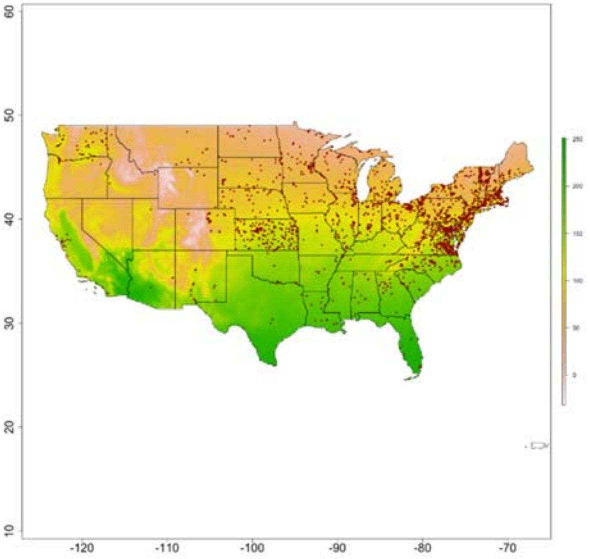
ISSUES : DATA SET

Painting turtles: an introduction to species distribution modeling in R

Anna L. Carter

Department of Ecology, Evolution & Organismal Biology, Iowa State University;
acarter1@iastate.edu



THE ECOLOGICAL QUESTION:

How are abiotic environmental conditions associated with a species' geographic distribution?

ECOLOGICAL CONTENT:

species distributions, habitat suitability

WHAT STUDENTS DO:

Students use R to search, download, and plot the distribution of painted turtles from the Global Biodiversity Information Facility (GBIF), then implement the BioClim modeling algorithm to build an occurrence-based species distribution model for painted turtles using the BioClim variables.

STUDENT-ACTIVE APPROACHES:

Cooperative learning: Students can work through the exercises in groups, using either the painted turtle data or group-specific taxa (i.e., each group selects a taxon to model). This module addresses multiple, complex concepts – online data availability & access, spatial query & analysis, global climate models, theoretical concepts of SDMs, model fitting – any of which can be expanded depending on the instructor's preference, group work can be especially beneficial for facilitating comprehension and may be more efficient.

Jigsaw: The module can be easily extended to allow each student/group to examine the distribution of a different species and explore data availability among taxa as well as different aspects of model fitting.

SKILLS:

Geospatial/mapping skills, R/command-line skills, data manipulation and plotting of large spatial datasets, introductory distribution modeling, searching public databases

TIEE, Volume 13 © 2017 – Anna L. Carter and the Ecological Society of America. *Teaching Issues and Experiments in Ecology* (*TIEE*) is a project of the Committee on Diversity and Education of the Ecological Society of America (http://tiee.esa.org).

ASSESSABLE OUTCOMES:

- In addition to using the provided student handout as an assessment tool, instructors can use the provided short-answer questions to build essay assessments, highlighting aspects of the material that are of particular interest to the course as a whole. Through completion of the module material students should be able to:

    1. Identify publically-accessible sources of occurrence records for different species and discuss the attribute data that are associated with those records.

    2. Visualize and describe spatial patterns in occurrence data.

    3. Identify the types and structure of data available in climate/bioclimate layers and discuss how those data are selected for building SDMs.

    4. Explain how the Bioclim algorithm fits species occurrence records.

    5. Discuss differences between presence-only and presence-absence SDMs and some of the strengths/limitations of each.

SOURCE:

- GBIF database [http://www.gbif.org/]
- WorldClim database [http://www.worldclim.org/bioclim]
- US Census Bureau [https://www.census.gov/geo/maps-data/data/cbf/cbf_state.html]

OVERVIEW OF THE ECOLOGICAL BACKGROUND

In this exercise, students will gain insight into the process of building an occurrence-based species distribution model (SDM) by investigating the distribution of painted turtles (*Chrysemys picta*), a semi-aquatic species that is widespread throughout North America. Students will gain familiarity with available sources of climate and distribution data and acquire basic skills for handling geospatial data and spatially-explicit ecological datasets in open-source software. The module contains copy-and-paste code for R as well as example questions that walk students through the exercise and prompt them to critically consider the limitations of spatially-explicit data and occurrence-based distribution models for answering ecological questions. As written, the module does not explicitly examine whether the SDM is a 'good' model for painted turtles. However, this exercise is a much-simplified version of the vignettes in "Species distribution modeling with R" by Hijmans & Elith (2017) and can be extended to include more advanced exercises, such as statistical exploration of the BioClim variables, model validation, or the effects of climate change. Additional references are listed in the Faculty Notes.

DATA SETS

All the data used in this module are freely available online.

- a folder containing a .shp file and associated metadata for plotting a 20m map of US states, downloaded from the US Census Bureau [https://www.census.gov/geo/maps-data/data/cbf/cbf_state.html]. This is the only file that students need to be provided. The remaining files are for instructor reference only.: usa.zip
- cleaned version of the painted turtle occurrence records that students download from the GBIF database [http://www.gbif/org]: occur.csv.
- a folder containing .png files of the five maps students create during the exercise: maps.zip
- R code for completing the exercises, copied from the student handout: painting_turtles.R

STUDENT INSTRUCTIONS

Painting turtles: an intro to species distribution modeling in R

Species distribution models (SDMs), mathematical representations of the environmental conditions correlated with the geographic distribution of a species, are one of the most commonly-used tools in ecology, evolution, and conservation biology. In this exercise, you will gain insight into the process of building an occurrence-based SDM by investigating the distribution of painted turtles (*Chrysemys picta*), a semi-aquatic species that is widespread throughout North America. You will also gain familiarity with available sources of climate and distribution data and acquire basic skills for handling geospatial data and spatially-explicit ecological datasets in open-source software.

Instructions

Work through the following exercises to build a basic, presence-only species distribution model (SDM) for painted turtles in R. Lines of code can be copied/pasted directly into your R console and run by hitting <Enter>. Answer the numbered questions and fill in other information as you go. **You must have an internet connection to complete this assignment.**

- **To complete this exercise, you will need to install some R packages for working with geospatial data:**

```
install.packages(c('dismo','maptools','raster','sp','rgdal','spocc'))
```

- **Load each of the above required R packages using the** `library()` **function:**

```
library(dismo)
library(maptools)
```

...etc.

- **Set your R working directory to the parent directory of the 'usa' file folder you downloaded for this exercise:**

```
setwd('/path')
```

EXERCISES

A. Exploring occurrence records

The first step in building a species distribution model is collecting data on a species' occurrence, that is, the specific locations where individuals of the same species have been observed in the wild. Data on thousands of species have been collected and organized in online databases, where data are publically available and freely downloadable. In this exercise, you will query online databases to locate records for a widespread species.

- **First, download occurrence records for painted turtles (*Chrysemys picta*) using the `gbif()` function in the `dismo` package (this might take a minute):**

```
dat <- gbif('Chrysemys', 'picta')
```

1. What does the `gbif()`function do? (Hint: use `?gbif` to open the R documentation for the function in the `Help` window.)

   What arguments did you supply to the `gbif()`function in order to find records for painted turtles?

   How would you use the `gbif()`function to find occurrence records for the green sea turtle?

- **Run the following line of code to check for georeferenced painted turtle records in other databases:**

```
occ(query='Chrysemys picta', from=c('gbif','bison','inat','ebird',
'ecoengine','vertnet'), has_coords=T)
```

2. How many total records did you find?

   Which databases could you use to download occurrence records for painted turtles?

   Why does the eBird database not contain records for painted turtles?

Run the `occ()`function for a different species and record the species you chose: ________________________, the number of georeferenced records you found: ___________, and which databases contained those records: ____.

Does the availability of georeferenced occurrence records for the species you chose differ from the availability of records for painted turtles? Speculate on the reason(s) for those differences, if any:

- **Look at the size of your painted turtle dataset using `dim(dat)`:**

3. How many occurrence records were downloaded for painted turtles?

   How many variables (columns) of data were downloaded?

- **Run the following line of code to remove painted turtle occurrence records that do not have latitude/longitude coordinates:**

```
dat <- subset(dat, !is.na(lon) & !is.na(lat))
```

4. How many occurrence records are now in the dataset?

   How many records did not contain coordinates?

   Why do you think species occurrence records might not include coordinates?

   How might a lack of coordinates affect a presence-only SDM for a given species?

How could you increase the number of occurrence records that have coordinates? (Hint: Run `colnames(dat)` and look at the names of columns 98-101.)

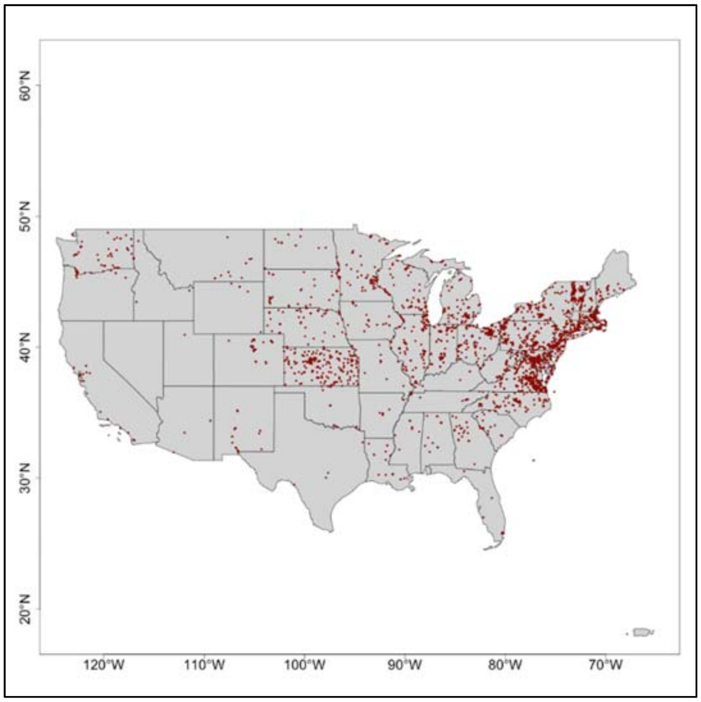
- **Import the shapefile containing a map of the US:**

```
usa <- readOGR('usa/cb_2016_us_state_20m.shp')
```

Plot the georeferenced occurrence records for painted turtles in the U.S.:

```
plot(usa, xlim=c(-125, -60), ylim=c(30,50), axes=T, cex.axis=2, col='light gray')
points(dat$lon[dat$country=='United States'], dat$lat[dat$country=='United States'], col='dark red', pch=20, cex=0.75)
```

Click the <Zoom> button in your Plots window. Your plot should look something like this:



5. In the `plot()` function, what do the `xlim` and `ylim` arguments indicate?

In the `plot()` code you just used to create the map, how were the plotted points limited to the U.S.?

6. Based on the plot you just made, describe the rough spatial distribution of occurrence records for painted turtles, noting whether they appear random, clustered, or dispersed in different states/areas within the continental U.S.

   Briefly discuss some possible explanations for the observed patterns.

   Do you think the plotted observations are a reasonable representation of the ecological niche of the painted turtle? Why or why not?

   Do you think the any of the plotted points are incorrect? Why or why not?

B. Exploring climate data

Once you have located occurrence records, you will need to access climate variables in order to build the environmental 'background' for your species of interest. In this exercise, you will access and download the bioclim dataset (not to be confused with the Bioclim SDM-fitting algorithm) and explore your species' relationship with its abiotic environment.

- **Download the 'bioclim' variables using the `getData()` function in the `raster` package:**

```
bioclim <- getData('worldclim', res=10, var='bio')
```

7. How many spatial resolutions are available for the bioclim variables? (Hint: try `?getData`.)

   How might the spatial resolution of climate data that are used to build SDMs affect model interpretation?

8. How many variables are included in the 'bioclim' dataset and what data, in general, do they represent? (Hint: see the variable descriptions at http://www.worldclim.org/bioclim.)

9. Use an internet search to locate two other sources of climate data that could be used in species distribution modeling and briefly describe the variables they contain. Do they differ from the variables in the bioclim dataset?

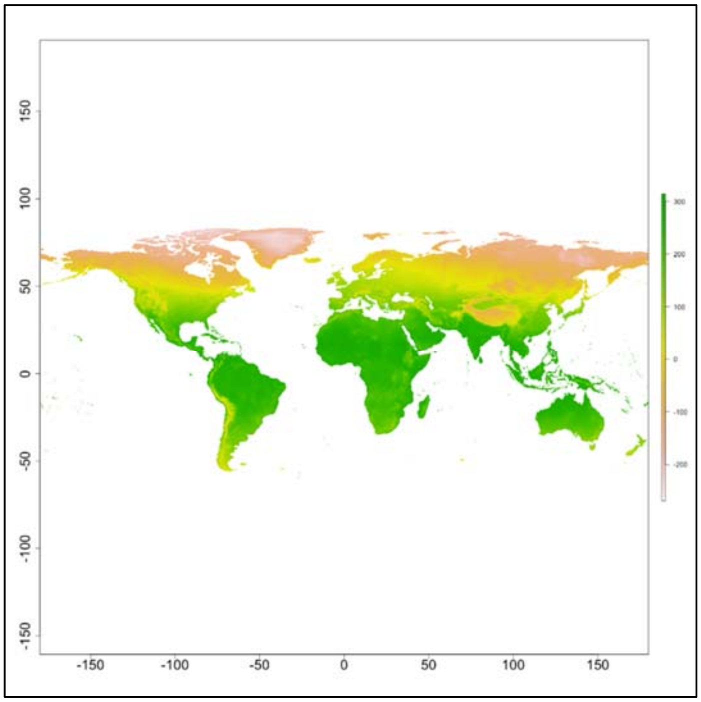
   How would you determine which climate variables to use in an SDM?

Briefly describe two different species that would require you to select different variables for building SDMs and explain your reasoning.

- **Plot one of the 'bioclim' variables (you can choose bio1 – bio19):**

```
plot(bioclim$bio1, cex.axis=2)
```

Click the <Zoom> button in your Plots window. Your plot should look something like this:



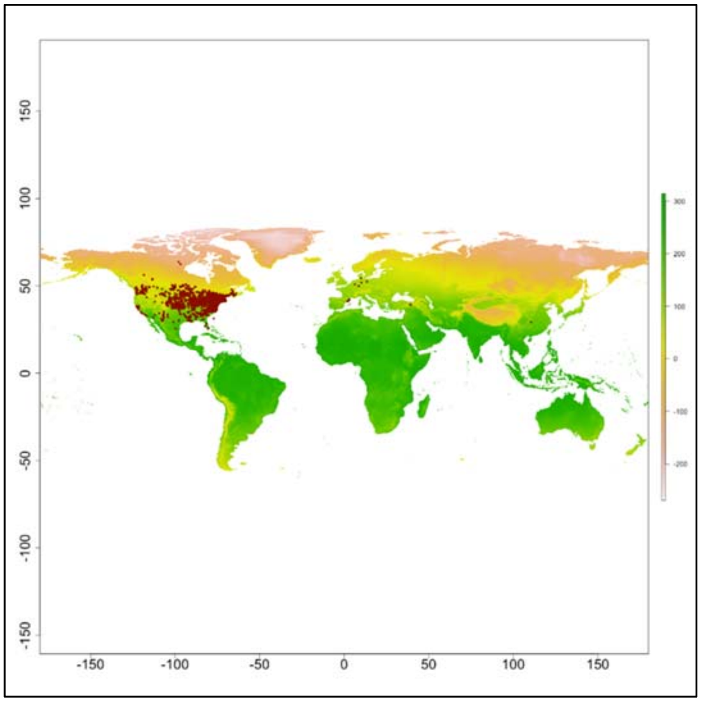
10. Which bioclim variable did you choose to plot, and what data does it contain?

What spatial extent do the bioclim variables cover?

- **Plot the painted turtle occurrence records on top of the same 'bioclim' variable:**

```
plot(bioclim$bio1, cex.axis=2)
points(dat$lon, dat$lat, col='dark red', pch=20, cex=0.75)
```

Click the <Zoom> button in your Plots window. Your plot should look something like this:



11. Describe the global distribution of painted turtles:

Assuming the occurrence records are correct, provide a possible explanation for the presence of painted turtles outside of North America:

Should occurrence records outside of North America be used to develop an SDM for painted turtles? Why or why not?

How might using global occurrence records affect and SDM?

- **Create a set of spatial points that contain only the painted turtle occurrence records for the U.S.:**

First, create a set of points based on the ‘country’ variable:

```
points.us <- SpatialPointsDataFrame(cbind(dat$lon[dat$country=='United
States'], dat$lat[dat$country=='United States']), dat[dat$country=='United
States',])
```

Define and set the coordinate system (CRS) of the points using the `crs()`**function in the** raster **package:**

```
crs_wgs84 <- ' +proj=longlat +datum=NAD83 +no_defs +ellps=GRS80
+towgs84=0,0,0'
crs(points.us) <- crs_wgs84
```

Use indexing to clip the points to the geographic boundaries of the U.S. to make sure all of the points labeled ‘United States’ are actually within the U.S.:

```
occur <- points.us[usa,]
```

Now, project the bioclim variables to match the coordinate system of the painted turtle occurrence records using the `projectRaster()`**function in the** `raster` **package (this might take a minute):**

```
bioclim.proj <- projectRaster(bioclim, crs= crs_wgs84)
```

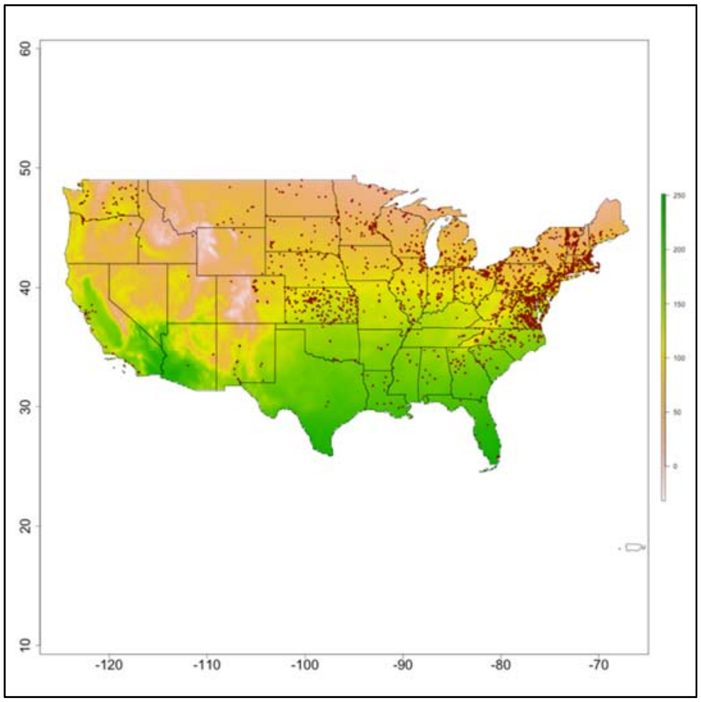
Finally, clip the bioclim variables to the geographic boundaries of the U.S. using the `mask()`**function in the** `raster` **package (this might also take a minute):**

```
bio.occur <- mask(bioclim.proj, mask=usa)
```

Plot the painted turtle occurrence records on top of one of the clipped bioclim variables (you can choose bio1 – bio19) and the U.S. state outlines:

```
plot(bio.occur$bio1, cex.axis=2, xlim=c(-127, -65), ylim=c(20,50))
lines(usa)
points(occur$lon, occur$lat, col='dark red', pch=20, cex=0.75)
```

Click the <Zoom> button in your Plots window. Your plot should look something like this:



12. Describe the distribution of painted turtles in the U.S. in reference to the bioclim variable you plotted. What values of the variable appear to be strongly positively/negatively associated with the presence of painted turtles?

Do you think the relationship between the distribution of painted turtles and the bioclim variable you plotted makes ecological sense? Why or why not? (Hint: See http://worldclim.org/bioclim for definitions of the bioclim variables.)

C. Fitting a species distribution model

In the final exercise, you will fit a climate envelope model to the occurrence records for painted turtles in the continental U.S. using the Bioclim algorithm (not to be confused with the bioclim variables). Bioclim is not as accurate as some other model-fitting methods (Elith et al 2006) and is not useful for predicting the impacts of climate change on species distributions (Hijmans & Graham 2006), but it is a relatively simple model and useful for learning about the process of fitting and evaluating SDMs (Hijmans & Elith 2017).

- **Extract the values of the bioclim variables at the occurrence points using the** `extract()`**function in the** `raster` **package:**

```
presence <- extract(bio.occur, occur)
```

Look at the first few rows and columns of data:

```
head(presence[,1:5])
```

13. What do the values in the 'presence' dataset represent?

- **Fit a Bioclim model to a subset of the data using the** `bioclim()`**function in the** `dismo` **package:**

```
bio.fit <- bioclim(presence[,c('bio1','bio2','bio3','bio7','bio8','bio12',
'bio15','bio18','bio19')])
```

14. How many presence points did the model fit? (Hint: look at the model by typing `bio.fit` in your Console window and hitting <Enter>.)

15. What do your chosen bioclim variables represent?

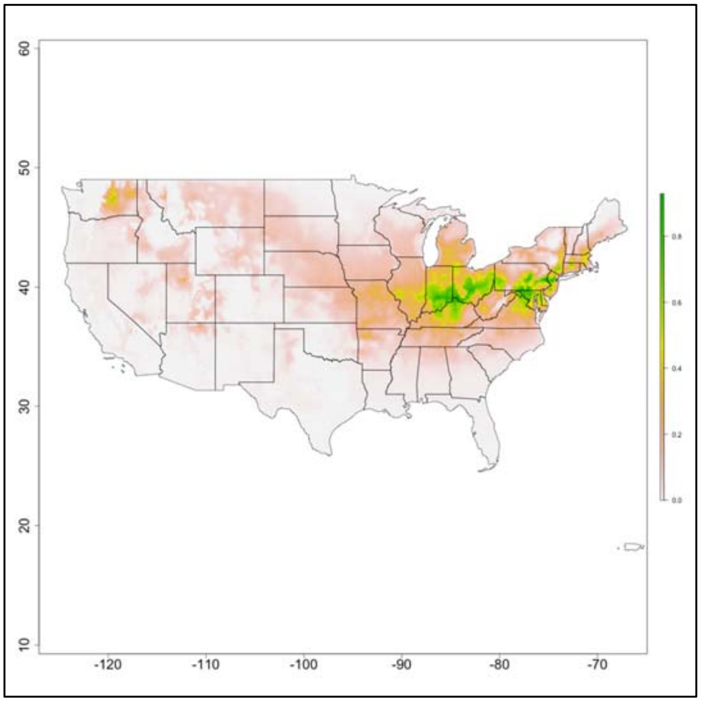
- **Create a predictive map of 'suitability scores' using the Bioclim model and bioclim rasters:**

```
predict.map <- predict(bio.occur, bio.fit)
```

Plot the suitability map:

```
plot(predict.map, cex.axis=2, xlim=c(-127, -65), ylim=c(20,50))
lines(usa)
```

Click the <Zoom> button in your Plots window. Your plot should look something like this:



16. What does the color scale of 'suitability scores' represent?

How does the Bioclim model compute suitability scores? (Hint: try `?bioclim`.)

Compare the suitability map to the first map of painted turtle occurrence records you created earlier. How do the two maps differ? How are they similar?

Do you think the suitability map above is a good prediction of habitat suitability for painted turtles? Why or why not?

17. The Bioclim model is a presence-only SDM and is simpler than other types of SDMs. Briefly explain the difference between presence-only and presence-absence SDMs.

What is ‘pseudo-absence’?

Do you think SDMs that use occurrence records to make predictions, like Bioclim, are good models for predicting species’ responses to climate change? Why or why not?

The painted turtle is a very well-studied species. Describe some ways that fitting an SDM might differ for a species with very few occurrence records and the implications of those differences for examining distributions.

References

Elith, J, CH Graham, RP Anderson, M Dudík, S Ferrier, A Guisan, RJ Hijmans, F Huettmann, JR Leathwick, A Lehmann, J Li, LG Lohmann, BA Loiselle, G Manion, C Moritz, M Nakamura, Y Nakazawa, JMcCM Overton, AT Peterson, SJ Phillips, K Richardson, R Scachetti-Pereira, RE Schapire, J Soberón, S Williams, MS Wisz, NE Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

Hijmans, RJ and J Elith. 2017. Species distribution modeling with R. Available from: http://rspatial.org/sdm/.

Hijmans RJ and CH Graham. 2006. Testing the ability of climate envelope models to predict the effect of climate change on species distributions. Global Change Biology 12:2272-2281.

NOTES TO FACULTY

In this exercise, students will gain insight into the process of building an occurrence-based species distribution model (SDM) – which may also be referred to as 'climate envelope models' – by investigating the distribution of painted turtles (*Chrysemys picta*), a semi-aquatic species that is native to and widespread throughout North America. Students will acquire skills for using geospatial data and large, spatially-explicit ecological datasets in open-source software. The provided exercises and questions increase familiarity with commonly-used data sources and prompt students to critically consider the limitations of spatially-explicit data and occurrence-based distribution models for answering ecological questions. This exercise is a much-simplified version of the sdm package modeling vignettes in Hijmans & Elith (2017) (freely downloadable – see citation at end of the Faculty Handout) that focuses on the process of fitting an SDM, not on whether or not the model is good or useful. In addition, painted turtles (along with many other reptile species) have temperature-dependent sex determination, which provides an opportunity to explore the distribution-limiting implications of climate change on both individual physiology and population demographics.

The BioClim model is one of several used to fit occurrence-based SDMs but is seldom used in the recent scientific literature, having been largely superseded by MaxLik (maximum likelihood), MaxEnt (maximum entropy, which has recently become available as the `maxnet` R package), and other machine-learning algorithms (see Pearson 2010 under 'Pre-requisites') that are more reliable for making predictions. The material in this module can be adapted to use any presence-based model-fitting algorithm. However, the advantage of using Bioclim for these exercises is its relative simplicity and accessibility. If students successfully complete the material in the module as-is, they should be able to understand and interpret outcomes of different models fairly easily. The simplicity of the Bioclim algorithm allows students to examine the process of building an SDM, while keeping the learning curve as shallow as possible.

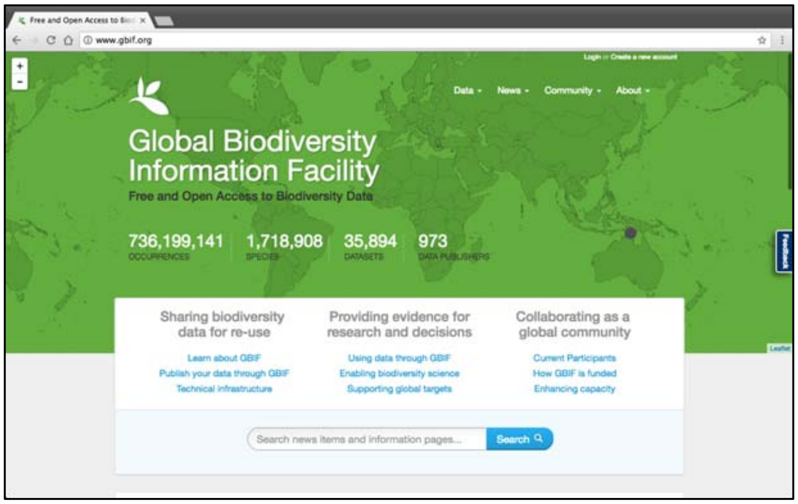
Teaching approaches

This module is a flexible, computer-based lab exercise for any number of students to complete, either individually or as a group exercise. Given that students have installed, or otherwise have access to, R and RStudio, the exercises can be completed during a single, 2-3 hour lab period. The exercises can act as a stand-alone assignment or as an introduction to concepts that are explored throughout a course.

Evaluations and extensions

The exercises and questions in the Student Handout can be supplemented using a variety of approaches:

1. Introduce students to the Global Biodiversity Information Facility (GBIF) and other databases through their online portals. Ask students to critically consider the pros/cons of querying occurrence records online vs. in R.



Global Biodiversity Information Facility

Free and Open Access to Biodiversity Data

736,199,141 OCCURRENCES | 1,718,908 SPECIES | 35,894 DATASETS | 973 DATA PUBLISHERS

Sharing biodiversity data for re-use

Providing evidence for research and decisions

Collaborating as a global community

Free and Open Access to Biod
www.gbif.org
Login or Create a new account
Data
News
Community
About
Learn about GBIF
Publish your data through GBIF
Technical infrastructure
Using data through GBIF
Enabling biodiversity science
Supporting global targets
Current Participants
How GBIF is funded
Enhancing capacity
Search news items and information pages...
Search
Feedback
Leaflet

2. Cooperative learning: Ask small groups of students to each develop an SDM for a different species and create a presentation, discussing the limitations of SDMs for their chosen species.
3. Jigsaw: Assign small groups of students a different aspect of SDMs to comprehensively research, e.g.: data availability or reliability, selection of explanatory variables, model validation, different methods of SDM fitting.
4. Essay: Have individual students fit an SDM to occurrence records of a species or taxonomic group of their choosing and compose hypotheses about the implications of climate change for that species. Extensions of the material include asking students to compare/contrast different model-fitting algorithms or to consider the requirements of organizing and hosting large data repositories.

Pre-requisites

- This is a computer-based exercise. Students should have downloaded and installed R from the Comprehensive R Archive Network and RStudio from rstudio.com for their operating systems and have basic familiarity with typing and executing commands in the R/RStudio console. (See included screenshots and R Resources at the end of this document).
- Students should have familiarity with basic spatial concepts like latitude/longitude.

- Instructors should have basic familiarity with using RStudio and troubleshooting common errors in R syntax. (See included screenshots and R Resources at the end of this document.)
- The following papers provide introductions to species distribution modeling, as well as general climate-dependent ecology of reptiles:
  - Pearson R.G. (2010) Species' distribution modeling for conservation educators and practitioners. Lessons in Conservation 3: 54-89. Download .pdf from ncep.amnh.org/linc/. **This guide is part of a teaching module that addresses the theory and application of SDMs at a graduate-professional level, including a practical using Maxent (perhaps the most popular method of building a presence-based SDM). Instructors without SDM-building experience will find this to be a particularly useful reference.**
  - Lozier J.D., Aniello P. & HIckerson M.J. (2009) Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. Journal of Biogeography 36: 1623-1627. Download .pdf from onlinelibrary.wiley.com. **This paper demonstrates some of the pitfalls of presence-based SDMs and is particularly useful for considering Questions 17-18 or for developing essay questions.**
  - Fourcade Y., Engler J.O., Rödder D. & Secondi J. (2014) Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. PloS ONE doi: 10.1371/journal.pone.0097122. Download .pdf from journals.plos.org. **This paper is an example application of Maxent that demonstrates some of the complexities that arise with biased occurrence data. One of the key points of this exercise is that building distribution models (at least to a level suitable for drawing reliable inferences) is an intensive process that takes place within a simple framework.**
  - Cunningham H.R., Rissler L.J., Buckley L.B. & Urban M.C. (2015) Abiotic and biotic constraints across reptile and amphibian ranges. Ecography 38: 001-008. Download .pdf from hydrodictyon.eeb.uconn.edu. **This paper provides information about the environmental constraints on herpetofauna (reptiles and amphibians) that will be useful for answering taxon-specific questions.**
- Students should have downloaded a copy of the 'usa' folder. The remaining data will be downloaded through the course of the exercise. However, a copy of georeferenced occurrence records from GBIF is in 'Instructor_files' if needed.
- **Text in blue contains notes for instructors, including example answers.**

Annotated student instructions

Work through the following exercises to build a basic, presence-only species distribution model (SDM) for painted turtles in R, using the relatively simple Bioclim algorithm. Lines of code can be copied/pasted directly into your R console and run by hitting <Enter>. Answer the numbered questions and fill in other information as you go. **You must have an internet connection to complete this assignment. Complete code for faculty use is in the file 'painting_turtles.R'.**

- **To complete this exercise, you will need to install some R packages for working with geospatial data:**

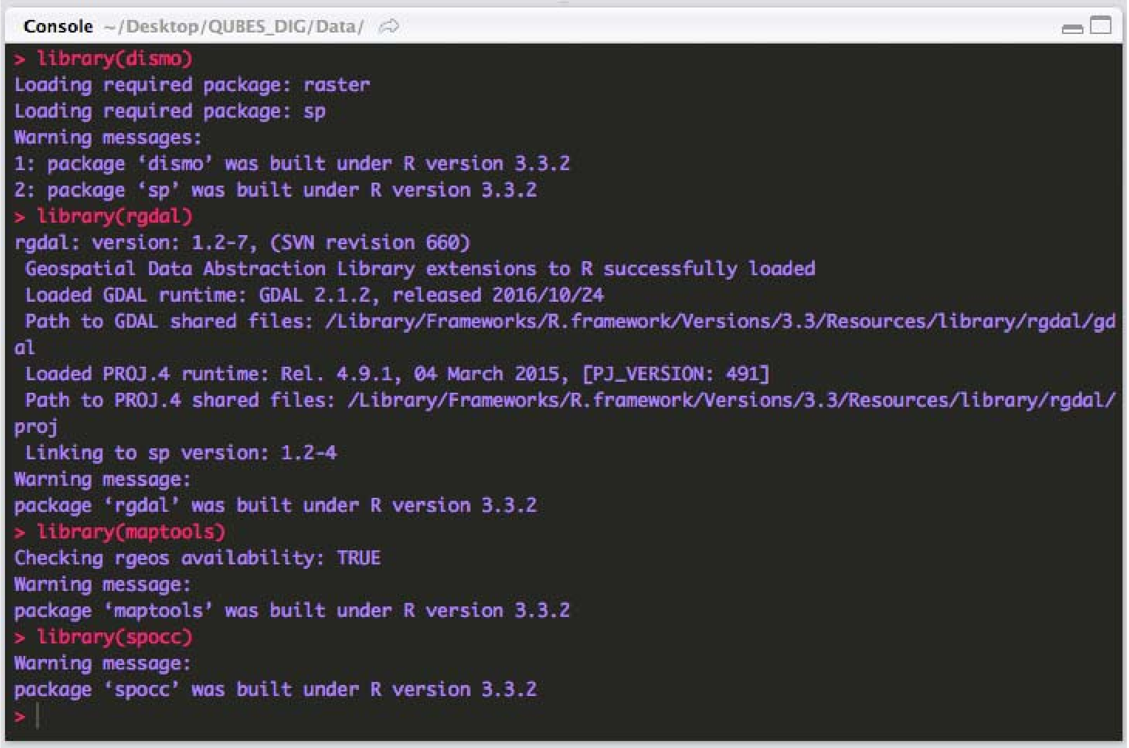
  NOTE: packages can also be installed through the Packages window in RStudio.

  ```
  install.packages(c('dismo','maptools','raster','sp','rgdal','spocc'))
  ```

- **Load each of the above required R packages using the** `library()` **function**:

  ```
  library(dismo)
  library(rgdal)
  library(maptools)
  library(spocc)
  ```

  Loading the `dismo` **package will also load** `raster` **and** `sp`. **The warning messages can be ignored.**



```
> library(dismo)
Loading required package: raster
Loading required package: sp
Warning messages:
1: package 'dismo' was built under R version 3.3.2
2: package 'sp' was built under R version 3.3.2
> library(rgdal)
rgdal: version: 1.2-7, (SVN revision 660)
 Geospatial Data Abstraction Library extensions to R successfully loaded
 Loaded GDAL runtime: GDAL 2.1.2, released 2016/10/24
 Path to GDAL shared files: /Library/Frameworks/R.framework/Versions/3.3/Resources/library/rgdal/gd
al
 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/Resources/library/rgdal/
proj
 Linking to sp version: 1.2-4
Warning message:
package 'rgdal' was built under R version 3.3.2
> library(maptools)
Checking rgeos availability: TRUE
Warning message:
package 'maptools' was built under R version 3.3.2
> library(spocc)
Warning message:
package 'spocc' was built under R version 3.3.2
> |
```

- **Set your R working directory to the parent directory of the 'usa' file folder you downloaded for this exercise:**

`setwd('/path')` **NOTE: the working directory can also be set from the menu bar via Session → Set Working Directory.**

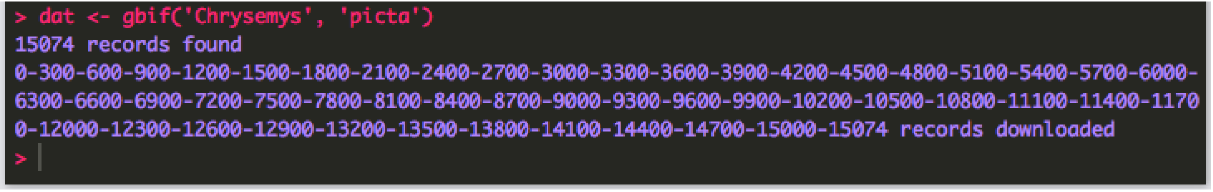
EXERCISES

A. Exploring occurrence records

The first step in building a species distribution model is collecting data on a species' occurrence, that is, the specific locations where individuals of the same species have been observed in the wild. Data on thousands of species have been collected and organized in online databases, where data are publically available and freely downloadable. In this exercise, you will query online databases to locate records for a widespread species.

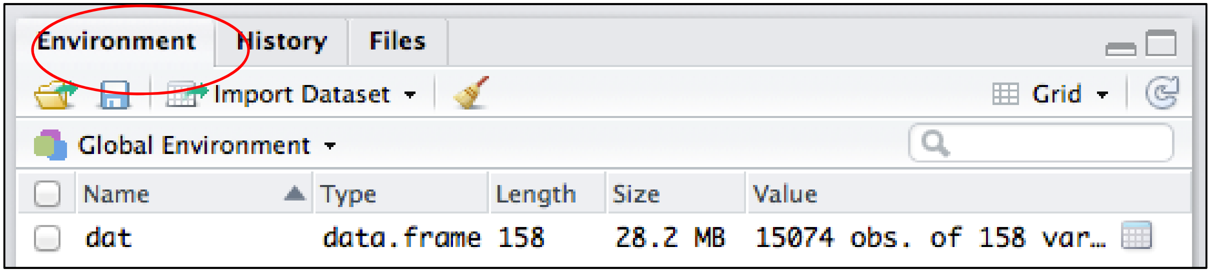
- **First, download occurrence records for painted turtles (*Chrysemys picta*) using the `gbif()` function in the `dismo` package (this might take a minute):**

```
dat <- gbif('Chrysemys', 'picta')
```



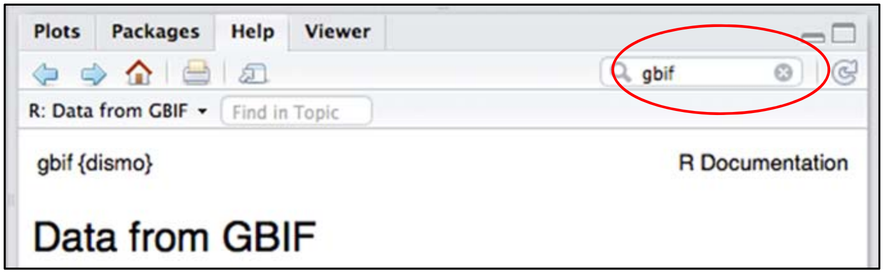
```
> dat <- gbif('Chrysemys', 'picta')
15074 records found
0-300-600-900-1200-1500-1800-2100-2400-2700-3000-3300-3600-3900-4200-4500-4800-5100-5400-5700-6000-
6300-6600-6900-7200-7500-7800-8100-8400-8700-9000-9300-9600-9900-10200-10500-10800-11100-11400-1170
0-12000-12300-12600-12900-13200-13500-13800-14100-14400-14700-15000-15074 records downloaded
> 
```

Students will now have a single definition in their Environment window.



Name	Type	Length	Size	Value
dat	data.frame	158	28.2 MB	15074 obs. of 158 var...

12. What does the `gbif()`function do? (Hint: use `?gbif` to open the R documentation for the function in the `Help` window or type `gbif` directly into the search bar in the `Help` window.)



gbif {dismo} R Documentation

Data from GBIF

The `gbif()` function downloads species, subspecies, or genus occurrence records from the Global Biodiversity Information Facility (GBIF).

13. What arguments did you supply to the `gbif()`function in order to find records for painted turtles?

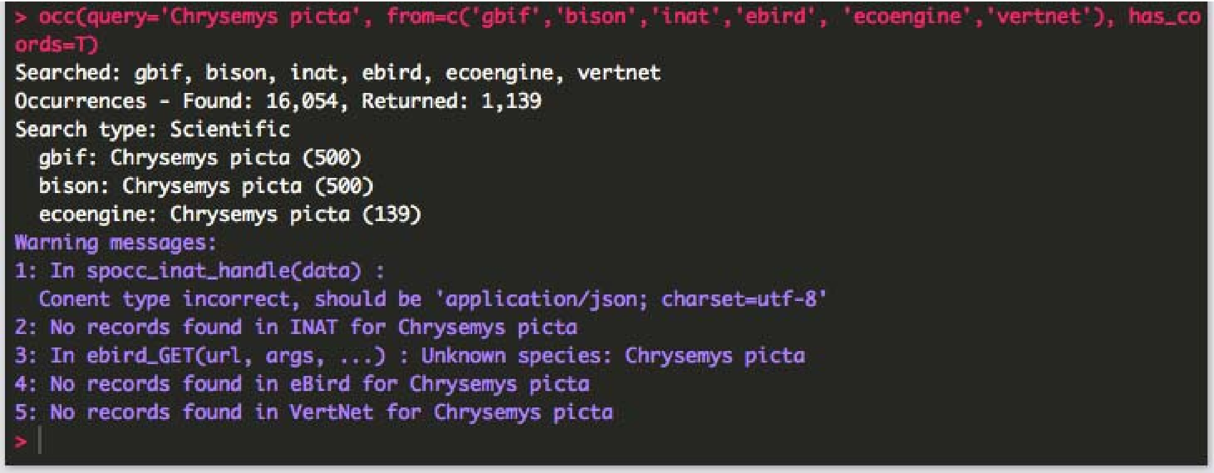
`gbif(genus= , species=)`

How would you use the `gbif()`function to find occurrence records for the green sea turtle?

`gbif('Chelonia', 'mydas')`

- **Run the following line of code to check for georeferenced painted turtle records in other databases:**

```
occ(query='Chrysemys picta', from=c('gbif','bison','inat','ebird',
'ecoengine','vertnet'), has_coords=T)
```



```
> occ(query='Chrysemys picta', from=c('gbif','bison','inat','ebird', 'ecoengine','vertnet'), has_coords=T)
Searched: gbif, bison, inat, ebird, ecoengine, vertnet
Occurrences - Found: 16,054, Returned: 1,139
Search type: Scientific
  gbif: Chrysemys picta (500)
  bison: Chrysemys picta (500)
  ecoengine: Chrysemys picta (139)
Warning messages:
1: In spocc_inat_handle(data) :
  Conent type incorrect, should be 'application/json; charset=utf-8'
2: No records found in INAT for Chrysemys picta
3: In ebird_GET(url, args, ...) : Unknown species: Chrysemys picta
4: No records found in eBird for Chrysemys picta
5: No records found in VertNet for Chrysemys picta
> 
```

14. How many total records did you find? **15,830 (numbers may vary over time)**

Which databases could you use to download occurrence records for painted turtles? **GBIF, BISON, EcoEngine**

Why does the eBird database not contain records for painted turtles? **eBird database contains records for birds only**

Run the `occ()`function for a different species and record the species you chose: ___________________________, the number of georeferenced records you found: ____________, and which databases contained those records: ____.

Does the availability of georeferenced occurrence records for the species you chose differ from the availability of records for painted turtles? Speculate on the reason(s) for those differences, if any: **E.g., Very common or well-studied species like *C. picta* are likely to have more records. Cryptic or rare species may have few, if any, records. Students can use the IUCN database to determine the conservation status of their chosen species.**

- **Look at the size of your painted turtle dataset using** `dim(dat)`:

15. How many occurrence records were downloaded for painted turtles? **14,859**

How many variables (columns) of data were downloaded? **158**

Note that students can use `?dim` to open the Help file for the `dim()` command.

- Run the following line of code to remove painted turtle occurrence records that do not have latitude/longitude coordinates:

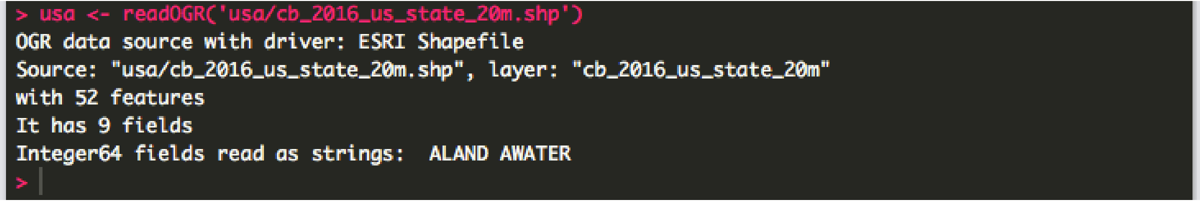
```
dat <- subset(dat, !is.na(lon) & !is.na(lat))
```

16. How many occurrence records are now in the dataset? **6,548**
How many records did not contain coordinates? **8,311**
Why do you think so many species occurrence records didn't include coordinates? How might a lack of coordinates affect a presence-only SDM for a given species? **Collection of coordinates has not been standard research practice. Older records, especially, are very unlikely to contain coordinates. Lack of coordinates or lack of accurate coordinates (e.g., if geographic information is only available at 'state' or 'county' level) will reduce reliability of any SDM.**
How could you increase the number of occurrence records that have coordinates? (Hint: Run `colnames(dat)` and look at the names of columns 98-101.) **Columns may provide alternate information about locality that could be used to georeference occurrence records with no coordinates provided.**

- **Import the shapefile containing a map of the US:**

```
usa <- readOGR('usa/cb_2016_us_state_20m.shp')
```

If the working directory is not set to the path of the 'usa' folder, an error will be encountered here. Either set the working directory OR provide the full path to 'usa' to the `readOGR()` function.



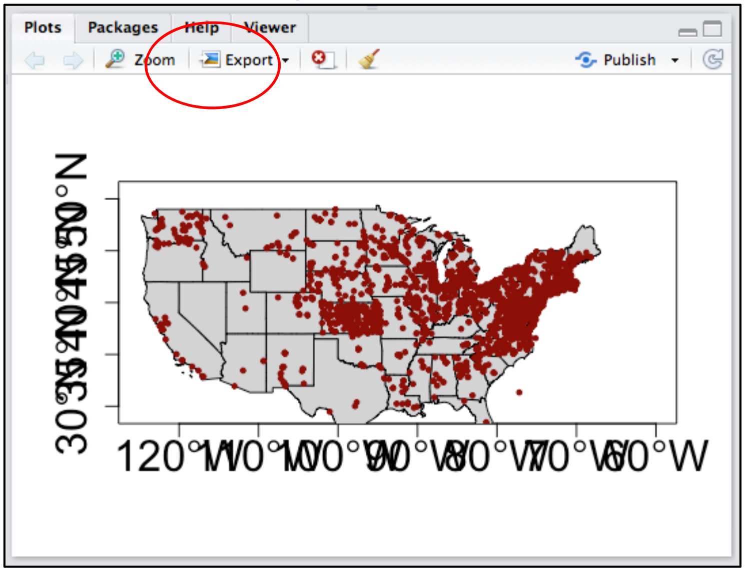
```
> usa <- readOGR('usa/cb_2016_us_state_20m.shp')
OGR data source with driver: ESRI Shapefile
Source: "usa/cb_2016_us_state_20m.shp", layer: "cb_2016_us_state_20m"
with 52 features
It has 9 fields
Integer64 fields read as strings:  ALAND AWATER
> 
```

Plot the georeferenced occurrence records for painted turtles in the U.S.:

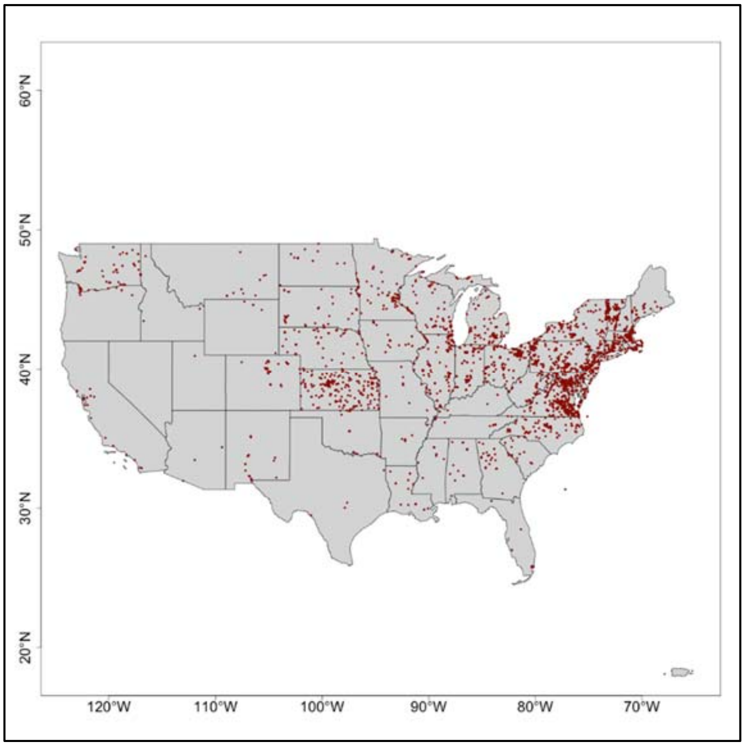
```
plot(usa, xlim=c(-125, -60), ylim=c(30,50), axes=T, cex.axis=2,
col='light gray')
points(dat$lon[dat$country=='United States'],
dat$lat[dat$country=='United States'], col='dark red', pch=20,
cex=0.75)
```

NOTE: Students can visit http://www.statmethods.net/management/subset.html to learn about different methods of subsetting data in R.

When plots are created in the `Plots` window, they may appear 'squished' as below, even when zoomed-in. Note that an error may result (Error in `plot.new() : figure margins too large`) if the `Plots` window has been made very small. To create the square plots, use the code for creating 'mapX.png' files in the 'painting_turtles.R' script.



Click the <Zoom> button in your Plots window. Your plot should look something like this:



In the `plot()` function, what do the `xlim` and `ylim` arguments indicate?
`xlim = longitude` (plotted on the x-axis) and `ylim = latitude` (plotted on the y-axis)

In the `plot()` code you just used to create the map, how were the plotted points limited to the U.S.?
Using `dat$country == 'United States'` to index the lat/lon vectors.

17. Based on the plot you just made, describe the rough spatial distribution of occurrence records for painted turtles, noting whether they appear random, clustered, or dispersed in different states/areas within the continental U.S.
Occurrence records are distributed throughout the continental U.S. (and into Canada) but are especially clustered in the NE and sparse in the western states.
Briefly discuss some possible explanations for the observed patterns. **E.g., true local suitability of habitat (especially the presence of freshwater habitats, since *C. picta* are semi-aquatic turtles) vs. sampling bias, duplication errors that lead to apparent clustering (e.g., more records will be present for locations where more sampling has occurred, which would be especially observed in locations that are relatively easy to access).**
Do you think the plotted observations are a reasonable representation of the ecological niche of the painted turtle? Why or why not?

E.g., discuss the presence of freshwater habitats, suitable temperature conditions in NE vs. western U.S.

Do you think the any of the plotted points are incorrect? Why or why not?

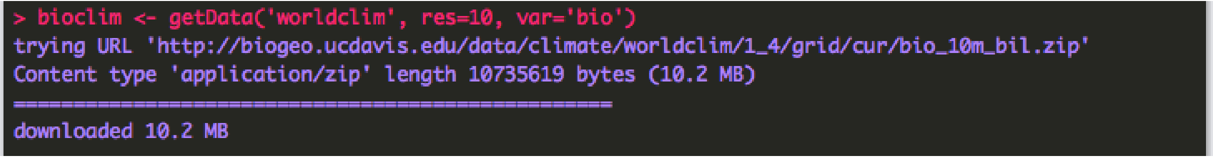
At least two points are incorrect because they're floating in the Atlantic Ocean (the coordinate system may have been recorded incorrectly for these points, either when they were collected or at some point during data entry).

B. Exploring climate data

Once you have located occurrence records, you will need to access climate variables in order to build the environmental 'background' for your species of interest. In this exercise, you will access and download the bioclim dataset (not to be confused with the Bioclim SDM-fitting algorithm) and explore your species' relationship with its abiotic environment.

- **Download the 'bioclim' variables using the** `getData()` **function in the** `raster` **package**:

```
bioclim <- getData('worldclim', res=10, var='bio')
```



```
> bioclim <- getData('worldclim', res=10, var='bio')
trying URL 'http://biogeo.ucdavis.edu/data/climate/worldclim/1_4/grid/cur/bio_10m_bil.zip'
Content type 'application/zip' length 10735619 bytes (10.2 MB)
==================================================
downloaded 10.2 MB
```

18. How many spatial resolutions are available for the bioclim variables? (Hint: try `?getData`.)

Available spatial resolutions are 0.5, 2.5, 5, and 10' (minutes of a degree of latitude).

How might the spatial resolution of climate data that are used to build SDMs affect model interpretation?

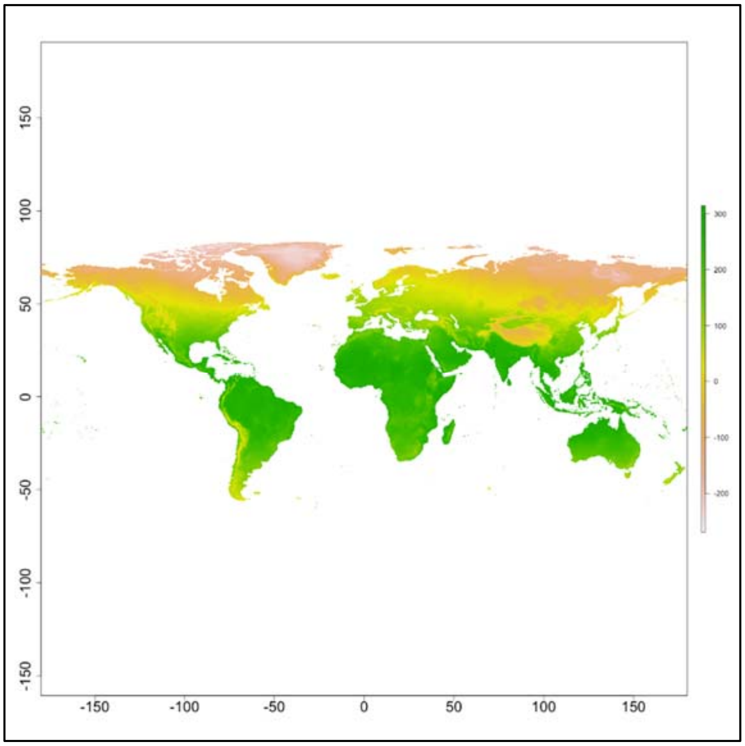
Organisms experience their environments at a spatial scale that is usually much finer than gridded climate layers like BioClim. The lower the spatial resolution, the greater the chance that a model will NOT correctly capture the environmental conditions that make a particular location suitable or unsuitable for a given organism.

19. How many variables are included in the 'bioclim' dataset and what data, in general, do they represent? (Hint: see the variable descriptions at http://www.worldclim.org/bioclim.)

bio1 – bio19 represent annual trends (e.g., mean annual temperature, annual precipitation), seasonality (e.g., annual range in temperature and precipitation), and extreme or limiting environmental factors (e.g., temperature of the coldest and warmest month, precipitation of the wet and dry quarters).

20. Use an internet search to locate two other sources of climate data that could be used in species distribution modeling and briefly describe the variables they contain. Do they differ from the variables in the bioclim dataset? **E.g., University of East Anglia Climatic Research Unit (CRU) has gridded precipitation, mean temperature, diurnal temperature range, wet-day frequency, vapour pressure, cloud cover at 0.5° resolution that covers all land areas except Antarctica. NOAA (National Oceanic and Atmospheric Administration) has a variety of daily/monthly/annual summaries of weather data.**
 How would you determine which variables to use in an SDM?
 E.g., accurately representing a species' ecology/physiology is more important than the ease of finding a dataset, potential explanatory variables should be examined for colinearity, etc. Some species' distributions might be more accurately represented by non-climate variables, e.g., elevation, vegetation, or soil maps.
 Briefly describe two different species that would require you to select different variables for building SDMs and explain your reasoning. **E.g., the habitat requirements for any species with a highly specialized niche might not be captured by long-term, low-resolution climate data. A desert-dwelling plant might have specific requirements for accumulation of precipitation, while occurrence of a temperate lizard could depend more on thermal tolerance.**

- **Plot one of the 'bioclim' variables (you can choose bio1 – bio19):**
  ```
  plot(bioclim$bio1, cex.axis=2)
  ```
  Click the <Zoom> button in your Plots window. Your plot should look something like this:



21. Which bioclim variable did you choose to plot, and what data does it contain?

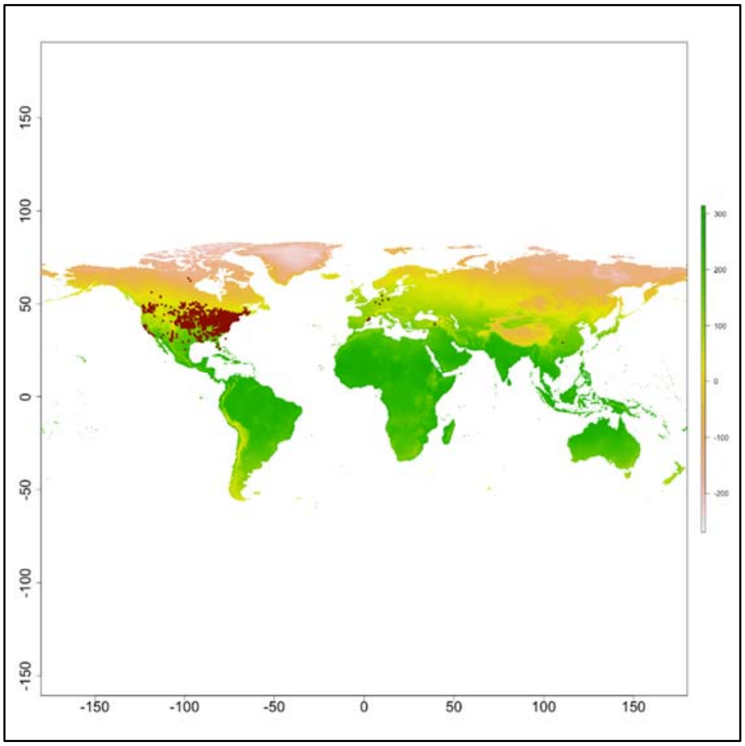
E.g., bio1 is 'annual mean temperature'

What spatial extent do the bioclim variables cover? **global**

- **Plot the painted turtle occurrence records on top of the same 'bioclim' variable:**

```
plot(bioclim$bio1, cex.axis=2)
points(dat$lon, dat$lat, col='dark red', pch=20, cex=0.75)
```

Click the <Zoom> button in your Plots window. Your plot should look something like this:



22. Describe the global distribution of painted turtles: **Present throughout the U.S. and the whole N American continent, with scattered records from Europe and Asia.**

Assuming the occurrence records are correct, provide a possible explanation for the presence of painted turtles outside of North America:

Pet trade releases have introduced painted turtles to Germany, Indonesia, the Phillipines, and Spain.

Should occurrence records outside of North America be used to develop an SDM for painted turtles? Why or why not? **If an introduced species survives and successfully reproduces in the wild, occurrence records in locations of its introduction could be informative. Scattered occurrence records from introductions should be used with caution, particularly if (as with *C. picta*) numerous records are available for its native range.**

How might using global occurrence records affect an SDM?

E.g., could decrease accuracy of an SDM if scattered occurrence records outside of the native range are used, and the environment in those locations does not reflect truly suitable habitat for the species.

Create a set of spatial points that contain only the painted turtle occurrence records for the U.S.:

In order to plot occurrence records on top of the bioclim variables, the points must be converted to a SpatialPointsDataFrame, and coordinate systems among all spatial layers must be defined (and must match). Thu, the procedure to create the following plots requires additional steps from the original plotting of occurrence points on top of the simple 'usa' map, which is already in the WGS_84 coordinate system (as defined below) and can be plotted underneath simple lat/lon values.

First, create a set of points based on the 'country' variable:

```
points.us <-
SpatialPointsDataFrame(cbind(dat$lon[dat$country=='United States'],
dat$lat[dat$country=='United States']), dat[dat$country=='United
States',])
```

Define and set the coordinate system (CRS) of the points using the `crs()`**function in the** `raster` **package**:

```
crs_wgs84 <- ' +proj=longlat +datum=NAD83 +no_defs +ellps=GRS80
+towgs84=0,0,0'
crs(points.us) <- crs_wgs84
```

Use indexing to clip the points to the geographic boundaries of the U.S. to make sure all of the points labeled 'United States' are actually within the U.S.:

```
occur <- points.us[usa,]
```

Now, project the bioclim variables to match the coordinate system of the painted turtle occurrence records using the `projectRaster()`**function in the** raster **package (this might take a minute):**

```
bioclim.proj <- projectRaster(bioclim, crs= crs_wgs84)
```

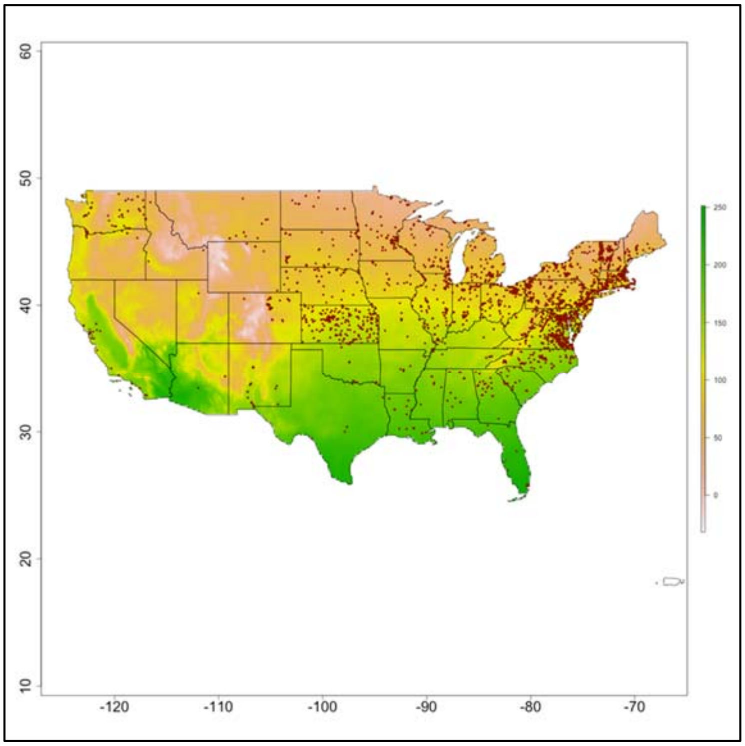
Finally, clip the bioclim variables to the geographic boundaries of the U.S. using the `mask()`**function in the** `raster` **package (this might also take a minute):**

```
bio.occur <- mask(bioclim.proj, mask=usa)
```

Plot the painted turtle occurrence records on top of one of the clipped bioclim variables (you can choose bio1 – bio19) and the U.S. state outlines:

```
plot(bio.occur$bio1, cex.axis=2, xlim=c(-127, -65), ylim=c(20,50))
lines(usa)
points(occur$lon, occur$lat, col='dark red', pch=20, cex=0.75)
```

Click the <Zoom> button in your Plots window. Your plot should look something like this:



18. Describe the distribution of painted turtles in the U.S. in reference to the bioclim variable you plotted. What values of the variable appear to be strongly positively/negatively associated with the presence of painted turtles? **Answer will depend on the variable chosen.**
    Do you think the relationship between the distribution of painted turtles and the bioclim variable you plotted makes ecological sense? Why or why not? (Hint: See http://worldclim.org/bioclim for definitions of the bioclim variables.) **Answer will depend on the variable chosen, but should be reflective of the basic habitat requirements of a semi-aquatic reptile.**

C. Fitting a species distribution model

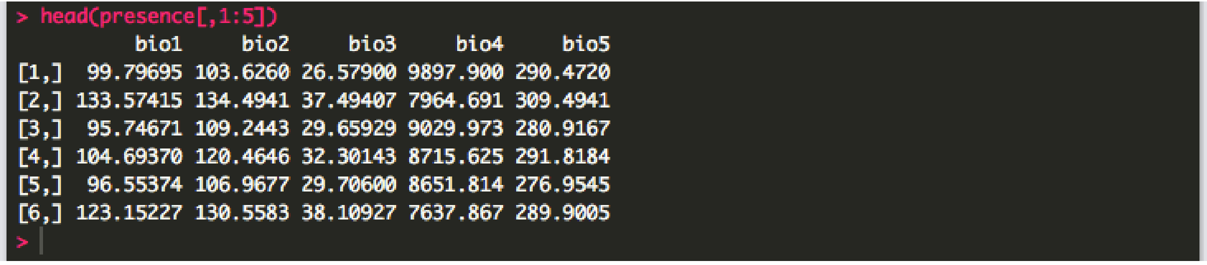
In the final exercise, you will fit a climate envelope model to the occurrence records for painted turtles in the continental U.S. using the Bioclim algorithm (not to be confused with the bioclim variables). Bioclim is not as accurate as some other model-fitting methods (Elith et al 2006) and is not useful for predicting the impacts of climate change on species distributions (Hijmans & Graham 2006), but it is a relatively simple model and useful for learning about the process of fitting and evaluating SDMs (Hijmans & Elith 2017).

- **Extract the values of the bioclim variables at the occurrence points using the** `extract()`**function in the** `raster` **package**:

```
presence <- extract(bio.occur, occur)
```

Look at the first few rows and columns of data:

```
head(presence[,1:5])
```



```
> head(presence[,1:5])
          bio1     bio2     bio3     bio4     bio5
[1,]  99.79695 103.6260 26.57900 9897.900 290.4720
[2,] 133.57415 134.4941 37.49407 7964.691 309.4941
[3,]  95.74671 109.2443 29.65929 9029.973 280.9167
[4,] 104.69370 120.4646 32.30143 8715.625 291.8184
[5,]  96.55374 106.9677 29.70600 8651.814 276.9545
[6,] 123.15227 130.5583 38.10927 7637.867 289.9005
> 
```

19. What do the values in the 'presence' dataset represent? **Values of the bioclim variables at the locations of *C. picta* occurrence records.**

- **Fit a Bioclim model to a subset of the data using the** `bioclim()`**function in the** `dismo` **package**:

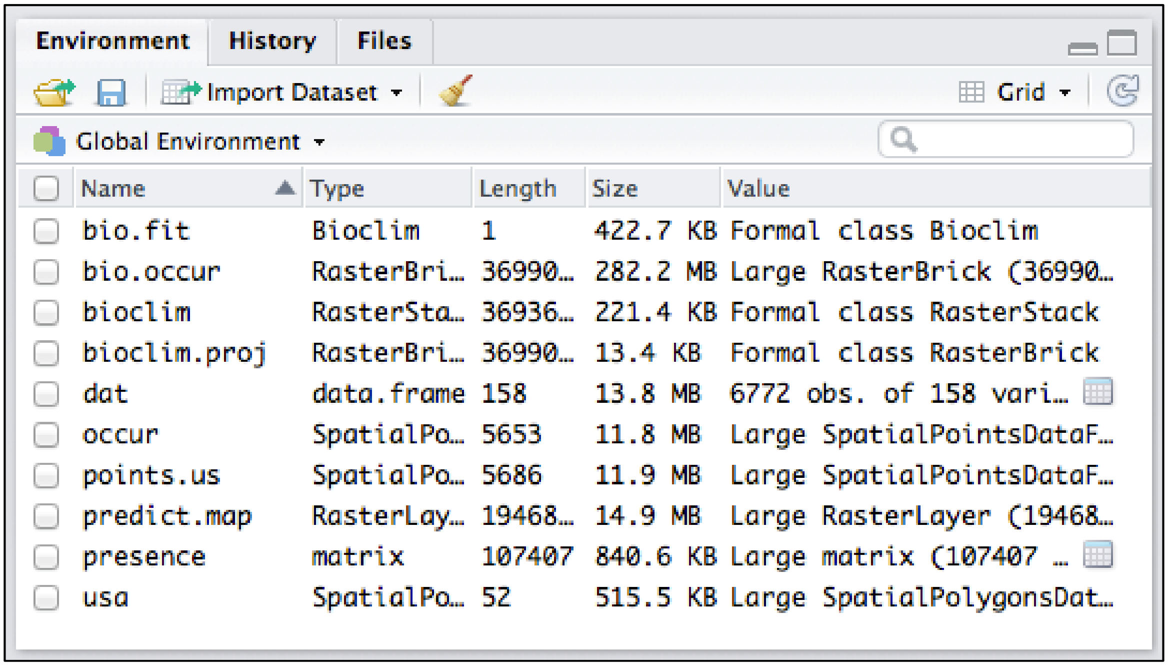
  `bio.fit <- bioclim(presence[,c('bio1','bio2','bio3','bio7','bio8','bio12', 'bio15','bio18','bio19')])` **You can also select any subset of bioclim variables or ask students to make/justify their own selection and discuss in their answer to Question 15.**

20. How many presence points did the model fit? (Hint: look at the model by typing bio.fit in your Console window and hitting <Enter>.) **5,331**

21. What do your chosen bioclim variables represent? **Students can reference the website in Question 12. NOTE: due to data storage requirements, the bioclim variables have been multiplied by 10 to remove decimal points. For example, the range of bio1 is -200 – 300, indicating a range of mean annual temperatures of -20 – 30°C.**

- **Create a predictive map of 'suitability scores' using the Bioclim model and bioclim rasters:**

  `predict.map <- predict(bio.occur, bio.fit)` **Once all code is executed, the `Environment` window will contain 10 definitions. In the `predict()` function, you can also choose to ignore one of the tails of the distribution (e.g, to make low rainfall a limiting factor, but not high rainfall).**

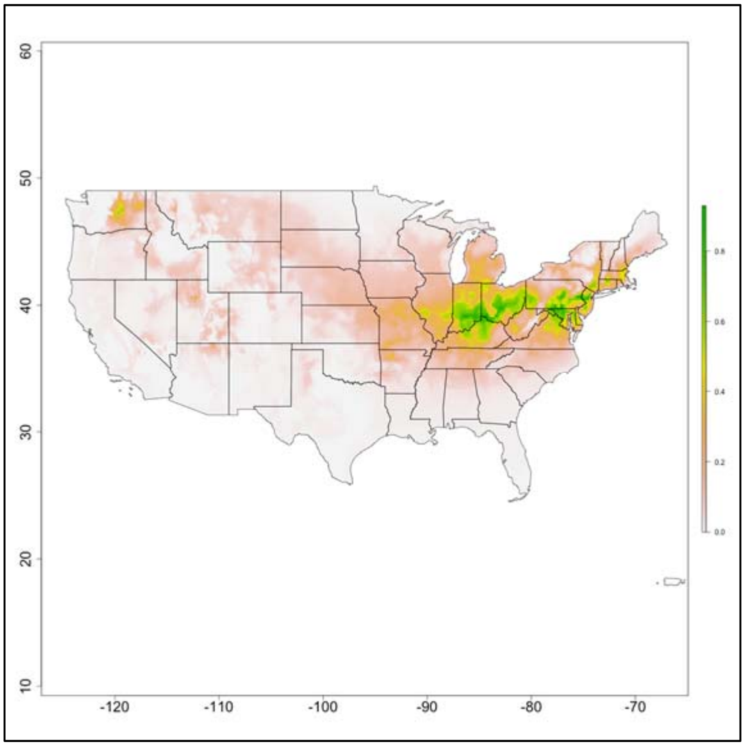


□	Name ▲	Type	Length	Size	Value
□	bio.fit	Bioclim	1	422.7 KB	Formal class Bioclim
□	bio.occur	RasterBri…	36990…	282.2 MB	Large RasterBrick (36990…
□	bioclim	RasterSta…	36936…	221.4 KB	Formal class RasterStack
□	bioclim.proj	RasterBri…	36990…	13.4 KB	Formal class RasterBrick
□	dat	data.frame	158	13.8 MB	6772 obs. of 158 vari…
□	occur	SpatialPo…	5653	11.8 MB	Large SpatialPointsDataF…
□	points.us	SpatialPo…	5686	11.9 MB	Large SpatialPointsDataF…
□	predict.map	RasterLay…	19468…	14.9 MB	Large RasterLayer (19468…
□	presence	matrix	107407	840.6 KB	Large matrix (107407 …
□	usa	SpatialPo…	52	515.5 KB	Large SpatialPolygonsDat…

Plot the suitability map:

```
plot(predict.map, cex.axis=2, xlim=c(-127, -65), ylim=c(20,50))
lines(usa)
```

Click the <Zoom> button in your Plots window. Your plot should look something like this:



22. What does the color scale of 'suitability scores' represent?

From `?bioclim`: percentile scores are between 0 and 1, but predicted values larger than 0.5 are subtracted from 1. Then, the minimum percentile score across all the environmental variables is computed (i.e. this is like Liebig's law of the minimum, except that high values can also be limiting factors). The final value is subtracted from 1 and multiplied with 2 so that the results are between 0 and 1. The reason for this transformation is that the results become more like that of other distribution modeling methods and are thus easier to interpret. The value 1 will rarely be observed as it would require a location that has the median value of the training data for all the variables considered. The value 0 is very common as it is assigned to all cells with a value of an environmental variable that is outside the percentile distribution (the range of the training data) for at least one of the variables.

How does the Bioclim model compute suitability scores? (Hint: try `?bioclim`.)

From `?bioclim`: The BIOCLIM algorithm computes the similarity of a location [to locations of occurrence] by comparing the values of environmental variables at any location to a percentile distribution of the values at known locations of occurrence ('training sites'). The closer to the 50th percentile (the median), the more suitable the location is. The tails of the distribution are not distinguished, that is, 10 percentile is treated as equivalent to 90 percentile.

Compare the suitability map to the maps of painted turtle occurrence records you created earlier. How do the maps differ? How are they similar? **E.g., The suitability map is probabilistic and based on the distribution of multiple environmental variables, rather than providing a value of a single variable at an occurrence point.**
Do you think the suitability map above is a good prediction of habitat suitability for painted turtles? Why or why not? **E.g., Note how well the suitability map 'matches' the occurrence records, relative to locations where there are no occurrence records. Are occurrence points differentiated from non-occurrence points? Are all locations with occurrence records 'equally suitable'?**

23. The Bioclim model is a presence-only SDM and is simpler than other types of SDMs. Briefly explain the difference between presence-only and presence-absence SDMs. **A presence-only SDM is based only on occurrence records, while a presence-absence SDM attempts to model both occurrence and non-occurrence.**
What is 'pseudo-absence'? **Pseudo-absence data are sampled 'background' points that represent the available environmental conditions in an area where a species occurs.**

24. Do you think SDMs that use occurrence records to make predictions, like Bioclim, are good models for predicting species' responses to climate change? Why or why not?
E.g., Generally, occurrence-based models are not ideal for predicting responses to climate change or other novel environments, because presence records only represent the environments where an organism exists under its current environment, not the range of possible environments where it *could* exist, based on its physiological limitations.
The painted turtle is a very well-studied species. Describe some ways that fitting an SDM might differ for a species with very few occurrence records and the implications of those differences for examining distributions.
E.g., Similar to above, reliably fitting an occurrence-based SDM for a species with few occurrence records would be more difficult because we would have limited information about the range of environmental conditions under which an organism *could* survive, relative to where it actually is (or where we know it is).

References

Elith, J, CH Graham, RP Anderson, M Dudík, S Ferrier, A Guisan, RJ Hijmans, F Huettmann, JR Leathwick, A Lehmann, J Li, LG Lohmann, BA Loiselle, G

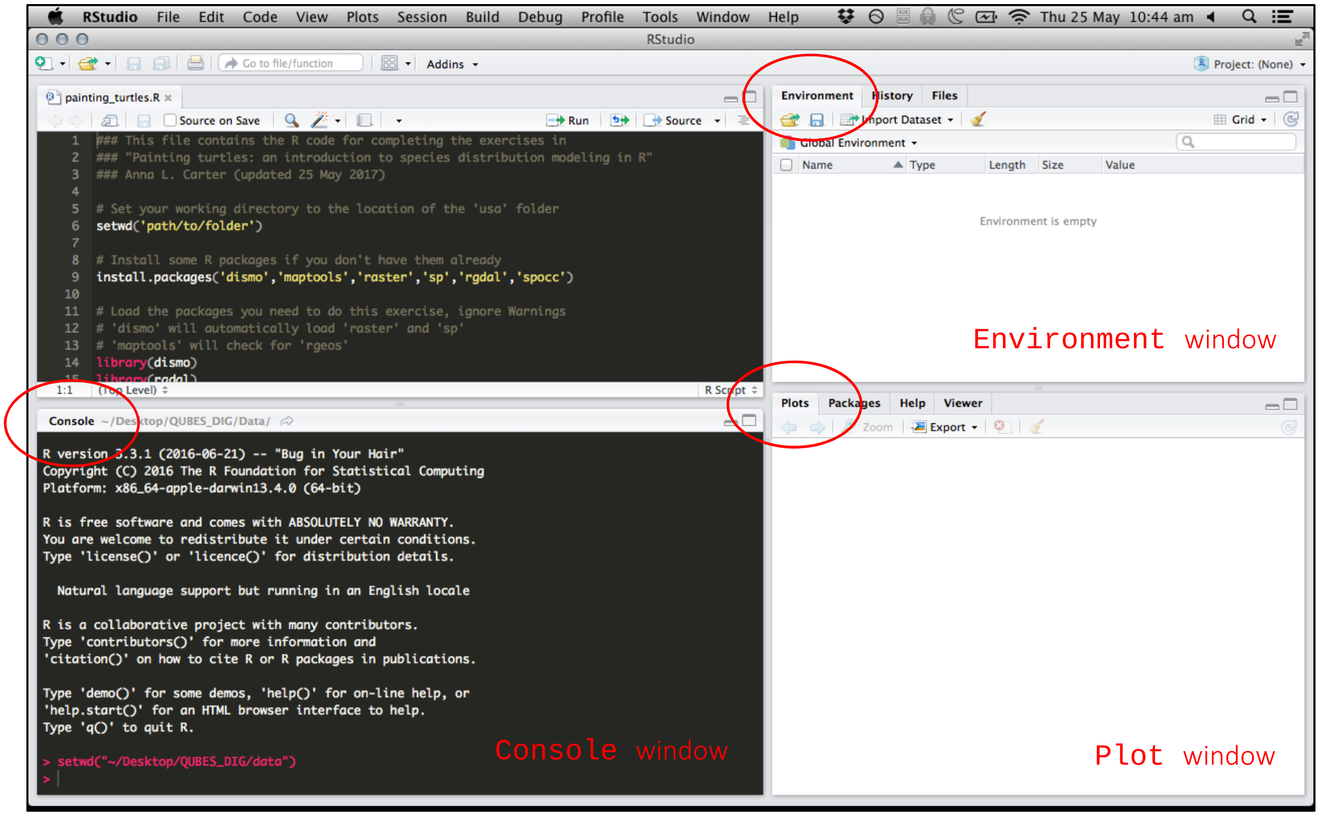
Manion, C Moritz, M Nakamura, Y Nakazawa, JMcCM Overton, AT Peterson, SJ Phillips, K Richardson, R Scachetti-Pereira, RE Schapire, J Soberón, S Williams, MS Wisz, NE Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

Hijmans, RJ and J Elith. 2017. Species distribution modeling with R. R package version 0.8-11 (2013). Accessed from https://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf.

Hijmans, RJ and CH Graham. 2006. Testing the ability of climate envelope models to predict the effect of climate change on species distributions. Global Change Biology 12: 2272-2281.

R resources

This module uses R software, run using the RStudio IDE (below). Students can copy/paste or type the provided lines of code into the Console window. The general appearance of RStudio may vary by user preference and operating system (the included screenshots are from R v 3.3.1 running in RStudio v 1.0.143 on Mac OSX). The relative locations of the Console and other windows can be modified from the menu bar using RStudio → Preferences → Pane Layout. Software should be installed prior to undertaking the exercises, as successful installation is a common roadblock for beginning R users.



```
### This file contains the R code for completing the exercises in
### "Painting turtles: an introduction to species distribution modeling in R"
### Anna L. Carter (updated 25 May 2017)

# Set your working directory to the location of the 'usa' folder
setwd('path/to/folder')

# Install some R packages if you don't have them already
install.packages('dismo','maptools','raster','sp','rgdal','spocc')

# Load the packages you need to do this exercise, ignore Warnings
# 'dismo' will automatically load 'raster' and 'sp'
# 'maptools' will check for 'rgeos'
library(dismo)
library(rgdal)
```

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> setwd("~/Desktop/QUBES_DIG/data")
> 
```

Preventing common syntax errors

- R is case-sensitive.
- Commands should be entered exactly as written, including all commas and quotation marks. Either single (") or double ("") quotes can be used.
- All parentheses must be matched.

Troubleshooting common error messages

1. **`Error in file(file, "rt") : cannot open the connection…`**
   This error occurs when R cannot locate a file because the working directory has not been set or is mis-specified. For this module, the working directory should be set to the location of the ‘usa’ folder.

2. **`Error: could not find function ""`**

   This error arises when the R package that contains the called function has not been loaded. Ensure the required package is installed and run `library(package)`. If the package has been loaded, check that the function name is spelled/capitalized correctly.

3. **`Error in eval(…): object "" not found`**

   This error indicates that the required object definition is not in the Environment. Go back through the code and determine if an object has not been defined using `object <-`

`definition`. For example, if you have not defined `dat <- gbif('Genus', 'species')`, the error message will state `object "dat"` not found.

Online resources for learning/using R

Students with no experience in R should become familiar with entering and executing basic commands in the Console window prior to undertaking this assignment.

1. **Impatient R | Burns Statistics** [http://www.burns-stat.com/documents/tutorials/impatient-r/]
2. **Quick-R | Statmethods** [http://www.statmethods.net/index.html]
3. **Getting used to R, RStudio, and R Markdown** [https://ismayc.github.io/rbasics-book/index.html]
4. **An Introduction to R | The R Manuals** [https://cran.r-project.org/manuals.html]
5. **TryR | Code School** [http://tryr.codeschool.com/]
6. **Online R resources for beginners (a list of resources, including books, videos, podcasts & webcasts)** [http://www.introductoryr.co.uk/R_Resources_for_Beginners.html]
7. **NEON Data Skills Tutorials for spatial data**
    a. **Raster 00: Intro to raster data in R** [http://neondataskills.org/R/Introduction-to-Raster-Data-In-R/]
    b. **Raster 01: Plot raster data in R** [http://neondataskills.org/R/Plot-Rasters-In-R/]
    c. **Vector 00: Open and plot shapefiles in R** [http://neondataskills.org/R/open-shapefiles-in-R/]

COPYRIGHT STATEMENT

GENERIC DISCLAIMER

Adult supervision is recommended when performing this lab activity. We also recommend that common sense and proper safety precautions be followed by all participants. No responsibility is implied or taken by the contributing author, the editors of this Volume, nor anyone associated with maintaining the TIEE web site, nor by their academic employers, nor by the Ecological Society of America for anyone who sustains injuries as a result of using the materials or ideas, or performing the procedures put forth at the TIEE web site, or in any printed materials that derive therefrom.