Using Digitized Collections-Based Data in Research: Applications for Ecology, Phylogenetics, and Biogeography
Botany 2023

Sponsored by iDigBio and BiotaPhy

Florida Museum of Natural History, University of Florida

The following are hands-on exercises to introduce the participants to the programs and protocols described during the workshop.

Table of Contents

SCHEDULE

8:00	Welcome and Overview of the Workshop – Pam (Makenzie) (01)
8:15	Data Standards and Data Fields – Pam (Makenzie) (02)
8:30	Data Downloads: Overview and Portals –Makenzie (Pam) (03)
9:00	***Activity**:* Manual Data Downloads, iDigBio Portal – Lauren
9:30	***Activity**:* R-based Data Downloads – Shelly (04)
9:45	Data Cleaning Overview – JT (05)
10:00	**Break**
10:30	***Activity***: Data Cleaning (A) Manual (JT) (B) R-based (Shelly)
11:00	Georeferencing Overview – Lauren (06)
11:15	***Activity*****:** Georeferencing Activity (A) Manual and (B) R-based - Lauren
12:15	**Lunch**
1:00	Climatic Processing Overview – Shelly (07)
1:20	***Activity***: Climatic Processing (A) Manual (Elizabeth) (B) R-based (Shelly)
2:00	Applications of Ecological Niche Modeling (ENMs)– Makenzie (08)
2:15	ENM Settings Overview – JT (09)
2:30	***Activity***: Ecological Niche Modeling (A) Manual (Elizabeth) (B) R-based (Shelly)
3:00	**Break**
3:30	Interpreting ENMs (Shelly) (10)
3:40	***Activity***: Interpreting ENMs (A) Manual (Elizabeth/Lauren/Makenzie) (B) R-based (Shelly)
4:00	Post-ENM Analysis – Shelly (11)
4:15	BiotaPhy: An Overview – Doug/Maria (12)
4:30	***Wrap-up and Q&A***
5:00	End

Workshop Leaders:
Makenzie Mabry: mmabry44@gmail.com
Lauren Gillett: laurenwhitehurst@ufl.edu
JT Miller: http://jtmiller@umail.ucsb.edu
Shelly Gaynor: michellegaynor@ufl.edu
Elizabeth White: elizabethwhite1@ufl.edu
Malu Ore Rengifo: m.orerengifo@ufl.edu
(Remote) Pam Soltis: psoltis@flmnh.ufl.edu
(Remote) Doug Soltis: dsoltis@ufl.edu
(Remote) Maria Cortez: mariacortez@ufl.edu

SET-UP

(1) Download the dropbox file locally (suggested "Desktop/")

(2) R and R Studio (demo built using R Versions 4.2.1)

https://www.rstudio.com/products/rstudio/download/
https://cran.rstudio.com/index.html

- Download and install R and the free desktop version of RStudio
  - Then in the shared dropbox folder (or the version of this folder that you downloaded):
    - Open the R project by double clicking the .Rproj file. This can be found under "Demo/Rbased/CrashCourse/CrashCourse.Rproj"
      - Navigate to 00_Setup.R. Click on 00_Setup.R; then, to install the files that you will need, go to Source in the upper left quadrant and select Source with Echo from the drop-down menu. The packages will be installed automatically.

(3) QGIS

- QGIS (version: 3.16)
  - MacOS: https://qgis.org/downloads/macos/qgis-macos-ltr.dmg
  - Windows: https://qgis.org/downloads/QGIS-OSGeo4W-3.16.8-4.msi
  - Other: https://qgis.org/en/site/forusers/download.html

(4) Install an Add-in on EXCEL

- Add-in the Analysis ToolPak:
  - **i.** Find instructions for this here: **https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4**

iDigBio Activities

DATA DOWNLOAD

(A) Manual – iDigBio web-portal (https://www.idigbio.org/portal/search)

- Download data from your web browser

(B) R-based

- Open the R project by double clicking the .Rproj file. This can be found under "*Demo/Rbased/CrashCourse/CrashCourse.Rproj"*
  - Navigate to 01*_Download_Occurrence_Data.R*
- Or follow along on the *CrashCourse_2023.html* file which can be found *"Demo/Rbased/CrashCourse_2023.html"*. This file can be opened in a web-browser. You can also view the HTML here: https://tinyurl.com/msdurpk

DATA CLEANING

**Depending on the size of your dataset and your comfort with R and RStudio, you may or may not want to use the R script that we provide. The basic steps are as follows:

1. Resolve taxon names
2. Clean localities
   a. Round up the latitude/longitude
   b. Remove coordinates at 0,0
   c. Remove coordinates in cultivated zones, botanical gardens, etc.
   d. Remove coordinates outside of the desired range
3. Remove duplicates
4. Spatial correction
5. Produce a csv

(A) Manual –

- Instructions are available in "Demos/Manual/Data_Cleaning/"
  - Activity from Gaynor, M. (2020). Cleaning Biodiversity Data: A Botanical Example Using Excel or RStudio. Biodiversity Literacy in Undergraduate Education, QUBES Educational Resources. doi:10.25334/DRGD-F069.

(B) R-based –

- Open the R project by double clicking the .Rproj file. This can be found under *"Demo/Rbased/CrashCourse/CrashCourse.Rproj"*
  - Navigate to 02*_Occurrence_Data_Cleaning.R*
- Or follow along on the *CrashCourse_2023.html* file which can be found *"Demo/Rbased/CrashCourse_2023.html"*. This file can be opened in a web-browser. You can also view the HTML here: https://tinyurl.com/msdurpk

GEOREFERENCING

(A) Manual –

Files for this activity can be found in the "Demo/Manual/Georeferencing/" folder.

Georeferencing Demo

Resources:

GEOLocate – Web application: http://www.geo-locate.org/default.html
Google Maps: https://www.google.com/maps
Falling Rain: http://www.fallingrain.com
Getty Thesaurus of Geographic Names (TGN): http://bit.ly/Getty-TGN
Fuzzy Gazetteer: http://dma.jrc.it/services/fuzzyg/

1. Use the **standard** GeoLocate client to identify the first three localities in the GeorefExamples_Florida.xls file.
    a. Enter the locality string, country, state, and county information from the Excel sheet.
    b. Click "Georeference."
    c. Inspect the "Possible Locations" by clicking on the "XX possible locations found" where XX is the number of locations GeoLocate identified.
    d. Use an alternative resource to double check the locality. Try Google Maps.
    e. Adjust the point location as you see fit. The green point is the active one.
    f. Click the green point on the map, then click "Edit uncertainty". Adjust the uncertainty radius by moving the grey arrow.
    g. Return to the "Workbench" and record the latitude, longitude, and uncertainty.
        i. If the uncertainty is >1000 then discards the points.
2. Optional.
   Use the **batch** GeoLocate client to upload the localities in the GeorefExamples.xls file.
    a. Copy and paste the appropriate information from the GeorefExamples.xls file into your own GeoLocateBatchFormat.csv.
        i. http://www.geo-locate.org/standalone/tutorial.html

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
Row1	locality string	country	state	county	latitude	longitude	correction status	precision	error polygon	multiple results

        ii. Do not label the columns (your first row = first sample)
        iii. **Make sure to save as a .csv**
        iv. The majority of the columns will be empty
    b. Go to the **batch** GeoLocate client and upload the formatted csv file
    c. "Page Georeference" will georeference all eight localities available at once. "Georeference" will do one at a time.
    d. Select a locality and go through **Steps 1c** to **1g**. Once you are pleased with the locality and uncertainty click "Correct" to note that you have gone through this georeference.
    e. Work through the remaining localities.

f. If you **do not** finish a batch georeferencing, you can click on "File Management" at the bottom of the screen to receive a retrieval code. This will allow you to re-access this file whenever you wish without the need to download and upload.
g. If you **do** finish a batch georeferencing, you can click on "File Management" and then "Export" to download the finished georeferenced file.

3. Use alternative resources to identify the localities in the example file. These are much more difficult and could use some historical maps and/or corrected spelling.

(B) R-based –

- Open the R project by double clicking the .Rproj file. This can be found under *"Demo/Rbased/CrashCourse/CrashCourse.Rproj"*
    - Navigate to 03_*Georeferencing.R*
- Or follow along on the *CrashCourse_2023.html* file which can be found *"Demo/Rbased/CrashCourse_2023.html"*. This file can be opened in a web-browser. You can also view the HTML here: https://tinyurl.com/msdurpk

CLIMATE LAYER PROCESSING

QGIS provides a much better understanding of the processes happening with this step, but the R script streamlines a largely repetitive process. We will demo both options.

(A) Manual – QGIS - Optional

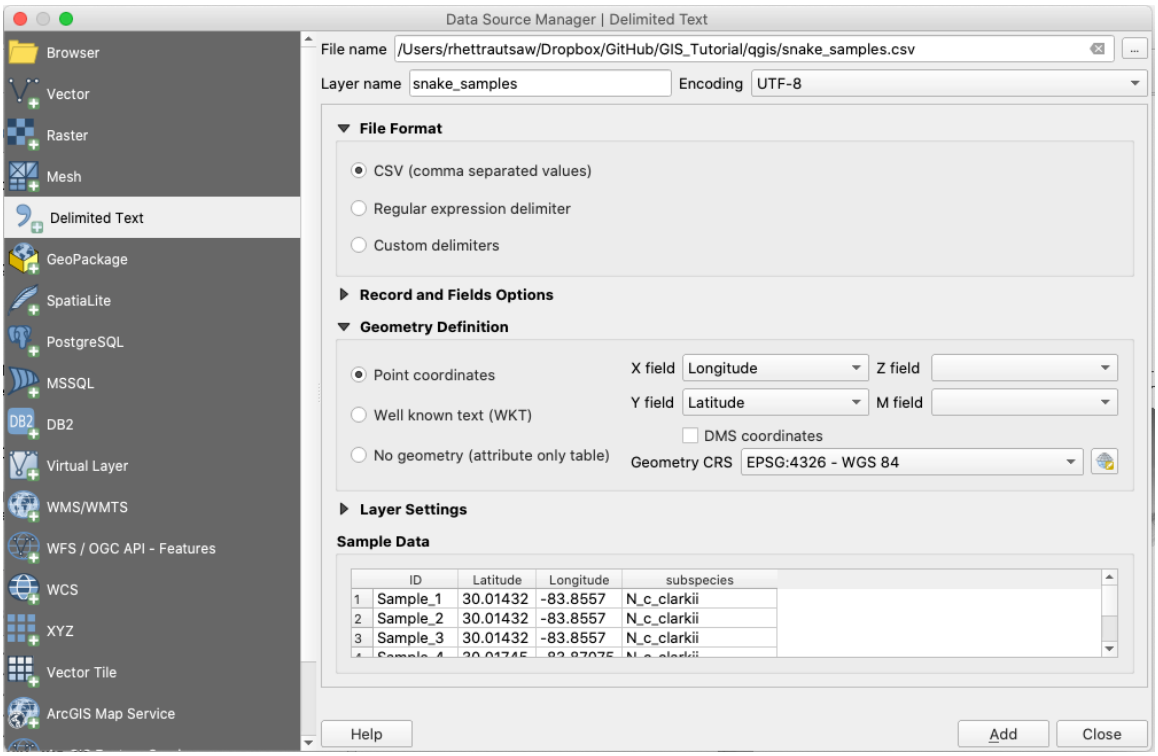
This activity was made by Rhett Rautsaw. Files for this activity can be found in *"Demo/Manual/Climate_Layer_Processing/"* folder.

QGIS version has to be 3.16; if not, this will not work

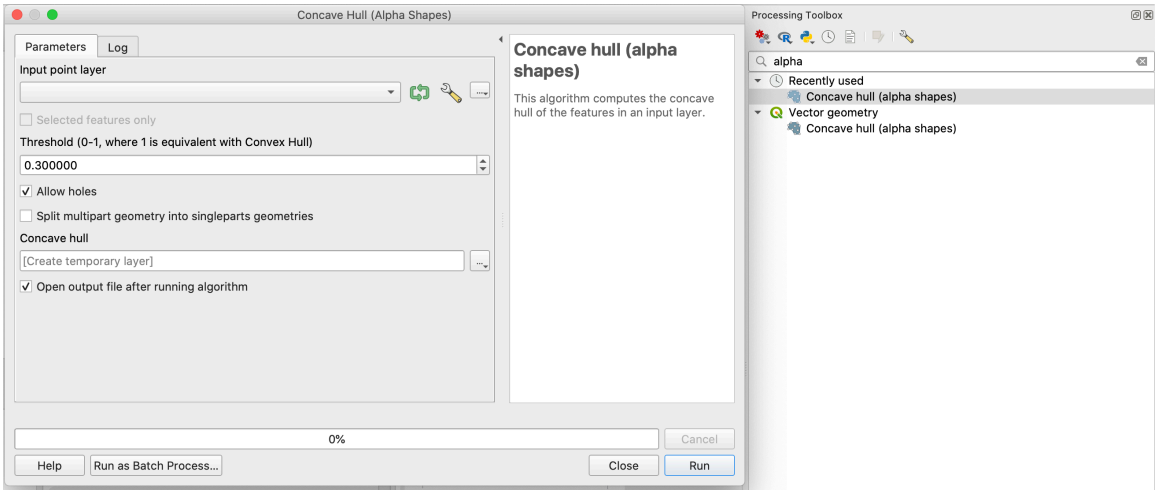
QGIS (version: 3.16)

1. Open QGIS.
2. Drag the layers (.tif files) found in the *"data/climate_processing/bioclim/"* folder into QGIS. They should automatically appear. The box on the left lists the different layers not the layer is displayed.
3. Add occurrence records from text-delimited file (Layer Menu > Add Layer > Add Delimited Text Layer…). Navigate to "*data/cleaning_demo/maxent_ready/diapensiaceae_maxentready_20220625.csv*". X field is "longitude" and Y field is "latitude". Make sure the CRS is

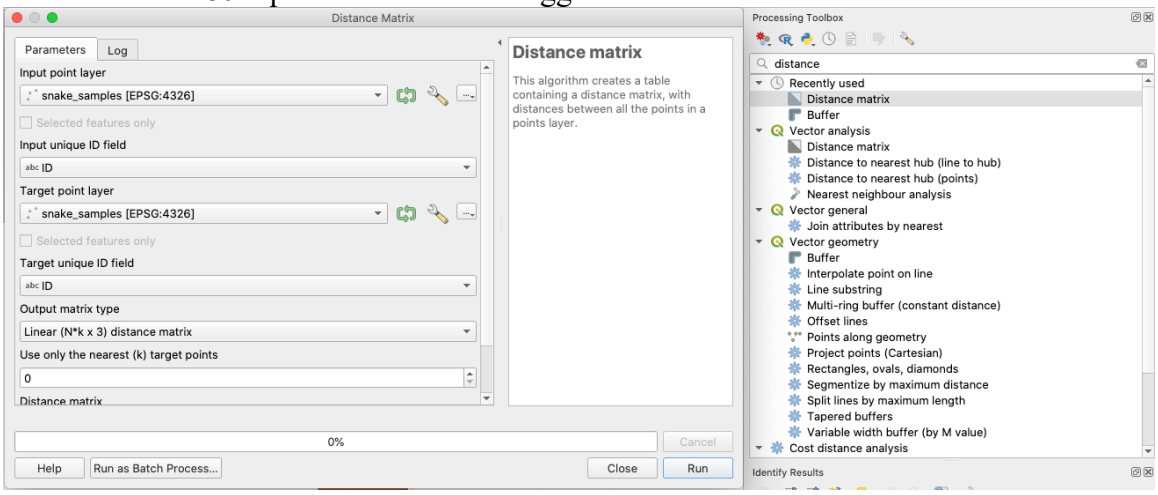
EPSG:4326 – WGS 84.



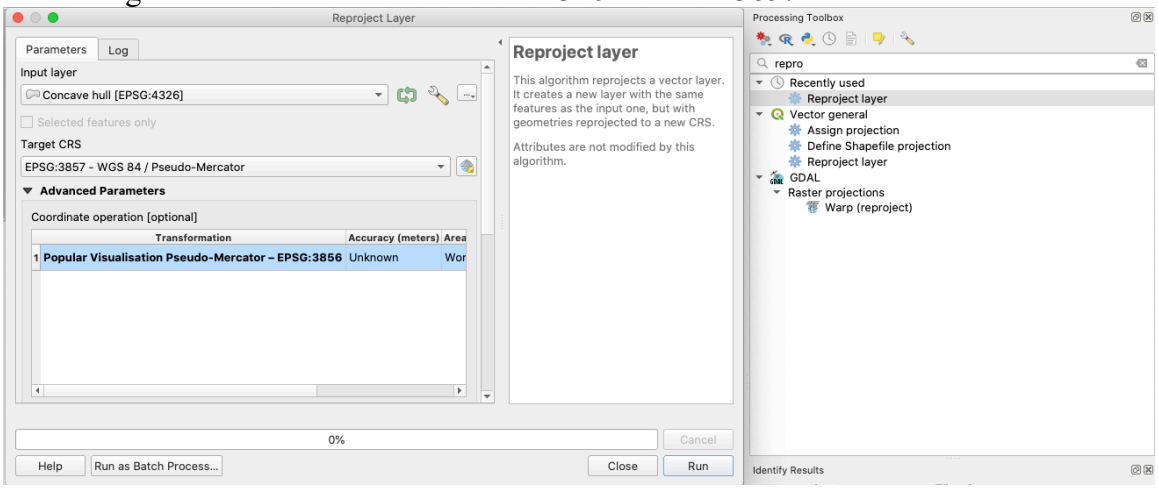
4. Create an alpha hull/shape, using the Processing Toolbox Concave Hull Tool. Set the threshold to 1.



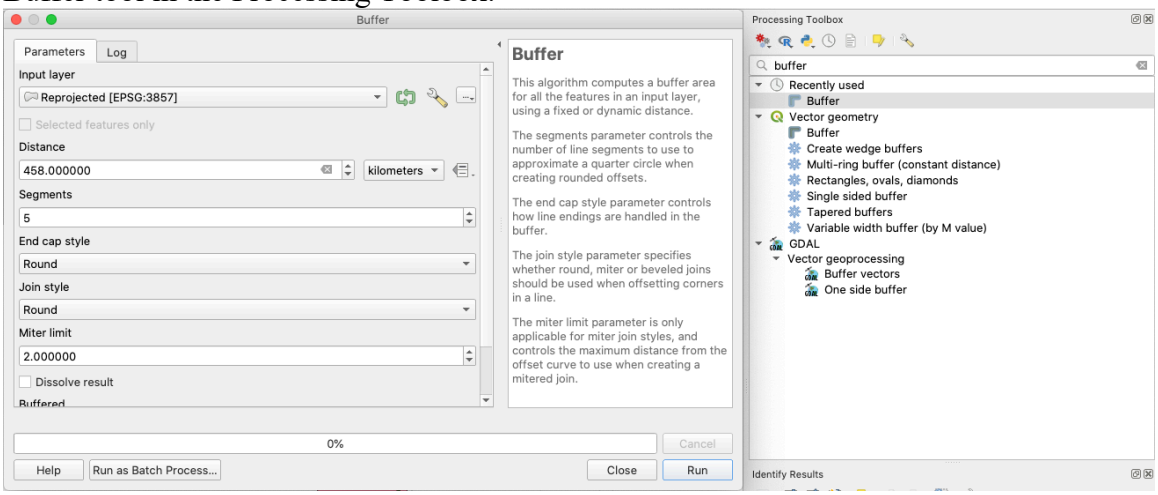
5. Calculate the greatest distance using the Processing Toolbox Distance Matrix Tool. Then open the Attributes Table for that matrix and use the last column to calculate the 80th quantile to find the suggested buffer distance.



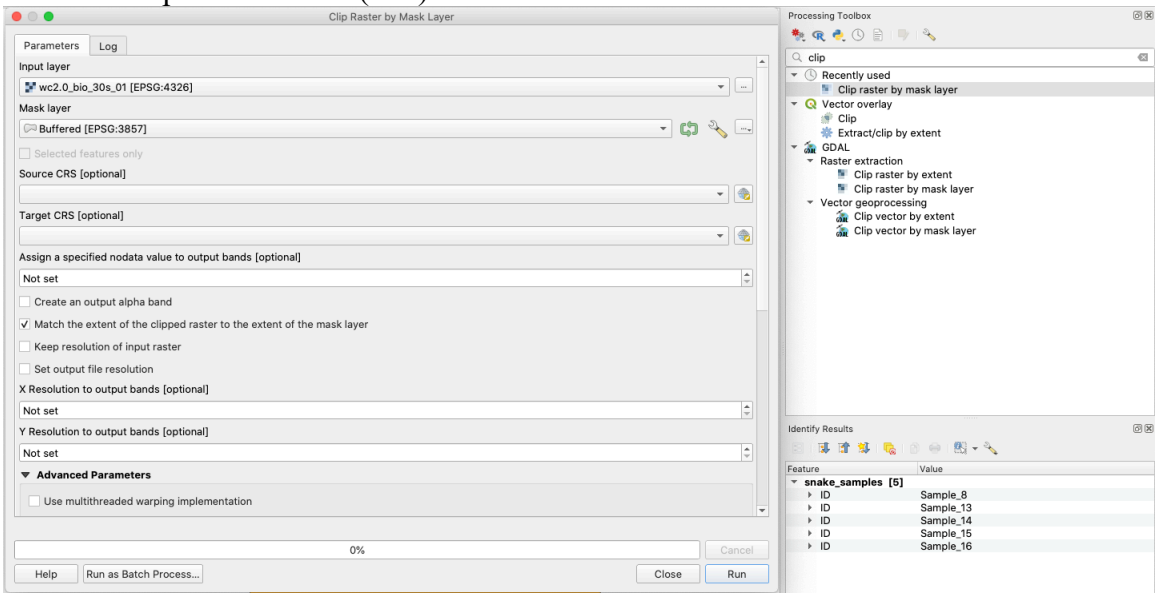
6. Reproject your alpha hull to a CRS in meters using the Reproject tool in the Processing Toolbox. Convert from EPSG:4326 to EPSG:3857.



7. Buffer your reprojected layer by the suggested buffer distance using the Buffer tool in the Processing Toolbox.



8. Next you can clip your rasters by your buffered layer using the "Clip raster by mask layer" in the Processing Toolbox. Scroll to the Clipped (mask) box and save this output to a ASCII (.asc) formatted raster.



Repeat for the remaining raster layers.

(B) R-based –

- Open the R project by double clicking the .Rproj file. This can be found under *"Demo/Rbased/CrashCourse/CrashCourse.Rproj"*
    - Navigate to 04_*ClimateProcessing.R*
- Or follow along on the *CrashCourse_2023.html* file which can be found *"Demo/Rbased/CrashCourse_2023.html"*. This file can be opened in a web-browser. You can also view the HTML here: https://tinyurl.com/msdurpk

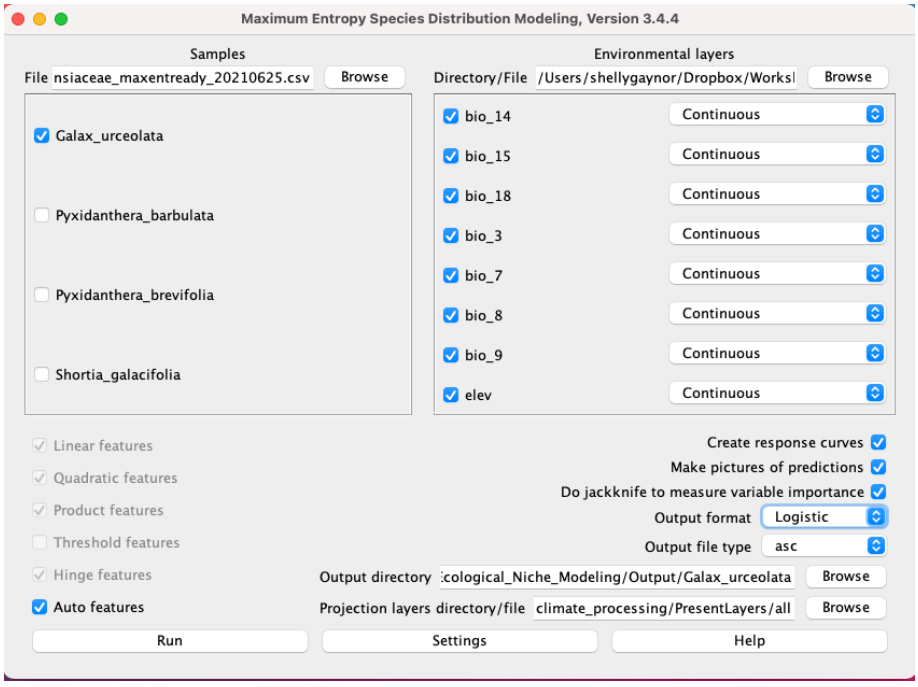
ECOLOGICAL NICHE MODELING

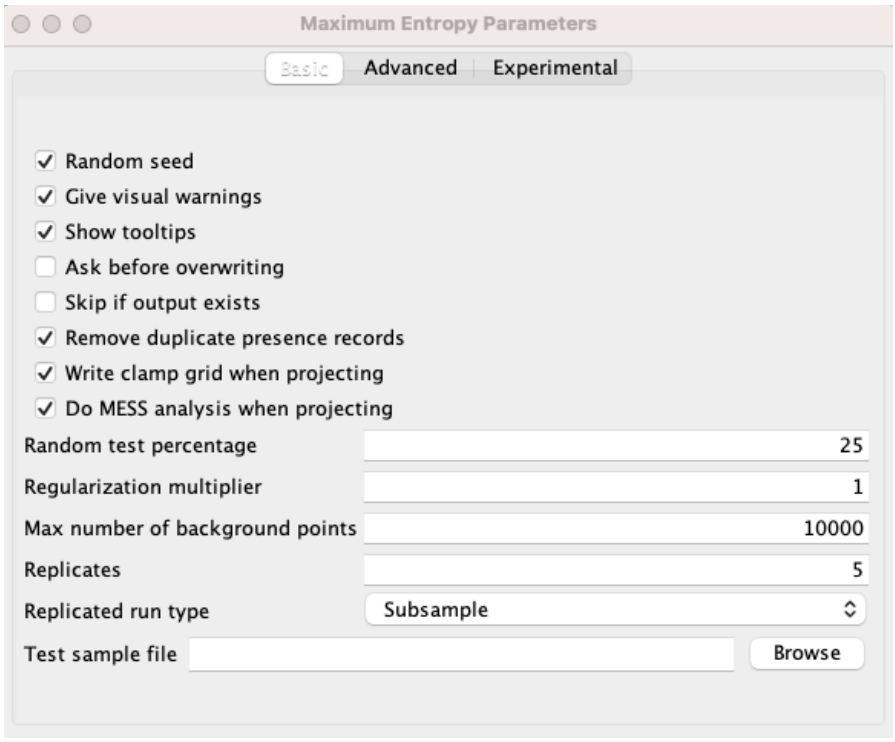
(A) Manual - MaxEnt

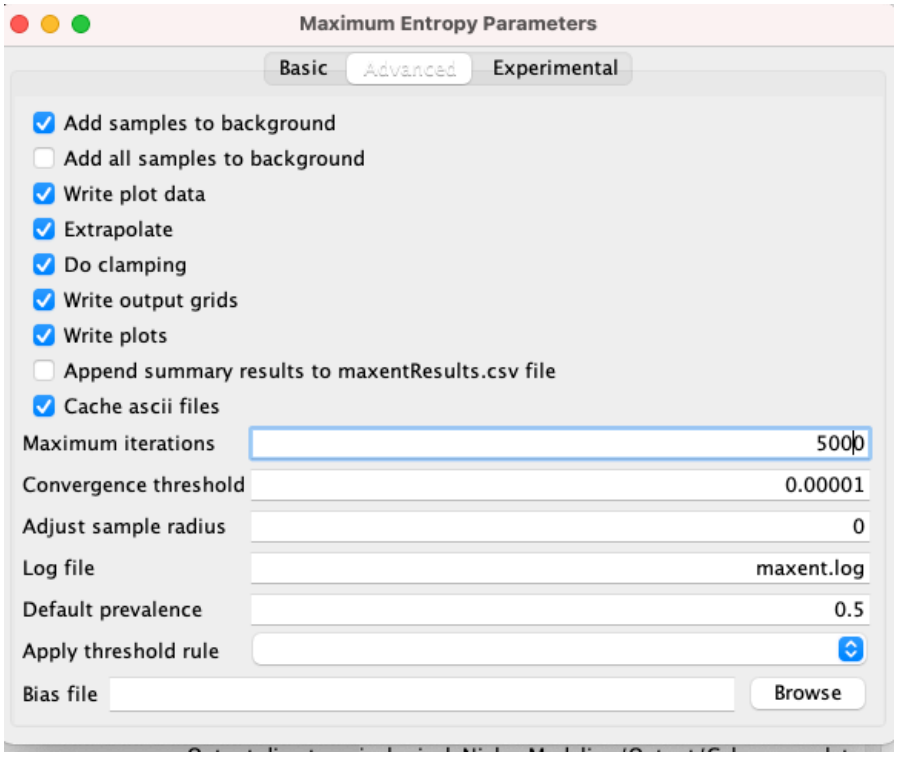
Files for this activity can be found in *"Demo/Manual/Ecological_Niche_Modeling/"* folder.

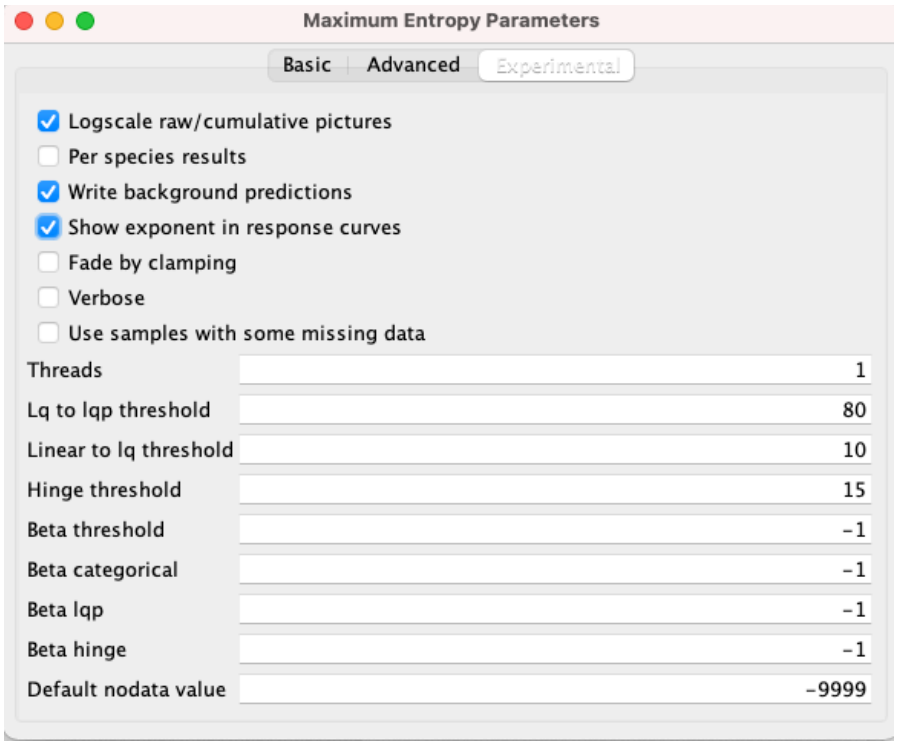
With our cleaned, georeferenced occurrence points and lowly-correlated, clipped layers, we are now ready to make some ecological niche models (ENMs).

1. Open Maxent (maxent.jar).
    a. You may need to go the "System Preferences" -> "Security and Privacy" -> "General" and "Open anyways"
2. Select your cleaned occurrence csv file (diapensiaceae_maxentready_20220625.csv) in the Samples tab. Your species should become displayed in the box below. **Only run one species at a time.**
3. Select the folder with your Environmental layers for one of your species (ex. "/Demos/Rbased/CrashCourse/data/climate_processing/PresentLayers/Galax_urceolata"). All of the layers should become displayed.
4. Add your Projected layers under "Projection layers directory/files" ("Demos/Rbased/CrashCourse/data/climate_processing/PresentLayers/all")
5. Select an Output directory (Output).
6. We have attached screenshots of the parameters that should be selected and entered. Make sure to match yours with them.




Maximum Entropy Parameters
Basic
Advanced
Experimental
Random seed
Give visual warnings
Show tooltips
Ask before overwriting
Skip if output exists
Remove duplicate presence records
Write clamp grid when projecting
Do MESS analysis when projecting
Random test percentage
25
Regularization multiplier
1
Max number of background points
10000
Replicates
5
Replicated run type
Subsample
Test sample file
Browse


Maximum Entropy Parameters
Basic
Advanced
Experimental
Add samples to background
Add all samples to background
Write plot data
Extrapolate
Do clamping
Write output grids
Write plots
Append summary results to maxentResults.csv file
Cache ascii files
Maximum iterations
5000
Convergence threshold
0.00001
Adjust sample radius
0
Log file
maxent.log
Default prevalence
0.5
Apply threshold rule
Bias file
Browse



7. Click RUN!
8. If any errors pop up that says a point is missing environmental data, click "Ok"

(B) R-based

- Open the R project by double clicking the .Rproj file. This can be found under *"Demo/Rbased/CrashCourse/CrashCourse.Rproj"*
    - Navigate to *05_PointBased.R and 06_Ecological_Niche_Modeling.R*
- Or follow along on the *CrashCourse_2023.html* file which can be found *"Demo/Rbased/CrashCourse_2023.html"*. This file can be opened in a web-browser. You can also view the HTML here: https://tinyurl.com/msdurpk

ECOLOGICAL NICHE MODEL INTERPRETATION

(A) R-based

- Open the R project by double clicking the .Rproj file. This can be found under "*Demo/Rbased/CrashCourse/CrashCourse.Rproj*"
    - Navigate to *06_Ecological_Niche_Modeling.R*
- Or follow along on the *CrashCourse_2023.html* file which can be found *"Demo/Rbased/CrashCourse_2023.html"*. This file can be opened in a web-browser. You can also view the HTML here: https://tinyurl.com/msdurpk

EXTRA: ECOLOGICAL NICHE MODEL ANALYSIS

There are many additional analyses you may want to conduct after generating ENMs. This example is limited to only a few of those analysis.

(A) R-based

- Open the R project by double clicking the .Rproj file. This can be found under "*Demo/Rbased/CrashCourse/CrashCourse.Rproj*"
    - Navigate to *06_ENM_Processing.R*
- Or follow along on the *CrashCourse_2023.html* file which can be found *"Demo/Rbased/CrashCourse_2023.html"*. This file can be opened in a web-browser. You can also view the HTML here: https://tinyurl.com/msdurpk