**Species Distribution Modeling**

As part of our discussion of biogeography, we took a look into the basics of species distribution modeling. This is also sometimes called species niche modeling, because it is based around an assessment of the fundamental niche of an organism. The fundamental niche is the set of environmental conditions necessary for the survival and reproduction of an organism. These conditions may include temperature, precipitation, soil acidity, and others. The fundamental niche is subsequently modified to form the realized niche, or actual places where the organism lives. The fundamental niche tends to be generated by abiotic forces and the realized niche tends to be generated by biotic forces and interactions.

Species distribution or niche modeling is an attempt to map the fundamental niche of an organism in order to determine all of the places on Earth that a particular organism can grow. There are a number of reasons to do this. Species distribution models and can help to identify potentially invasive species before they arrive. Species distribution models can help us understand the distribution or rare species, such that we might better preserve them. Species distribution models can also be used to predict past or future species distributions to understand how climate may impact species distributions.

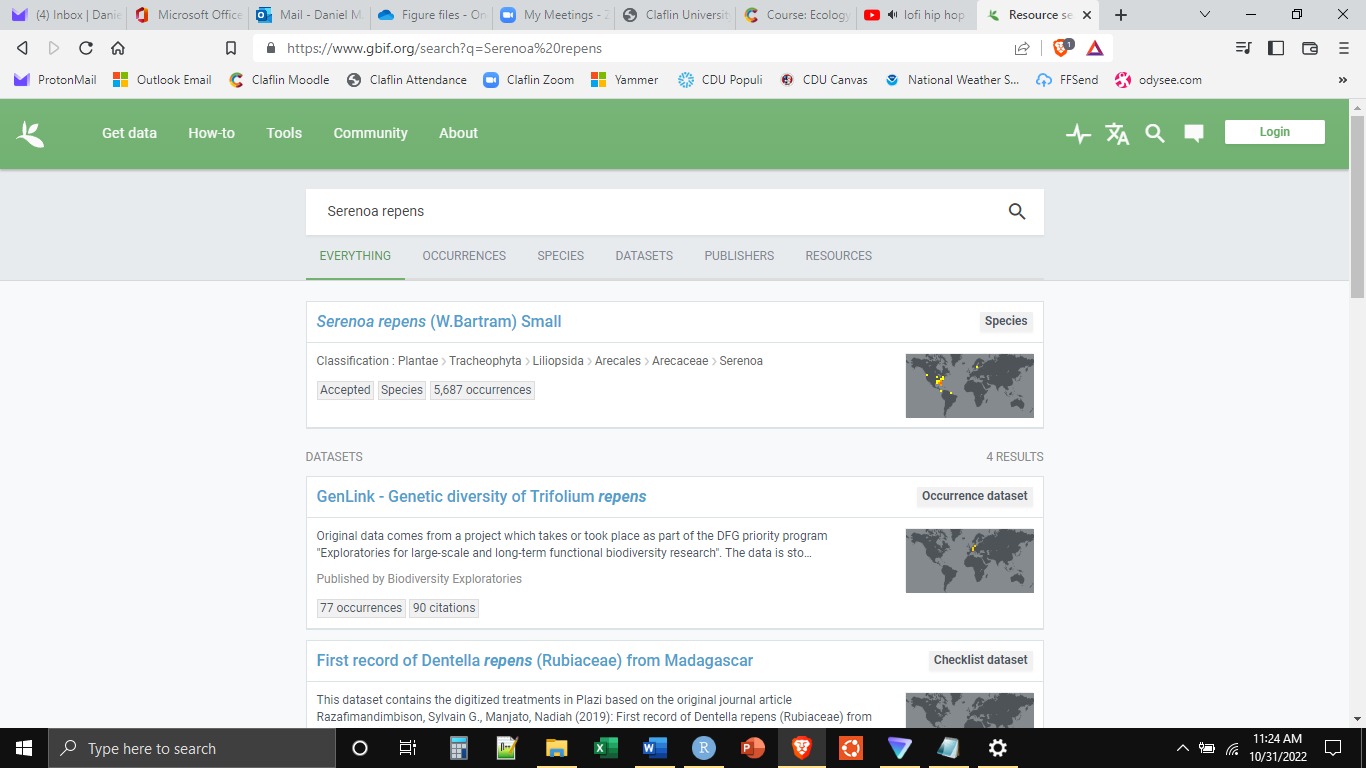
To reinforce the concepts of species distribution models, we will today construct a distribution model for the saw palmetto (*Serenoa repens*, Arecaceae).



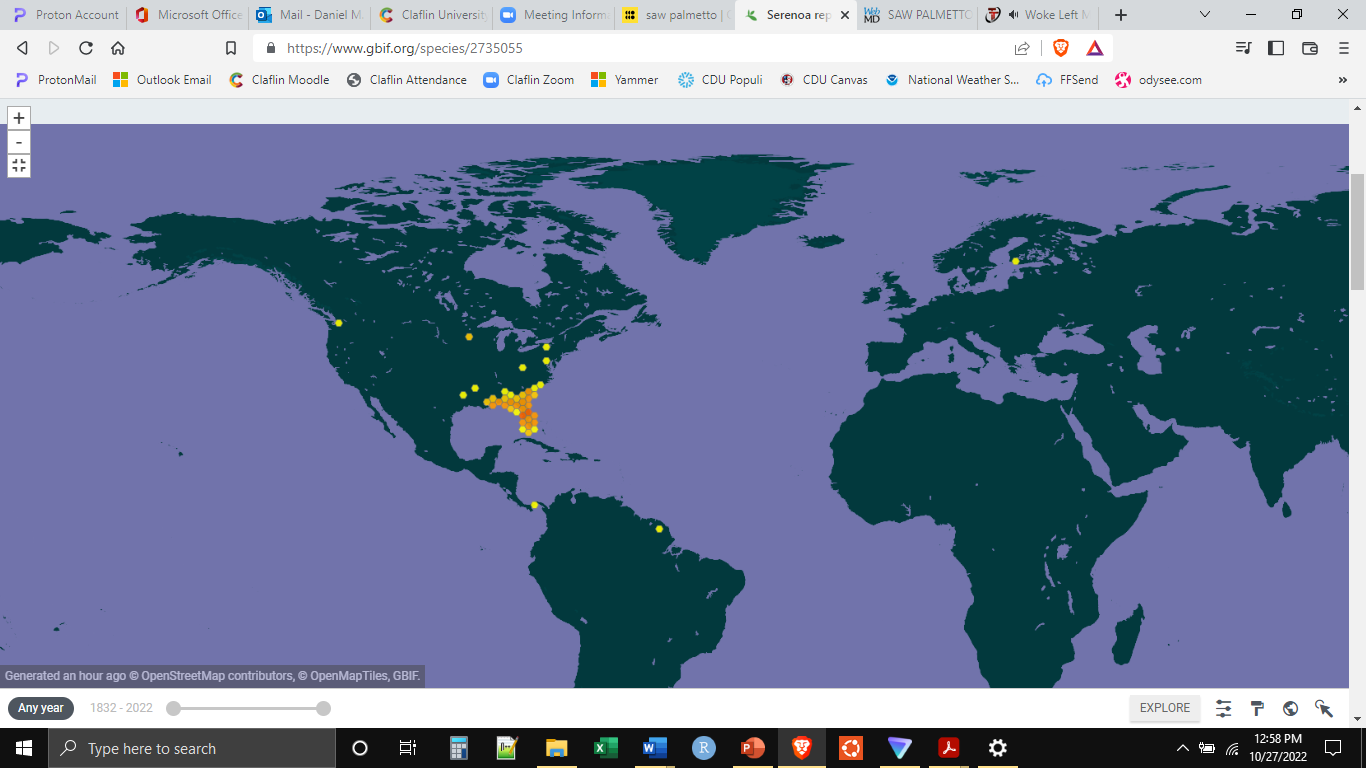
Any species distribution model is built using two major components 1) a collection of known growing or living locations for the species in question, known as “occurrence records”, and 2) environmental data for all of the places where the species is known to live. This environmental data, again, can include climate as well as soil information.

Occurrence record data includes longitude and latitude of the location where the species has been observed. There are a variety of ways to get this information. It is possible to derive this information from personal observations. It is also possible to use museum specimens to obtain this information. Each museum specimen includes information for where that specimen was collected. Webtools, such as GBIF have been developed to aggregate this information from museums throughout the world.

**Go to www.GBIF.org. In the search field type the scientific name for the saw palmetto (*Serenoa repens*). Then select the top result.**



**What you will see are the longitude and latitude points for occurrences of the saw palmetto.**



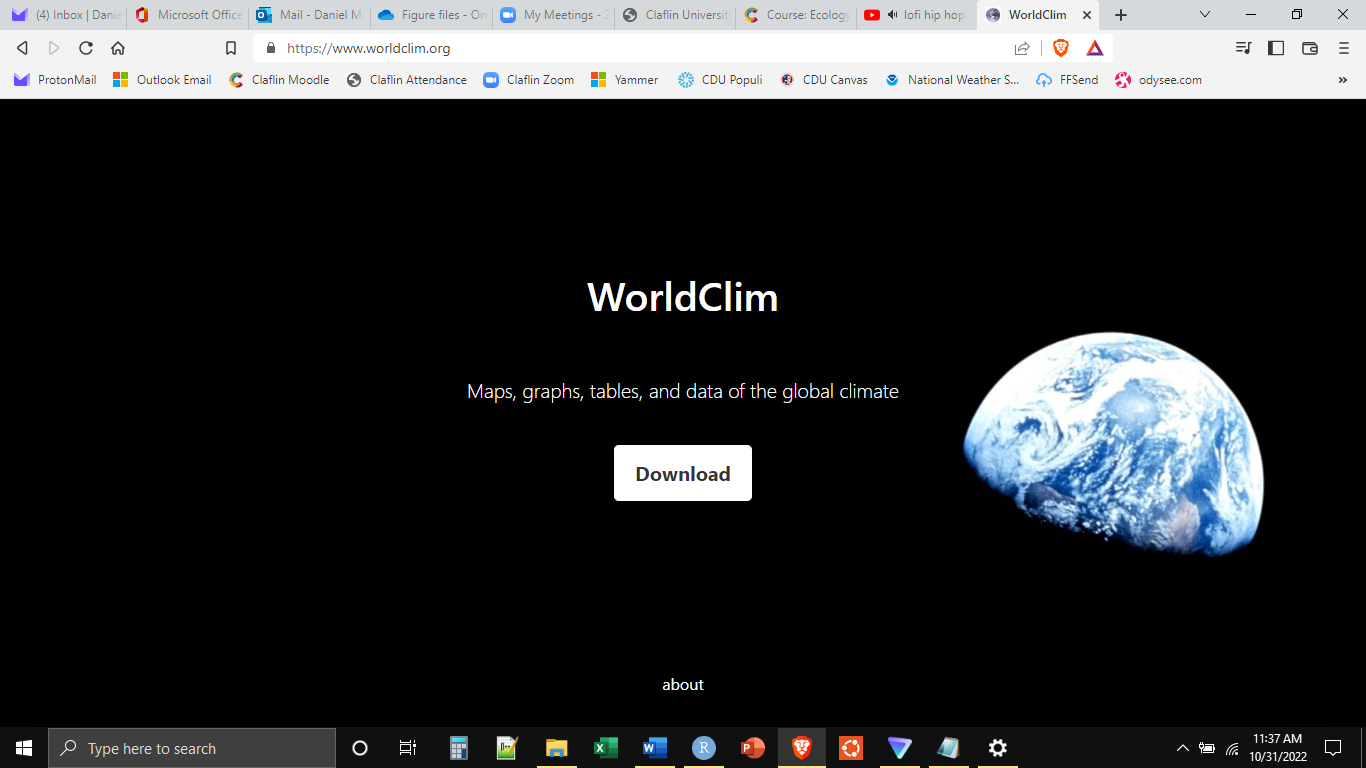
**https://www.gbif.org/species/2735055**

You will notice that it is not a perfect system. There are some records that are clearly incorrect, or at least not natural. For example, there is one point in Finland, in the arctic. That is either an error, or it is a record from a botanic garden of some sort. Thus, one of the important aspects of dealing with occurrence record data is “cleaning” it so that all the points represent genuine occurrences.

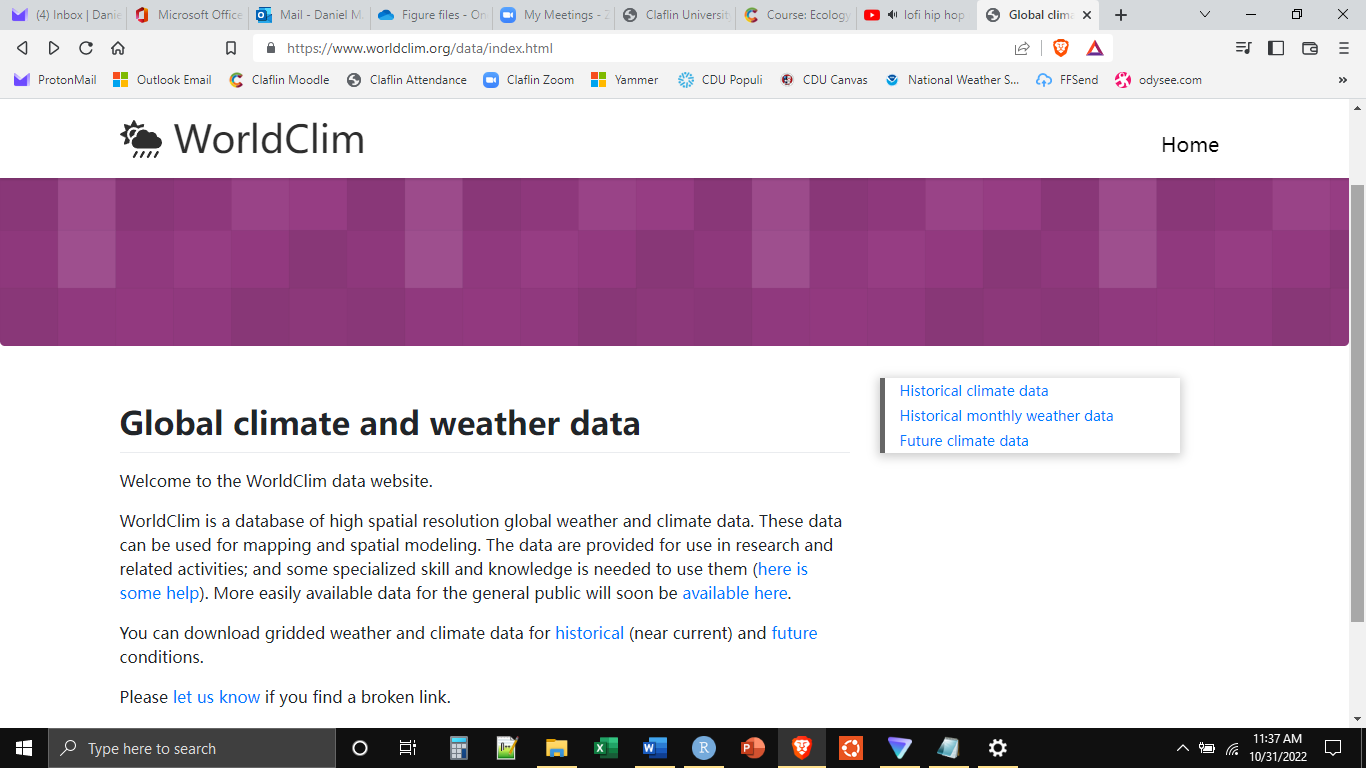
A second aspect of dealing with occurrence record data is making sure that the plant or animal identified in the record was identified correctly. For example, a researcher could record a sighting of a downy woodpecker at a particular location but it was actually a hairy woodpecker. This could be corrected by an examination of the specimen attached to the record. This requires some expertise, and we will not be doing this portion of the work today.

In addition to the occurrence record data, the environmental information for all of the occurrence locations is also necessary. Depending on the environmental data being used, these data will likely come from different sources. For example, climate data such as temperature and rainfall, can often bet determined through the use of satellites. Thus, reasonably detailed maps can be generated for the entire globe. Soil data, such as the depth of the topsoil layer, would need to be generated manually and the detail associated with these records is often reduced.

Today we will use the BioClim variables. These are 19 standard climate variables that tend to influence niche breadth, or the geographical size of the fundamental niche. **Go to the WorldClim website, www.worldclim.org, and click on “Download”.**



**Click on “historical (near current)”, this will take you to the page with the download links for the different**



You will notice that there are different options for the download and that these differences are indicated in minutes and seconds. These are the different resolutions. Below the level of the degree are first minutes and then seconds. Thus, the shorter the time the smaller the squares on the Earth’s surface and the finer the resolution of the data.



The 19 standard BioClim variables are (https://www.nccs.nasa.gov/services/data-collections/land-based-products/bioclim):

BIO1 Annual Mean Temperature

BIO2 Mean Diurnal Range

BIO3 Isothermality

BIO4 Temperature Seasonality

BIO5 Max Temperature of Warmest Month

BIO6 Min Temperature of Coldest Month

BIO7 Temperature Annual Range

BIO8 Mean Temperature of Wettest Quarter

BIO9 Mean Temperature of Driest Quarter

BIO10 Mean Temperature of Warmest Quarter

BIO11 Mean Temperature of Coldest Quarter

BIO12 Annual Precipitation

BIO13 Precipitation of Wettest Month

BIO14 Precipitation of Driest Month

BIO15 Precipitation Seasonality (Coefficient of Variation)

BIO16 Precipitation of Wettest Quarter

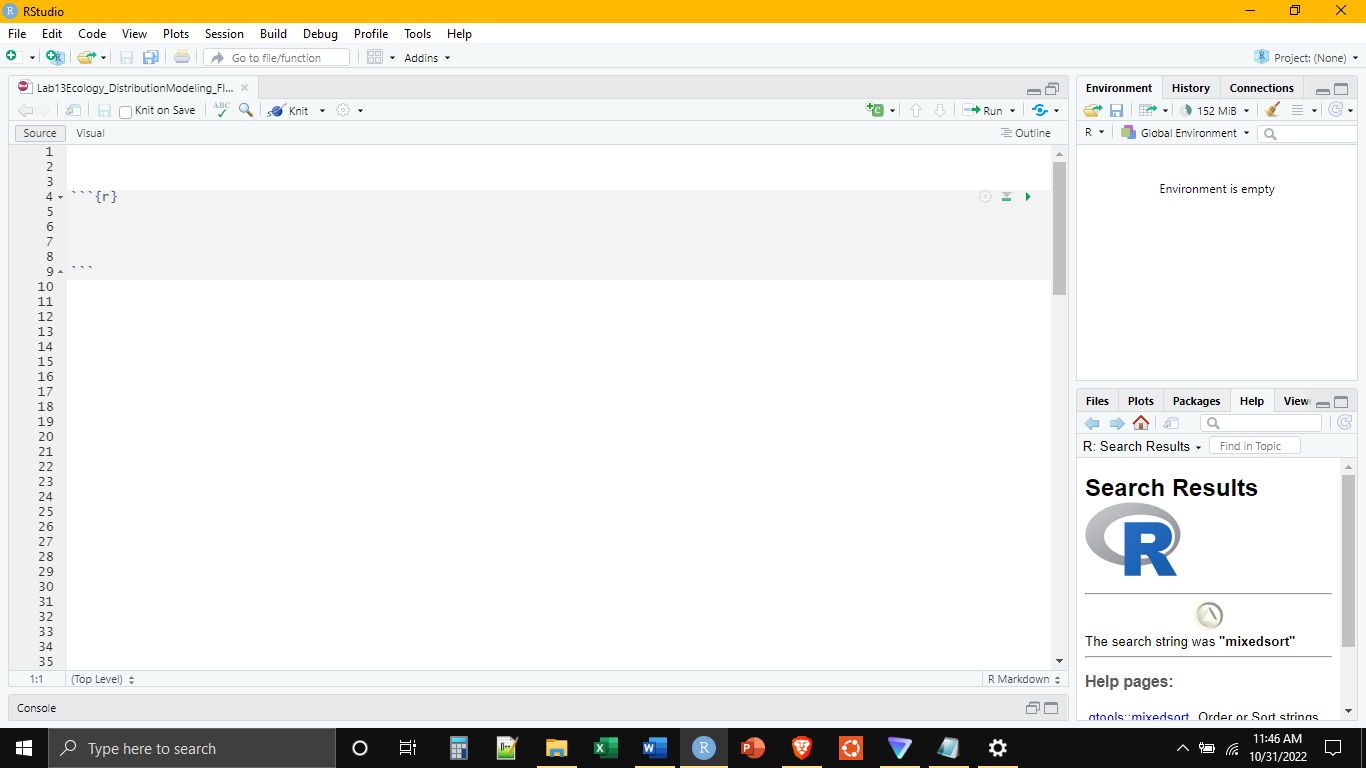
BIO17 Precipitation of Driest Quarter

BIO18 Precipitation of Warmest Quarter

BIO19 Precipitation of Coldest Quarter

For our work today, we will use these 19 standard climate variables. We will ignore soil variables for today, though for plants this can be very important. As stated above, we also will not spend the time to clean the occurrence record data as much as might be done in an actual study.

We will use R to generate the species distribution model. Open R Studio, open a new notebook file and save it in a folder for lab today.



The first thing that we will need to do is to install a few packages specific to species distribution modeling. In the console, execute each of the following lines of code.

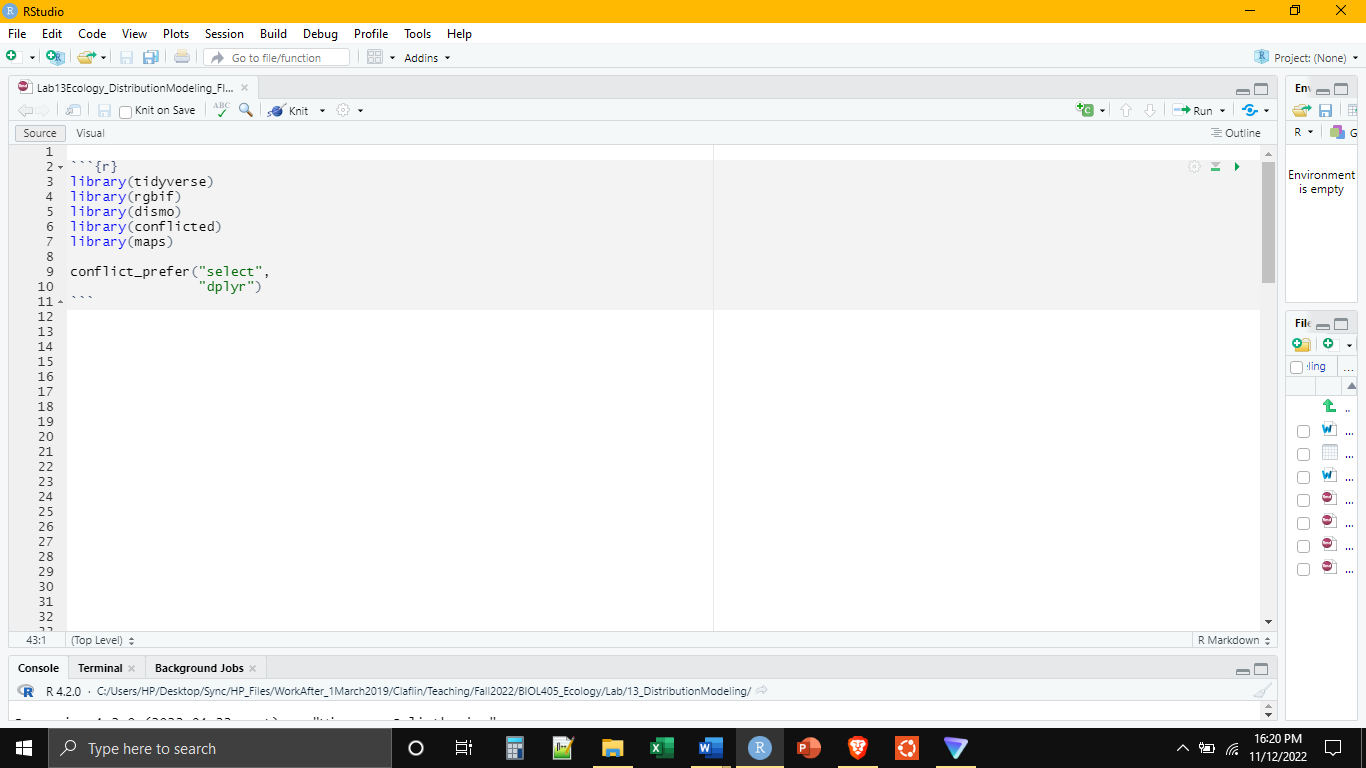
install.packages(“rgbif”)

install.packages(“dismo”)

install.packages(“conflicted”)

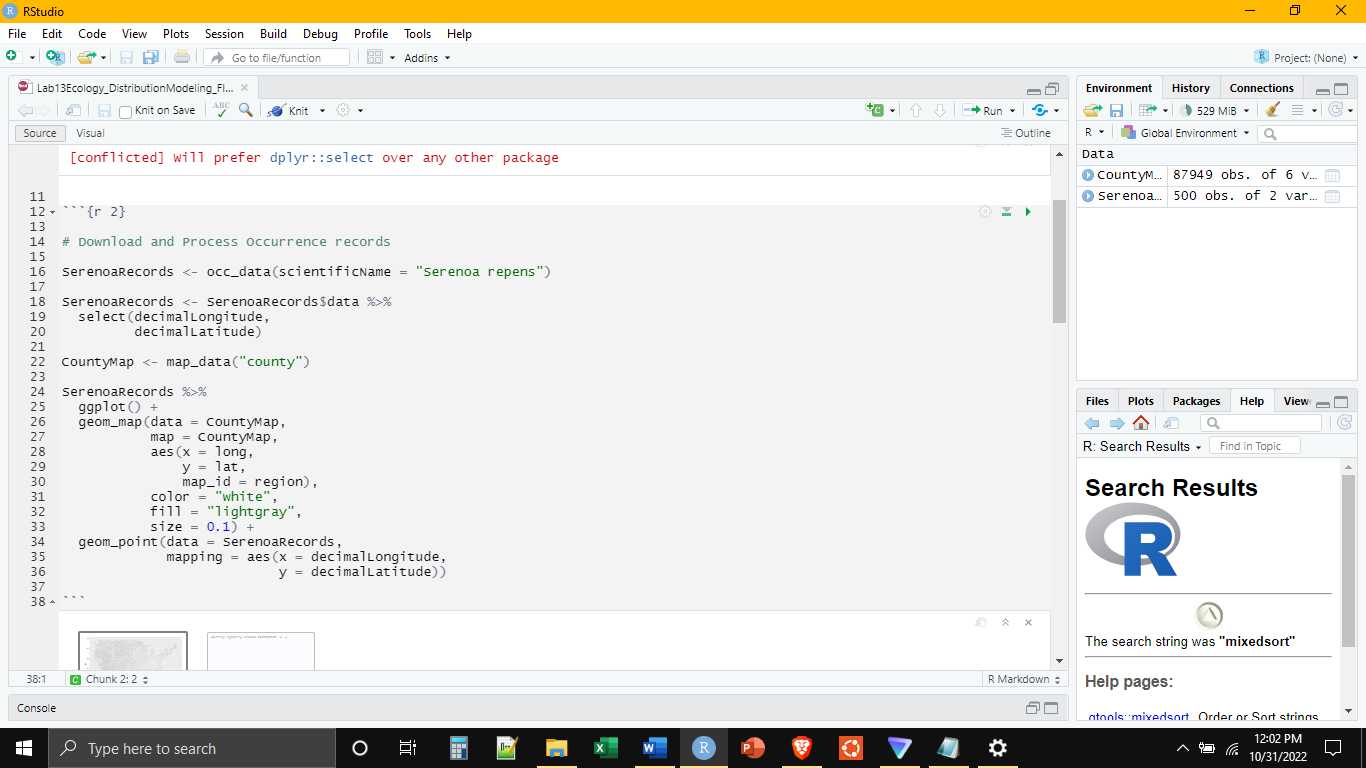
install.packages(“maps”)

After this is done, reduce the console (if desired), and use the first chunk to load libraries.

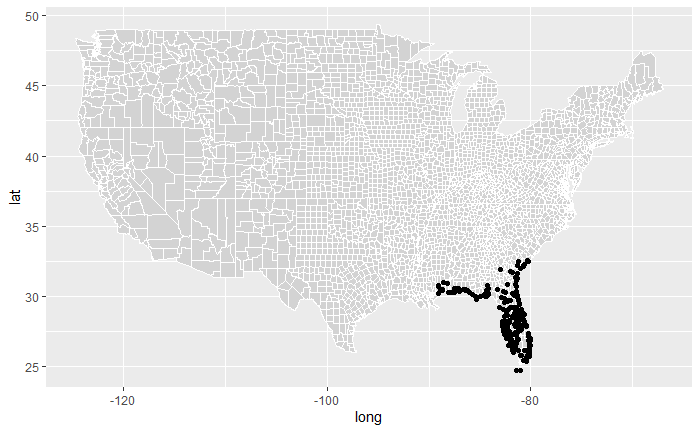


The conflict\_prefer() command is there because the “dismo” package and the “tidyverse” package both have a command called select(), and we are telling R to use the version of select from tidyverse as the default. Execute all of these lines of code.

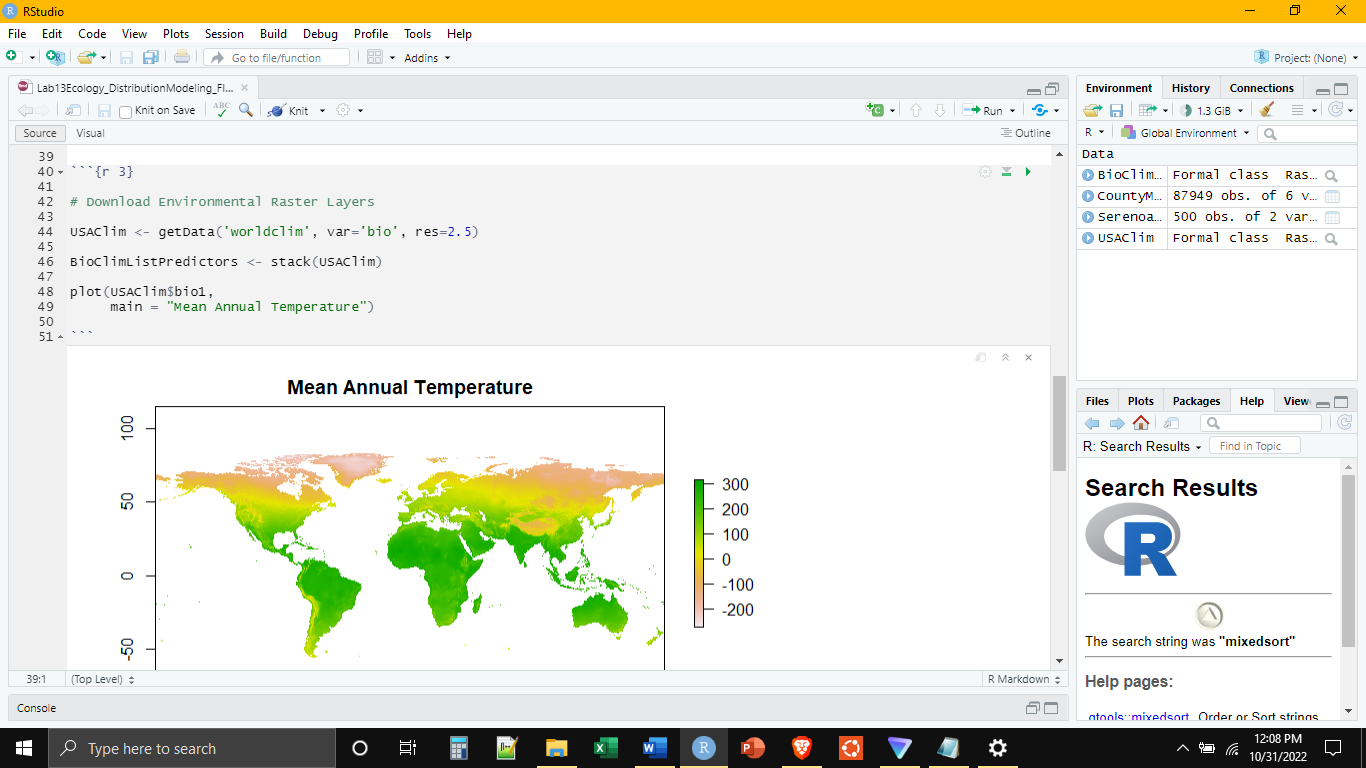
Start another chunk, remember to name it something different than the first one. We will use this chunk to read in the occurrence record data. We can actually pull in the occurrence records directly from GBIF, without needing to download them separately.



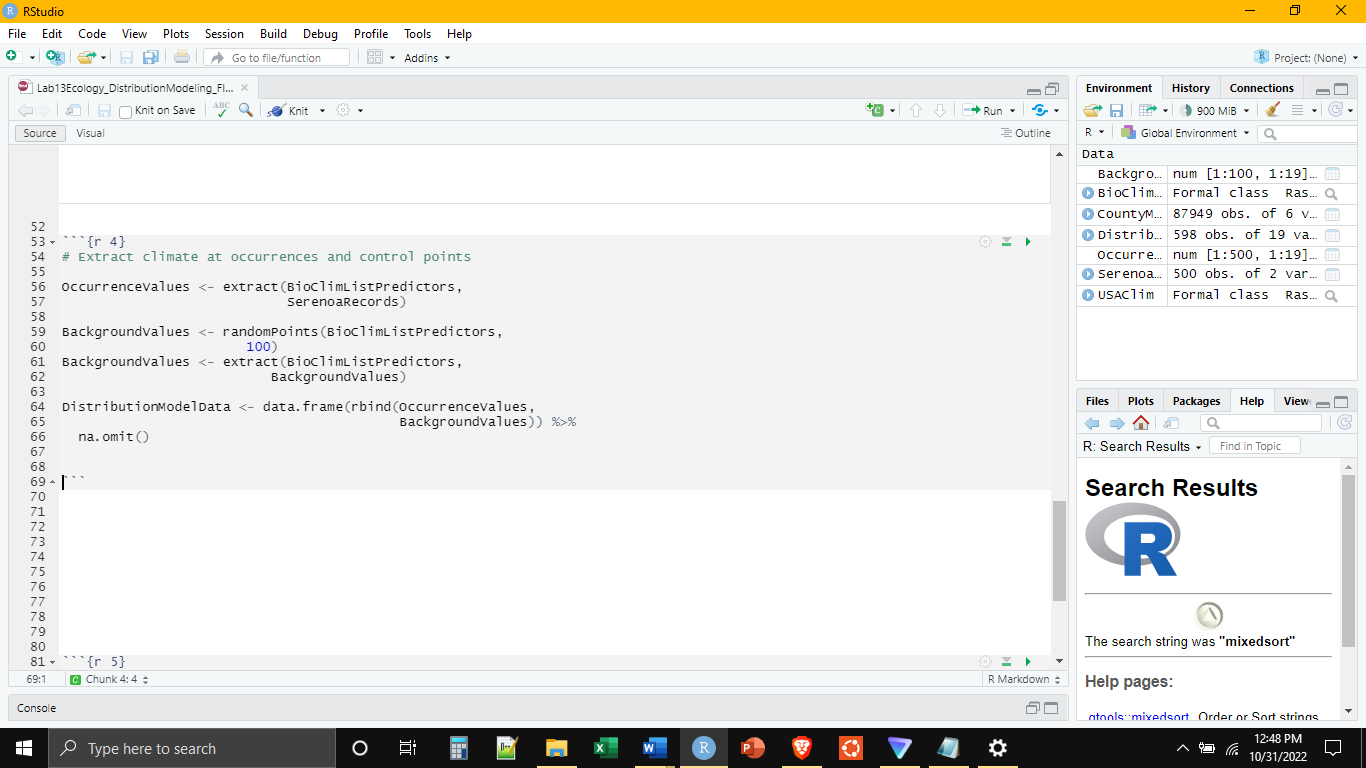
GBIF will actually only read in the first 500 occurrence records unless you have an account, an added complexity unnecessary for our purposes. We can plot the records that did get downloaded using a combination of geom\_point() and geom\_map(). The result will look like this.



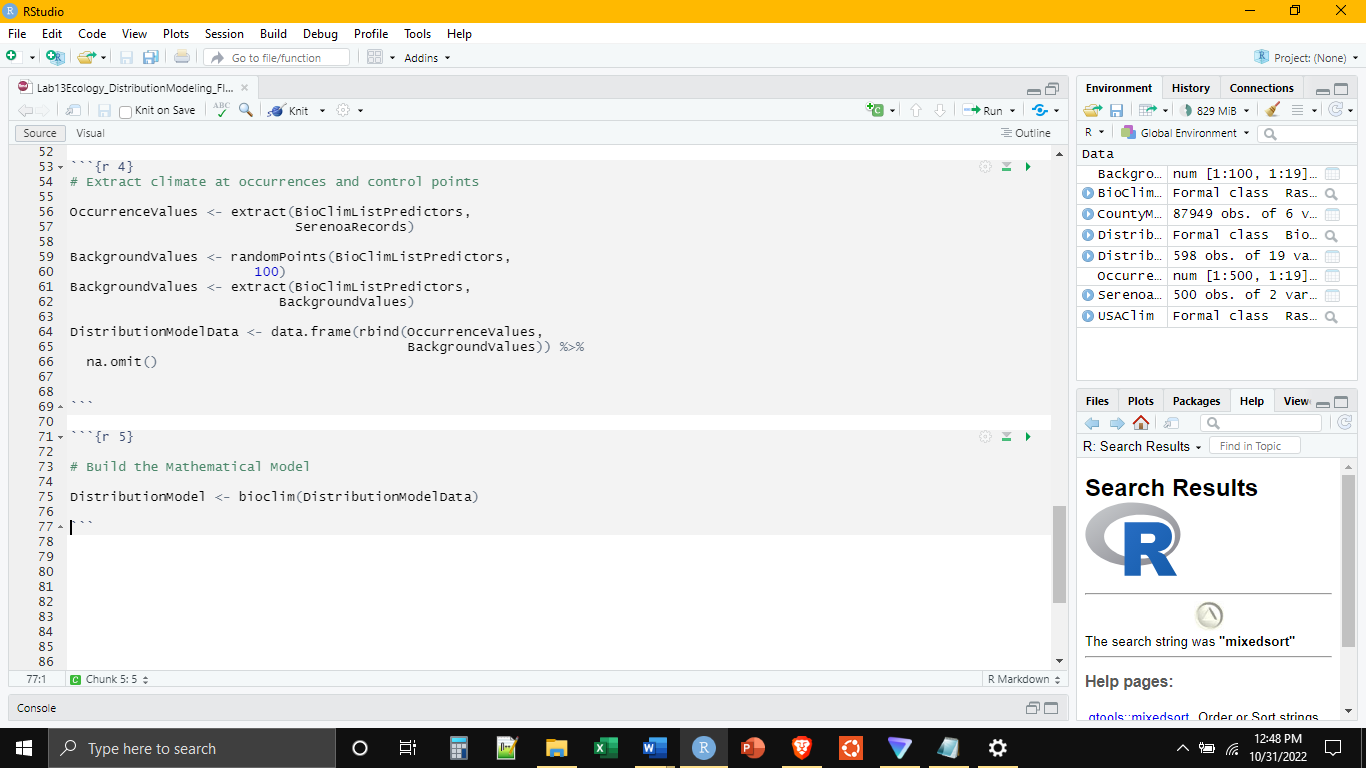
The next step is to download the environmental data. This can also be done directly from R Studio, and will come in the form of “raster” layers. Rasters are basically image files used to describe the geographic distribution of some variable. These are the type of data that satellites produce. Notice that we have to select a resolution. We are selecting 2.5 minutes, files that aren’t too big but have enough resolution to give us what we want. Nevertheless, the downloading and unpacking of the files may take a little while. Be patient. When it is done, we can stack all the rasters together and then plot one of them to see what it looks like.



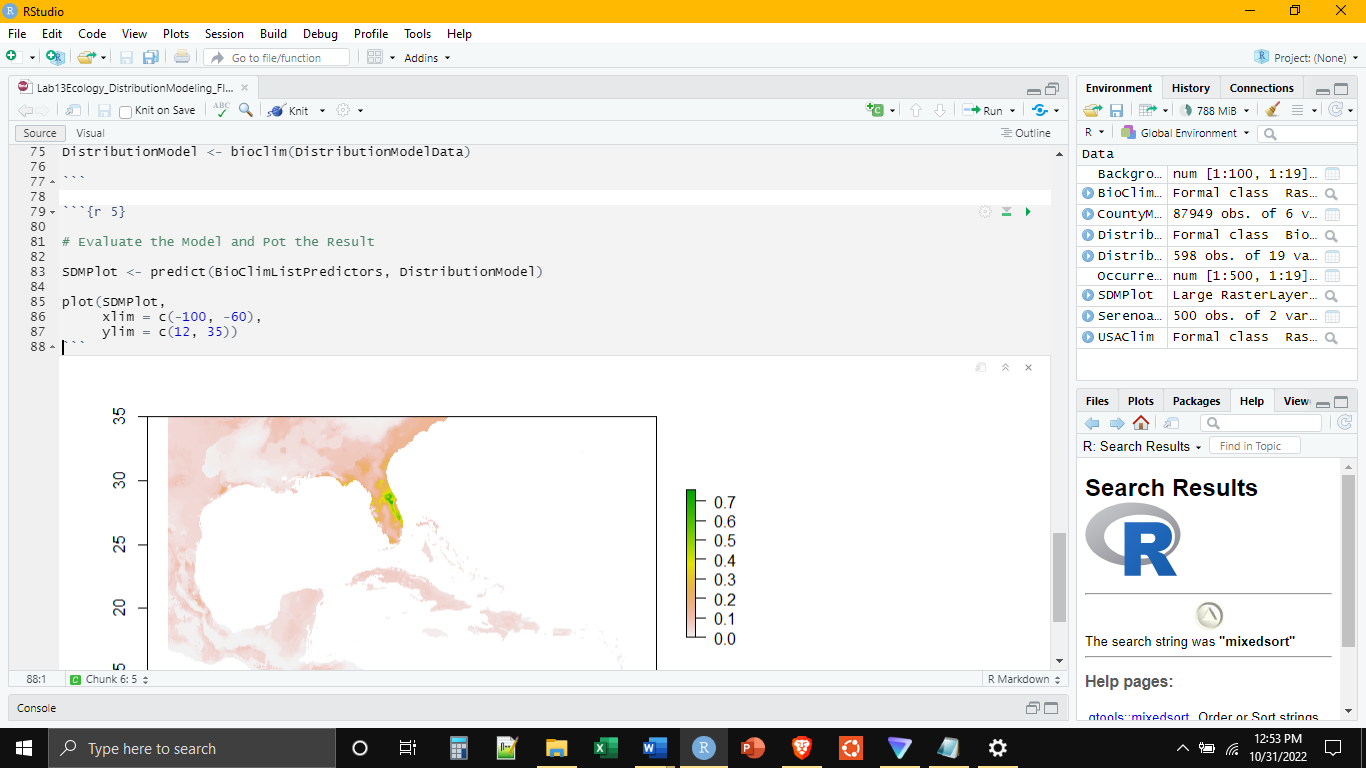
The next step is to “extract” the environmental data at the points that correspond to the occurrence records. In other words, for each pair of longitude and latitude, we want the values for each of the 19 variables at that point. This is the information we will use to build our model. Additionally, we are going to take environmental data from 100 random points as background environmental data to help fill out the environmental values. We will then combine these two datasets together. Executing these commands may take some time. Be patient.



Now that we have all of the data (the environmental conditions at each of the occurrence sites), the next step is to build the model.



Finally, we are going to run the model (which also may take some time), and then plot the model to see where it is most likely for the saw palmetto to live.



In any given model exercise, this one included, there is always a final step which is the evaluation of the model. Is the result of my model reasonable or unreasonable? Is it different from what we would normally expect or somewhat uninteresting. This evaluation would involve, at the very least, a comparison of the model result to a null model. This would then allow statistical evaluation of the model and its results. This is an essential part of modeling, but is somewhat beyond the scope of our work here.