Exercise 4. Loading environmental data, partitioning occurrence data, and calibrating a niche model

|  |  |
| --- | --- |
| Skills Acquired | Data Required |
| * Process environmental data layers to within your training region * Sample background points within your training region * Partition a dataset for an ecological niche model | * A set of biologically-relevant environmental covariates that are minimally correlated |

Step 1. Loading environmental data in Wallace

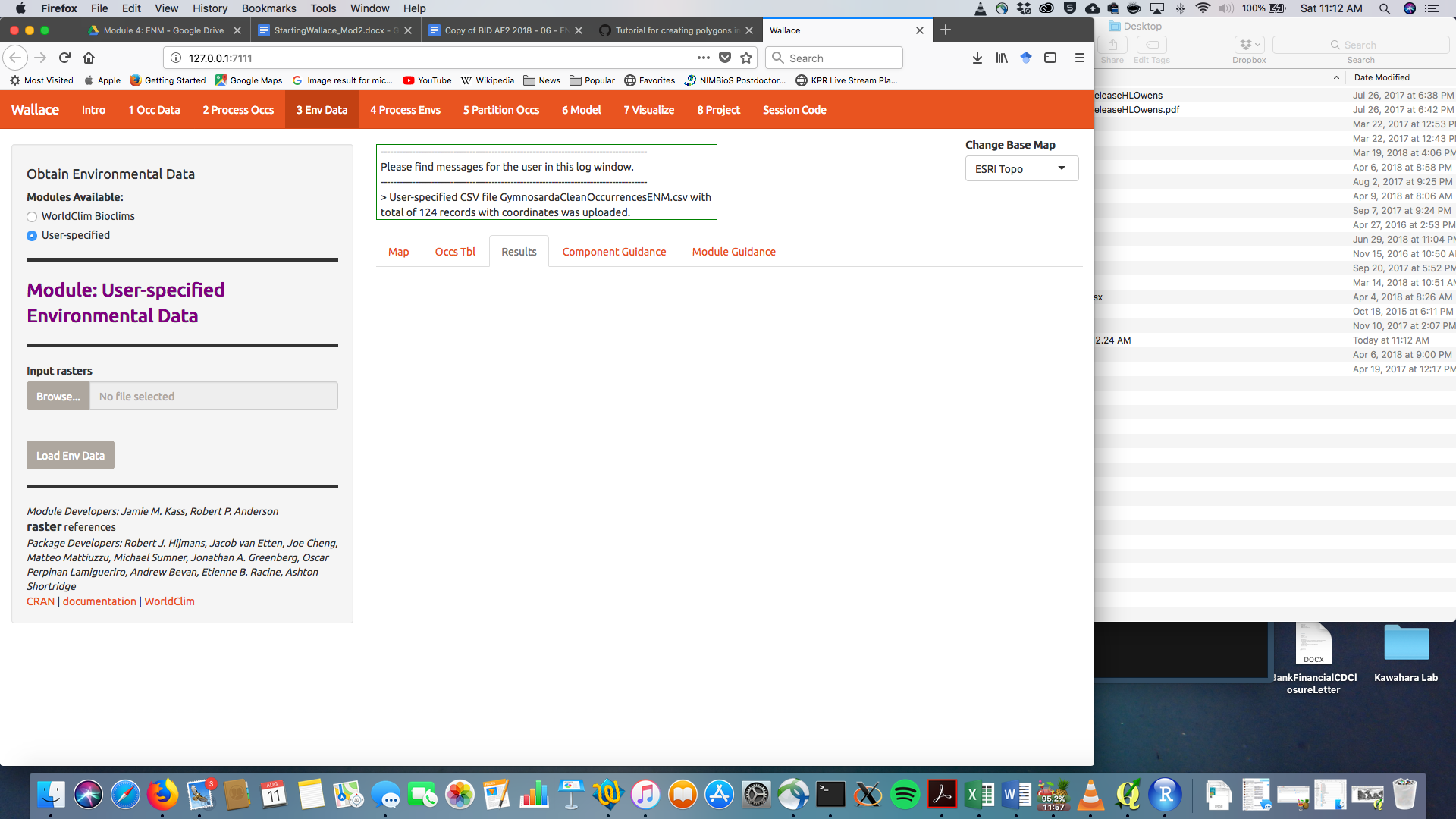
*Now that you know how to determine a training region, you will load your environmental data into Wallace and trim it to a training region we have made for you. Training region shapefiles are best made by creating a polygon shapefile in QGIS (or the GIS program of your choice). This is beyond the scope of this workshop, but here’s a tutorial:*<https://github.com/mtop/speciesgeocoder/wiki/Tutorial-for-creating-polygons-in-QGIS>*. We have provided you with a series of environmental variables from the* [*MARSPEC*](http://www.marspec.org/) *dataset - a set of high resolution climatic and geophysical GIS data layers for the world ocean from 1955 - 2010. The variables have provided have been selected because they are biologically relevant and covary minimally. There is a very nice resource for distribution modeling at* [*http://rspatial.org/sdm/*](http://rspatial.org/sdm/)*; for more information on determining covariate correlations, see Chapter 4 on that website.*

● We are skipping ‘2 Process Occs’ because we processed the occurrence data in Exercise 2.

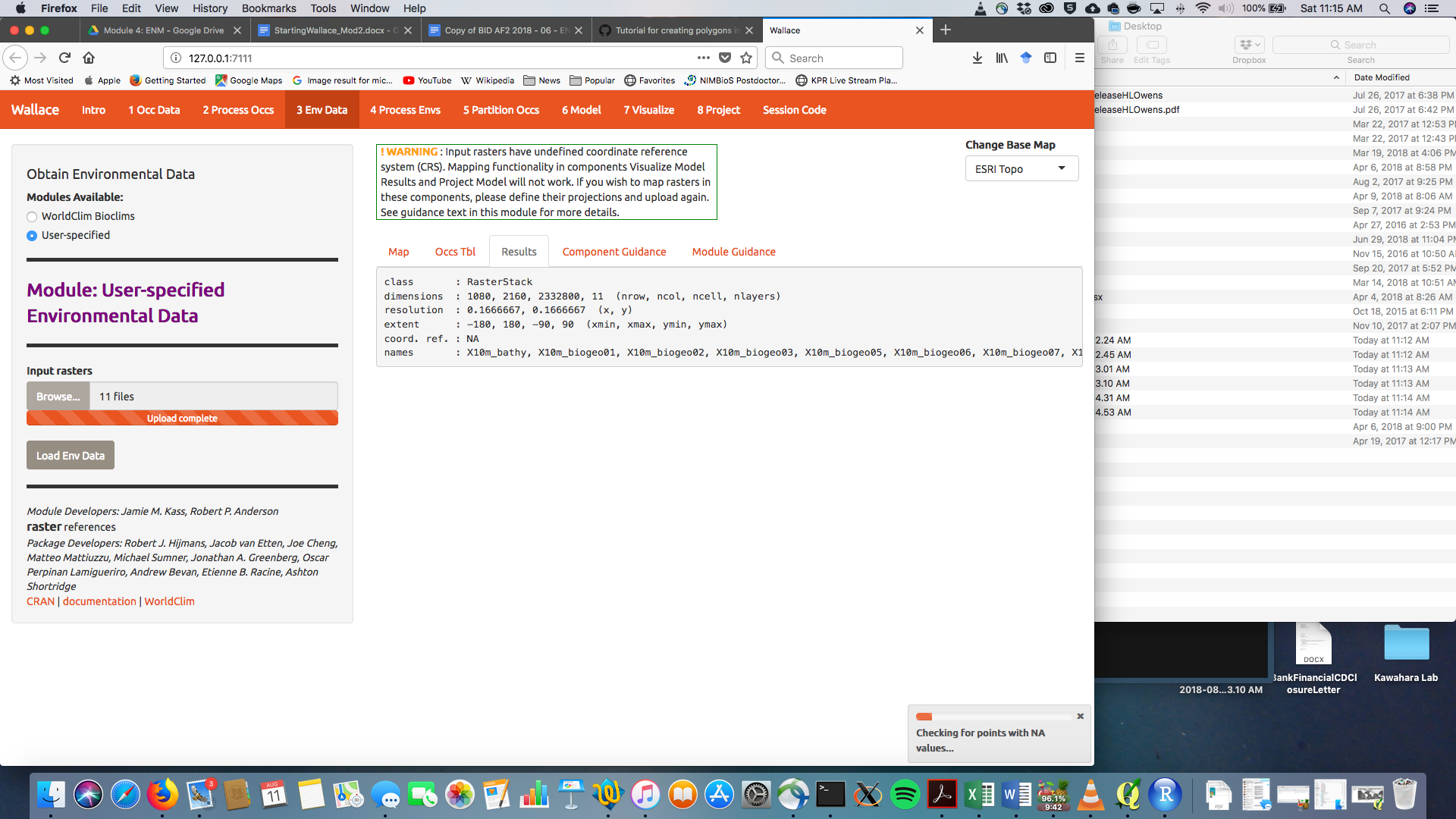
● If you need a review on environmental data, click on the “Component Guidance” tab.

● If you need additional information about the environmental data functions in Wallace, click on the “Module Guidance” tab.

1. *Click on “3 Env Data”* in the browser window in which Wallace is running.
2. *Select the “User-specified”* radio button, and *click on the “Browse”* button.



1. *Navigate to the folder* with the environmental data, *select all the “.tif” files*, and *click “Open”.* After a few moments, you should see an orange bar under the “Browse” button that says “Upload Complete”.
2. *Click the “Load Env Data” button*. Your view should change to something similar to that shown below. The gray box will show metadata on the environmental data you have uploaded.

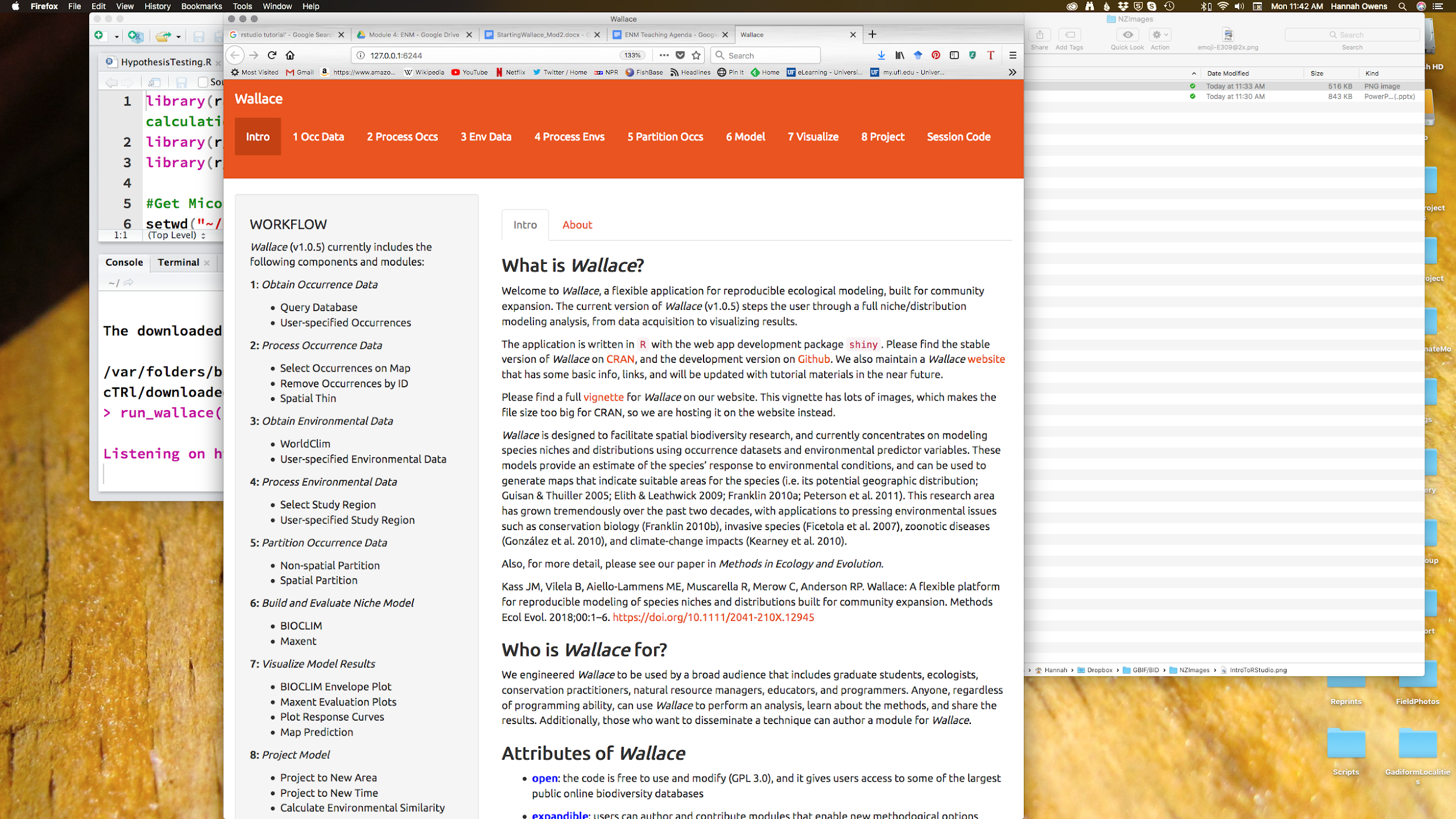


Step 2. Processing environmental data in Wallace

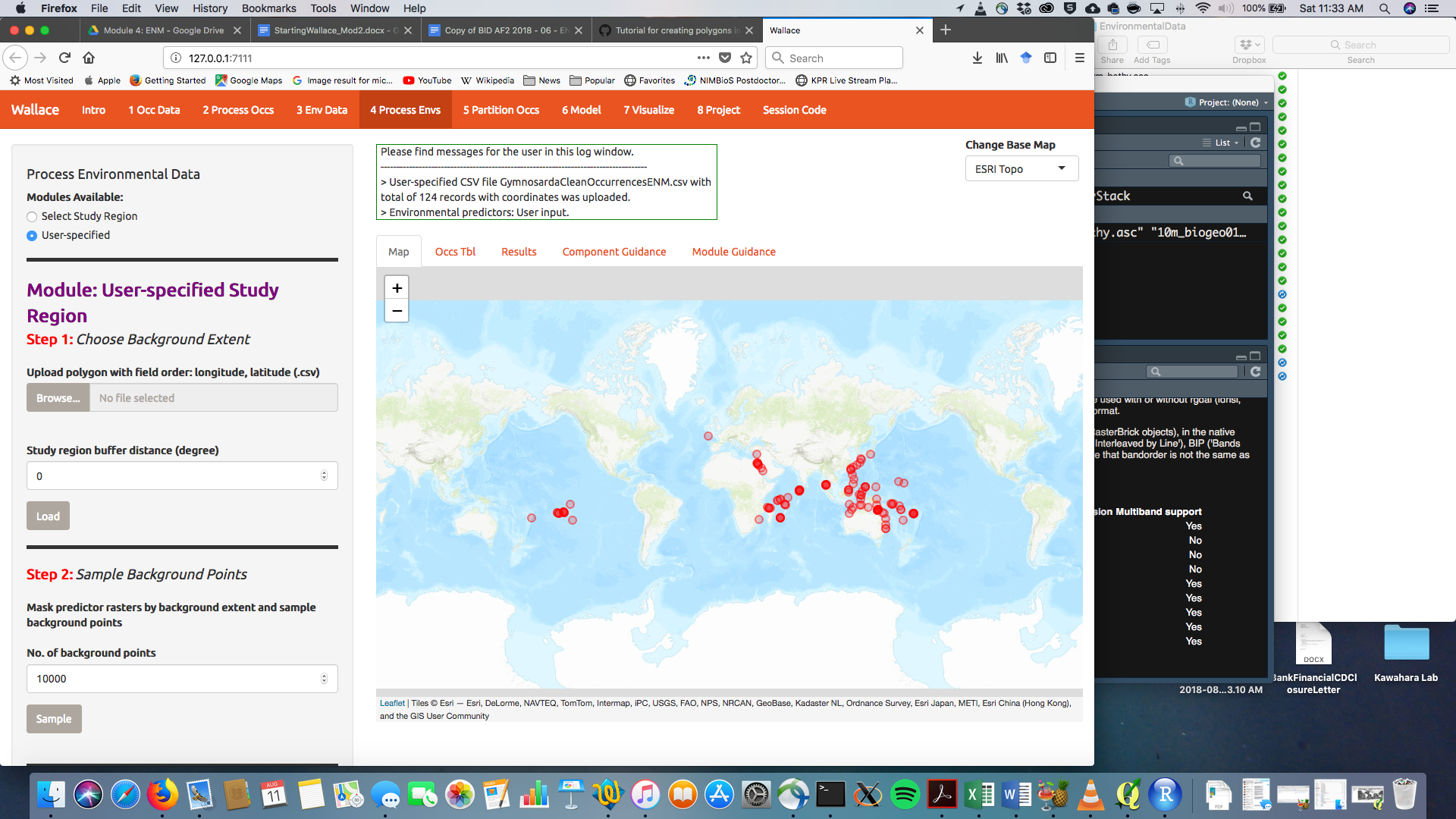
● To review background information on the processing of environmental variables, click the “Component Guidance” tab.

● If you need additional information about the environmental variable processing functions in Wallace, click on the “Module Guidance” tab.

1. *Click on “4 Process Envs”* in the browser window in which Wallace is running.

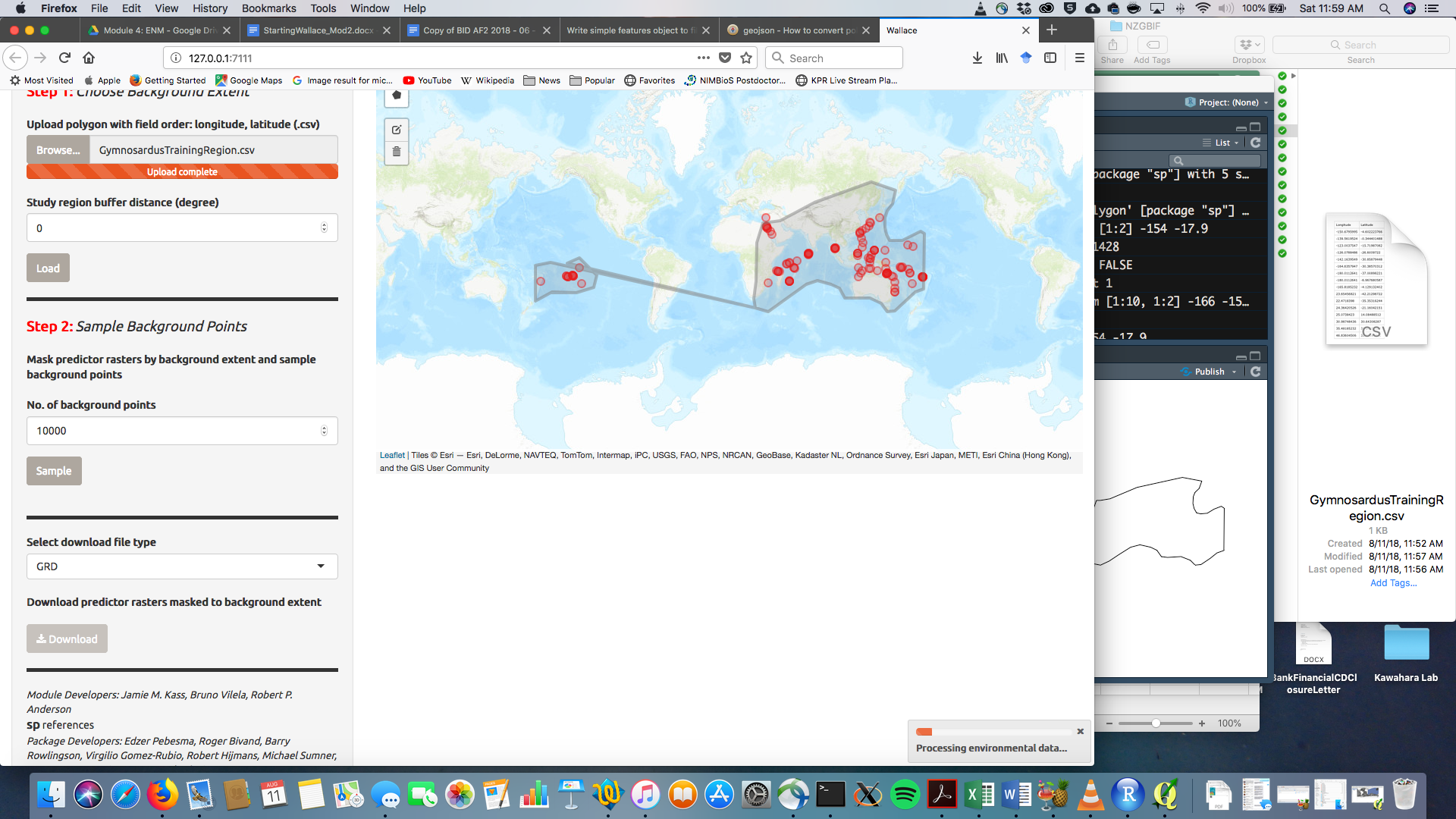


1. Load the training region.

* *Select the “User-specified” radio button*, and *click on the “Browse” button*.
* *Select the “GymnosardusTrainingRegion.csv” file* and *hit “Open”.* This is a CSV file with the coordinates of a polygon representing the training region.
* *Click “Load”*. You should now see the training region on the map. You will see something that looks like the map below. The training region looks weird because the polygon is split by the meridian (these tools were designed by people that do not work in the Pacific). No worries--it will function correctly moving forward.

1. Sample background points.

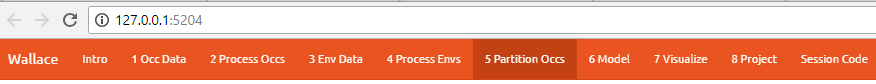
* The number of background points shown (10,000) is fine. This is the number of points that will be sampled randomly from the training region you have uploaded. Values of predictor variables for these background points can then be compared to those at the occurrence points to improve model fit. *Click “Sample”,* and be patient. This takes a little time.

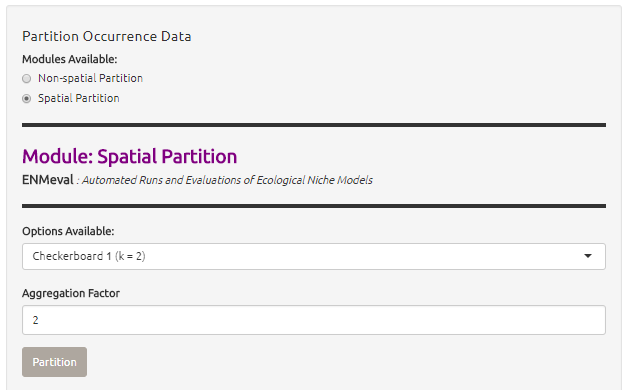


Step 3. Partitioning Occurrence Data.

*Ideally, you will have two completely independent occurrence datasets in order to determine the strength of the model's predictive ability. Unfortunately, this rarely reality. When no independent datasets exist, one solution is to partition your data into subsets we assume are independent of each other, then sequentially build a model on all the subsets but one and evaluate this model on the left-out subset. This is known as k-fold cross-validation (where k is the total number of subsets). After this sequential model- building step is complete, Wallace summarizes (averages) the statistics over all of the partitions and then builds a consensus model using* ***all*** *the data.*

· Click on the “Component Guidance” tab for information on the theoretical basis of partitioning occurrence data and the “Module Guidance” for information on about spatial partitioning.

1. *Click on “5 Partition Occs”* in the browser window in which Wallace is running. 

b) Choose a spatial partition method. 

* Coordinate with your group to select different partition methods. You can then compare the results of the different methods to better see how the methods differ.
* *Select the “Spatial Partition*” radio button.
* From the “Options Available” dropdown menu, *select your method*.
* *Click on “Partition”.* This may take a few minutes depending on the amount of occurrence data you have and the partition option selected.

ANSWER QUESTION 1 FOR EXERCISE 6

1. Save your partitioned data. Download your partitioned occurrence background localities with partition values as a .csv file. To do this, *click on “Download”* and select your working project folder. *Save the file as: “Gymnosarda-partitioned.csv”.*

* Although you won’t be using this separate .csv file for your modeling, it is good practice to maintain copies of data as it is processed, just as you should always keep detailed notes of each step taken during research.