the fundamental physiological tolerances of a species, in an attempt to describe the climatic or environmental niche of the species: the full set of climatic conditions under which the species is likely to occur. Once the "climatic space" occupied by a species is characterized, it can be mapped into geographic space, based either on current climatological data or projections based on models of future climate change. These approaches, known as Environmental Niche Models or Species Distribution Models have matured to the point that they are one of the foundational methods that ecologists use to make climate change predictions.

Our goal in this exercise is to develop species distribution models for a set of species native to our local environment here in Ohio, to analyze the relative importance of different climatic factors in determining the species' geographic range, and to apply our model to predict how the species will respond to climate change over the next half century based on the latest generation of climate change scenarios.

This exercise adapts code from Fiona Spooner at University College London, from the Zarnetske Spatial Ecology lab at Michigan State, and from the Species Distribution Modeling vignette that accompanies the 'dismo' package from Hijmans and Elith (2011). We have also installed the java executable file for the maxent model.

Case Study: The Timber Rattlesnake, *Crotalus horridus*

As an example, we will model the geographic distribution of the timber rattlesnake, *Crotalus horridus*, which is one of two native rattlesnakes in Ohio. Based on field observations of *C. horridus* at specific locations, we will use climate data to model the distribution of the species. After analyzing the model to see which climate variables are the strongest predictors of the species' range, we will then use climate change predictions for 2070 to assess how the suitability of *C. horridus* habitat will change over the region.

Timber Rattlesnake

To get started, we need several libraries, which all need to be installed up front.

```r
library(raster)
```

```
## Loading required package: sp
```

```r
library(rgdal)
```

```
## rgdal: version: 1.1-10, (SVN revision 622)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 1.11.4, released 2016/01/25
##  Path to GDAL shared files: /Library/Frameworks/R.framework/Versions/3.3/Resources/
library/rgdal/gdal
##  Loaded PROJ.4 runtime: Rel. 4.9.1, 04 March 2015, [PJ_VERSION: 491]
##  Path to PROJ.4 shared files: /Library/Frameworks/R.framework/Versions/3.3/Resource
s/library/rgdal/proj
##  Linking to sp version: 1.2-3
```

```r
library(maps)
library(mapdata)
library(dismo)
library(rJava)
library(maptools)
```

```
## Checking rgeos availability: TRUE
```

```
library(jsonlite)
```

Data Access and Preliminary Visualization

Our modeling effort is going to be based on data from a couple of large "ecoinformatic" efforts. The current and future environmental data are drawn from the Worldclim project (http://worldclim.org (http://worldclim.org), Hijmans et al. 2005), which has collected globally standardized climate coverages at very high resolution. Specifically, we will be using their data at a resolution of 2.5 arc seconds of latitude and longitude. Specifically, we will be examining "bioclimatic" variables that are likely to influence species distributions. We can download the data directly using the 'getData()' function from the 'raster' package

```
currentEnv=getData("worldclim", var="bio", res=2.5)
```

Worldclim also compiles matched future climate projections for a large collection of global climate models (GCM's) used in the phase 5 Coupled Model Intercomparison Project (CMIP5). Specifically, we are going to use estimates from the Hadley Centre Global Environmental Model version 2, with Earth System components (HADGEM2-ES). Earth System models include a bit more ecology, allowing landcover (vegetation) and hydrology to change more dynamically as the climate changes. To estimate change, we are going to use the most extreme climate change scenario (RCP8.5) which assumes that greenhouse gas emissions will continue to increase through 2100. We are downloading the predictions for 2070.

```
futureEnv=getData('CMIP5', var='bio', res=2.5, rcp=85, model='HE', year=70)
names(futureEnv)=names(currentEnv)
```

The climate data have lots of different variables, many of which our correlated with one another. To make things a bit simpler, we will limit ourselves to a bit smaller set of bioclimatic predictors.

```
currentEnv=dropLayer(currentEnv, c("bio2", "bio3", "bio4", "bio10", "bio11", "bio13",
"bio14", "bio15"))
futureEnv=dropLayer(futureEnv, c("bio2", "bio3", "bio4", "bio10", "bio11", "bio13", "b
io14", "bio15"))
```

Our remaining predictors are:

- BIO1 = Annual Mean Temperature
- BIO5 = Max Temperature of Warmest Month
- BIO6 = Min Temperature of Coldest Month
- BIO7 = Temperature Annual Range (BIO5-BIO6)
- BIO8 = Mean Temperature of Wettest Quarter
- BIO9 = Mean Temperature of Driest Quarter
- BIO12 = Annual Precipitation
- BIO16 = Precipitation of Wettest Quarter
- BIO17 = Precipitation of Driest Quarter
- BIO18 = Precipitation of Warmest Quarter
- BIO19 = Precipitation of Coldest Quarter

Now that we have our environmental data we need to get records of our species occurrences, so we can know what the climate is like where they are known to occur. Such occurrence data have been compiled for millions of species worldwide by the Global Biodiversity Information Facility (GBIF, http://gbif.org (http://gbif.org)). The

'dismo' package has a 'gbif()' function for accessing data directly. All you need is the scientific name for the species.

```
rattler=gbif('crotalus','horridus')
```

```
## 5851 records found
```

```
## 0-
```

```
## 300-
```

```
## 600-
```

```
## 900-
```

```
## 1200-
```

```
## 1500-
```

```
## 1800-
```

```
## 2100-
```

```
## 2400-
```

```
## 2700-
```

```
## 3000-
```

```
## 3300-
```

```
## 3600-
```

```
## 3900-
```

```
## 4200-
```

```
## 4500-
```

```
## 4800-
```

```
## 5100-
```

```
## 5400-
```

```
## 5700-
```

```
## 5851 records downloaded
```

Before we run our model, we need to validate the observational data to make sure that they have geographic locations and that all the observations are reasonable. First, we will remove all the observations without latitudes and longitudes.

```
rattler=subset(rattler, !is.na(lon) & !is.na(lat)) # the ! symbol can be used to mean
  'not'
```

Next, we want to find and eliminate any duplicate locations.

```
rattlerdups=duplicated(rattler[, c("lon", "lat")])
rattler <-rattler[!rattlerdups, ]
```

Now let's look at a preliminary map to see how our *C. horridus* observations are distributed. We'll make the map extend 1 degree around the observations

```
data(wrld_simpl)
plot(wrld_simpl, xlim=c(min(rattler$lon)-1,max(rattler$lon)+1), ylim=c(min(rattler$lat
)-1,max(rattler$lat)+1), axes=TRUE, col="light yellow")
points(rattler$lon, rattler$lat, col="orange", pch=20, cex=0.75)
```
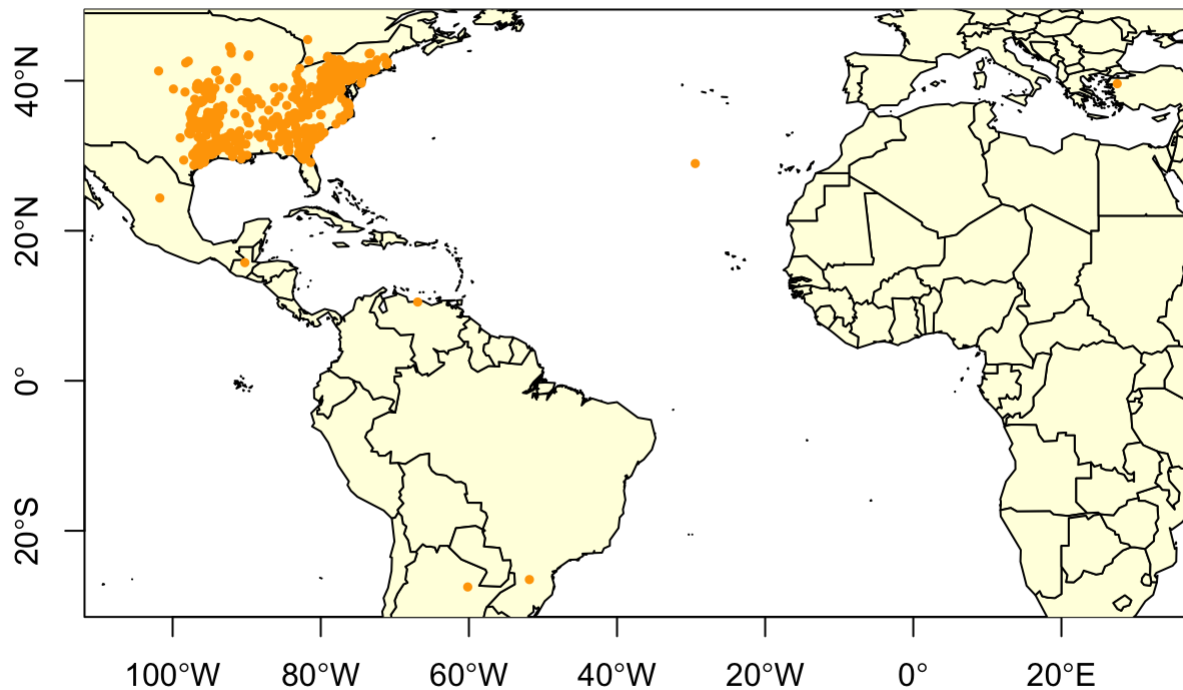
Notice that we have some questionable observations, *e.g.* in the Middle East and the middle of the Atlantic Ocean. These could represent data transcription errors such as reversing the sign of the geographic coordinates. To be conservative we will limit our data to the what appears to be the 'native' range of the species in eastern North America. **For different species, this area will be different.**

```
rattler <- rattler[rattler$lon < -40 & rattler$lat > 25 , ]
```

To check our data again, we can use a higher resolution world map.

```
map('worldHires', xlim=c(min(rattler$lon)-1,max(rattler$lon)+1), ylim=c(min(rattler$lat)-1,max(rattler$lat)+1), fill=TRUE, col="light yellow")
points(rattler$lon, rattler$lat, col="orange", pch=20, cex=0.75)
```
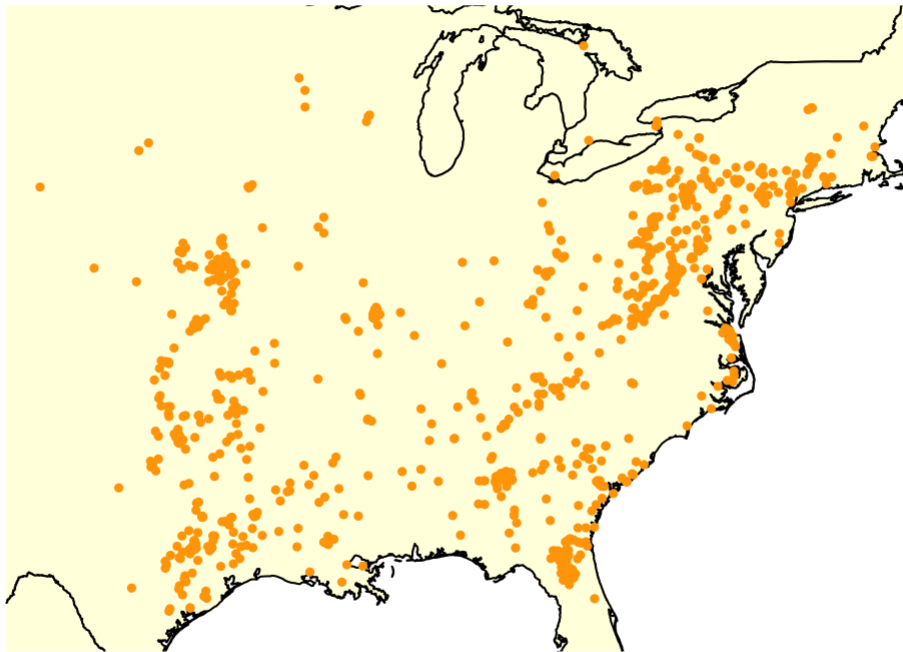
That looks reasonable. Now, before we model the climatic niche of the species, we need to consider a reasonable geographic area, because like most other approaches in ecology, the results of SDMs are *scale-dependent*. We want to capture the range of climates that define both the most suitable *core* habitat of the species where it predictably occurs, but also capture environmental conditions in which it is reliably *absent* so we can reasonably estimate the boundaries. However, if we range too widely, the 'signal' of the favorable environments could be lost in the 'noise' of the widespread poor environments. For the purposes of this necessarily preliminary approach, we will make a ten degree buffer around the extremes of the species range. To trim all of our predictor data to this region we can define a geographic *extent* then crop the climate raster data to this extent.

```
model.extent<-extent(min(rattler$lon)-10,max(rattler$lon)+10,min(rattler$lat)-10,max(r
attler$lat)+10)
modelEnv=crop(currentEnv,model.extent)
modelFutureEnv=crop(futureEnv, model.extent)
```

To get an idea of how the environment varies across the species range, let's look at the mean annual temperature (the 'bio1' variable)

```
plot(modelEnv[["bio1"]]/10, main="Annual Mean Temperature")
map('worldHires',xlim=c(min(rattler$lon)-10,max(rattler$lon)+10), ylim=c(min(rattler$l
at)-10,max(rattler$lat)+10), fill=FALSE, add=TRUE)
points(rattler$lon, rattler$lat, pch="+", cex=0.2)
```

Annual Mean Temperature
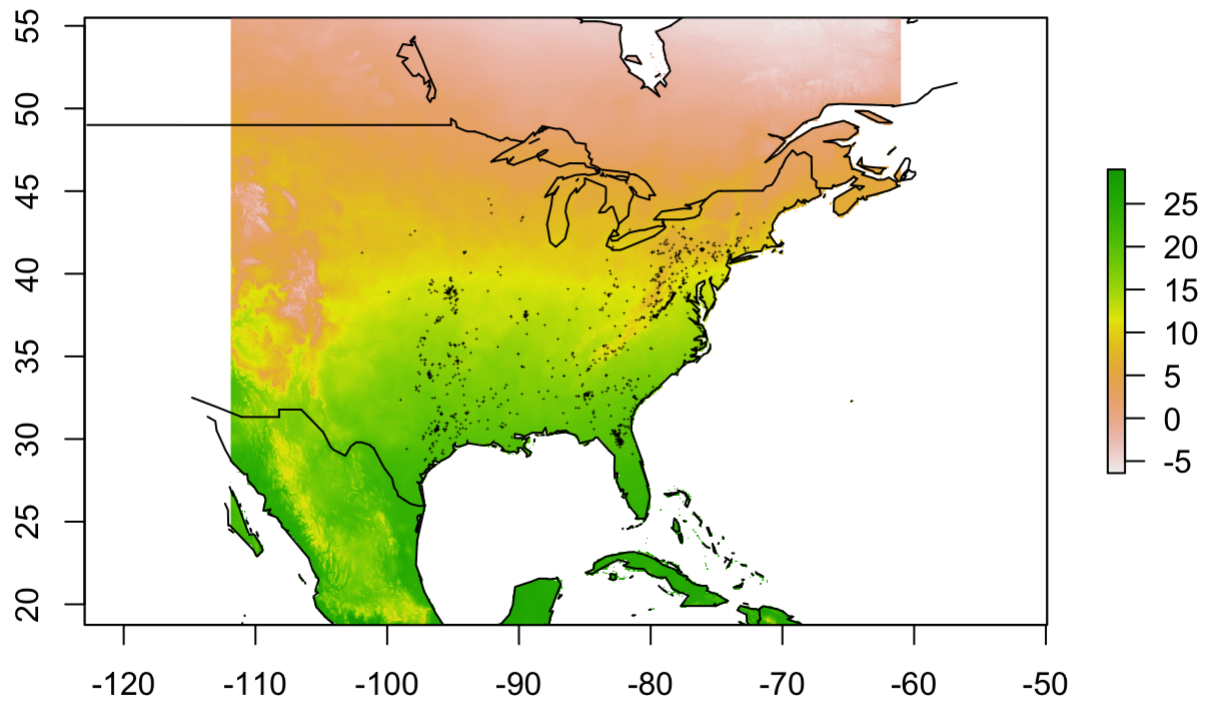
To see how the climate is predicted to change, let's compare that to the 2070 mean annual temperature projection.

```
plot(modelFutureEnv[["bio1"]]/10, main="Future Annual Mean Temperature")
map('worldHires',xlim=c(min(rattler$lon)-10,max(rattler$lon)+10), ylim=c(min(rattler$l
at)-10,max(rattler$lat)+10), fill=FALSE, add=TRUE)
points(rattler$lon, rattler$lat, pch="+", cex=0.2)
```

Future Annual Mean Temperature
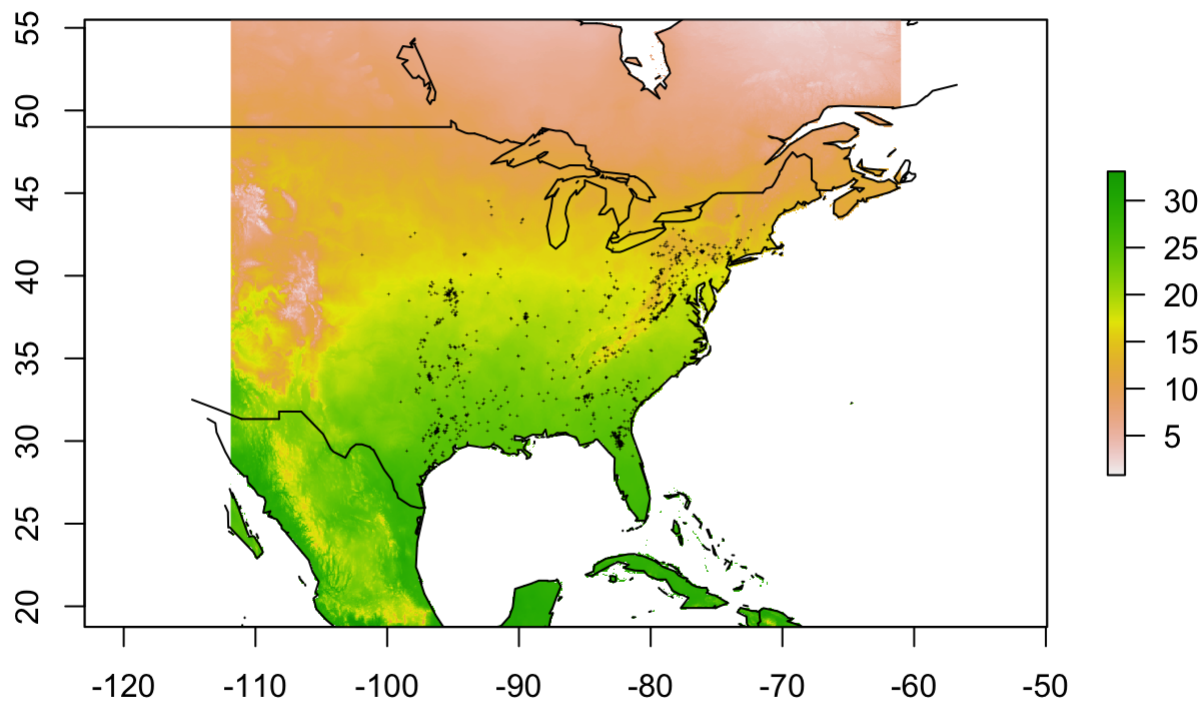
Species Distribution Model Construction and Assessment

Now that we have our data together, we are ready to construct our species distribution model. But before we think about using the model to project the potential response of the species' range to climate change, we need to assess how well the model can actually predict whether we will find the species in a particular location. To make this assessment, will use 'cross-validation', setting aside some of our species locations, and using them to *test* the model, which will be constructed using the remaining data points. In our case, we will randomly withhold 20% of the observations as *test data* and retain the other 80% as *training data*.

```
rattlerocc=cbind.data.frame(rattler$lon,rattler$lat) #first, just make a data frame of
latitudes and longitudes for the model
fold <- kfold(rattlerocc, k=5) # add an index that makes five random groups of observa
tions
rattlertest <- rattlerocc[fold == 1, ] # hold out one fifth as test data
rattlertrain <- rattlerocc[fold != 1, ] # the other four fifths are training data
```

In real applications, since the particulars of the model depend on the data used to fit it, we would actually fit the model multiple times, withholding each fifth of the data separately, then average the results. This is called *k-fold cross-validation* (in our case 5-fold). However, for our purposes here, we will just fit the model once.

Now we can fit the SDM using the Maximum Entropy (Maxent) algorithm, which tries to define the combination of environmental responses that best predicts the occurrence of the species.

```
rattler.me <- maxent(modelEnv, rattlertrain) #note we just using the training data
```

Model visualization and evaluation

To explore the model and use it to understand the climatic determinants of the species range, we can first compare the relative importance of the different predictors in the final model.

```
plot(rattler.me)
```

Variable contribution
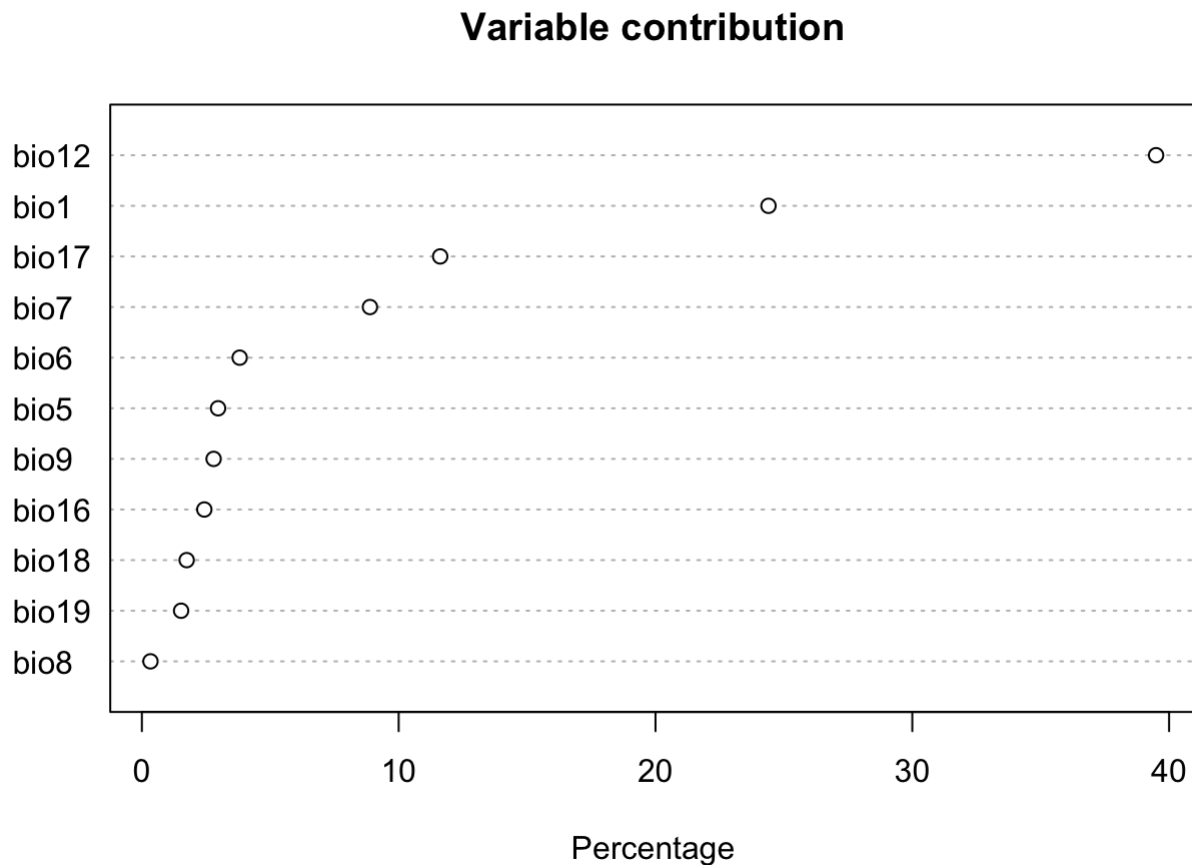
From this plot, we can see that the model is most sensitive to variation in mean annual precipitation ('bio12'), with additional contributinos from mean annual temperature ('bio1'), the precipitation of the driest quarter ('bio17'), the minimum temperature of the coldest month ('bio6'), and the annual temperature range ('bio7').

So how does the likelihood of species occurrence respond to variation in these climatic conditions? To see the shape of the response curves estimated by the model, we can use the cleverly named 'response()' function.

```
response(rattler.me)
```
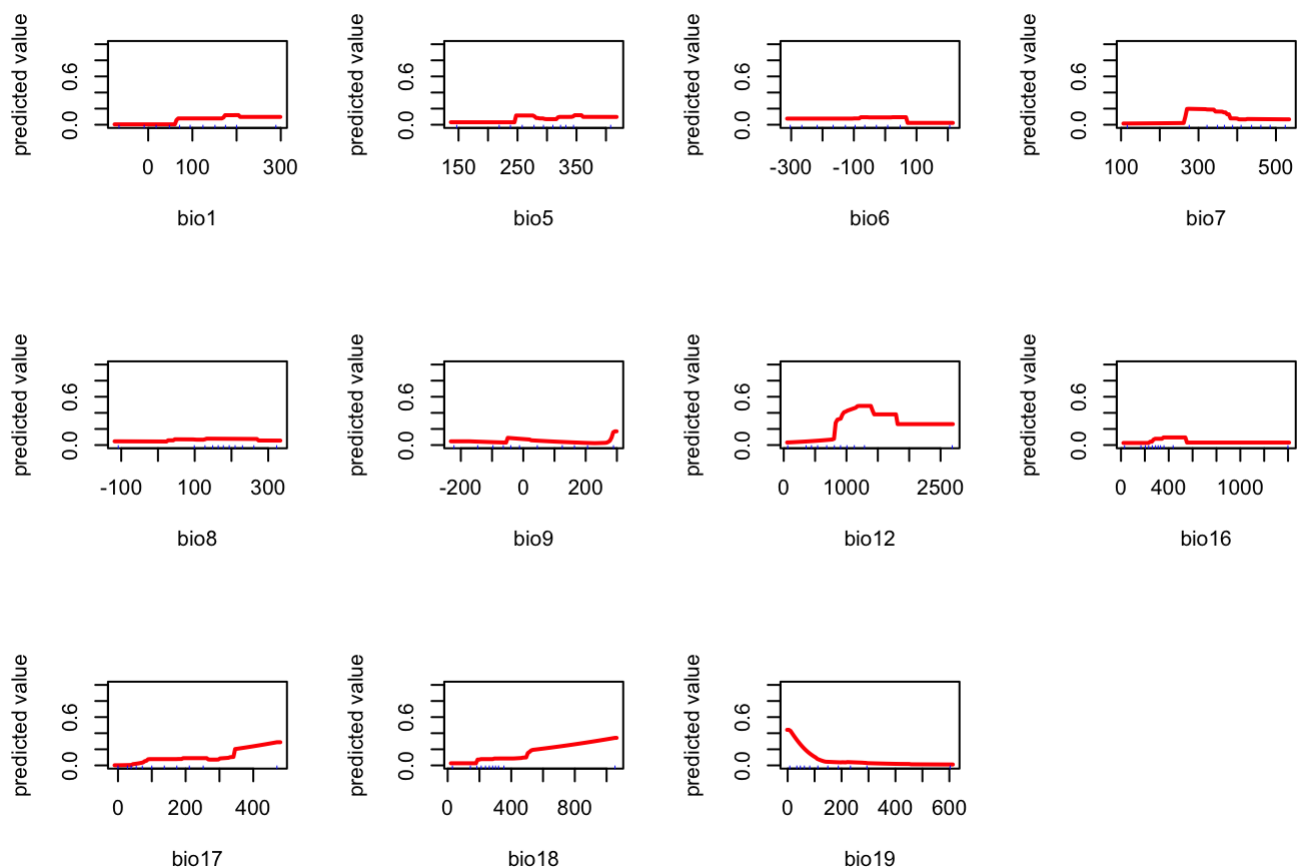
In the response plots, we are looking at how the probability of the species occurring (Y-axes, from zero to one) varies with each the environmental predictors (X-axes). From these plots, we can see that the Maxent models can include complex environmental responses including thresholds, linear, and nonlinear shapes. For example, if we look at mean annual precipitation ('bio12'), we can see that below a threshold of about 750 mm of rain per year, the probability of *C. horridus* occurrence is about zero, and that it increases nonlinearly then plateaus in wetter regions.

So what does this model predict about the distribution of favorable habitats across eastern North America? To map the predicted probability of occurrence, we must first generate the predicted values for every cell in our region.

```
rattler.pred <- predict(rattler.me, modelEnv)
```

Then, we can map those predicted values and superimpose our original locations.

```
plot(rattler.pred, main="Predicted Suitability")
map('worldHires', fill=FALSE, add=TRUE)
points(rattler$lon, rattler$lat, pch="+", cex=0.2)
```

Predicted Suitability
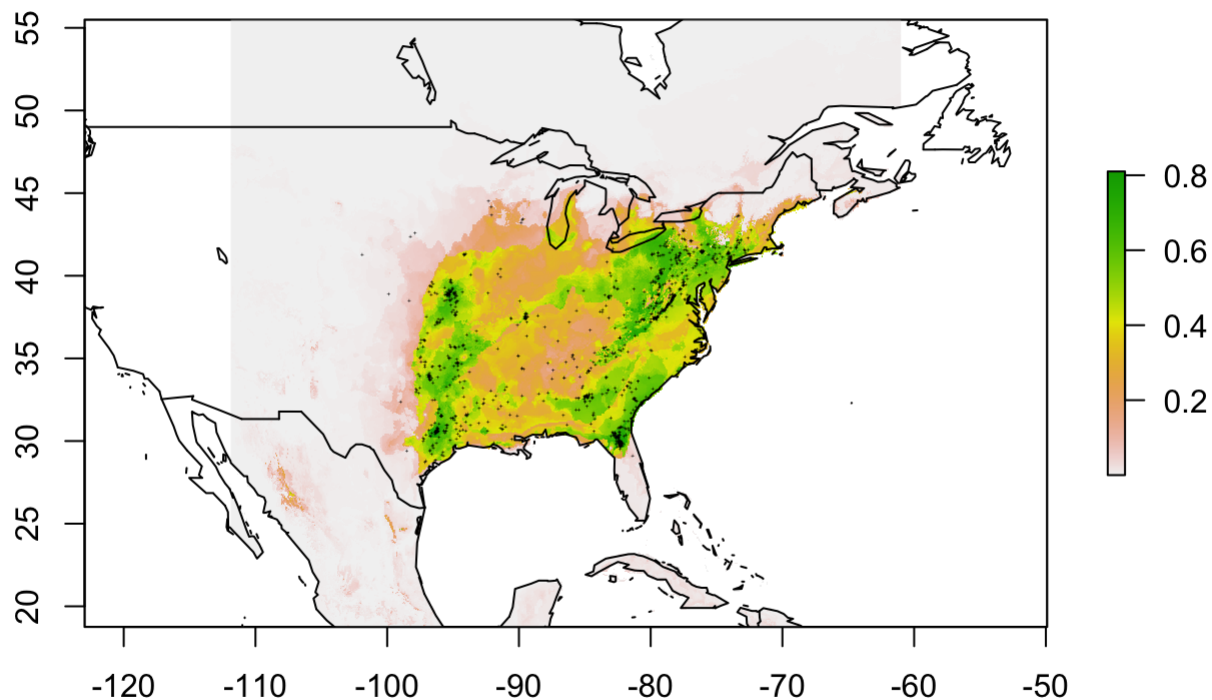
There are a few things we can notice from this model output. First, comfortingly, habitat is predicted to be good in areas where there are lots of species occurrences. This is probably not too surprising, since the data points were used to fit the model. However, note that there are also large expanses of habitat that is predicted to be quite good, but where few or no observations occur (*e.g.*, northern Ohio). Moreover, there are actual observations in several places where the probability of occurrence is predicted to be quite low, below even 0.2 or 0.1. Finally, it is interesting to note that there seem to be three distinct regions of favorable habitat, one in the north, one in the south, and one in the west. Patterns like this in model output can be used to generate hypotheses about the genetic structure across the species range, about local phenotypic adaptation, and even to suggest areas where additional surveys might be beneficial for finding populations of endangered species.

But how well does the model actually do at predicting the occurrence of *C. horridus*? We must answer this critical question before we think about using the model to predict responses to climate change. After all, the predictions from a crappy model (even if it is the 'best' model), are still crappy!

To evaluate the predictive accuracy of the model, we turn back to our test data, and use cross-validation to test the model. Our evaluation will use an approach called the Area Under the Receiver Operator Curve (AUC). Basically, the process is to set thresholds on the prediction to generate different levels of the false postive rate (predicting presence for places where the species is absent), then assess the true postive rate (successfully predicting presence) as a function of the false positive rate. The area under this curve, which varies from zero to one, provides an assessment of the model. An AUC value of 0.5 is the same as random guessing of presence/absence, while values towards one mean our predictions are more reliable.

Of course, we only have presence data, not absences. Luckily, the AUC procedure can be adapted for presence only data by generating 'pseudoabsences' from random points within the region. There are many caveats to this process, but for our purposes, it is unlikely to lead us too far astray.

To generate and evaluate the AUC for our model, we first generate background points for pseudoabsences. The 'randomPoints()' function even makes sure that the points occur only in areas where the predictor variables exist. That is, none of our points will occur in the ocean.

```
bg <- randomPoints(modelEnv, 1000)
```

Then we can use 'evaluate()' to generate several diagnostics as well as the AUC, using our test data as our presences against our pseudoabsences.

```
e1 <- evaluate(rattler.me, p=rattlertest, a=bg, x=modelEnv)

plot(e1, 'ROC')
```
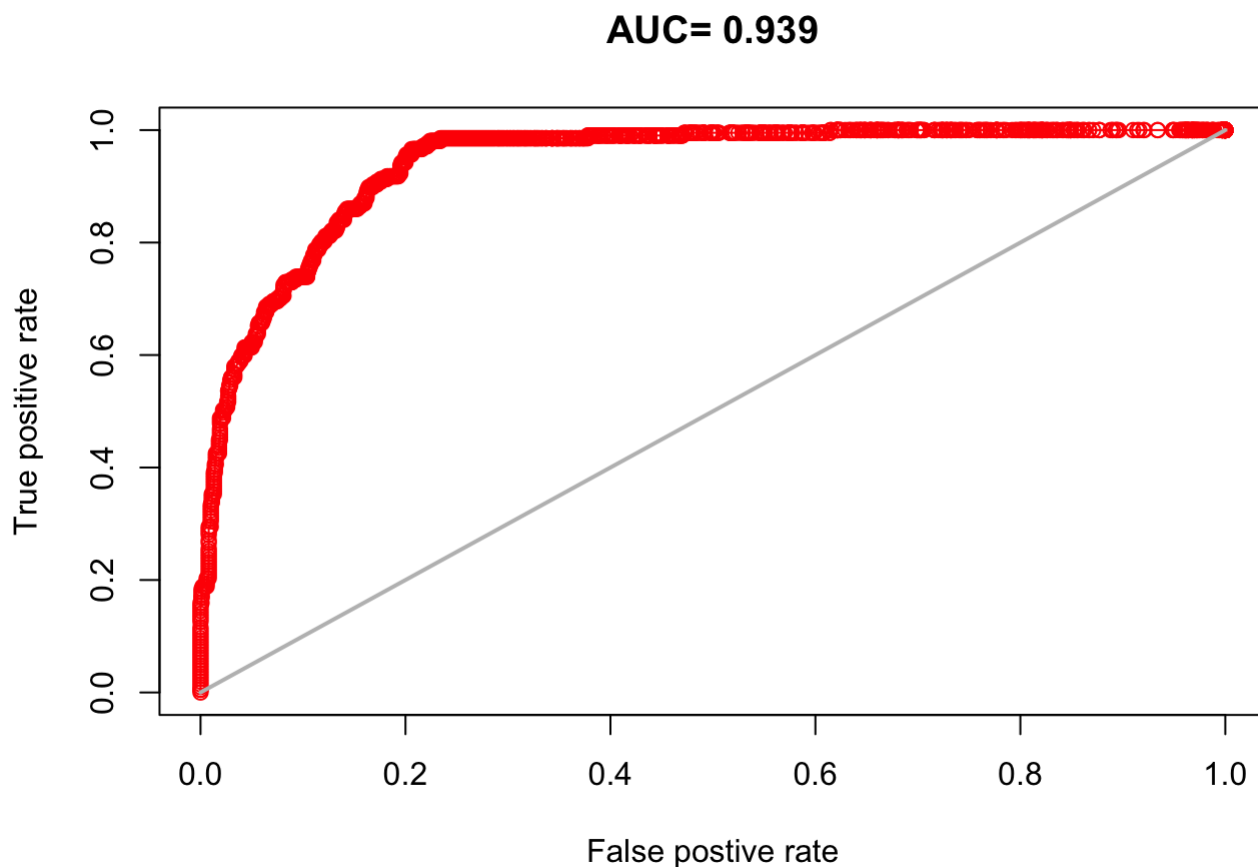

On the receiver operator curve, the 1:1 line give AUC of 0.5. From our curve and the AUC, it is clear that our model appears to do substantially better than random guessing. But again, we want to be cautious since we are not using actual absence data. More careful assessments can be made by incorporating other factors, such as measures of sampling intensity, into the generation of the pseudoabsences.

Projecting responses to climate change

So now that we have provisionally validated our species distribution model for *C. horridus*, we can ask how the distribution of *potential habitat* will shift under climate change. To do so, all we need to do is use our 2070 climate estimates to predict habitat suitability using the model.

```
rattler.2070 = predict(rattler.me, modelFutureEnv)
```

Just as before, we can then map the predicted values. As we did last time, we will superimpose the current distribution of observations.

```
plot(rattler.2070, main="Predicted Future Suitability")
map('worldHires', fill=FALSE, add=TRUE)
points(rattler$lon, rattler$lat, pch="+", cex=0.2)
```


Predicted Future Suitability
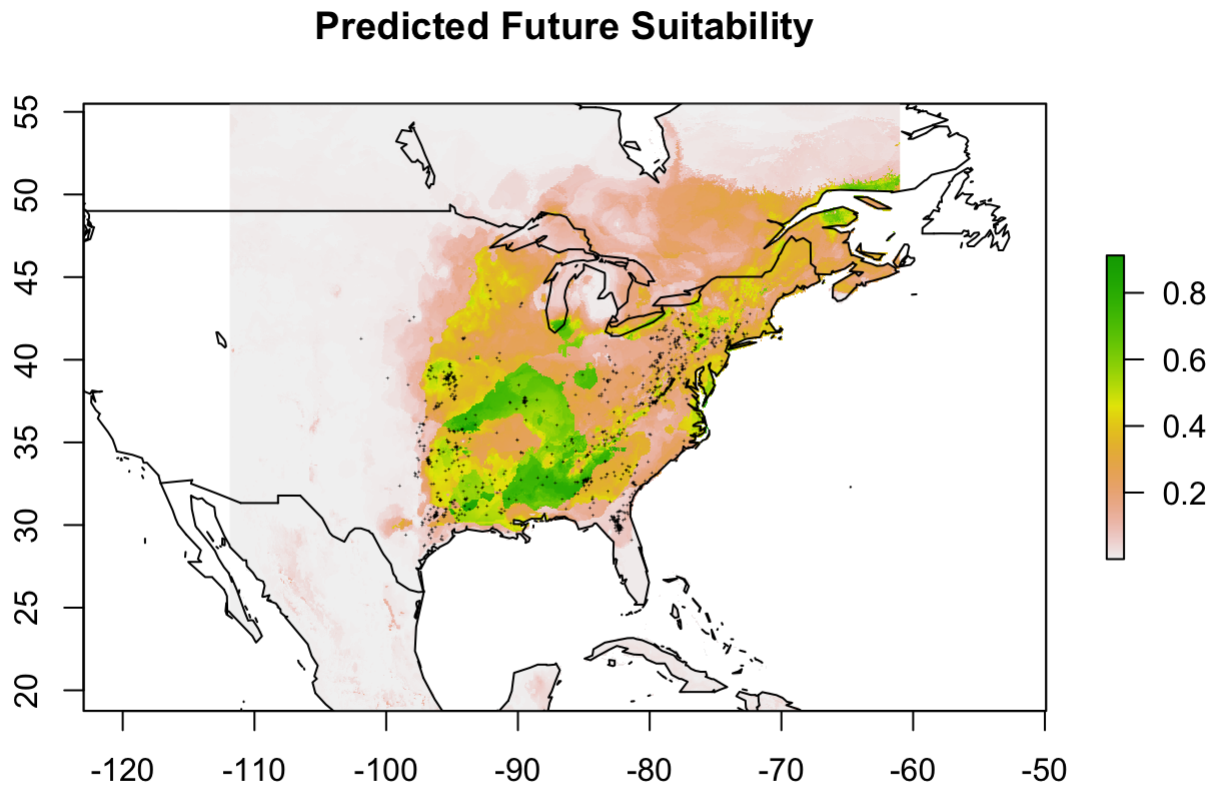
Compared to our current predictions above, we can see subtle, but substantial changes. In particular, even though we do observe the poleward shift in available habitat, the effects of climate change vary substantially across the three areas of favorable habitat we observed above. The western region of the range appears to exhibit an expansion of favorable habitat largely overlapping or adjacent to currently favorable regions. However, the southern and northern areas of favorable habitat are substantially fragmented and reduced, which could lead to smaller, more isolated populations in these areas.

It is, of course, critically important to keep in mind that these predicted changes are based only on climate, ignoring many other important factors like dispersal ability and the distribution of human-dominated land use like agriculture and urban areas.

Another way to visualize the response to climate change is to map the difference in habitat suitability between current conditions and those in 2070, so that positive values represent regions where conditions are better for *C. horridus* and negative values represent declines in habitat quality.

```
rattler.change=rattler.2070-rattler.pred
plot(rattler.change, main="Predicted Change in Suitability")
map('worldHires', fill=FALSE, add=TRUE)
points(rattler$lon, rattler$lat, pch="+", cex=0.2)
```

Predicted Change in Suitability
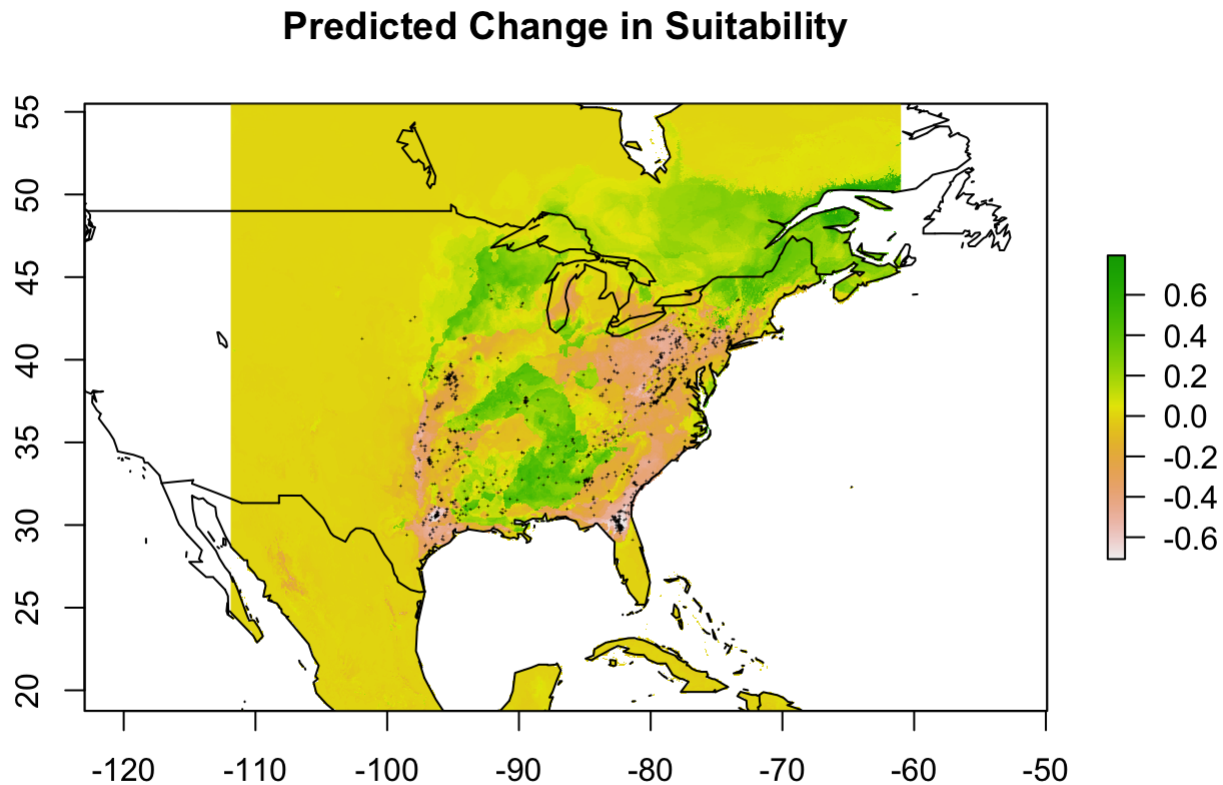
Note here that the color scale has changed, with yellow representing 'no change', greener colors representing improvement, and browner colors representing degradataion. Distressingly, at many of the locations where *C. horridus* currently occurs, the suitability of climatic conditions appears likely to decline. We can visualize these more specific changes by extracting the values of the 'change map' for each occurrence and examining a histogram.

```
rattlerChangePoints = extract(rattler.change, rattlerocc)
hist(rattlerChangePoints, main="")
abline(v=0, col="red")
```

rattlerChangePoints
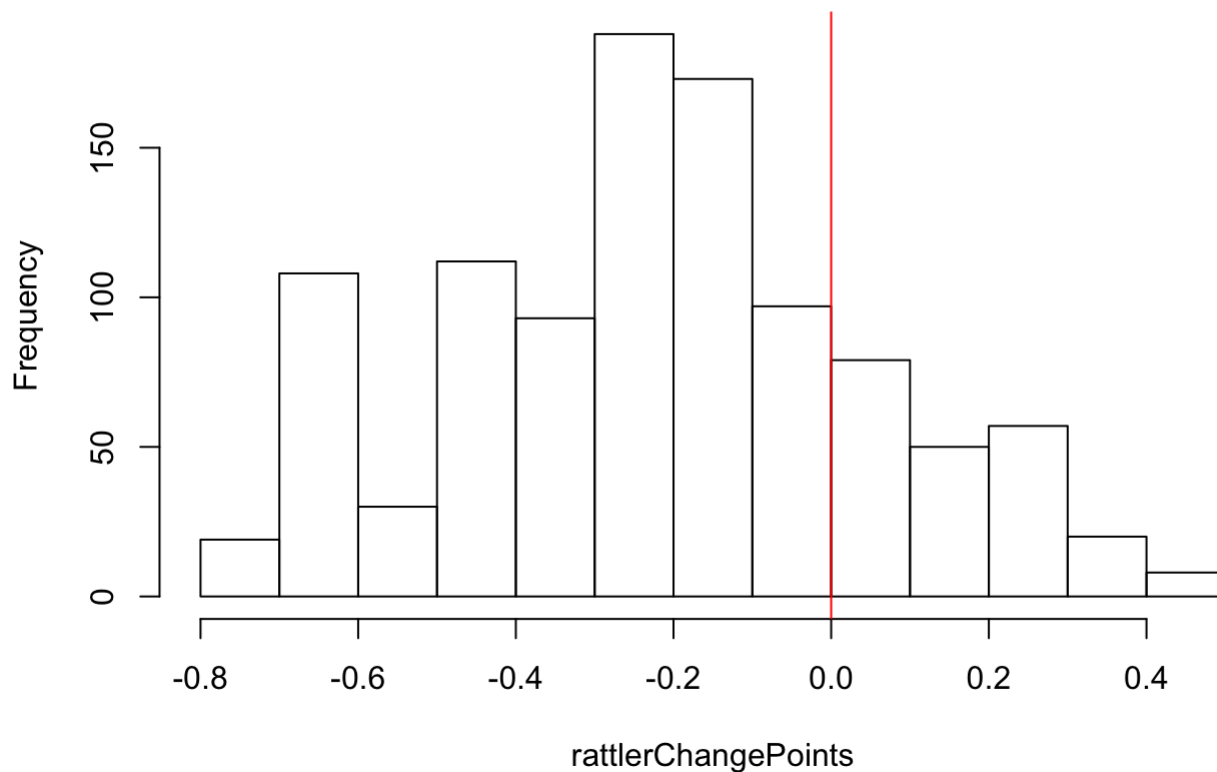
From this figure, we can see that in most areas where it occurs, continued greenhouse gas emissions are likely to make climatic conditions less suitable for *C. horridus* populations. At the same time, there are some areas where conditions will become more favorable climatically.

Predictions like these are of course, subject to uncertainty on many levels, from the climate model predictions themselves to the potential idiosyncrasies of wildlife sampling to issues of statistical power and overfitting. Much of the current research in the field is focused on identifying and estimating the impact of those uncertainties, in an attempt to improve our ability to forecast responses to climate change. But even from a preliminary analysis like this, we can see that ecological responses to climate change cannot simply be boiled down to 'species need to move north as the climate warms.' The pattern of climate variation over the continental landscape, together with the environmental tolerances of different species, leads to contingencies and complexity in the response of terrestrial ecosystems to climate change.