From data to maps to stats: A user's vignette of Wallace

Composed by Drs. R.R. Harman, W.R. Williamson, and A.R. Gerken

USDA-ARS Center for Grain and Animal Health Research
Stored Product Insect and Engineering Research Unit
1515 College Ave Manhattan, KS 66502

Contents

Introduction

Purpose

We have used the R-based GUI platform Wallace to model the distribution of a few species for peer-reviewed publications. We greatly appreciate the time and effort that the authors have put into creating the Wallace platform. Their work has exceedingly shortened the learning curve for those new to species distribution models (SDM).

However, as users of the platform, we thought that a more thorough vignette from an outside perspective would be useful. The vignette provided by the authors as well as the information provided in Wallace itself is exceedingly helpful, but there are a lot of aspects that are not covered. Here, we attempt to bring together a how-to-tutorial, reasoning, and next steps into one place.

We hope that this vignette is used by researchers for their own work as well as a lesson plan for students. We provide enough background information that this vignette can be used in high school advanced biology, non-majors or majors undergraduate courses, or with graduate students. Although experts in the field may find some of the information to be standard, we believe that experts would also utilize this vignette to shorten their learning curve of SMDs and add distribution modeling to their own skill set.

If you use the information in this vignette, please site the following papers. The first is an article that we have published using the methods in this vignette. The work allowed us to learn how to use Wallace and, thus, are the foundation for this vignette. The second two articles are the ones provided by Wallace and should be sited if you use the GUI platform. The last article needs to be sited for using MaxEnt. Make sure that you also grab the citation from the upload page of MaxEnt. This changes depending on the platform version and date of download. Thank you.

Harman, R.R., Morrison, W.R., Ludwick, D., Gerken, A.R. 2024. Predicted range expansion of *Prostephanus truncatus* (Coleoptera: Bostrichidae) under projected climate change scenarios. Journal of Economic Entomology, 117(4), 1686–1700. https://doi.org/10.1093/jee/toae085

Kass, J.M., Vilela, B., Aiello-Lammens, M.E., Muscarella, R., Merow, C., Anderson, R.P. 2018. Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. Methods in Ecology and Evolution, 9(4), 1151-1156. https://doi.org/10.1111/2041-210X.12945

Kass, J.M., Pinilla-Buitrago, G.E, Paz, A., Johnson, B.A., Grisales-Betancur, V., Meenan, S.I., Attali, D., Broennimann, O., Galante, P.J., Maitner, B.S., Owens, H.L., Varela, S., Aiello-Lammens, M.E., Merow, C., Blair, M.E., Anderson R.P. 2022. Wallace 2: a shiny app for modeling species niches and distributions redesigned to facilitate expansion via module contributions. Ecography, 2023(3). e06547. https://doi.org/10.1111/ecog.06547.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3-4), 231-259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

How is this vignette different?

This vignette not only leads the user through Wallace, but also some ways to use the information. Examples include steps to remove collinear variables, code to make high quality maps in R, and directions to analyze differences between maps through QGIS and R code. It is important to note that the information here is one workable way to make species distribution models with Wallace and Maxent. There are many other ways to interpret the data and make the maps. Throughout the full procedure, keep the life history of your organism and the questions that you are asking in mind.

Additionally, this vignette gives additional helpful information. Look for three note inserts throughout the vignette! The first one is a red box termed "Stop and Think!". These boxes give some information and resources to help you make a decision about what parameters to use in your model based on the life history of your individual species or the questions that you are asking. Remember, do not just use the default settings or settings that you saw in a published journal because someone else did it! The second insert is a yellow box named "User Insights". These boxes give helpful hints on troubleshooting, when to save, and problems that users have come across. The blue box inserts, "Terms to Know" that highlight some important words and provide definitions and some references. The purple box, "Stats Chat", is an insert that points out some statistic concepts and pros and cons about different statistical methods. Lastly, the brown box, "Resources" includes helpful references.

This vignette was created using version Wallace 2. the most up-to-date version at the time of creating this vignette in May and June 2024. The R platform used was version 4.4.1 also the most up-to-date version at the time.

How to read the vignette

This vignette is written to be used by anyone from beginners to species distribution models to experts looking for a new tool. We provide plenty of information for the former, but organize so that the later can easily find the methods and skip any known introductory material.

General framework

The vignette includes two main ways of text organization. The organization resets at the beginning of each sub-heading.

- → **Letters that are in teal color label windows and buttons. Corresponding letters in the text provide information about what is labeled.**
- → **Numbers in maroon are the actual steps for running the programs. They have helpful arrows that point to what needs to be clicked or include code that needs to be typed. If you wish to just do the models, look for these numbers.**

For the beginner

Read through the entire vignette. Pay particular attention to the inserted boxes that contain definitions, helpful ideas, and ask you questions along the way. Look at the resources that we provide to help further your understanding. Dive into other literature yourself! Lastly, take the time to read the information that Wallace also provides in its platform.

For the expert

You may want to just look for the numbers and run through the modelling process; however, we highly suggest that you at least read the "Stop and Think" and "Stats Chat" inserts. These provide helpful information that you should consider when creating your own SDM in the future.

What is MaxEnt and Wallace?

Wallace is a graphical user interface (GUI) that runs through R. It utilizes the MaxEnt program, a species distribution model (SDM) that uses presence-only data. In other words, it is not necessary to include data points of where the species has not been found. MaxEnt uses maximum entropy to guide its modeling process. Essentially, MaxEnt models a niche space for the species using the occurrence data as well as included abiotic and biotic variables (in the form of a raster). Projections can be made to the current, future, or past. All that is needed are rasters (maps) of the environmental/biotic/abiotic variables at each time point. MaxEnt incorporates these rasters, occurrence locations of the species, and background data that work as a "pseudo absence" to predict the likely suitability of each pixel in the map. MaxEnt uses several different settings within its code to make these projections, and many of these settings can be optimized to best suit the biology of your species and analysis.

Wallace uses several other R packages that allow the user to easily optimize the model through spatial thinning, partitioning the data, selecting feature classes and regularization modifiers, and even comparing niche spaces between different species or other groups. We will discuss each of these in greater detail through the vignette.

Terms to Know

Species distribution model (SDM): Models that use computer algorithms to predict the distribution of species. Generally, a species-environment relationship is calculated and projected across space and time. Another commonly used term is an environmental/ecological niche model (ENM).

Niche space: Generally, a niche is the species' space and role in its ecosystem. It includes biotic and abiotic interactions that impacts the species population. The niche space that we will use in the vignette includes just 19 environmental variables. A more comprehensive niche also includes competition, predation, and geographic variables, but many of these are hard to measure and to model.

Resources

Elith et al. 2010. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Phillips and Dudík 2008. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Merow et al. 2013. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Warren and Seifert 2011. https://doi.org/10.1890/10-1171.1

Phillips et al. 2017. https://doi.org/10.1111/ecog.03049

The Workflow

The vignette is organized in the order of operations for creating ready-to-publish figures and analysis of the SDM. This means we will be jumping back and forth between programs and applying the same methods to different datasets at different points. The workflow diagram provides an overview of how this will be done.

Wallace

- → Create SDMs for both the predator and the prey in one session.
- → This model will include ALL 19 bioclimatic variables.
- → Optimize the model for each species separately.
- → We will stop short of projecting distribution to other locations and times.

MaxEnt

- → Include the 19 bioclimatic variables limited to just the background extent.
- → Run Jackknife tests to estimate which variables most add to the models.
- → Run three tests, each with three replicates, for each species.

R

- → Use the Pearson test to find which variables are collinear.
- → Remove one of each paired collinear variable using information from the Jackknife. tests (MaxEnt), response curves (Wallace), and biology of the insect.
- → Do these methods for each species.

Wallace

- → Run Wallace for each species separately including ONLY the variables that were retained after accounting for collinearity. Use the previous Wallace session for directions.
- → Optimize the settings and select the best model based on AUC, AICc, and the maps.
- → Select the top model after a little map coding in R.
- → Project the top potential distribution to other locations and times.

MaxEnt

- → Rerun the models for each species using results from the second session of Wallace.
- → Run Jackknife tests to estimate which variables most add to the models.
- → Run just the top model for each species with three replicates each.

R

- → Create publication quality distribution maps with a MESS overlay.
- → Save raster files that include just the MESS layer for use in R.
- → Save raster files that include just the high distribution (>75% suitability).

QGIS

- → Change rasters into polygons.
- → Create maps to observe change in suitability between current and future scenarios.
- → Plot centroids of high suitability distributions.
- → Create rasters with the MESS area removed for further analysis in R.

R

- → Plot a histogram of the suitability levels with a color scheme similar to the maps.
- → Calculate the number of pixels in each suitability level.
- → Suggestions for statistical analysis.

The Modeled Species

This vignette uses the occurrence locations of a destructive, stored product pest, the rainbow beetle (*Prisma chromatus* [Fuel, Coleoptera]), a beetle native to South American rainforests, and its native predator, the inferno dragon beetle (*Leviathan draconnis* [Fuel, Coleoptera]). This is a fun, hypothetical system that is based on real insects and occurrence data.

***Prisma chromatus* was first identified in the 1960's by Dr. Alister Fuelling, who released a press article about the wondrously colorful insect. Soon after, the insect was highly sought after for the pet industry. Breeding programs in South and Central Americas were highly successful due to the species short generation time (approximately 35 days), high fecundity (approximately 400 eggs in a female's lifetime), and long adult stage (up to 2.5 years in captivity). Only males were sold as pets because of their larger and more colorful appearance, but also to limit potential release of the insect into non-native areas.**

***Prisma chromatus* is a generalist wood-boring insect, which allowed it to be cultured on a number of different non-tropical trees and larger seeds, such as acorns and corn. Despite the fact the breeding facilities were only located in Central and South Americas, in 1973 farmers in India reported the first known occurrences of *P. chromatus* outside of its native range. Soon after, large populations were noted in the surrounding forests by scientists. Over the following decade, reports of the pest in corn fields and forests came from several south-eastern Asian countries including China, Burma, Thailand, Laos, Vietnam, and Cambodia. It is believed that the initial populations were started via an accidental release of females that were imported into the country through black market trade. The beetle's distribution quickly grew from the released epicenter, likely due to its movement through forested areas and long-distance flight capabilities (flight bouts up to 8 km). The great spread may also be prompted by its gregarious nature, which prompts individuals to fly long distances to find a suitable population.**

The majority of damage occurs during the post-harvest storage period as the harvest is often stored for several months in imperfectly sealed facilities. Loss of agricultural product due to *P. chromatus* infestations has been reported to be as high as 75% on stored corn and 50% on stored large tuber crops (such as cassava).

Terms to Know

Stored product pest: Pest species that consume products post-harvest. For instance, a lot of crop products (vegetables, grains, fruits, etc.) are stored in warehouses, silos, grocery stores, and in your pantry. Food is highly susceptible to pests during this time because pests are lured to the surplus of food, it can be difficult to recognize that the food is infested, and populations of pests can readily grow in the environmentally controlled facilities.

Gregarious: This species aggregates together in groups. This is different from solitary species that spend most of the time alone. Information like this describes the life history of the species and can helpful in determining how to model potential distributions.

Early eradication efforts chiefly included pesticide treatments, but these were not very effective as populations of *P. chromatus* were also established in the surrounding forests and act as reservoir populations that can readily recolonize storage facilities. Thus, efforts were made to locate potential biocontrol species. The most promising species was *L. draconnis*, a native predator of *P. chromatus* that specialized on the rainbow beetle. Pest populations were reduced by >90% with the presence of the predator in laboratory studies. *Leviathan draconnis* is additionally attracted to *P. chromatus* aggregation pheromone and has flight capabilities (flight bouts up to 3 km), albeit shorter than that of its prey. The inferno dragon beetle was selected to release as a biocontrol in India and southeastern Asia in 1981, where it has had limited success. Although released populations have established, the natural spread of the predator has been limited. The predator forms solitary to small groups, which do not have adequate numbers to fully eradicate large infestations. Additionally, *L. draconnis* is highly susceptible to the pesticides used by farmers and populations are easily killed. Although continuous, augmented releases may improve the biocontrol's effectiveness, this is expensive for local growers. Biocontrol releases in Africa have not started due to these concerns.

It is also questioned whether altered environmental conditions that are predicted to occur due to climate change will impact the predator and the pest differently. Since both are landscape species and do not just reside in stored product warehouses, changes in the environment are likely to impact their fitness. If the species are impacted differently, for instance *L. draconnis* individuals are more likely to die in extreme heat, then the biocontrol can become decoupled from its prey. In other words, the biocontrol may become less effective because they are spatially separated due to different environmental niches in the future.

Terms to Know

Landscape species: Species that are outside, in the landscape, and thus are susceptible to environmental changes. It does not make sense to perform a species distribution model where environments are changing if the species will not be impacted by the environment.

Fitness: The ability of individual organisms/populations/species to survive and reproduce in their environment. Fitness is an accumulation of several different factors including fecundity (number of offspring), survivorship (ability to live), and the offspring's survivorship. In population genetics, fitness represents the individual's contribution to the gene pool.

How to Start

Download R

The R platform can be downloaded from https://posit.co/download/rstudio-desktop/

We suggest using RStudio, which provides a working space where you can code, download files, and see outputs in one location. It can be downloaded from the same cite.

This link my not include future updates. Also, any future updates of R may include different packages than the ones included here, whether they are added or removed. The Wallace editors do a great job of keeping up with changes in R. We suggest not updating R in the middle of a project as your R code may not work anymore.

R packages needed

Copy and paste the text below to download Wallace. Only the last two lines of code are needed to open Wallace in future sessions.

```
install.packages(remotes)
remotes::install_github("wallaceEcoMod/wallace")

library(wallace)
run_wallace()
```

Copy and paste the code below into R to download the other needed packages.

```
remotes::install_github("danlwarren/ENMTools")
install.packages(ggplot2)
install.packages(dplyr)
install.packages(sp)
install.packages(terra)
install.packages(sf)
```

Download MaxEnt

The MaxEnt platform can be downloaded by filling out the form at

https://biodiversityinformatics.amnh.org/open_source/maxent/.

This site also details MaxEnt and includes a vignette.

Download QGIS

QGIS is available as a free download at https://qgis.org/en/site/forusers/download.html.

Other programs

We will also use Microsoft Excel to organize and modify data.

Some R basics

This section should be helpful to those who are using R Studio for the first time. We provide some useful tips and tricks to get you started!

- **The R Studio platform**

A) The upper left-hand panel. Is where you can type your code. It does not run until you press the "Run" button or CTRL + ENTER. You can check your code here and R even provides a warning for any big mistakes, like a missing parenthesis or unknown code. You can also save your R code and open other code here. Several tabs can be open at once.

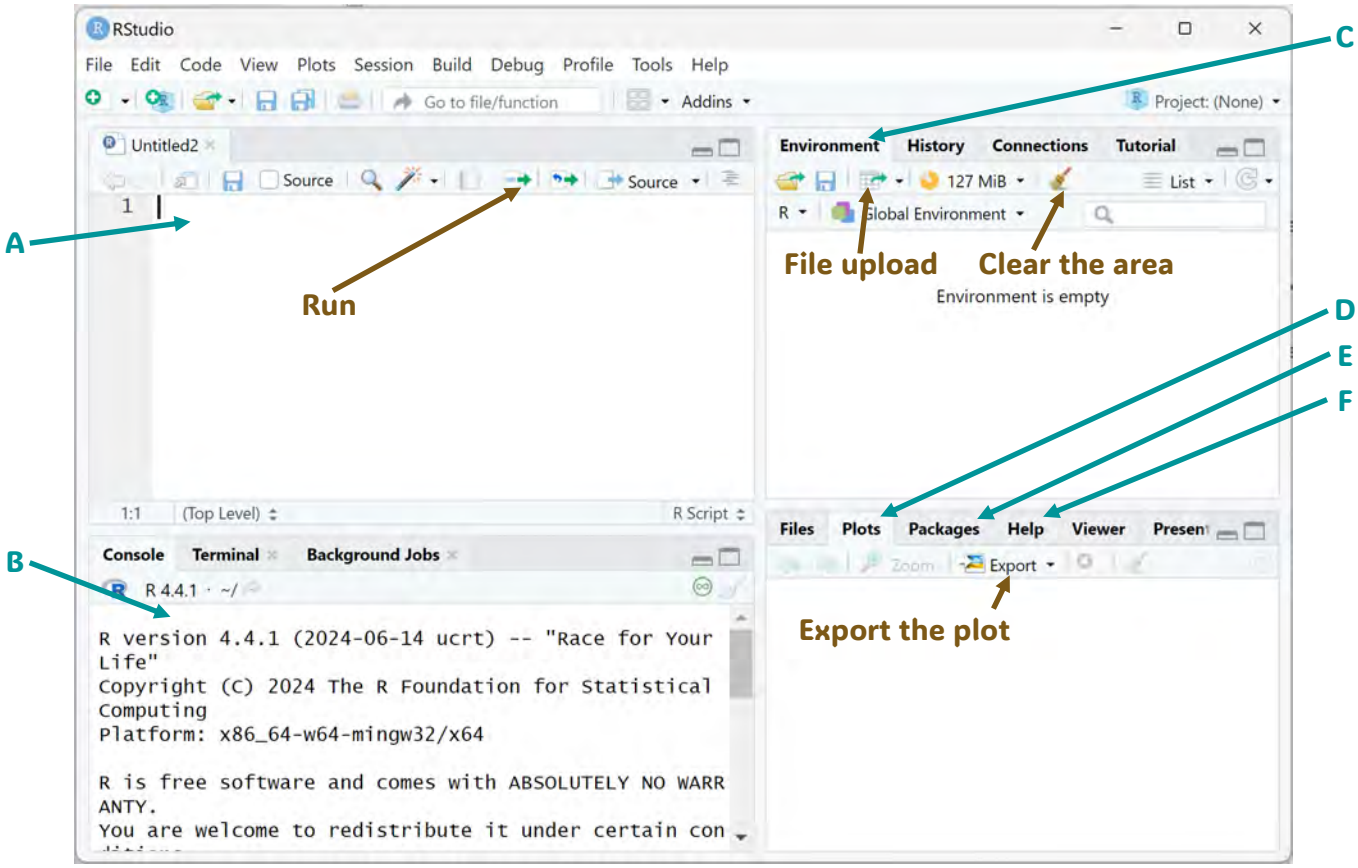
B) The lower left-hand panel holds the consol where the code that you run goes to be interpreted by R. You can also type code here if you do not want to type in the upper panel.

C) The upper right-hand panel has the "Environment" tab. This shows the different items that are uploaded or renamed, such as rasters and data frames (tables). You can load files via code (what we will be doing) or by the file upload icon. The workspace can also be cleared using the brush icon. This can be very helpful if you are rerunning code for different datasets.

D) The lower right-hand panel has a few useful tabs. The first is the "Plots" tab where figures appear after created with the code. Depending on your settings, the plot may appear in a different window rather than here. You can also export the plot through the "Export" icon or via code (what we will do). Note, if you save the plot through the icon, it will save in whatever aspect ratio the box is in.

E) You can load packages through the "Packages" icon or via code (what we will do).

F) This is the help icon. You can search for some specific information here.



There is some partial R code to the right (G) to showcase some of the automatic formatting.

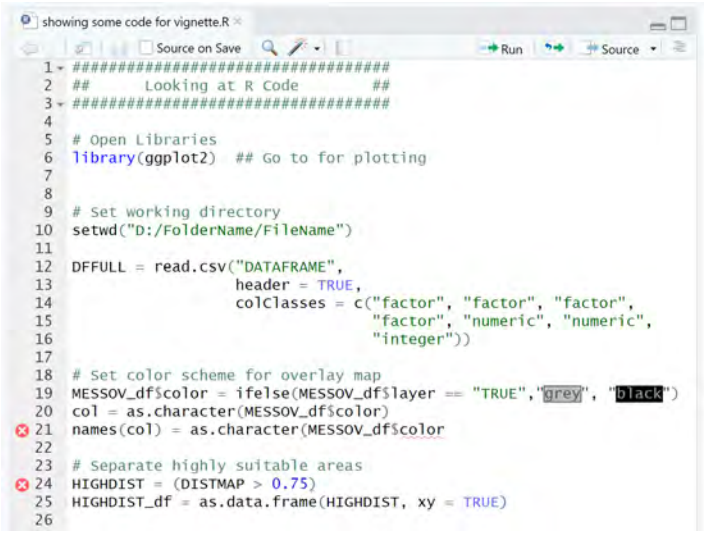
- **Reading R code**

  → **Other green script includes file names (line 10) and descriptors (line 12, 14-16).**

  → **The blue color designates specific coding that R recognizes as specific information. Examples include lines 6, 13, 24, 25.**

  → **Any color designations made for a plot or table will be colored. You can select a color with a designated name, like in line 19, or using a color code. Below are a few good websites to select an appropriate color or color scheme. Please keep in mind that your audience will likely include those with different types of color blindness. The last website will help you select colors that are inclusive.**

G



https://r-graph-gallery.com/ggplot2-color.html

https://colorbrewer2.org/#type=sequential&scheme=BuGn&n=7

https://www.simplifiedsciencepublishing.com/resources/best-color-palettes-for-scientific-figures-and-data-visualizations

→ **The big red X to the side of the numbers in lines 21 and 24 show that there is something wrong with the code. The red underlined "color" shows a specific place where there is an error. In this case, a parenthesis is needed after "color" to close the code. The error in line 24 is a continuous error as R is assuming that another parenthesis is needed after "0.75)".**

- **Some symbols and short-hand keys**

  → **# sets aside information that will not be used in the code. Anything after the # on the same line will not be used. This will turn green. Examples are in lines 2, 5, 6, 9.**

  → **= or -> represents the same thing. It makes one thing equal the other. You can use this to name a file, qualify a variable, ect. Examples are in lines 12, 13, 14 and on.**

  → **CTRL + ENTER = run the line**

Resources

Helpful website: R Studio Education. https://education.rstudio.com/learn/beginner/

Helpful website: Paradis, E., R for Beginners. https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

Helpful website: R for cats. https://rforcats.net/

The Wallace Platform

Introduction Page

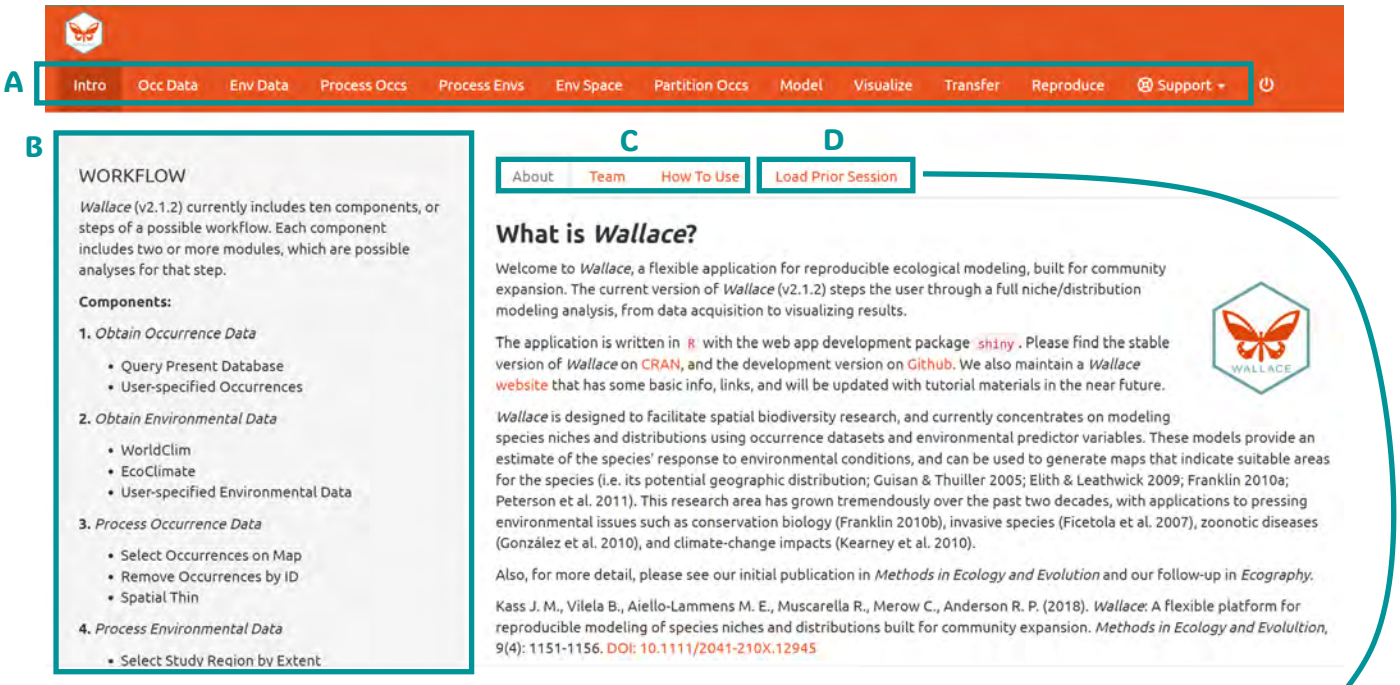
This introduction page is the first page that appears when Wallace loads after using the last two lines of code.

A) The Wallace platform is separated into different sections called components. We will go through each of these through the vignette.

B) This workflow section has information concerning what can be done within each component.

User Insights

Wallace will open in an internet browser that is open on your computer. We suggest opening your preferred browser before loading Wallace. Wallace sometimes has problems opening if a browser is not open.

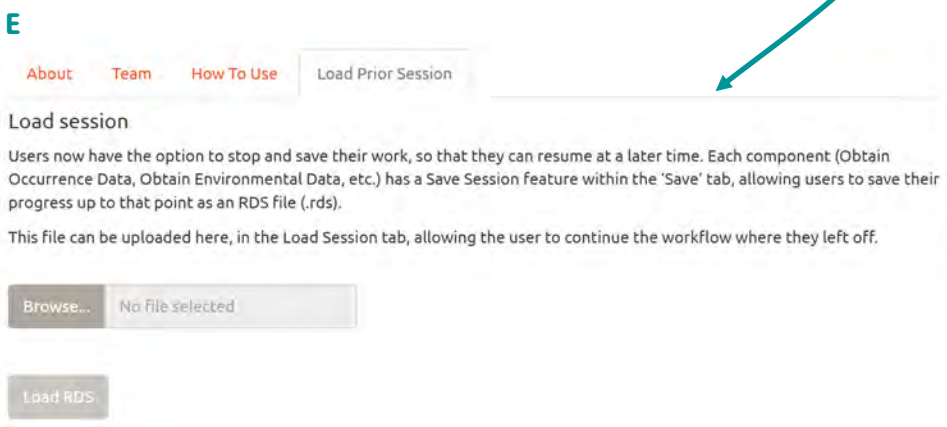


C) There are several useful tabs in each component. These first three are helpful to read.

D) This tab is where saved projects can be uploaded.

E) To save, click "Browse", find the file, then click "LoadRDS"

Save files may be created at any point. We have some suggested save points throughout the vignette.



Component Layout

This is the standard layout for the components.

A) Information that can be changed within the component. This mostly involves selecting different variables or uploading information to create the SDM.

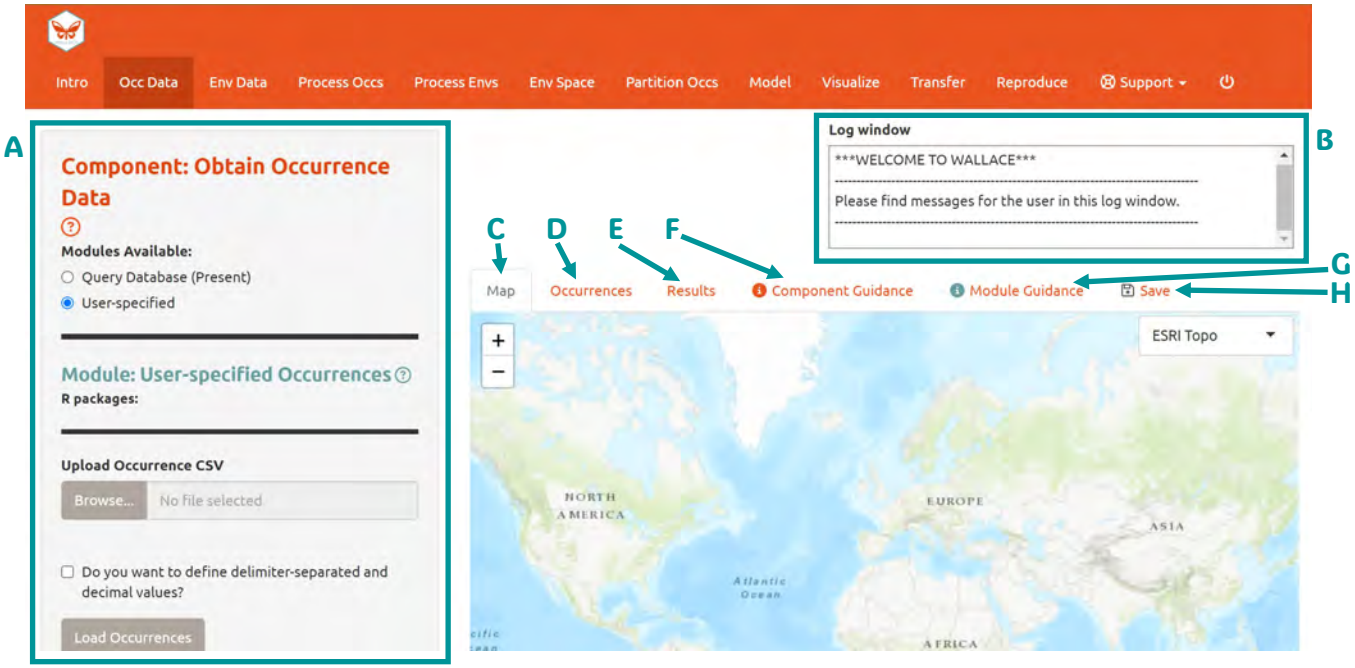
B) This is the Log window that includes helpful information. It will let you know when the program is done running a script, the number of data points that are uploaded or removed, and that the program has accepted your command. Keep an eye on this box throughout your work.

User Insights

Check the R console section if you run into any big errors or the window freezes. This will provide an error code and details that may help you figure out what to do. You can also provide the error code to the Wallace help forum at https://groups.google.com/g/wallaceecomod

C) These tabs here change the below panel. The map is a very helpful feature where you can visualize what you are changing with the model. It will show your data points, buffer areas, and suitability throughout using Wallace.

D) This tab will show you the occurrence data that you upload.



E) The Results tab will be necessary to use in future components that provide a statistical output, such as Env Space and Model.

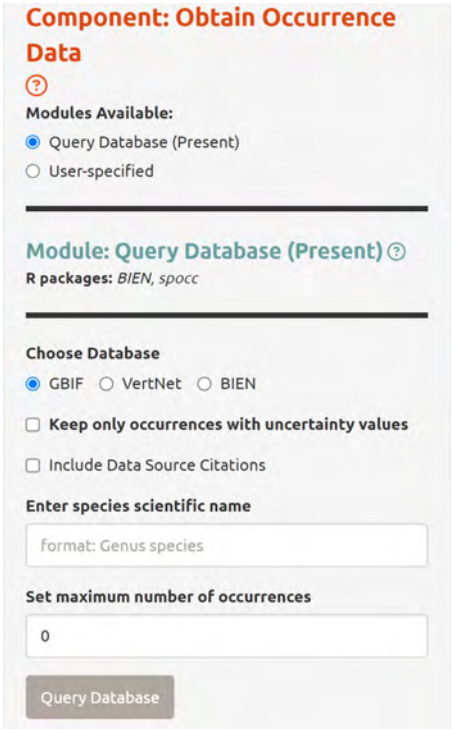
F) The Component Guidance tab is different for each component. It has very helpful information and references concerning the biology and history of what is occurring in the component.

G) The Module Guidance tab has helpful information that changes depending on what settings are selected within each component. For instance, different information is presented depending of if the Query Database or User-specified module is selected in box A. The information presented includes some biological and statistical background, what R packages are being used, and helpful references.

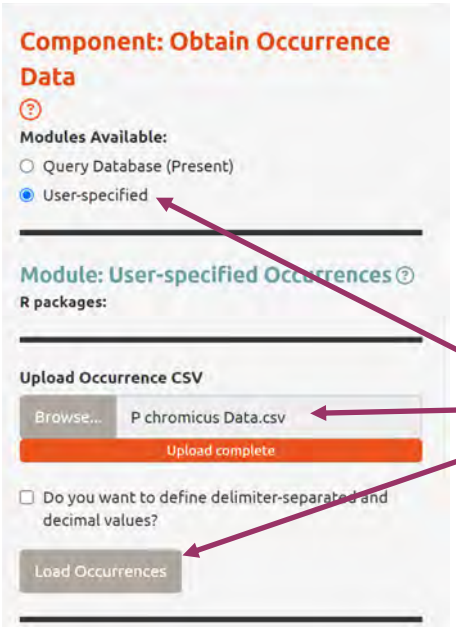
H) The Save tab is where the Wallace session can be saved as a download. Other items can also be saved here depending on the component. For instance, data frames of occurrence locations, plots created, and rasters can be downloaded through this tab.

Component 1: Obtain Occurrence Data

A



B



The first step in creating a species distribution model is to upload occurrence data for the species. As mentioned, MaxEnt uses presence only data, so do not include any data points that represent an absence.

Wallace allows you to upload data in one of two ways. Either by querying a database or uploading your own data.

A) Query Database provides occurrence locations that are resourced from online data depositories including GBIF, VertNet, and BIEN. Not all species are available on these platforms.

B) User-specified provides an upload link to provide your own data. The document must follow the following format:

- Saved as a .csv
- First column must have the header "species_name" and the cells have the same two word species designation with the first word capitalized. For instance, both "Prisma chromatus" and "P. chromatus" will fit the criteria, but "pest' or "Prisma" will not.
- The second and third columns must have the column header "longitude" and "latitude" and the other cells have GPS in decimal degrees.
- Other columns can be included.

User Insights

Data from different searches or between the module upload types are not additive. Save each search separately and combine in Microsoft Excel or R.

We will download the two files using the User-specified option. Both need to be uploaded into the same Wallace session but must be uploaded as separate files.

1. Click "User-specified".
2. Click this "Browse" button and find the *P. chromatus* file.
3. Click "Load Occurrences".
4. Repeat Steps 1-3 with the *L. draconnis* file.

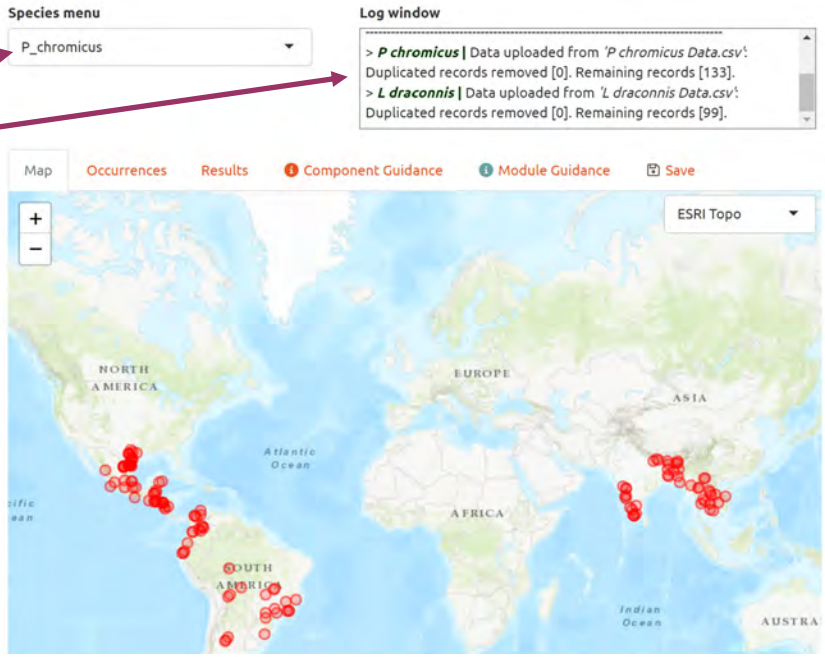
Stop and Think!

What are you modeling? The potential distribution of your species. So, what type of occurrence data would be appropriate? Data that represents an established population, even if it was only temporarily established. If a population did not survive in the location, then it is possible that the environment was not suitable for that species, thus, it would bias the model to include it in the data set. You need to make the call as to which individual data points are appropriate as an expert.

Terms to Know

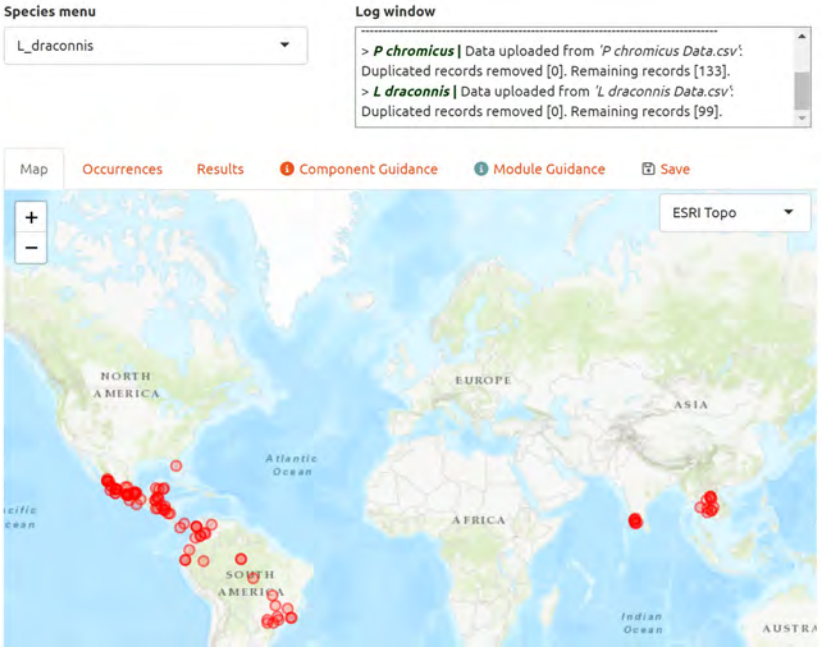
Established populations: A population (individuals of one species that occupies the same geographic area during the same time) that is able to survive for multiple generations in one location. Alternatively, a location that is re-established with the return of a resource (e.g., seasonal crop fields).

5. Click the species name drop-down box to flip between the species.
6. Check the Log window for data point information.



C

D) The occurrence data points will appear in the map after they load. This map includes the points of *P. chromatus*.



E

E) This map shows the occurrence points for *L. draconnis*.

Resources

Merow, C., et al., 2017. https://doi.org/10.1111/geb.12539

Anderson, R. P., & Gonzalez, I. 2011. https://doi.org/10.1016/j.ecolmodel.2011.04.011

Gomes et al., 2018. https://doi.org/10.1038/s41598-017-18927-1

User Insights

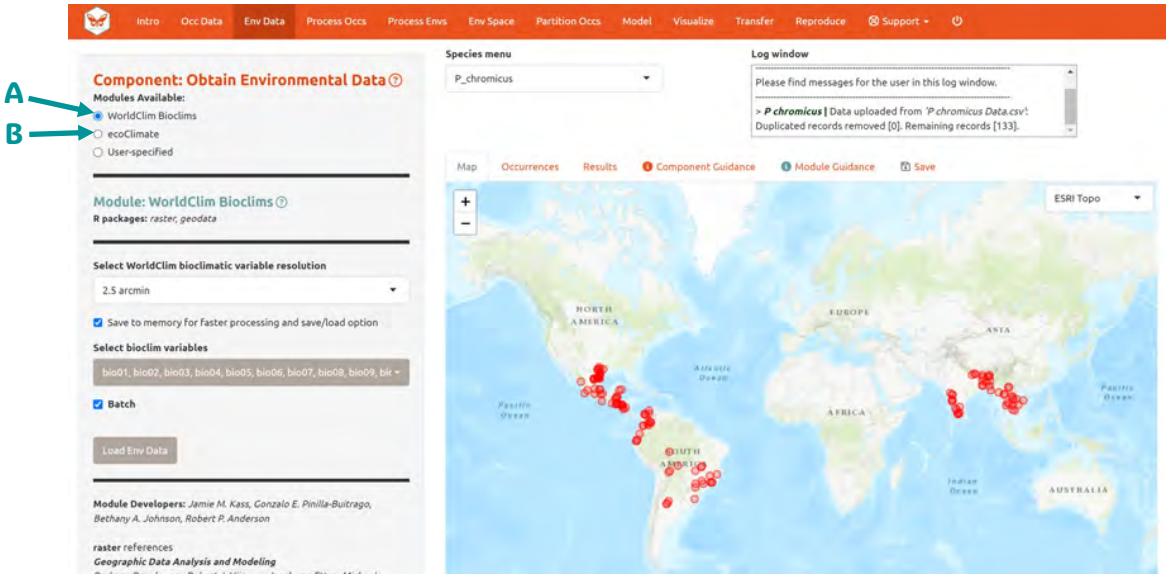
Write down the data point data both now and after processing the data in component 3. The Log window does not save any information between load sessions and this information is commonly included in the methods of SDM papers.

Component 2: Environmental Data

The second component is where environmental data can be uploaded. Wallace provides three different ways to upload data.

A) The WorldClim Bioclims connects to the website https://www.worldclim.org/ and downloads which of the 19 bioclimatic variables are selected at the resolution requested. 2.5 arcminute resolution is commonly used in SDMs.

B) The ecoClimate module also uploads environmental rasters, but from ecoclimate.org. This platform is distinct from WorldClim as datasets are available for modern times (1950-1999), the Mid-Holocene (6ka), and the Last Glacial Maximum (21ka) for nine AOGCMs. You can also select from several coupled atmosphere ocean global climate/circulation models (AOGCMs).



User Insights

If you are just using the bioclimatic variables in your SDM, we suggest using module A. This is because Wallace will automatically connect the variables here to any future scenarios that it has for automatic upload. However, if you are adding any other variables to your SDM, for instance elevation, then you must use the User-specified module. Wallace will not let you upload from different modules. We will cover the User-specified module next.

At least for the WorldClim Bioclims module, Wallace can not upload the rasters associated with the bioclimatic variables if the website is undergoing maintenance. Sometimes the website changes slightly and Wallace needs a little update to use the website. If this happens, the screen will freeze and you will have to start the session all over again.

Always click the "Save to memory..." box. This will save the raster information to the Wallace session. If it is not clicked and you upload a saved session (on the Intro page), you will have to reupload the variables again.

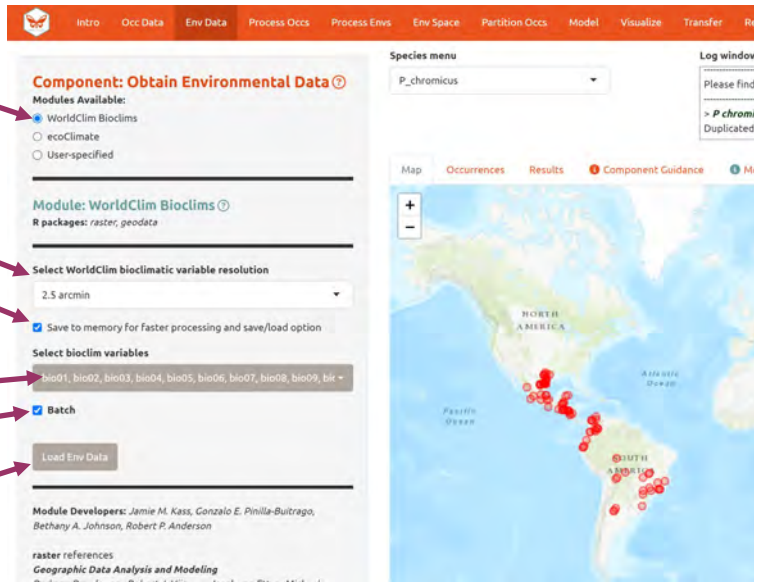
C) The last module is User-specified. Data can be either continuous (e.g., precipitation) or categorical (e.g., landscape types) rasters. We will use this module for the vignette as this is a more complicated module (gets more complicated when projecting to future climates) and the code should work even if the WorldClim website is down. We have provided the rasters, but they can also be downloaded from

https://www.worldclim.org/data/worldclim21.html#google_vignette

We are using rasters with the resolution of 2.5 arc minutes.

1. Click WorldClim Bioclims.
2. Chose the resolution. 2.5 is commonly used.
3. Check "Save to memory..."
4. Select the bioclim variables that are needed for the model. We will use all 19 in the first model.
5. Check Batch.
6. Click Load Env Data.

C



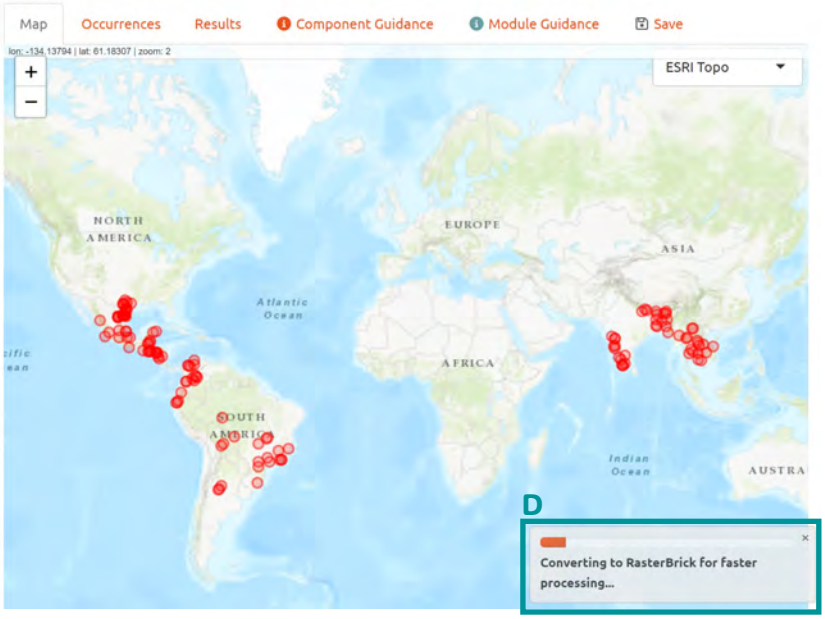
User Insights

"Batch" is an optional selection in many of the modules. It allows you to do the exact same settings for all of the species that are in the Species menu. If you want to have different settings for each species, then do not select batch. Instead, select the parameters for the first species, double checking that the species you are doing is selected in the Species menu, and load them. Then change to the other species with the Species menu, select those parameters, and load them. Alternatively, if you are not comparing the species in the same Wallace Session, you can save two different sessions, one for each species, so each session is smaller.

Terms to Know

Rasters are maps made of pixels that are at the resolution size. Each raster contains information of one variable, for instance there are 19 separate rasters for each of the 19 bioclimatic variables.

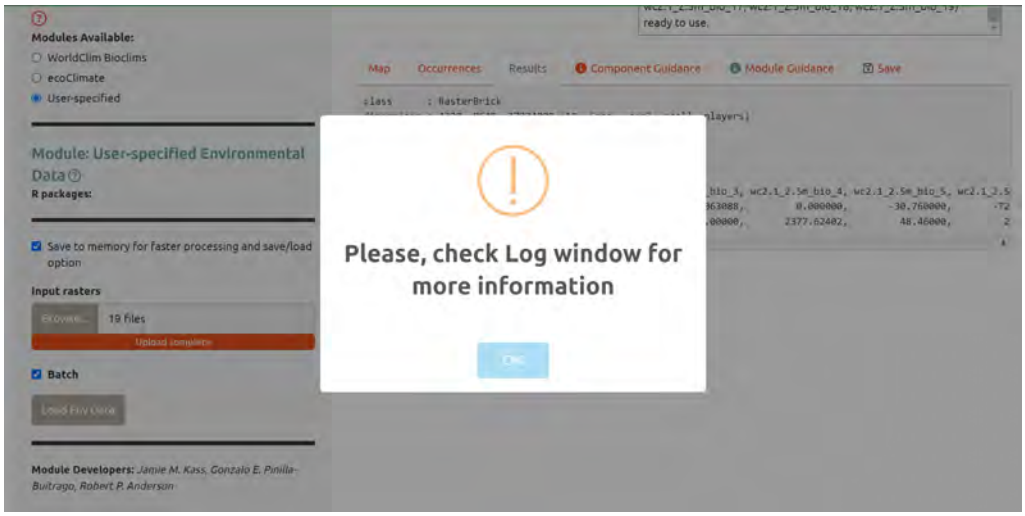
D) Wallace includes a very helpful loading bar that appears in the lower-left corner of the map screen.



D

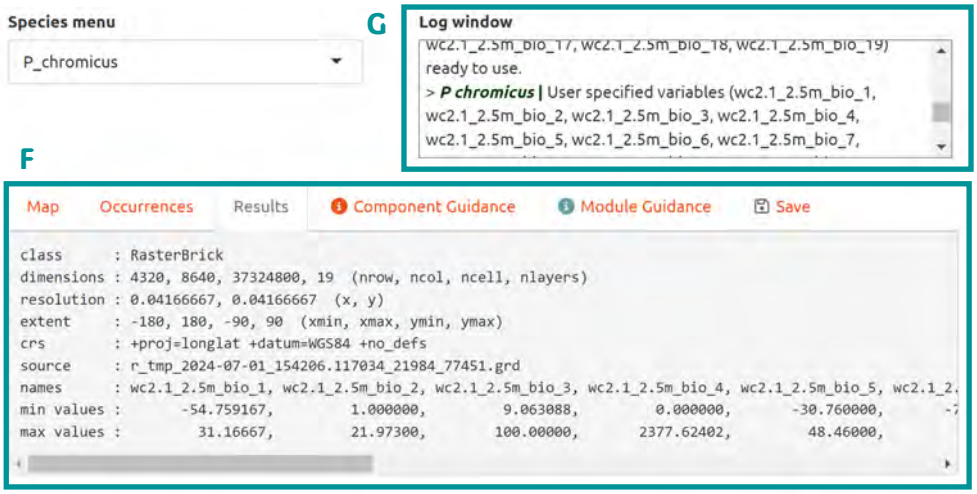
E) A check screen will automatically appear. Don't worry, this is not an error screen.

E



F) The page will automatically load to the results tab.

G) The log window now shows additional occurrence data points that were removed. These were removed due to either being in a pixel (2.5 arcminute square) that does not contain environmental data or shared a pixel with other points and Wallace kept one.



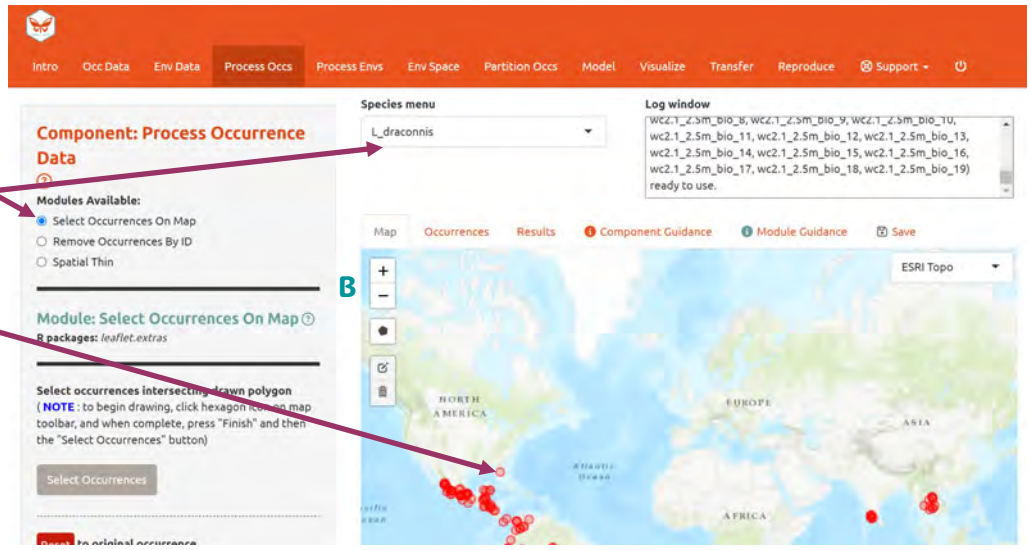
Component 3: Process Occurrence Data

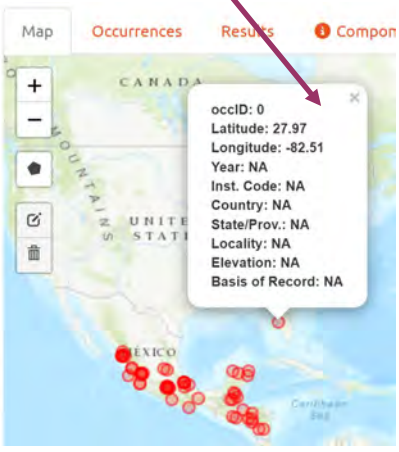
The third component provides modules that remove data points. Three different methods are provided. We will use two of them in the vignette. First, we will remove the *L. draconnis* datapoint that is in Florida. This is a museum sample, not an established population. Then, We will thin the data points.

A) The first module allows you to see occurrence IDs and select a group of data points by drawing a polygon around the ones that you want to keep. Simply click the polygon Icon (B), click around the map to encircle the datapoints, and press "Finish" in the options that pop up when the polygon icon is selected.

We need to remove one point from the *L. draconnis* data. To do so:

1. **Click "Select Occurrences on Map".**
2. **Select L_draconnis.**
3. **Click on the point located in Florida, US.**
4. **A pop-up bubble will appear with some information. Write down the occID. The ID is 0 for this point.**



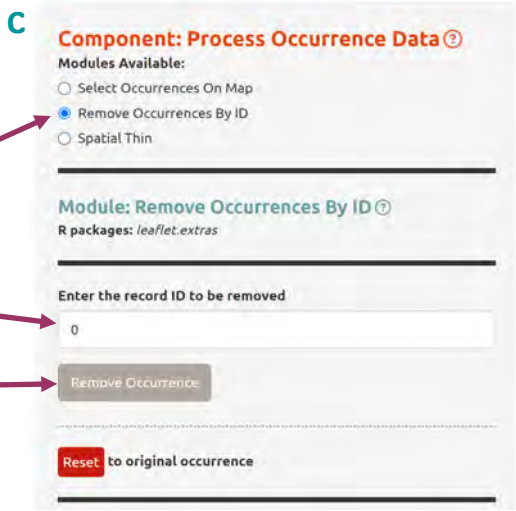


STOP AND THINK!

Look at all of the occurrence points. Does your data look plausible to you? Random data points may be a part of your species' data set as they are museum samples, misidentified, or mislabeled. Not every point may mean an established population. You may be able to check the source for this information. Did you get the data point from GBIF, an article, or a museum? The source might have relevant information to help you decide if the point is worth keeping.

C) The second module "Remove Occurrences BY ID" allows you to type in the occurrence ID to remove individual points. To do so:

5. **Click "Remove Occurrences By ID".**
6. **Type the occID code in the box. The ID is 0 in this vignette.**
7. **Click Remove Occurrence.**
8. **Note the removal of the data point for the methods section.**



The third module allows you to spatially thin the data. This is an important step that helps remove some sampling bias. Geographical biases in sampling can result in a distorted species niche space as areas that are oversampled are weighed more within environmental space and importance. Consider the biology of your species when determining the distance between points.

D) Wallace provides a thinning mechanism to maintain one datapoint within a selected distance in km. We will thin both species by 10 km as records within that radius may be individuals from the same population.

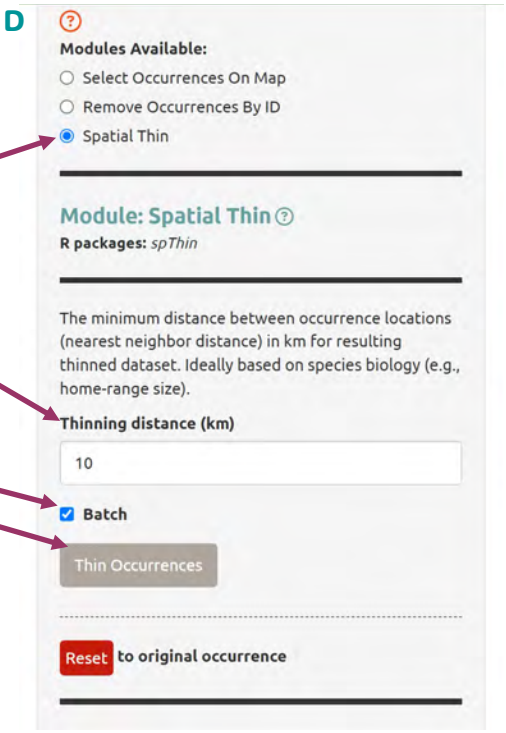
5. Select "Spatial Thin".

6. Type "10" into the box.

7. Click "Batch" so both species are thinned.

8. Click "Thin Occurrences".

9. Note how many points have been retained for the methods section.



Terms to Know

Sampling bias occurs when the methods done to collect data creates a bias in some way, such as more locations are recorded located near easily accessible areas such as roads, agricultural fields, or research stations rather than within natural areas. For SDMs, this bias can lead to over-representation of populations in some areas. This can change how the model calculates environmental suitability and the niche space, thus altering the outputted maps.

Stop and Think!

What should you consider when deciding the distance for spatially thinning? There are several life history traits to consider, but the species' long-distance movement abilities and meta-population dynamics are key. Consider the likely distance between established populations and thin so that each population is distinct and not overlapping.

Resources

Steen et al., 2020. https://doi.org/10.1111/2041-210X.13525

Inman et al., 2021. https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecs2.3422

Helpful website: https://www.scribbr.com/research-bias/sampling-bias/

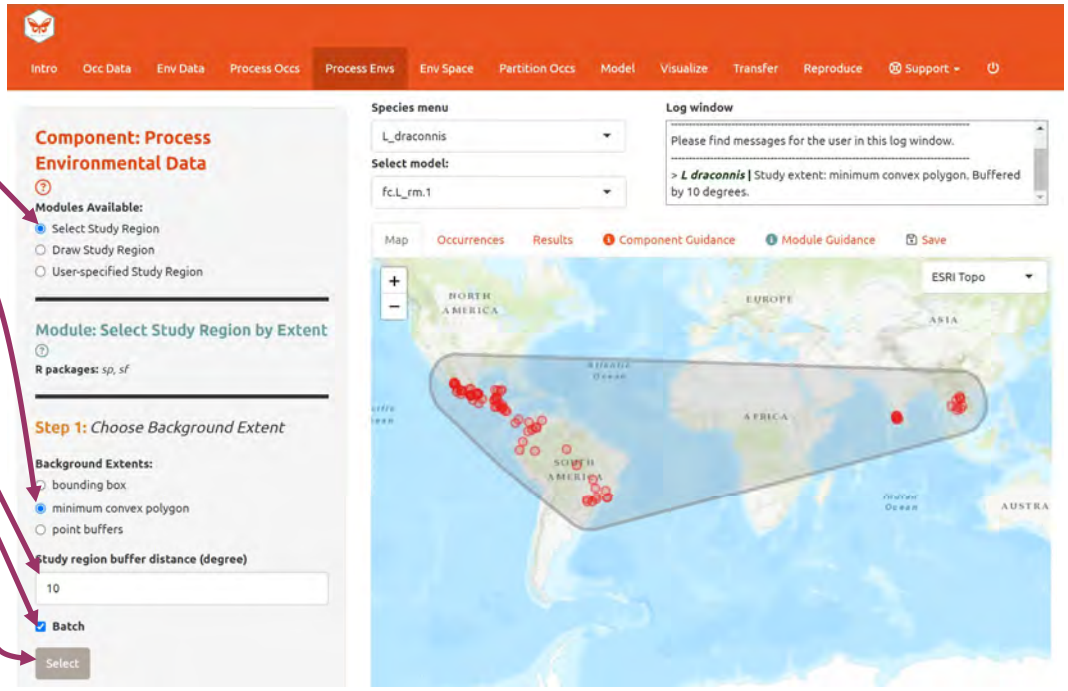
Component 4: Process Environmental Data

The fourth component includes a two step process to collect background environmental data. This module is very helpful and sets Wallace apart from MaxEnt as it allows you to easily select an area from which the background data can be collected. MaxEnt will generally collect data from across the raster (global raster if the full download from BioClim is used); however, this is not usually appropriate.

A) Wallace provides three modules to create an area extent to collect background data. We will use the first one in which Wallace draws shapes around the occurrence locations. The second and third modules allow you to draw a polygon and upload a polygon, respectively. These are helpful if there is an area that you want to specifically include or avoid.

1. Click "Select Study Region".
2. For now, select "minimum convex polygon".
3. Type a buffer distance in degree. For now, type 10.
4. Check "Batch".
5. Click "Select".
6. Look at your map. Does it make sense to collect background data from this polygon?

A



STOP AND THINK!

What does background data represent? These account for absence data and are often called "pseudo-absences". This data is collected from the background extent that is set in this component. It is important to choose a background extend that is appropriate for your data.

What type of background extent should I chose? There has been some debate about this. It boils down to what is biologically important to your species. Absence data should be in areas that are likely sampled for the species. For instance, a background point should not be collected from Canada for these tropical species. Researchers are not looking for the insect in Canada and a background point in this region will bias the model's calculation of the species-environment relationship. Limiting the extend to what is reasonable for the species (such as areas where the species is likely to have established and come to an equilibrium) helps narrow down the niche space.

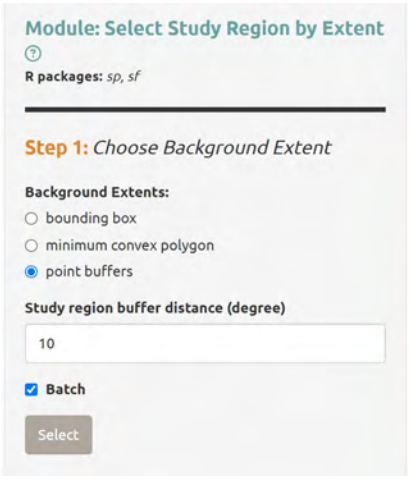
So, is the minimum convex polygon above appropriate for our study system? No. Background points can readily be collected from the African continent that neither species inhabits. We do not know if this area would allow for colonization or not. It should not be included in our model.

The point buffer background extent creates a buffer bubble around each point with a radius set in the box. This is more appropriate for our system.

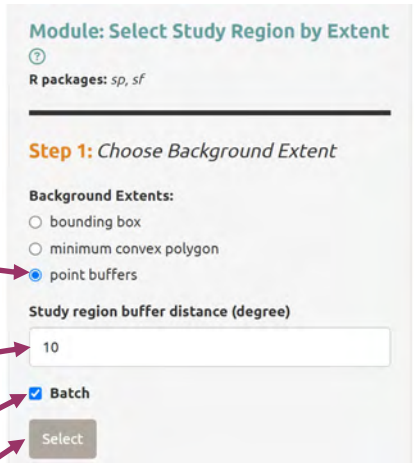
B) The buffered background extent around *P. chromicus*

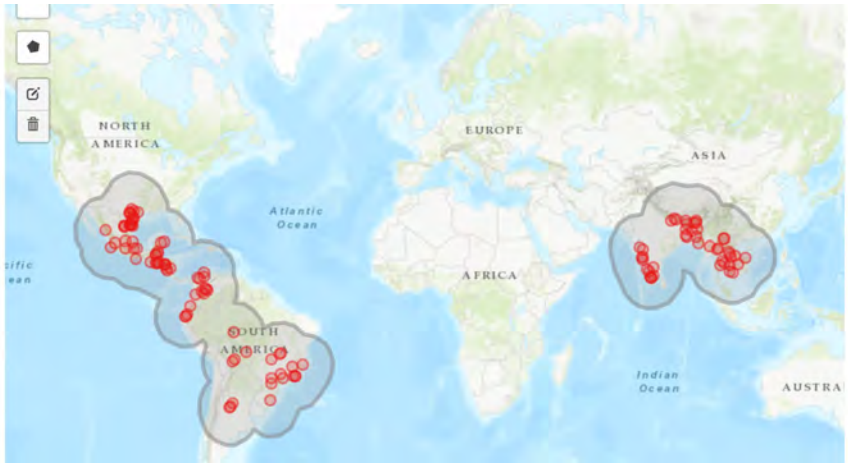
C) The buffered background extent around *L. draconnis*

B



C





7. Click "Select Study Region" if it is unselected (not shown again).

8. Select "point buffers".

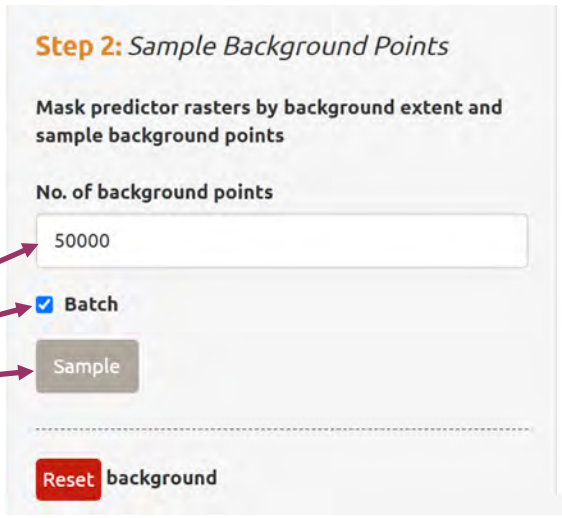
9. Type a buffer distance in degree. Type 10 if needed.

10. Check "Batch" if it is not already selected.

11. Click "Select".

D) The second step in the component allows you to change the number of background points that are collected in the buffered extent. The MaxEnt default is 10,000 points. However, we are collecting background from several continents, so it makes sense to increase this number.

D



12. Type 50000 into the box.

13. Check "Batch".

14. Click "Sample".

15. Wait for it to complete. It can take some time.

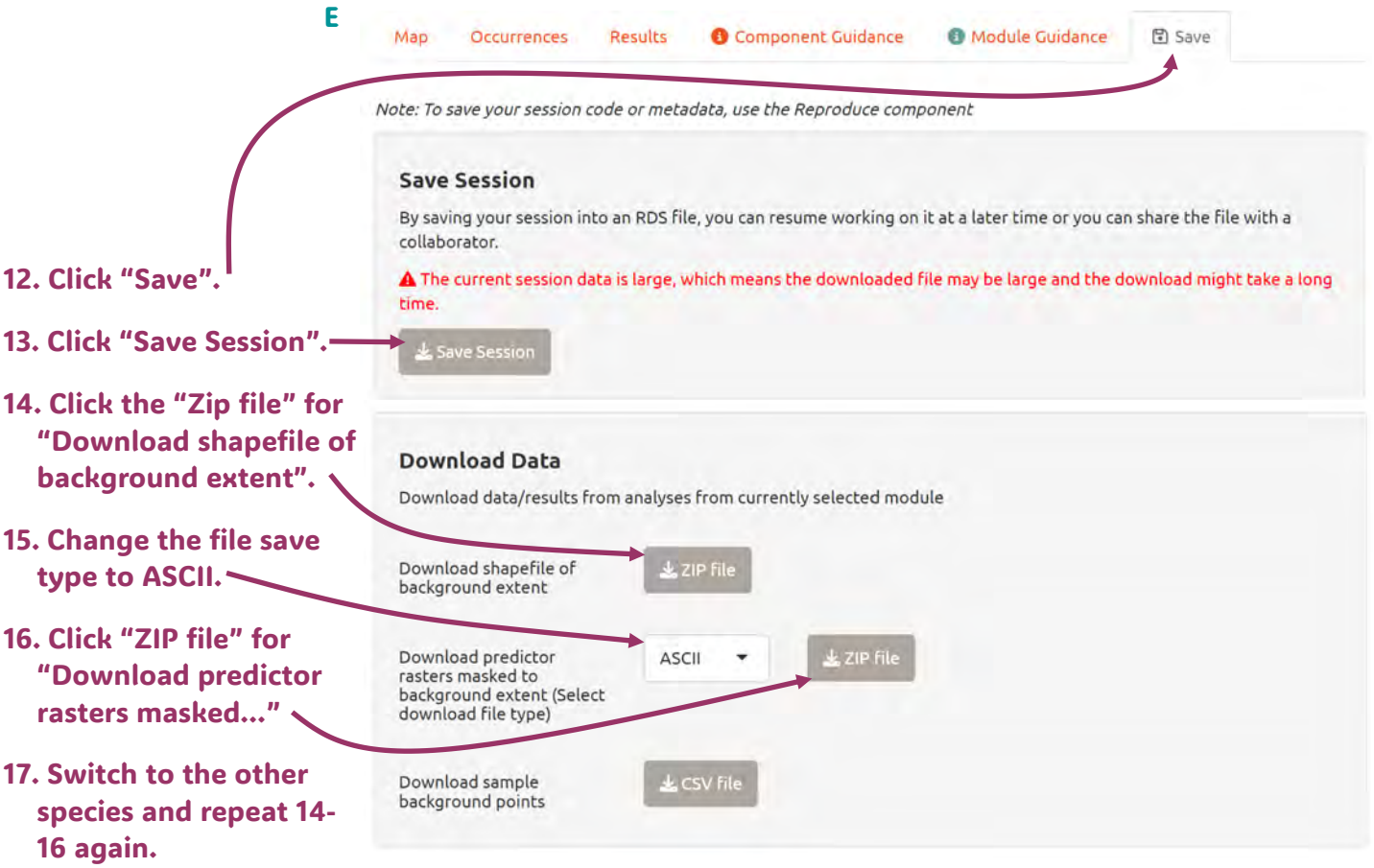
We will need to save a few things in this module to use later.

E. The Save tab allows you to save the session at any time as well as anything that was made in the module. If you ever need to save something from a previous module, don't worry, you can always go back to previous modules and save.

First, we will save the session just in case something happens. This way we will not have to redo any of the previous modules with all of the climate data.

Second, we will save a shapefile of the background extent. This is the buffer line around the data points that we created. This can be useful to make maps in QGIS later. These maps can be used in the published article or for posters so that the methods is more visible.

Lastly, we will save the predictor rasters masked to the background extent. Essentially, this is a ZIP folder of all 19 bioclimatic variables just within the background extent. We will use this in MaxEnt to gather information about variable importance. MaxEnt uses ASCII files.



User Insights

It would be good to save the predictor rasters as both a ASCII (required for MaxEnt) and GeoTIFF (for R). GeoTIFFs are easier to work with in R as they take less time to analyze.

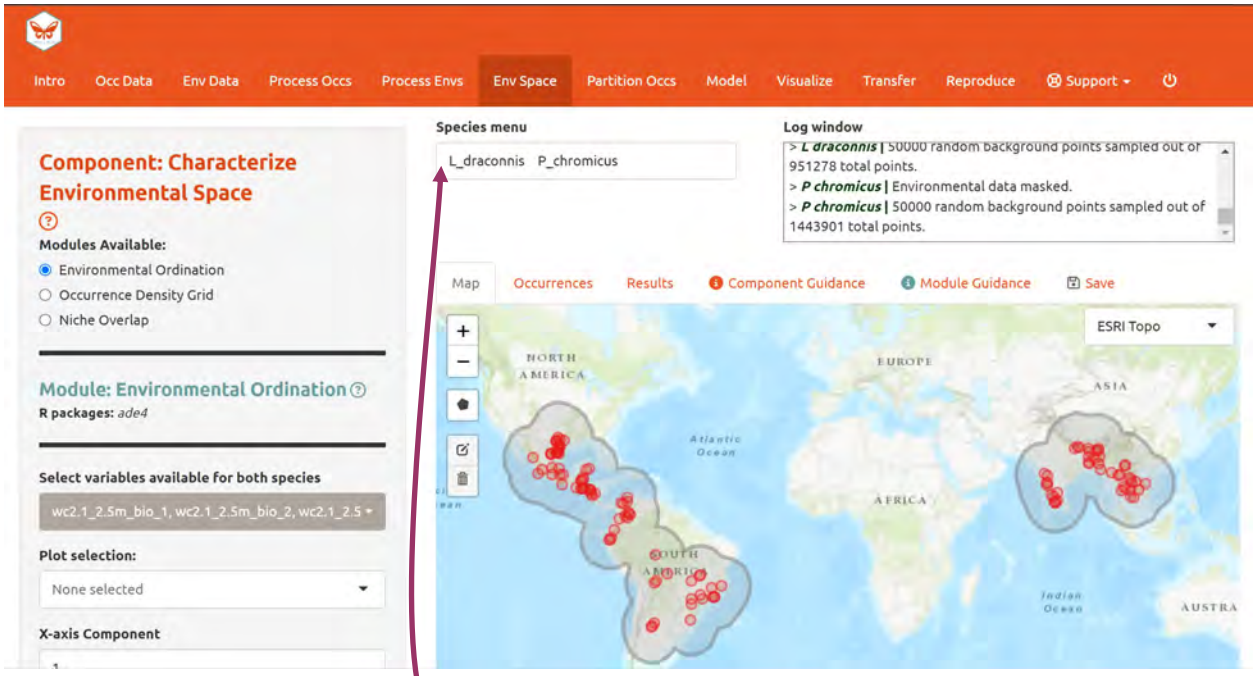
If you are having trouble downloading the predictor rasters, keep trying and click "resume" for downloads. Alternatively, you can use the shapefile downloaded from here and clip the full rasters to the background using QGIS. Clipping will be covered in the sections concerning map making.

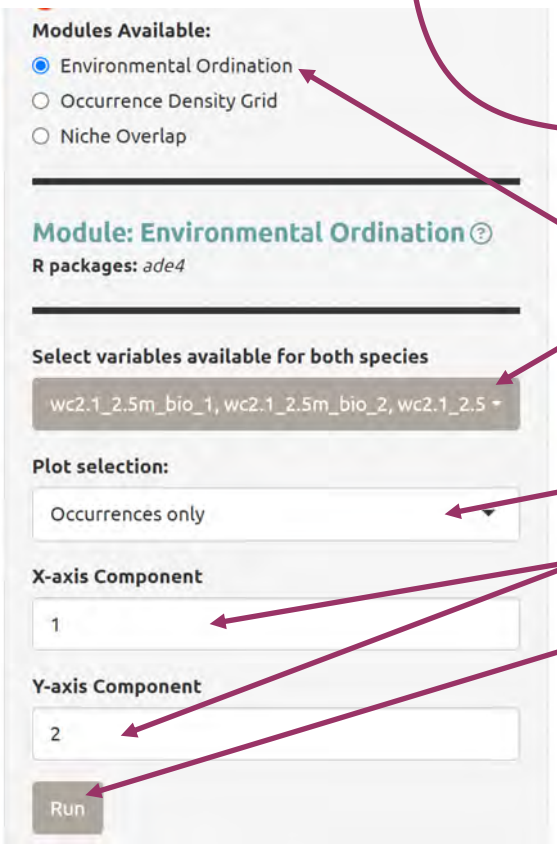
Component 5: Characterize Environmental Space

The fifth component compares environmental and niche space of two species. This can be skipped if you are not interested in comparing the species directly or if you are using Wallace for only one species. Here, we will compare the species, which makes sense for a predator and prey. The modules for this component need to be done in order to work.

A) The first module, Environmental Ordination, will generate four PCA plots that compare the environmental data at each of the points (both occurrence and background).

A



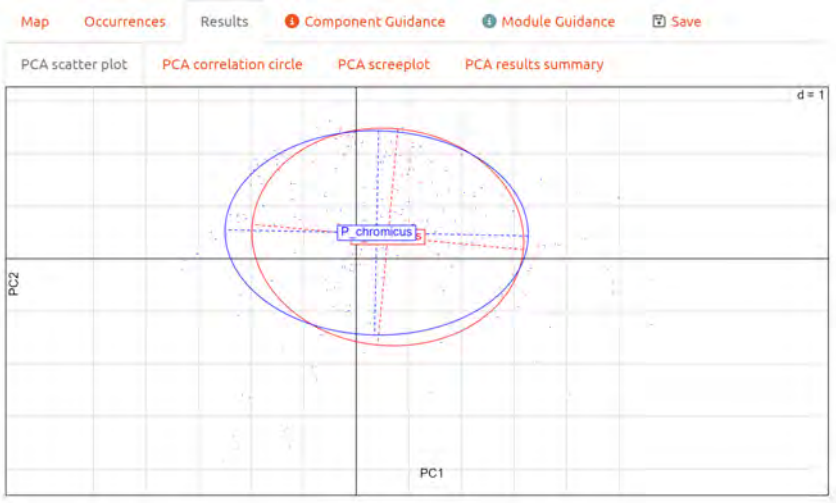


1. Click the species button and select the second species so that both appear in the select menu.
2. Click "Select Occurrences on Map".
3. Select only the variables that you want to maintain in the comparison. We will keep all of the variables for the comparisons.
4. Select "Occurrences only".
5. Keep the X and Y Components.
6. Click "Run".

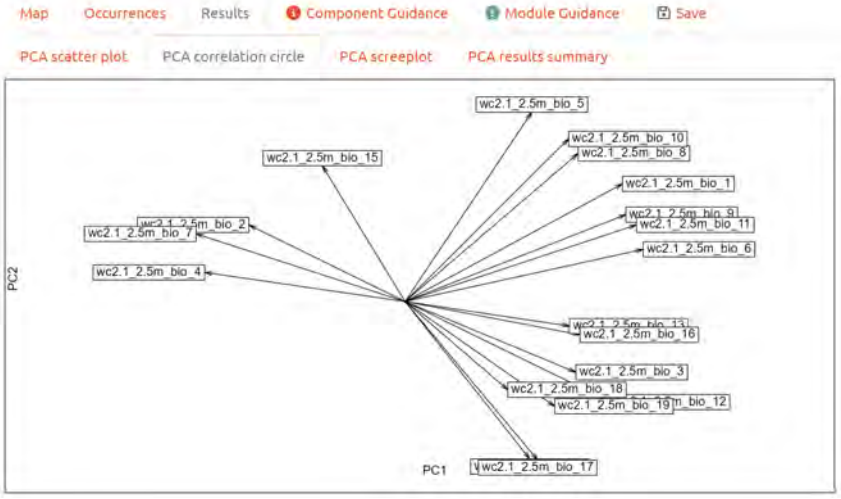
B) This is the PCA scatter plot under the Environmental Ordination module. The dots only show the occurrence locations as that was what we selected under "Plot selection". This plot shows high overlap of the environmental space between the species.

C) This is the PCA correlation circle, also under the Environmental Ordination module. The arrows length and direction show how related the variable is to either PC1 or PC 2 axis. Thus, if the variables are close, they are closely correlated with each other. Variables in opposite directions are negatively correlated with each other.

B



C



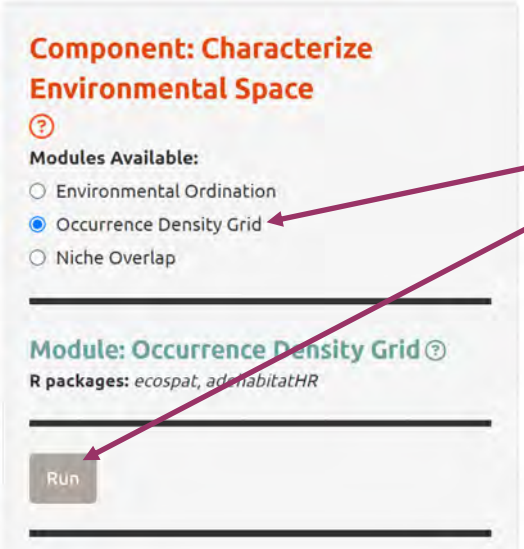
Stats Chat

PCA stands for Principal Component Analysis. In general, the Hutchinsonian nitche is reduced to just the variables that we included in the model. The PC1 and PC2 axes are combinations of these niche spaces that are highly variable, thus the resulting plots can show how similar (close) or different (far) data points or variables are in statistical space.

D) The second module "Occurrence Density Grid" produces an overlapping grid. This can be helpful as it shows where in environmental space the species most occupies (dark colors).

E) The Occurrence density grid

D

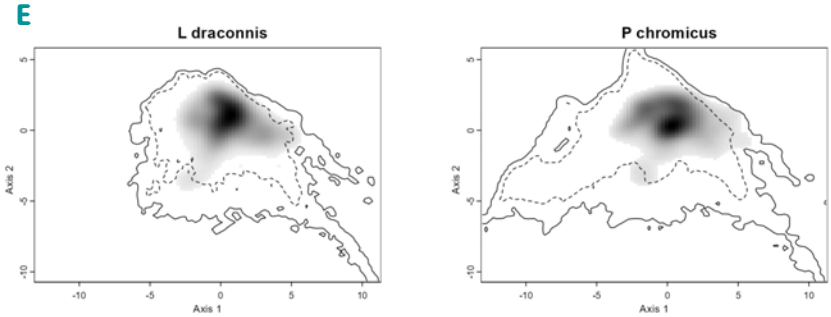


7. Click "Occurrence Density Grid".

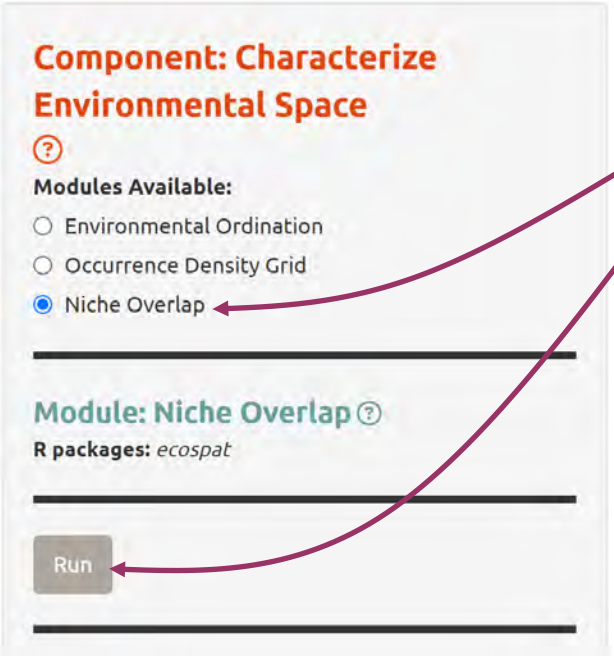
8. Click "Run".

9. Look at the resulting plot in the "Results" tab.

E



F



F) The third module, Niche Overlap, produces two figures as well as some helpful statistics that can be reported in the manuscript.

10. Click "Niche Overlap".

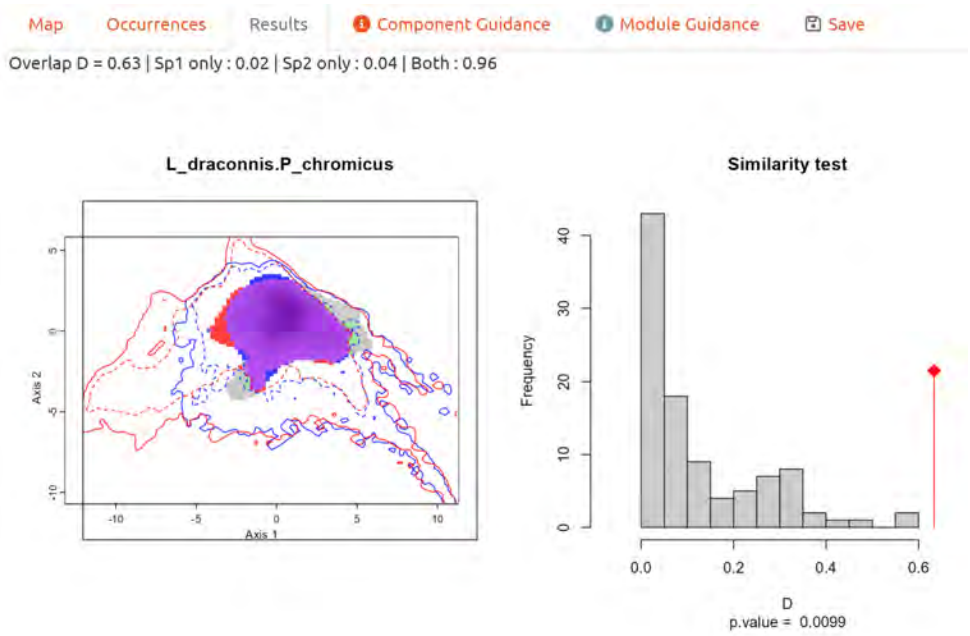
11. Click "Run".

12. Look at the resulting plot in the "Results" tab.

User Insights

Your statistic results and maps may be slightly different than the ones here. That is fine. Wallace only does one replicate of the test at a time. The statistics includes all of the occurrence data, 50,000 background points, and 19 variables. That is a lot of information, so there is some variability. If you want to account for this, run the test several times.

G



G) The Niche Overlap outputs.

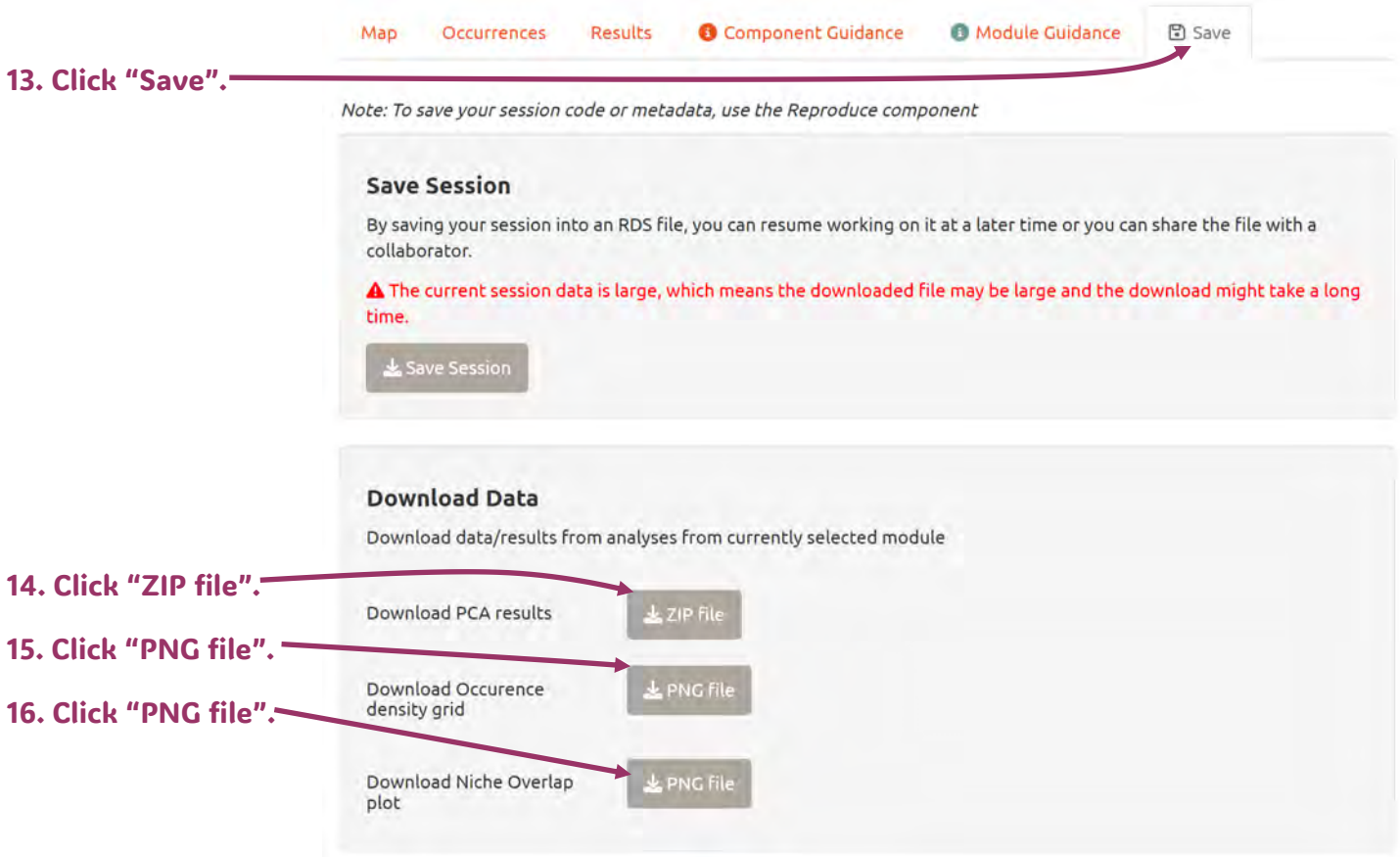
Stats Chat

The Niche Overlap module produces two different statistics.

Overlap D is a measurement of created by Schoener in 1968. The value reflects the relative use of the environmental space. The value ranges from 0 to 1 with greater values representing a greater overlap of niche space. The value that represents "Both" provides the combined overlap of the species.

The p value is initially confusing. The niche plots show high overlap, but the p value is < 0.05, which usually means that the variables are significantly different. However, we are testing for significant similarity, not significant differences. Thus, the p value showcases significant similarity.

G) Next, we will save all of the results. These can be used in the manuscript results section.



RESOURCES

Helpful website on Schoener's D: https://plantarum.ca/2021/12/02/schoenersd/

Bates and Bertelsmeier 2021. https://doi.org/10.1016/j.cub.2021.08.035

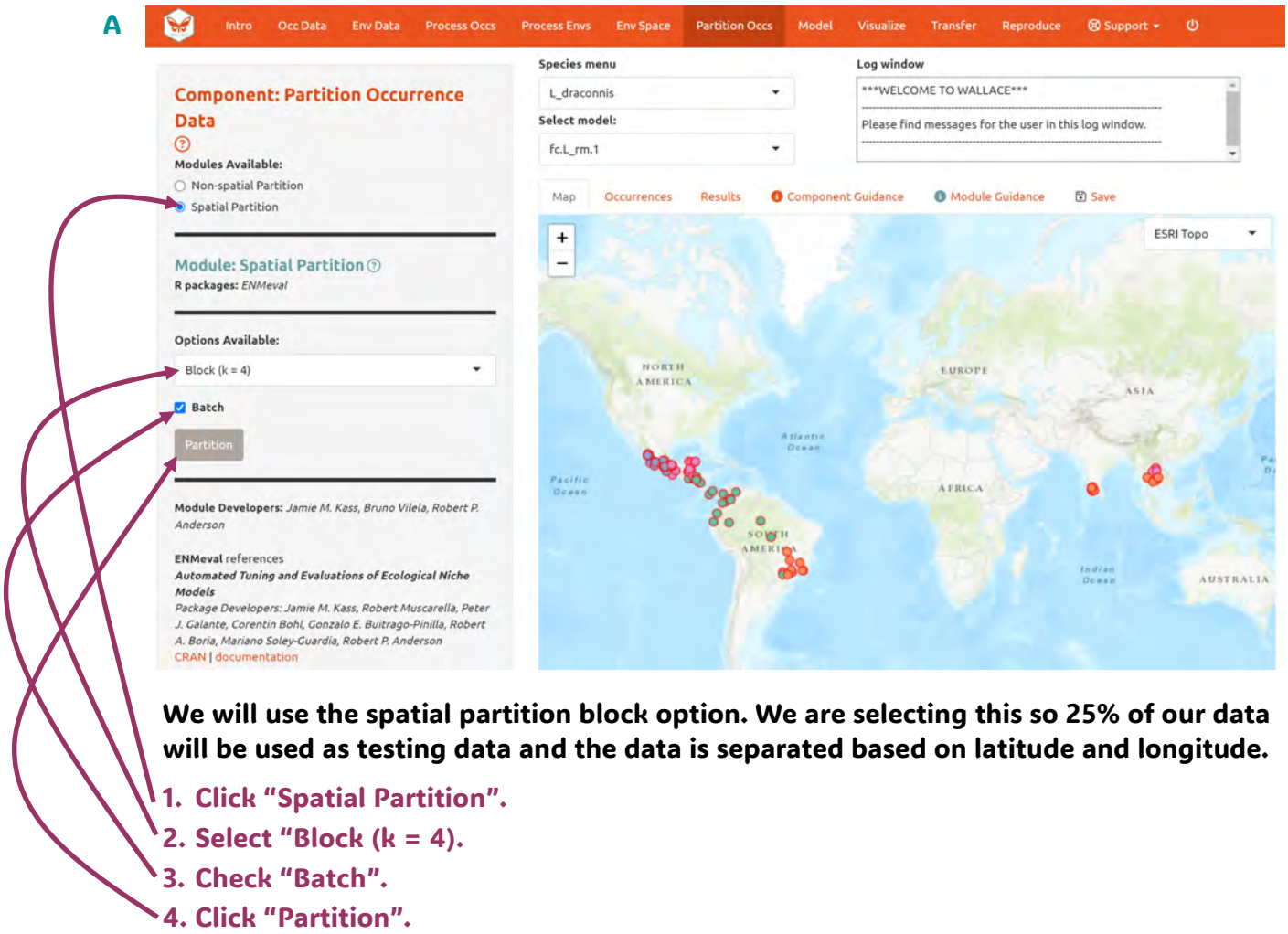
Geange S.W., et al., 2010. https://doi.org/10.1111/j.2041-210X.2010.00070.x

Suárez-Mota, M.E. and Villaseñor, J.L., 2020. https://www.jstor.org/stable/26958712

Component 6:

The sixth component provides different options as to how to separate the data into test and training data. This is done to evaluate the model as it is being made. The training data (e.g., 75% of the points) is used to make the model whereas the testing data (e.g., the other 25% of the points) is used to analyze the output. This is done over many iterations so that the points are placed into each of the different groups. What combination of points are in each group depends on how they are partitioned.

A) Wallace provides two different modules with several options in each one for partitioning. The first one separates points in a non-spatially. The second one, which we will use, partitions spatially.



We will use the spatial partition block option. We are selecting this so 25% of our data will be used as testing data and the data is separated based on latitude and longitude.

1. **Click "Spatial Partition".**
2. **Select "Block (k = 4).**
3. **Check "Batch".**
4. **Click "Partition".**

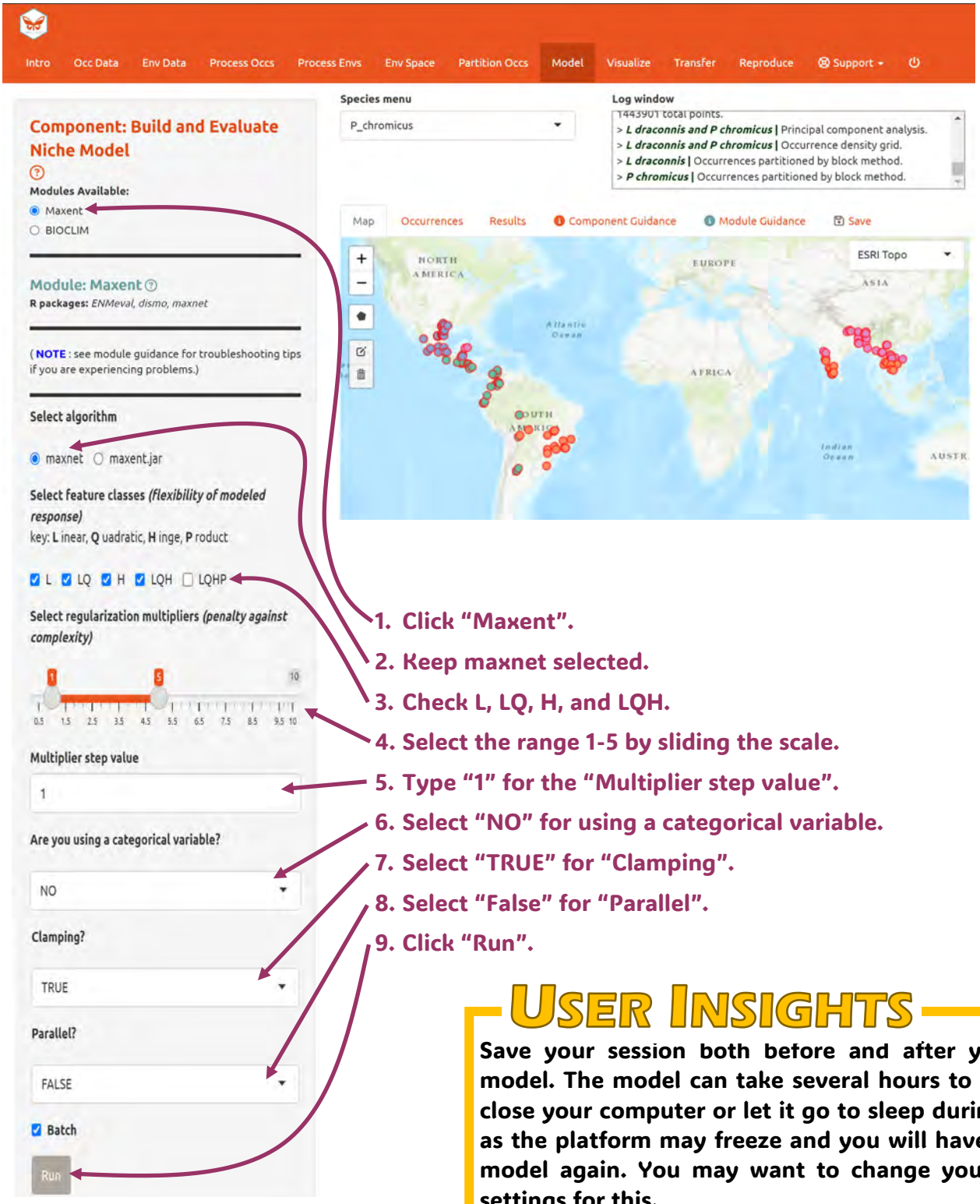
Stats Chat

How should you partition the data? It depends on what you want to do with the data. If you plan to transfer across space or time, then spatial partitions are more appropriate than non-spatial partitions. This is also the case with suspected sampling bias.

Component 7: Build and Evaluate Niche Model

A) The Model component is where the settings for the MaxEnt model are set. There are several different optimization options that reflect the MaxEnt program.

We will use the maxnet algorithm, although the maxent.jar can be used through the Java program.



User Insights

Save your session both before and after you run the model. The model can take several hours to run. Do not close your computer or let it go to sleep during this time as the platform may freeze and you will have to run the model again. You may want to change your computer settings for this.

Stats Chat

Feature classes allow for complex relationships to be modeled by differently transforming covariates. The MaxEnt feature classes, in order of complexity, include linear (L), quadratic (Q), hinge (H), product (P) and threshold (T). These can be used separately or together as a group (e.g., LQH) and more feature classes create a more complex model.

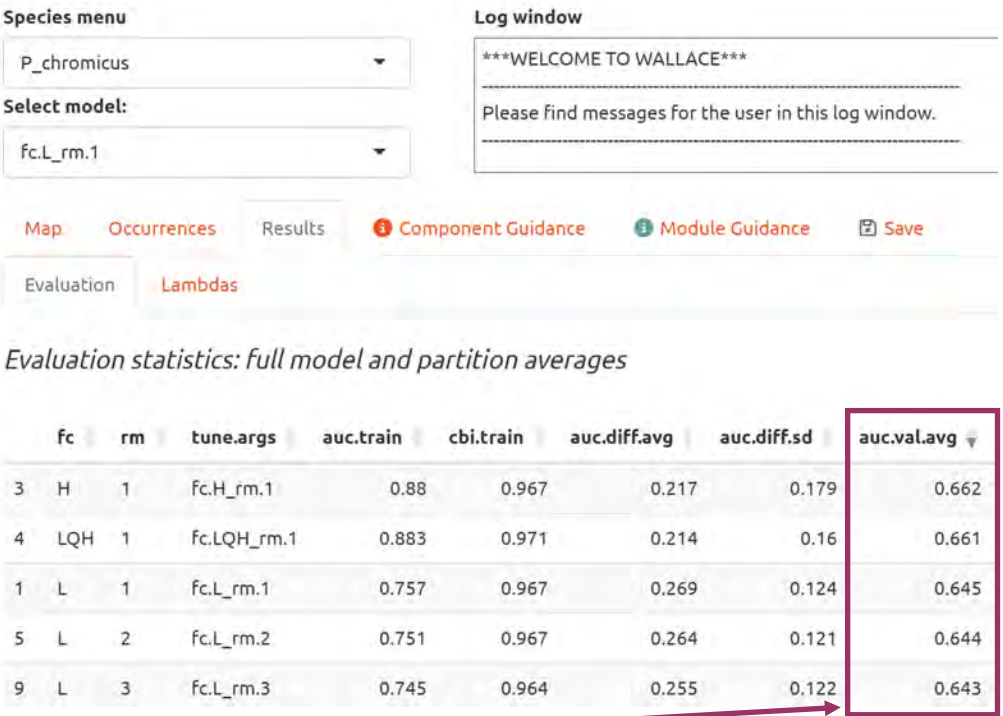
Regularization multiplier (RM) adds constraints to the model and penalizes for complexity. Too much complexity can create an overfitted models that are not very transferable to other locations or times. Optimizing the model with both feature classes and RM allow for a better balance.

Clamping constrains the environmental values in the projected range to the values found in the background data collected. This reduces the chances of unrealist patterns that might emerge because of the complex models being used.

B) *Prisma chromatus* models (each feature class and regularization combination) arranged by the area under the curve.

C) *Prisma chromatus* models (each feature class and regularization combination) arranged by Akaike information criterion (AIC).

B



	fc	rm	tune.args	auc.train	cbi.train	auc.diff.avg	auc.diff.sd	auc.val.avg
3	H	1	fc.H_rm.1	0.88	0.967	0.217	0.179	0.662
4	LQH	1	fc.LQH_rm.1	0.883	0.971	0.214	0.16	0.661
1	L	1	fc.L_rm.1	0.757	0.967	0.269	0.124	0.645
5	L	2	fc.L_rm.2	0.751	0.967	0.264	0.121	0.644
9	L	3	fc.L_rm.3	0.745	0.964	0.255	0.122	0.643

10. The results will automatically appear when the model is complete.
11. Arrange the models by pressing the arrows beside the category you want to organize. First organize auc.val.avg by high values. Note the top models.
13. Organize delta.AICc values by the lowest values. Note the top models. This table is broken to show both the models and the delta AICc values.

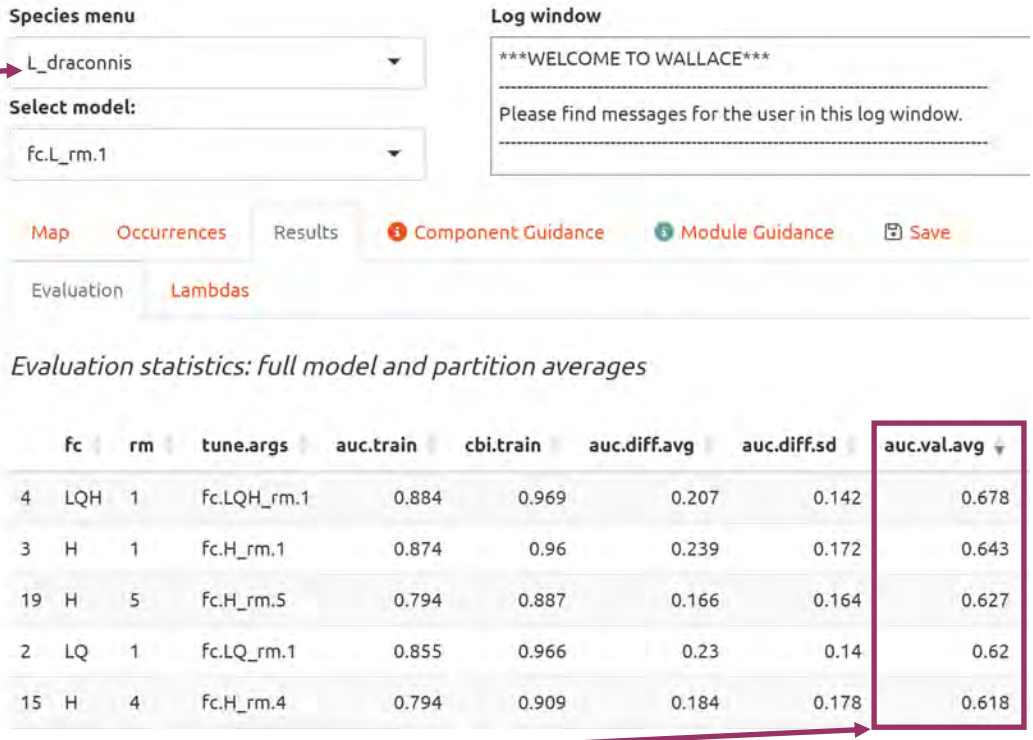
C

	fc	rm	tune.args	AICc	delta.AICc	w.AIC
7	H	2	fc.H_rm.2	3341.896	0	0.904
12	LQH	3	fc.LQH_rm.3	3346.529	4.634	0.089
8	LQH	2	fc.LQH_rm.2	3351.892	9.996	0.006
3	H	1	fc.H_rm.1	3356.811	14.916	0.001
11	H	3	fc.H_rm.3	3357.098	15.203	0

B) ***Leviathan draconnis*** **models (each feature class and regularization combination) arranged by the area under the curve.**

C) ***Leviathan draconnis*** **models (each feature class and regularization combination) arranged by Akaike information criterion (AIC).**

D



Evaluation statistics: full model and partition averages

	fc	rm	tune.args	auc.train	cbi.train	auc.diff.avg	auc.diff.sd	auc.val.avg
4	LQH	1	fc.LQH_rm.1	0.884	0.969	0.207	0.142	0.678
3	H	1	fc.H_rm.1	0.874	0.96	0.239	0.172	0.643
19	H	5	fc.H_rm.5	0.794	0.887	0.166	0.164	0.627
2	LQ	1	fc.LQ_rm.1	0.855	0.966	0.23	0.14	0.62
15	H	4	fc.H_rm.4	0.794	0.909	0.184	0.178	0.618

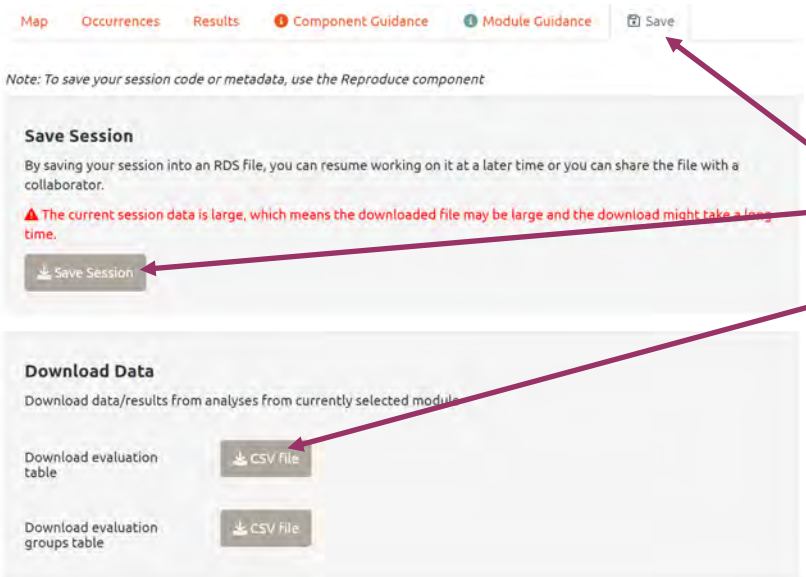
13. Switch to the *L. draconnis* data under the species menu tab.

14. Again, organize auc.val.avg by high values. Note the top models.

15. Organize delta.AICc values by the lowest values. Note the too models. This table is broken to show both the models and the delta AICc values.

E

	fc	rm	tune.args	AICc	delta.AICc	w.AIC
2	LQ	1	fc.LQ_rm.1	2252.504	0	0.99
12	LQH	3	fc.LQH_rm.3	2261.831	9.327	0.009
8	LQH	2	fc.LQH_rm.2	2267.417	14.913	0.001
16	LQH	4	fc.LQH_rm.4	2271.411	18.906	0
7	H	2	fc.H_rm.2	2277.829	25.325	0



16. Click the Save tab.

17. Save the session so you do not have to run the long model again.

18. Click CSV file to download the evaluation table.

19. Change the species and Click CSV file to download the evaluation table of the other species.

Stats Chat

AUC stands for the area under the receiver operating characteristic (ROC) curve. This is a measurement of the predictive accuracy of the model. The value ranges from 0 to 1. "Good" models and "Great" models are often designated as 0.8 - 0.9 and > 0.9, respectively. Although many SDM papers state that their model is very appropriate for their system because of this value, however, this reliability has been questioned. Thus, we suggest using this value as one of the values that you use to decide which model is the most appropriate.

AICc is the corrected version of the Akaike Information Criterion (AIC). It is a score given to individual models when comparing to the other models. It tests the model's performance using machine learning and selects the models that are the best fit on the data, giving it a lower score. The delta AICc provides the difference in the AIC scores so that the best model (lowest value) receives a 0 and each of the remaining models have a >0 value.

Which model should I use? Plenty of papers just use the model selected by AICc or AUC; however, these are not always appropriate. Each of the models can project the potential distribution a little differently. You should also look at the projection maps (under current conditions) of the top models and use your expertise to come to a decision as to which map is the most appropriate. For instance, a top model for one of our tropical species for a previous paper projected highly suitable areas in Canada and the other two top models did not. We discarded the one that had predictions that did not fit the biology of the insect and used one of the ones that was more reasonable. Take everything into account when deciding how to proceed with your model selection.

We will not be doing this quite yet. We first need to remove bioclimatic variables that are collinear with each other.

Resources

Lobo et al., 2007. https://doi.org/10.1111/j.1466-8238.2007.00358.x

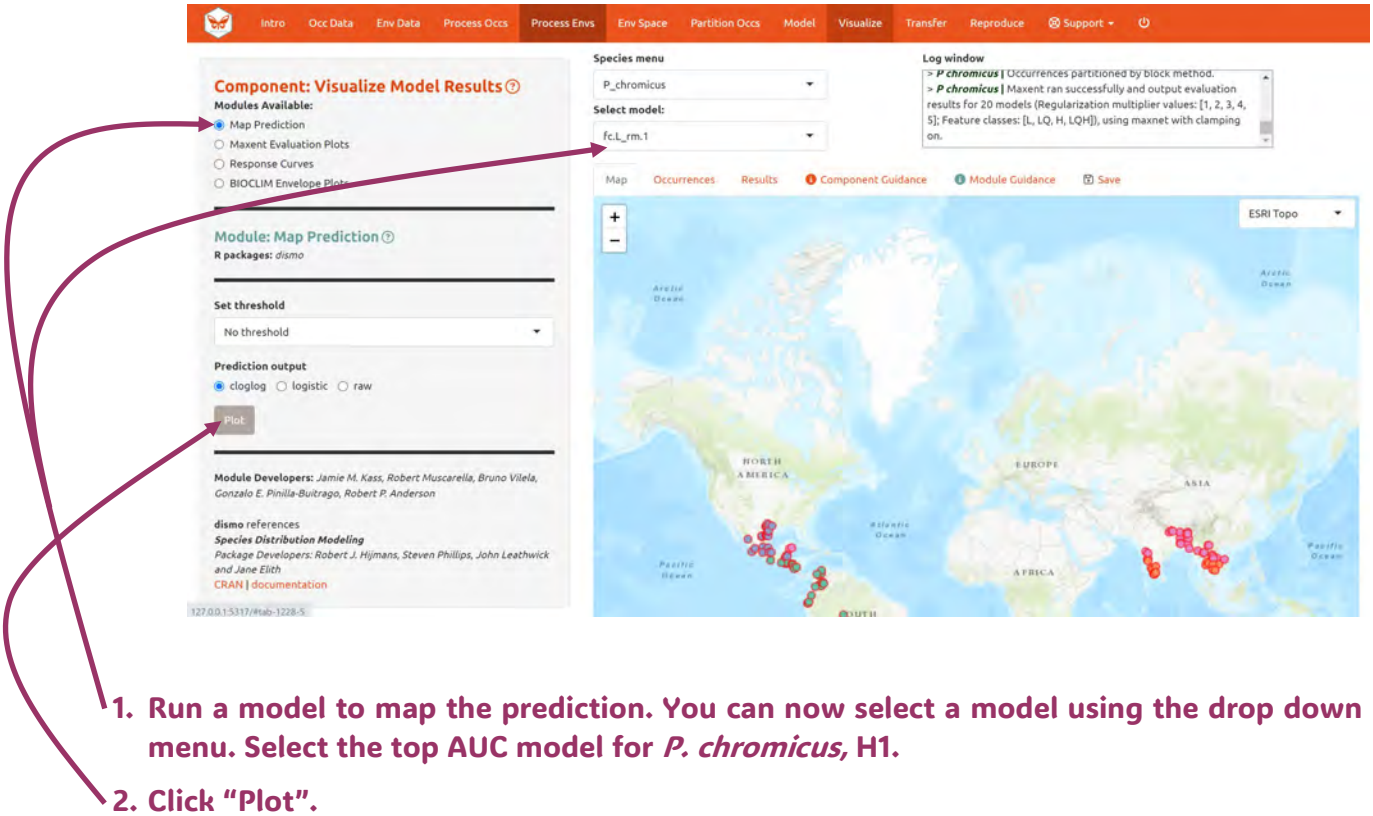
Jiménez-Valverde, A., 2011. https://doi.org/10.1111/j.1466-8238.2011.00683.x

Information on AIC: https://builtin.com/data-science/what-is-aic

Component 8:

A) The eighth component, Visualize, provides statistical information that we will use to assess the importance of the bioclimatic variables. For this initial model, we will use the response curves to select which highly correlated variable to remove with a Pearson test. After we create the final model with the remaining variables, the information in this component will help us tie the species -environment relationship to the SDM and the life history of the species.

The items available for download include MaxEnt evaluation plots, response curves, and BIOCLIM envelope maps. These modules are made from what is modeled in the buffer zones. This is another reason why selecting a buffer zone that is biologically sound is important.



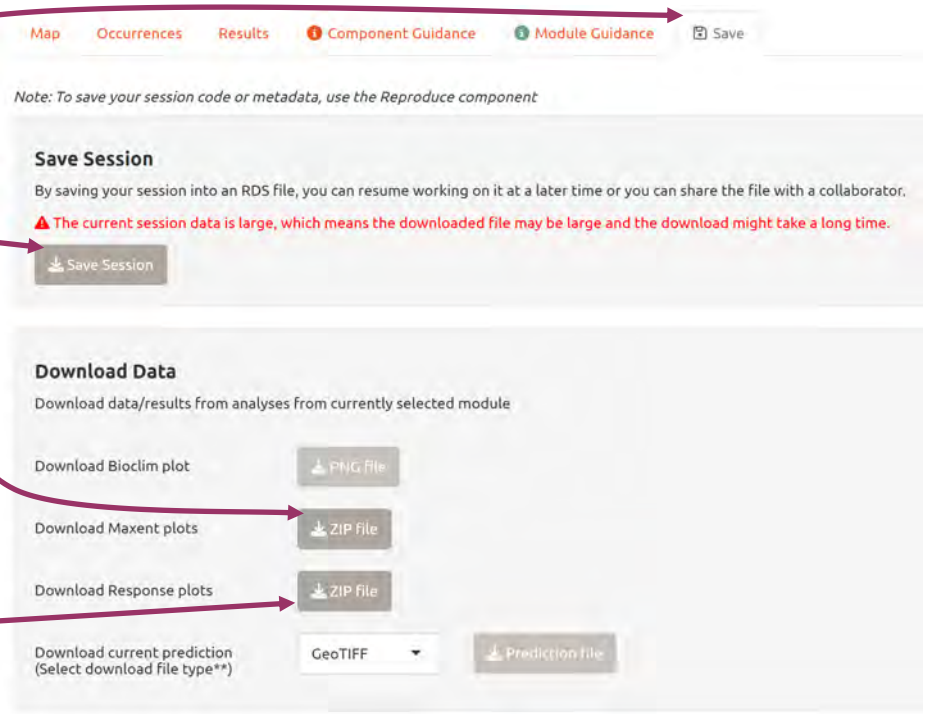
1. **Run a model to map the prediction. You can now select a model using the drop down menu. Select the top AUC model for *P. chromicus,* H1.**
2. **Click "Plot".**

You can view the model's Maxent evaluation plot and response curves using the next two modules. Alternatively, you do not need to view the plots in order to download them and use later. We will use the "Save" tab to download the plots as file folders.

User Insights

When you save the response curves, you will need to change the tabs for each species and model. You do not need to run the "Map prediction" module between each one. The response curves will automatically change. We suggest that you rename the downloaded file before downloading the next one. They are named similarly and it is easy to mix them up.

3. Click the "Save" tab.
4. Click "Save Session".
5. Download the MaxEnt Plots. You will need to do this for each species.
6. Save the Response Curves for the top two models for each species. *Prisma chromicus* top models were H1 and H2. *Leviathan draconnis* top models were LQH1 and LQ1.



Stats Chat

What are response curves? **Response curves show the relationship between the environmental variables and the potential distribution of the species. Wallace plots the curves by changing one response variable and holding the others constant at their mean value. The curves can provide interesting biological information about your system; for instance, if the species is likely to persist in warmer temperatures. Wallace only provides response curves for the variables that add information to the model.**

The MaxEnt Platform

Introduction

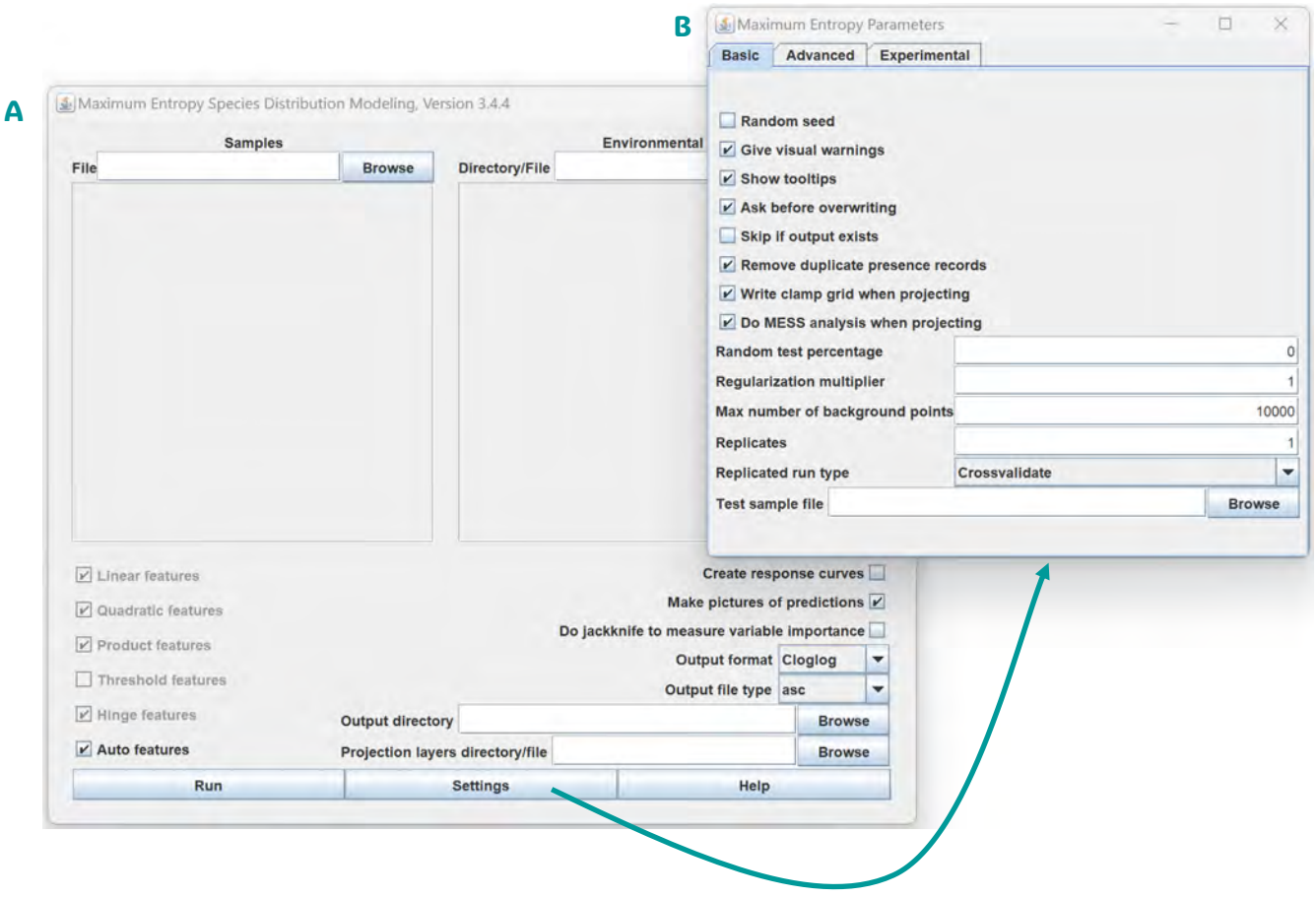
The MaxEnt platform is a program run through JAVA (you will need to download JAVA if you do not have it). This is the platform that Wallace runs in the background to create the distribution maps. We are using MaxEnt to obtain Jackknife tests, which rank the variables in order of their contribution to the model. We will run several replicates of each of the top models. This will allow us to assess the importance with some variability.

The Default Settings

Many authors creating SDM though MaxEnt use the default settings. This has been shown to not be appropriate as the resulting maps are often different than one that has been optimized for the specific species and model that is being created. We will select settings that are similar to those that we used in Wallace.

A) The MaxEnt main window includes areas to upload the data, environment layers, and change some settings.

B) The Settings tab that appears when clicking "Settings". Settings can be changed in three tabs. We will work in the "Basic" tab.

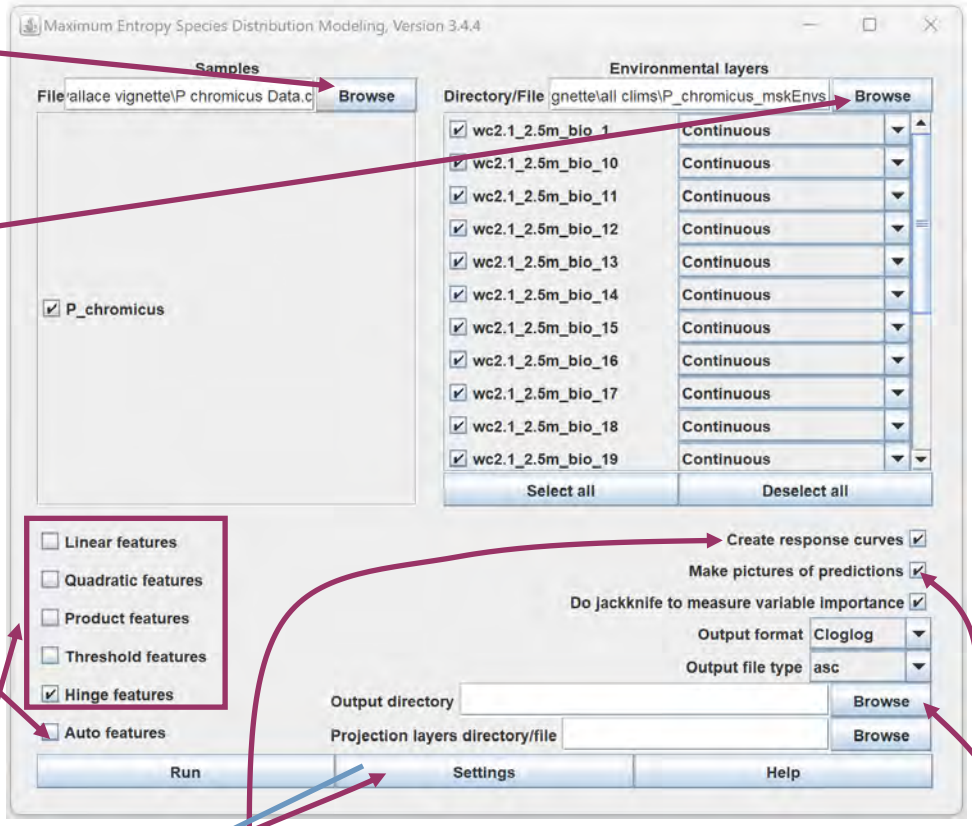


Prisma chromatus settings

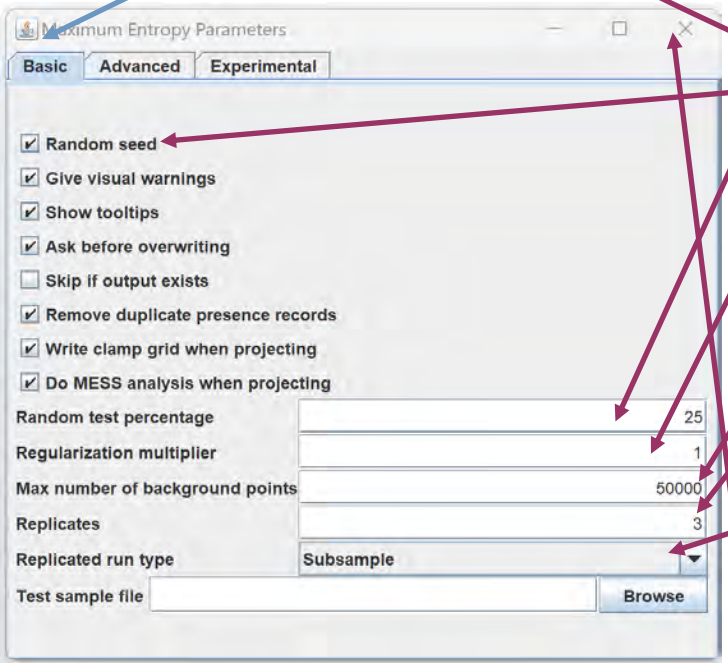
A) We will first run a few models of *P. chromicus.*

A

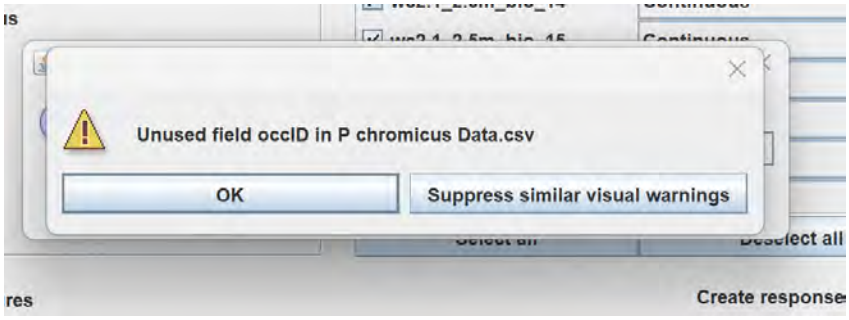
1. Click "Browse and find the *P. chromatus* data.
2. Click browse and find the FOLDER with just the *P. chromatus* ASCII bio-climatic variables within the buffered background extent.



3. Unclick "Auto features".
4. We will use the top AUC model, H1. Unclick all feature classes but H.
5. Click "Create response curves".
6. Click "Do jackknife to measure variable importance".
7. Select an output directory. This will be a folder where all of the files will be moved to. Click "Browse" and find or create such a folder.
8. Click "settings". The screen B. will appear.



9. Click "Random seed".
10. Change the "Random test percentage" to 25.
11. This time, we will keep the "Regularization multiplier" at 1, however, this is where this will be changed for other RM (e.g., LQ2, you will type 2 here).
12. Change the "Max number of background points" to 50000.
13. Change "Replicates" to 3.
14. Change "Replication run type" to "Subsample".
15. Click the X to escape. The other settings will remain the default.



16. **Read any warnings. Warnings like this are fine. Here, one of the columns is not being used in MaxEnt.**
17. **Click "Suppress similar visual warnings" for any warnings that will not impact the model. Another example is any warnings that a pixel does not include data.**
18. **After the first model is finished, run it again with the H2 model by changing the RM number and changing the output directory/file. It is easiest to have a different folder for each MaxEnt run.**
19. **Run two models for *L. draconnis*. We will run the top models LQH1 and LQ1. Change the Samples and Environmental layers to the *L. draconnis* files. Change the features and regularization multiplier to the appropriate settings for the two models. Keep the other settings.**

B) The file that you select will now be full of documents. Those marked with 0, 1, and 2 have the results of each of the replicates. The different documents include a website document that is an accumulation of the results, a couple of explanation files, and three tables that include statistic outputs.

C) The website document labeled with just the species name has the averaged results from all three replicates. We will look at this document for each of the four models.

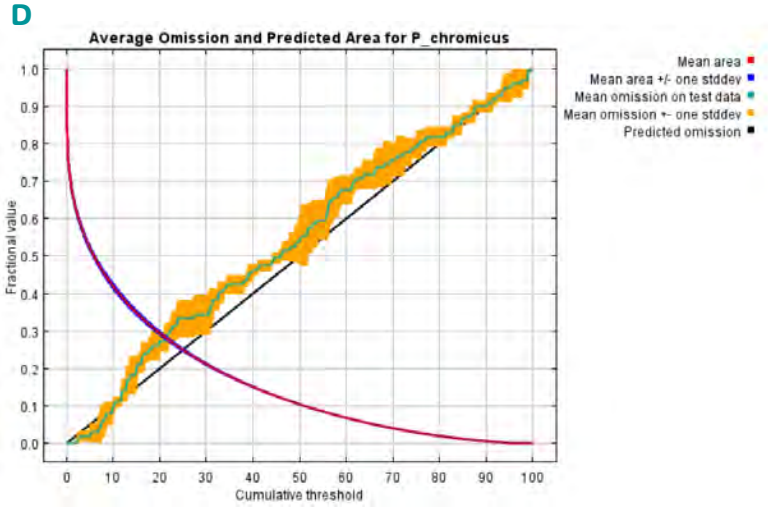
B

Name	Date modified	Type	Size
plots	7/11/2024 11:33 AM	File folder	
maxent	7/11/2024 12:00 PM	Text Document	175 KB
maxentResults	7/11/2024 12:00 PM	Microsoft Excel Co...	16 KB
C P_chromicus	7/11/2024 12:00 PM	Chrome HTML Doc...	12 KB
P_chromicus_0.asc	7/11/2024 11:33 AM	ASC File	72,698 KB
P_chromicus_0	7/11/2024 11:42 AM	Chrome HTML Doc...	16 KB
P_chromicus_0.lambdas	7/11/2024 11:33 AM	LAMBDAS File	10 KB
P_chromicus_0_explain	7/11/2024 11:33 AM	Windows Batch File	1 KB
P_chromicus_0_omission	7/11/2024 11:33 AM	Microsoft Excel Co...	27 KB
P_chromicus_0_sampleAverages	7/11/2024 11:33 AM	Microsoft Excel Co...	1 KB
P_chromicus_0_samplePredictions	7/11/2024 11:33 AM	Microsoft Excel Co...	11 KB
P_chromicus_1.asc	7/11/2024 11:42 AM	ASC File	72,700 KB
P_chromicus_1	7/11/2024 11:51 AM	Chrome HTML Doc...	16 KB
P_chromicus_1.lambdas	7/11/2024 11:42 AM	LAMBDAS File	10 KB
P_chromicus_1_explain	7/11/2024 11:42 AM	Windows Batch File	1 KB
P_chromicus_1_omission	7/11/2024 11:42 AM	Microsoft Excel Co...	27 KB
P_chromicus_1_sampleAverages	7/11/2024 11:42 AM	Microsoft Excel Co...	1 KB
P_chromicus_1_samplePredictions	7/11/2024 11:42 AM	Microsoft Excel Co...	11 KB
P_chromicus_2.asc	7/11/2024 11:51 AM	ASC File	72,699 KB

MaxEnt Outputs

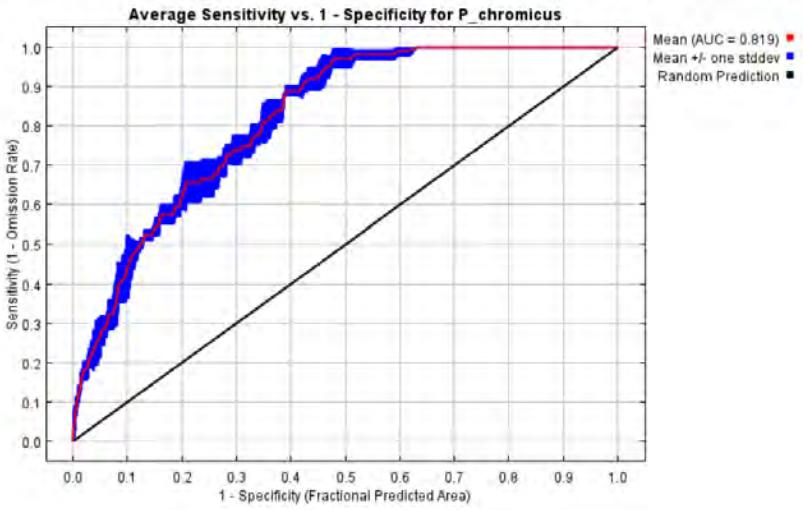
The website link has a lot of great information about the model. Let's take a moment and look at some of the information. The data that we will use is the H1 *P. chromicus* data. If your data looks slightly different, that is fine. As you will see, each replicate is slightly different, so comparing just three replicate will likely be different as well.

D



D) The omission rate curve shows the cumulative threshold as an averaged line (green) and the standard deviation (orange). This line should be close to the predicted omission line (black).

E



E) The next figure shows the receiver operating characteristic (ROC) curve. The red line provides the mean AUC value, which is 0.819. The blue area shows the deviation around the mean provided by the three replicates.

STATS CHAT

Why are these numbers different than what Wallace provided?

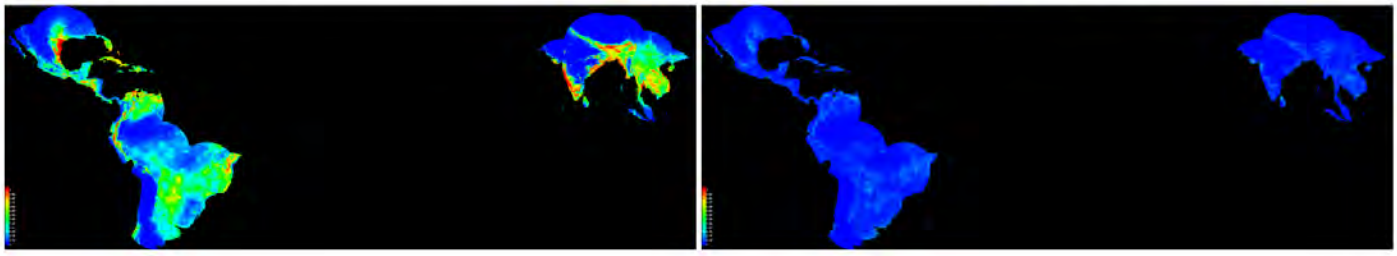
There are a few reasons why the results are a little different. First, the settings may be slightly different as Wallace is not clear on which exact settings are used. Second, replicates are variable, even using the same platform, as the figures clearly show. Look at the differences and see if any of them will change the overall results or discussion; for example, if high distribution occurs in different areas or important variables are now low on the list. These major differences do not often occur.

F) MaxEnt provides a projection of the model in the environmental layer that the background data is collected from.

F

Pictures of the model

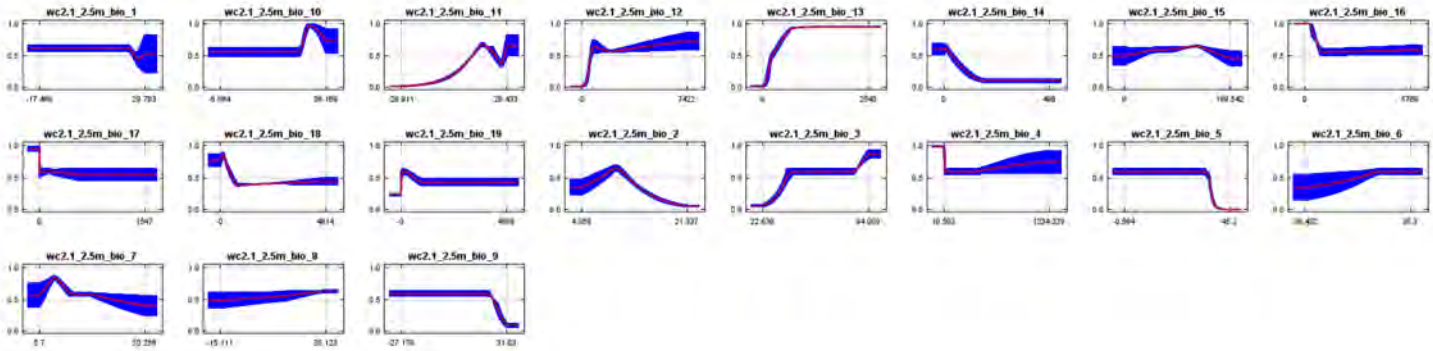
The following two pictures show the point-wise mean and standard deviation of the 3 output grids. Other available summary grids are min, max and median.



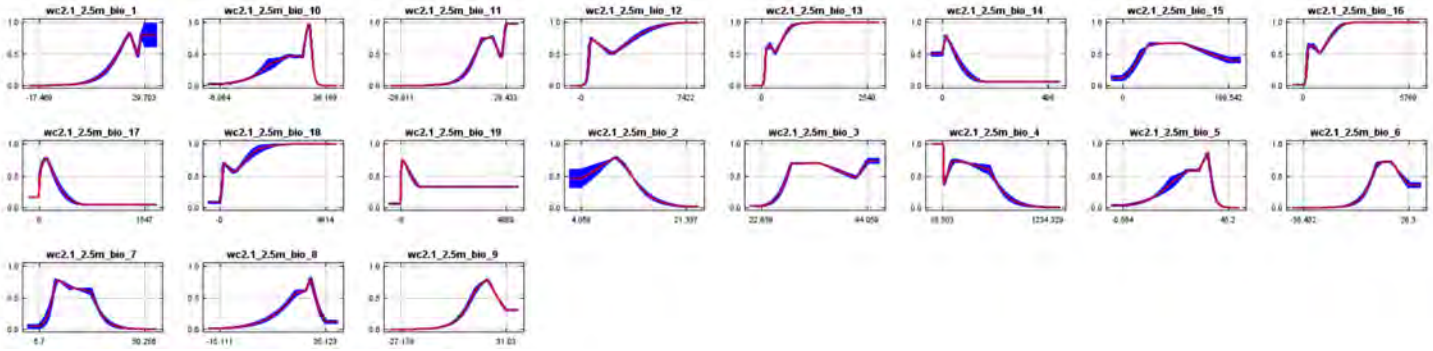
G) The response curves below show how each environmental variable (x axis) impacts the projected suitability (y axis) on a 0-1 scale with higher numbers representing greater suitability. The first set of curvesallows fro just the single variable to be varied while all others are maintained at their average sample value.

H) The second set of curves shows the response if only the single variable is used. This helps remove any biases with highly correlated variables.

G



H



I

Variable	Percent contribution	Permutation importance
wc2.1_2.5m_bio_12	18.9	8
wc2.1_2.5m_bio_5	12.7	7.4
wc2.1_2.5m_bio_6	12.6	1
wc2.1_2.5m_bio_2	8.1	7
wc2.1_2.5m_bio_14	7.7	4.6
wc2.1_2.5m_bio_10	6	7.6
wc2.1_2.5m_bio_16	4.6	18.3
wc2.1_2.5m_bio_11	4.4	5.3
wc2.1_2.5m_bio_18	4.3	4.6
wc2.1_2.5m_bio_13	4.1	18.7
wc2.1_2.5m_bio_19	3.3	1.9
wc2.1_2.5m_bio_4	3	0.5
wc2.1_2.5m_bio_9	2.5	4.8
wc2.1_2.5m_bio_3	2.4	4.1
wc2.1_2.5m_bio_15	1.7	0.6
wc2.1_2.5m_bio_7	1.3	2.4
wc2.1_2.5m_bio_17	1.3	1.4
wc2.1_2.5m_bio_1	1	1.9
wc2.1_2.5m_bio_8	0	0

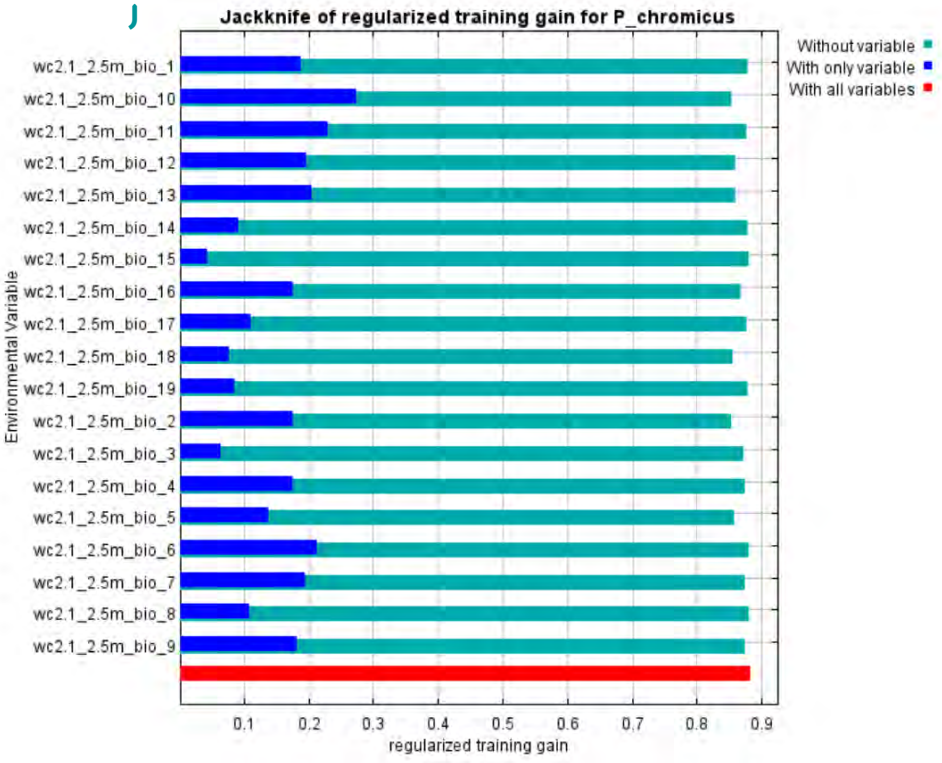
I) **The table shows the relative contributions of each variable to the model through a Jackknife test. This is calculated in two different ways.**

The first is the percent contribution, which is measured by the change in the regularization gain.

The permutation importance is measured by randomly permuting the training and background data and any difference in AUC is normalized to a percentage.

J

I) **The jackknife test results are visually representing in the next three graphs. We ill only look at the first graph. Variable importance is shown by the dark blue line, with only variable, and teal line, without variable. The further the blue line reaches, the more important that variable is. The shorter the teal line is, the more important the variable is.**



Removing Co-linear Variables

Pearson test in R

We will now use R to test if any of the bioclimatic variables are highly correlated with one another. This is done via the Pearson test.

1. **First, we will open the libraries that will allow us to run the Pearson test.**

```
#Libraries
library(ENMTools)
library(virtualspecies)
library(expss)
library(openxlsx)
```

2. **Next, we will set the working directory to the folder with the mskEnvs for the *P. chromatus.* These are the raster files limited to the areas where the background data was collected. Change the directory to where the file is, separating each subfolder by a /.**

```
#set working directory
setwd("C:/DIRECTORY)
```

3. **We will then upload the rasters and change the name to something easier to type. Note that the name in parenthesis is the name given by Wallace and the BioClim website download. If your files are named differently, change the name accordingly.**

   Additionally, the code below is for the ASCII files that we downloaded. Although R can handle these, the process is a lot slower than if the files are .tiff. If you saved the files as a .tiff, use those instead. Make sure that you name them with .tiff at the end rather than .asc so R can find the file.

```
#Load raster tif files
bio01 = raster("wc2.1_2.5m_bio_1.asc")
bio02 = raster("wc2.1_2.5m_bio_2.asc")
bio03 = raster("wc2.1_2.5m_bio_3.asc")
bio04 = raster("wc2.1_2.5m_bio_4.asc")
bio05 = raster("wc2.1_2.5m_bio_5.asc")
bio06 = raster("wc2.1_2.5m_bio_6.asc")
bio07 = raster("wc2.1_2.5m_bio_7.asc")
bio08 = raster("wc2.1_2.5m_bio_8.asc")
bio09 = raster("wc2.1_2.5m_bio_9.asc")
bio10 = raster("wc2.1_2.5m_bio_10.asc")
bio11 = raster("wc2.1_2.5m_bio_11.asc")
bio12 = raster("wc2.1_2.5m_bio_12.asc")
bio13 = raster("wc2.1_2.5m_bio_13.asc")
bio14 = raster("wc2.1_2.5m_bio_14.asc")
bio15 = raster("wc2.1_2.5m_bio_15.asc")
bio16 = raster("wc2.1_2.5m_bio_16.asc")
bio17 = raster("wc2.1_2.5m_bio_17.asc")
bio18 = raster("wc2.1_2.5m_bio_18.asc")
bio19 = raster("wc2.1_2.5m_bio_19.asc")
```

4. We will then create a raster stack, which combines all of the rasters into one data name so we can treat them the same.

```
bio_stack <- stack(bio01, bio02, bio03, bio04, bio05,
                   bio06, bio07, bio08, bio09, bio10,
                   bio11, bio12, bio13, bio14, bio15,
                   bio16, bio17, bio18, bio19)
```

5. Next, we will run the test on the entire stack so each variable is tested against the other. This code will produce to figures.

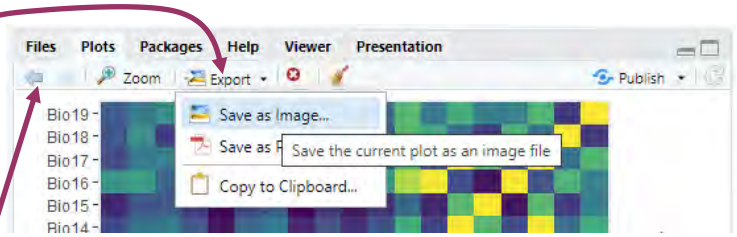
A) The first is an output of MSD space that shows the relatedness of the variables with variables that are more correlated located closer together. The results here are for *P. chromatus.*

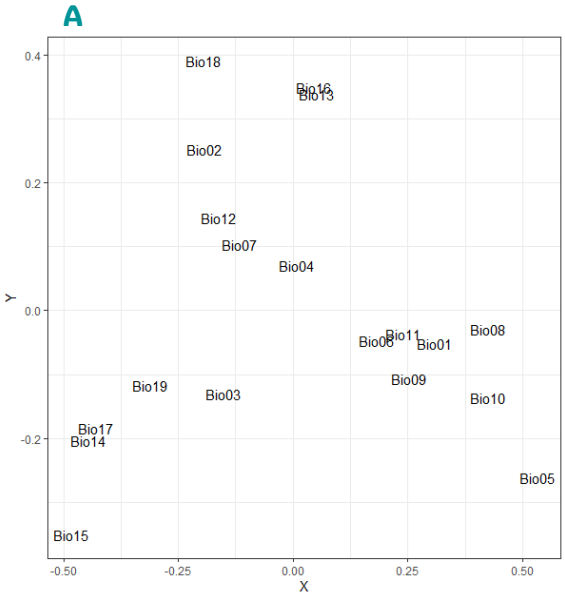
B) The second output is a grid with values showing how closely related each is. This is represented in colors and may be a nice figure to include in a Supplemental section of a paper. The results here are for *P. chromatus.*

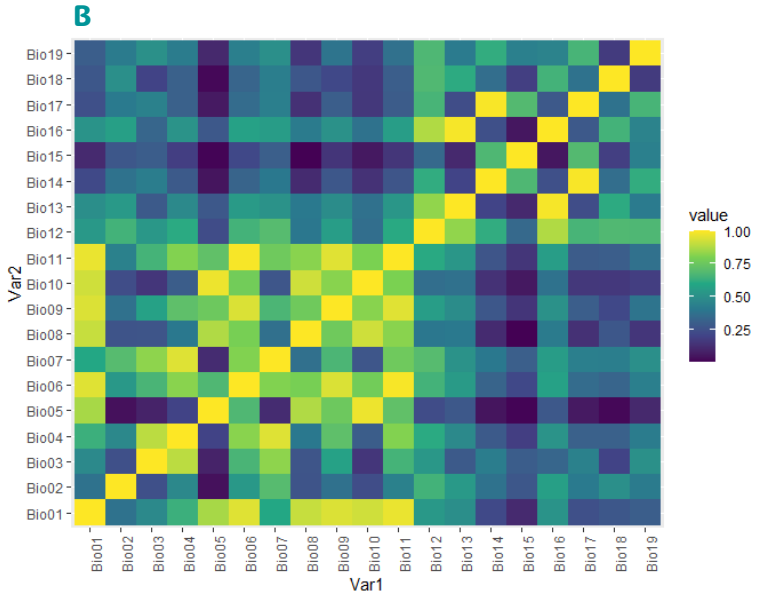
```
raster.cor.plot(bio_stack, method = "pearson")
```

6. Save the figures by pressing the "Export" button under the "Plots" tab. Then "Save as Image". Name your file and change the picture to as big as you want it. You will need to do this with each figure.

7. You can change between the figures using the arrow buttons.







User Insights

Any figures or data frames that you save will go to your working directory unless you change this during the save process.

You can change the names of the variables, the size of font, colors, ect. in R. Some of this will be covered with graph making in a later section, but we will not provide a comprehensive overview of this.

Resources

Helpful website: https://www.wallstreetmojo.com/pearson-correlation-coefficient/

7. We will calculate the Pearson test one more time, this time to obtain a data frame with actual numbers. We will then create a workbook book sheet and save this to our working directory.

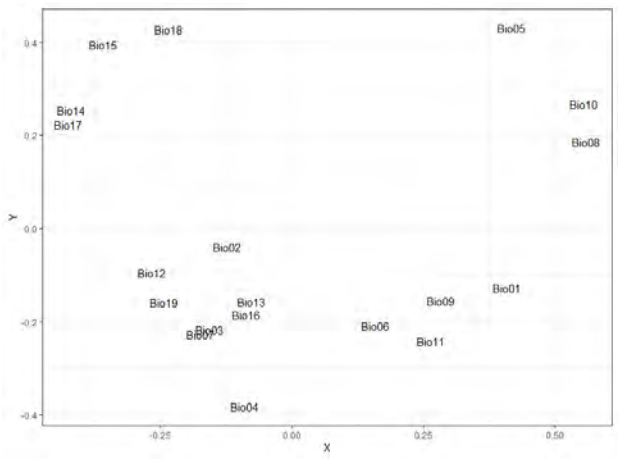
```
# Obtain a data frame containing the correlation coefficients of the raster stack
Bio_Pearson_df = raster.cor.matrix(bio_stack, method = "pearson")


#create workbook and worksheet in R
wb = createWorkbook()
sh = addWorksheet(wb, "Pearson")

xl_write(Bio_Pearson_df, wb, sh)


# this will save as a table that can be opened in excel
saveWorkbook(wb, "P chromicus Pearson data.csv", overwrite = TRUE)
```

8. Clear the workspace and rerun steps 1-7 with the *L. draconnis* data.

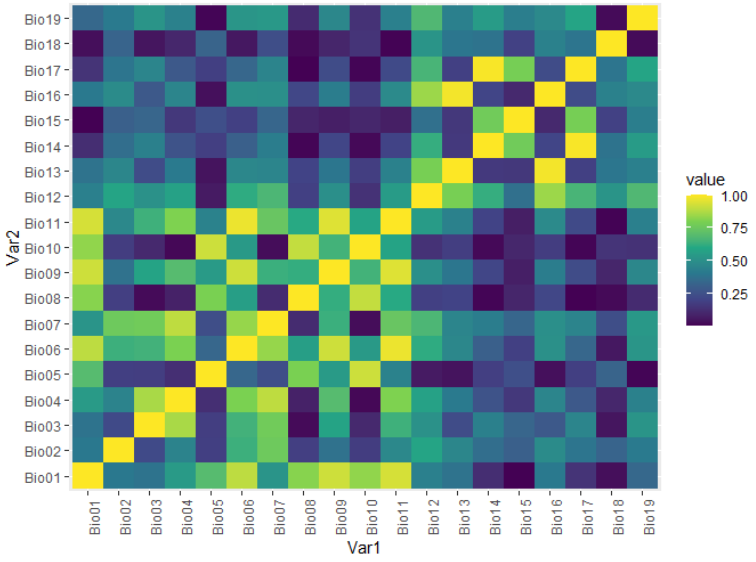
C) The MSD space figure for *L. draconnis.*

D) The visual representation of *L. draconnis* environment comparison.

C



D



9. Open the csv *P. chromicus* Pearson data that we just saved. Excel may give you a warning, but tell it to open the file.

E) The *P. chromicus* Pearson table. The cells show the Pearson results of each variable compared to one another (e.g., Bio06 compared to Bio16). The Pearson statistic is on a 0-1 scale with greater numbers indicating a higher correlation with each other. A value of 1 indicates that they are 100% correlated.

The first row designates what bioclimatic variable is used in the comparisons. The lower rows (2-20) include data that in the same order as that listed for the columns (e.g., cell A2 is Bio01 compared to Bio01, Cell G1-8 is Bio07 compared to Bio17).

There is a diagonal line where each cell has a value of 1. This makes sense as these are the cells where the variable is compared to itself. Each comparison is represented twice in this table, just in a different direction (e.g., Bio06 compared to Bio16 vs Bio16 compared to Bio06).

E

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Bio01	Bio02	Bio03	Bio04	Bio05	Bio06	Bio07	Bio08	Bio09	Bio10	Bio11	Bio12	Bio13	Bio14	Bio15	Bio16	Bio17	Bio18	Bio19
2	1	-0.3797365	0.474183599	-0.63664954	0.86147342	0.950011697	-0.5966092	0.908853524	0.942759441	0.924241281	0.967314605	0.5379045	0.49622347	0.22120005	-0.11132733	0.516320561	0.24716583	0.273669476	0.303121309
3	-0.3797365	1	-0.24994478	0.468696127	-0.03695448	-0.5357776	0.694380164	-0.26357309	-0.37682429	-0.23727381	-0.44008939	-0.65047589	-0.53895505	-0.38053321	0.27700214	-0.56551508	-0.40936659	-0.49767455	-0.41693401
4	0.474183599	-0.24994478	1	-0.89185585	0.087193649	0.663085643	-0.82631867	0.269063554	0.572948338	0.153630325	0.65425822	0.532733731	0.287652689	0.422790422	-0.29982316	0.334177639	0.442431135	0.20093746	0.504610076
5	-0.63664954	0.468696127	-0.89185585	1	-0.20025033	-0.81995571	0.94811009	-0.40207887	-0.7148757	-0.29704256	-0.81041161	-0.61198733	-0.47320369	-0.28708325	0.183221902	-0.51568379	-0.31304022	-0.31554869	-0.42185017
6	0.86147342	-0.03695448	0.087193649	-0.20025033	1	0.678135689	-0.11884947	0.875377025	0.757365501	0.973293998	0.721946764	0.234767579	0.279865939	0.046819264	0.011721678	0.277433465	0.060874576	0.019601364	0.106384549
7	0.950011697	-0.5357776	0.663085643	-0.81995571	0.678135689	1	-0.81032377	0.792071579	0.941511764	0.775264954	0.990687425	0.651788604	0.543123716	0.326495545	-0.21990766	0.574453031	0.356493603	0.330092934	0.432680768
8	-0.5966092	0.694380164	-0.82631867	0.94811009	-0.11884947	-0.81032377	1	-0.37214368	-0.66813335	-0.27136612	-0.76281072	-0.69339199	-0.51062585	-0.40377105	0.306445042	-0.55489177	-0.43309231	-0.43033266	-0.49973632
9	0.908853524	-0.26357309	0.269063554	-0.40207887	0.875377025	0.792071579	-0.37214368	1	0.764173722	0.92263405	0.820276955	0.398776981	0.405131674	0.114582803	-0.00490239	0.416955083	0.134169818	0.274632509	0.153702577
10	0.942759441	-0.37682429	0.572948338	-0.7148757	0.757365501	0.941511764	-0.66813335	0.764173722	1	0.818188069	0.950896931	0.556653272	0.485561473	0.275243359	-0.15025063	0.507944201	0.302251374	0.220229406	0.386140066
11	0.924241281	-0.23727381	0.153630325	-0.29704256	0.973293998	0.775264954	-0.27136612	0.92263405	0.818188069	1	0.799962591	0.365016983	0.377219056	0.139847193	-0.05859007	0.380872295	0.159025205	0.161038288	0.181037158
12	0.967314605	-0.44008939	0.65425822	-0.81041161	0.721946764	0.990687425	-0.76281072	0.820276955	0.950896931	0.799962591	1	0.608975635	0.529225821	0.265758708	-0.15180129	0.558329826	0.293890306	0.302590164	0.374613954
13	0.5379045	-0.65047589	0.532733731	-0.61198733	0.234767579	0.651788604	-0.69339199	0.398776981	0.556653272	0.365016983	0.608975635	1	0.830258535	0.621349262	-0.34426902	0.875940577	0.659453283	0.680863291	0.674673992
14	0.49622347	-0.53895505	0.287652689	-0.47320369	0.279865939	0.543123716	-0.51062585	0.405131674	0.485561473	0.377219056	0.529225821	0.830258535	1	0.204956603	0.10671648	0.987479251	0.238088218	0.614314881	0.414141687
15	0.22120005	-0.38053321	0.422790422	-0.28708325	0.046819264	0.326495545	-0.40377105	0.114582803	0.275243359	0.139847193	0.265758708	0.621349262	0.204956603	1	-0.67598743	0.247070403	0.992188552	0.363401021	0.623001811
16	-0.11132733	0.27700214	-0.29982316	0.183221902	0.011721678	-0.21990766	0.306445042	-0.00490239	-0.15025063	-0.05859007	-0.15180129	-0.34426902	0.10671648	-0.67598743	1	0.052313303	-0.68767311	-0.18477499	-0.43600316
17	0.516320561	-0.56551508	0.334177639	-0.51568379	0.277433465	0.574453031	-0.55489177	0.416955083	0.507944201	0.380872295	0.558329826	0.875940577	0.987479251	0.247070403	0.052313303	1	0.281956479	0.651577035	0.448902779
18	0.24716583	-0.40936659	0.442431135	-0.31304022	0.060874576	0.356493603	-0.43309231	0.134169818	0.302251374	0.159025205	0.293890306	0.659453283	0.238088218	0.992188552	-0.68767311	0.281956479	1	0.380934008	0.659211999
19	0.273669476	-0.49767455	0.20093746	-0.31554869	0.019601364	0.330092934	-0.43033266	0.274632509	0.220229406	0.161038288	0.302590164	0.680863291	0.614314881	0.363401021	-0.18477499	0.651577035	0.380934008	1	0.171752079
20	0.303121309	-0.41693401	0.504610076	-0.42185017	0.106384549	0.432680768	-0.49973632	0.153702577	0.386140066	0.181037158	0.374613954	0.674673992	0.414141687	0.623001811	-0.43600316	0.448902779	0.659211999	0.171752079	1

10. Although not necessary, we like to organize the file a little. Place a column to the left of Bio01 and label it with the 19 variables. As the numbers above the diagonal 1s show the same comparisons as those below, color in one of the sides with black to remove them from analysis.

F) The table after formatting it for better readability.

F

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		Bio01	Bio02	Bio03	Bio04	Bio05	Bio06	Bio07	Bio08	Bio09	Bio10	Bio11	Bio12	Bio13	Bio14	Bio15	Bio16	Bio17	Bio18	Bio19
2	Bio01	1																		
3	Bio02	-0.3797365	1																	
4	Bio03	0.474183599	-0.24994478	1																
5	Bio04	-0.63664954	0.468696127	-0.89185585	1															
6	Bio05	0.86147342	-0.03695448	0.087193649	-0.20025033	1														
7	Bio06	0.950011697	-0.5357776	0.663085643	-0.81995571	0.678135689	1													
8	Bio07	-0.5966092	0.694380164	-0.82631867	0.94811009	-0.11884947	-0.81032377	1												
9	Bio08	0.908853524	-0.26357309	0.269063554	-0.40207887	0.875377025	0.792071579	-0.37214368	1											
10	Bio09	0.942759441	-0.37682429	0.572948338	-0.7148757	0.757365501	0.941511764	-0.66813335	0.764173722	1										
11	Bio10	0.924241281	-0.23727381	0.153630325	-0.29704256	0.973293998	0.775264954	-0.27136612	0.92263405	0.818188069	1									
12	Bio11	0.967314605	-0.44008939	0.65425822	-0.81041161	0.721946764	0.990687425	-0.76281072	0.820276955	0.950896931	0.799962591	1								
13	Bio12	0.5379045	-0.65047589	0.532733731	-0.61198733	0.234767579	0.651788604	-0.69339199	0.398776981	0.556653272	0.365016983	0.608975635	1							
14	Bio13	0.49622347	-0.53895505	0.287652689	-0.47320369	0.279865939	0.543123716	-0.51062585	0.405131674	0.485561473	0.377219056	0.529225821	0.830258535	1						
15	Bio14	0.22120005	-0.38053321	0.422790422	-0.28708325	0.046819264	0.326495545	-0.40377105	0.114582803	0.275243359	0.139847193	0.265758708	0.621349262	0.204956603	1					
16	Bio15	-0.11132733	0.27700214	-0.29982316	0.183221902	0.011721678	-0.21990766	0.306445042	-0.00490239	-0.15025063	-0.05859007	-0.15180129	-0.34426902	0.10671648	-0.67598743	1				
17	Bio16	0.516320561	-0.56551508	0.334177639	-0.51568379	0.277433465	0.574453031	-0.55489177	0.416955083	0.507944201	0.380872295	0.558329826	0.875940577	0.987479251	0.247070403	0.052313303	1			
18	Bio17	0.24716583	-0.40936659	0.442431135	-0.31304022	0.060874576	0.356493603	-0.43309231	0.134169818	0.302251374	0.159025205	0.293890306	0.659453283	0.238088218	0.992188552	-0.68767311	0.281956479	1		
19	Bio18	0.273669476	-0.49767455	0.20093746	-0.31554869	0.019601364	0.330092934	-0.43033266	0.274632509	0.220229406	0.161038288	0.302590164	0.680863291	0.614314881	0.363401021	-0.18477499	0.651577035	0.380934008	1	
20	Bio19	0.303121309	-0.41693401	0.504610076	-0.42185017	0.106384549	0.432680768	-0.49973632	0.153702577	0.386140066	0.181037158	0.374613954	0.674673992	0.414141687	0.623001811	-0.43600316	0.448902779	0.659211999	0.171752079	1

Stats Chat

How correlated is too correlated? Different papers use different thresholds for "highly correlated", such as 70%, 75%, 80%, 90%. The higher the threshold is, the fewer variables need to be removed. However, you still risk some bias in your maps and response curves if you prune too lightly, but your model might not be informative if you prune too much. Keep this in mind when choosing the threshold.

R can provide a list that removes a highly correlated variable. Why are we not using this? Yes, R can very easily provide you a list with variables that are maintained after running a Pearson Test and providing a threshold. This will take moments whereas the methods we are doing will take a while.

However, R does not consider the biological or statistical importance of the variables when removing a correlated variable. It is our job as scientists to keep the biology in statistics. In this vignette, we are comparing each variable and selecting which one to keep based on the response curves, jackknife results, and biology of the species.

Remember: We are optimizing the model for our species, not just based on statistics.

Both of the highly correlated variables are important to my system. Can I keep both? Maybe. Think about why the variables are both important and highly correlated.

Are they important in just the Jackknife tests and is this importance overly generalized because of their correlation? If so, you probably should remove one.

Is one known to be important biologically and the other one important according the Jackknife test? You now have an argument for keeping both in the model, particularly if the value is close to the set threshold.

11. We now need to find any comparisons that are highly correlated. Lets use a Pearson value ≥ 0.8. Change the font color to red of any cells that are ≥ | 0.8 |. Note: this is the absolute value, so change the font whether the number is positive or negative.

G) The table after changing the text color for values ≥ 0.8. We're not going to worry about the values of 1 that are the variable compared to itself.

G

	Bio01	Bio02	Bio03	Bio04	Bio05	Bio06	Bio07	Bio08	Bio09	Bio10	Bio11	Bio12	Bio13	Bio14	Bio15	Bio16	Bio17	Bio18	Bio19
Bio01	1																		
Bio02	-0.3797365	1																	
Bio03	0.4741836	-0.2499448	1																
Bio04	-0.6366495	0.46869613	-0.8918558	1															
Bio05	0.86147342	-0.0369545	0.08719365	-0.2002503	1														
Bio06	0.9500117	-0.5357776	0.66308564	-0.8199557	0.67813569	1													
Bio07	-0.5966092	0.69438016	-0.8263187	0.94811009	-0.1188495	-0.8103238	1												
Bio08	0.90885352	-0.2635731	0.26906355	-0.4020789	0.87537703	0.79207158	-0.3721437	1											
Bio09	0.94275944	-0.3768243	0.57294834	-0.7148757	0.7573655	0.94151176	-0.6681333	0.76417372	1										
Bio10	0.92424128	-0.2372738	0.15363032	-0.2970426	0.973294	0.77526495	-0.2713661	0.92263405	0.81818807	1									
Bio11	0.96731461	-0.4400894	0.65425822	-0.8104116	0.72194676	0.99068743	-0.7628107	0.82027695	0.95089693	0.79996259	1								
Bio12	0.5379045	-0.6504759	0.53273373	-0.6119873	0.23476758	0.6517886	-0.693392	0.39877698	0.55665327	0.36501698	0.60897563	1							
Bio13	0.49622347	-0.538955	0.28765269	-0.4732037	0.27986594	0.54312372	-0.5106259	0.40513167	0.48556147	0.37721906	0.52922582	0.83025853	1						
Bio14	0.22120005	-0.3805332	0.42279042	-0.2870833	0.04681926	0.32649555	-0.403771	0.1145828	0.27524336	0.13984719	0.26575871	0.62134926	0.2049566	1					
Bio15	-0.1113273	0.27700214	-0.2998232	0.1832219	0.01172168	-0.2199077	0.30644504	-0.0049024	-0.1502506	-0.0585901	-0.1518013	-0.344269	0.10671648	-0.6759874	1				
Bio16	0.51632056	-0.5655151	0.33417764	-0.5156838	0.27743346	0.57445303	-0.5548918	0.41695508	0.5079442	0.3808723	0.55832983	0.87594058	0.98747925	0.2470704	0.0523133	1			
Bio17	0.24716583	-0.4093666	0.44243113	-0.3130402	0.06087458	0.3564936	-0.4330923	0.13416982	0.30225137	0.15902521	0.29389031	0.65945328	0.23808822	0.99218855	-0.6876731	0.28195648	1		
Bio18	0.27366948	-0.4976746	0.20093746	-0.3155487	0.01960136	0.33009293	-0.4303327	0.27463251	0.22022941	0.16103829	0.30259016	0.68086329	0.61431488	0.36340102	-0.184775	0.65157704	0.38093401	1	
Bio19	0.30312131	-0.416934	0.50461008	-0.4218502	0.10638455	0.43268077	-0.4997363	0.15370258	0.38614007	0.18103716	0.37461395	0.67467399	0.41414169	0.62300181	-0.4360032	0.44890278	0.659212	0.17175208	1

12. Analyze the tables to remove one of the variables for each comparison that is highly correlated.

Consider the Jackknife tests, response curves, and biology of the insect when choosing. You may also decide to remove one variable that is highly correlated with several other variables to maintain the most variables, particularly if it is not greatly important to the model.

Let's work through *P. chromicus* together.

13. First, let's look at the Jackknife results from both of the top models, H1 and H2.

We are looking at both models because the ranking can be very different, so it is good to take both into account. In this case, the order of importance is very similar, likely because both top models include just the hinge feature classes.

H) The Jackknife test results for the *P. chromicus* H1 model with variables listed in order of importance.

I) The Jackknife test results for the *P. chromicus* H2 model with variables listed in order of importance.

H

Variable	Percent contribution	Permutation importance
wc2.1_2.5m_bio_12	18.9	8
wc2.1_2.5m_bio_5	12.7	7.4
wc2.1_2.5m_bio_6	12.6	1
wc2.1_2.5m_bio_2	8.1	7
wc2.1_2.5m_bio_14	7.7	4.6
wc2.1_2.5m_bio_10	6	7.6
wc2.1_2.5m_bio_16	4.6	18.3
wc2.1_2.5m_bio_11	4.4	5.3
wc2.1_2.5m_bio_18	4.3	4.6
wc2.1_2.5m_bio_13	4.1	18.7
wc2.1_2.5m_bio_19	3.3	1.9
wc2.1_2.5m_bio_4	3	0.5
wc2.1_2.5m_bio_9	2.5	4.8
wc2.1_2.5m_bio_3	2.4	4.1
wc2.1_2.5m_bio_15	1.7	0.6
wc2.1_2.5m_bio_7	1.3	2.4
wc2.1_2.5m_bio_17	1.3	1.4
wc2.1_2.5m_bio_1	1	1.9
wc2.1_2.5m_bio_8	0	0

I

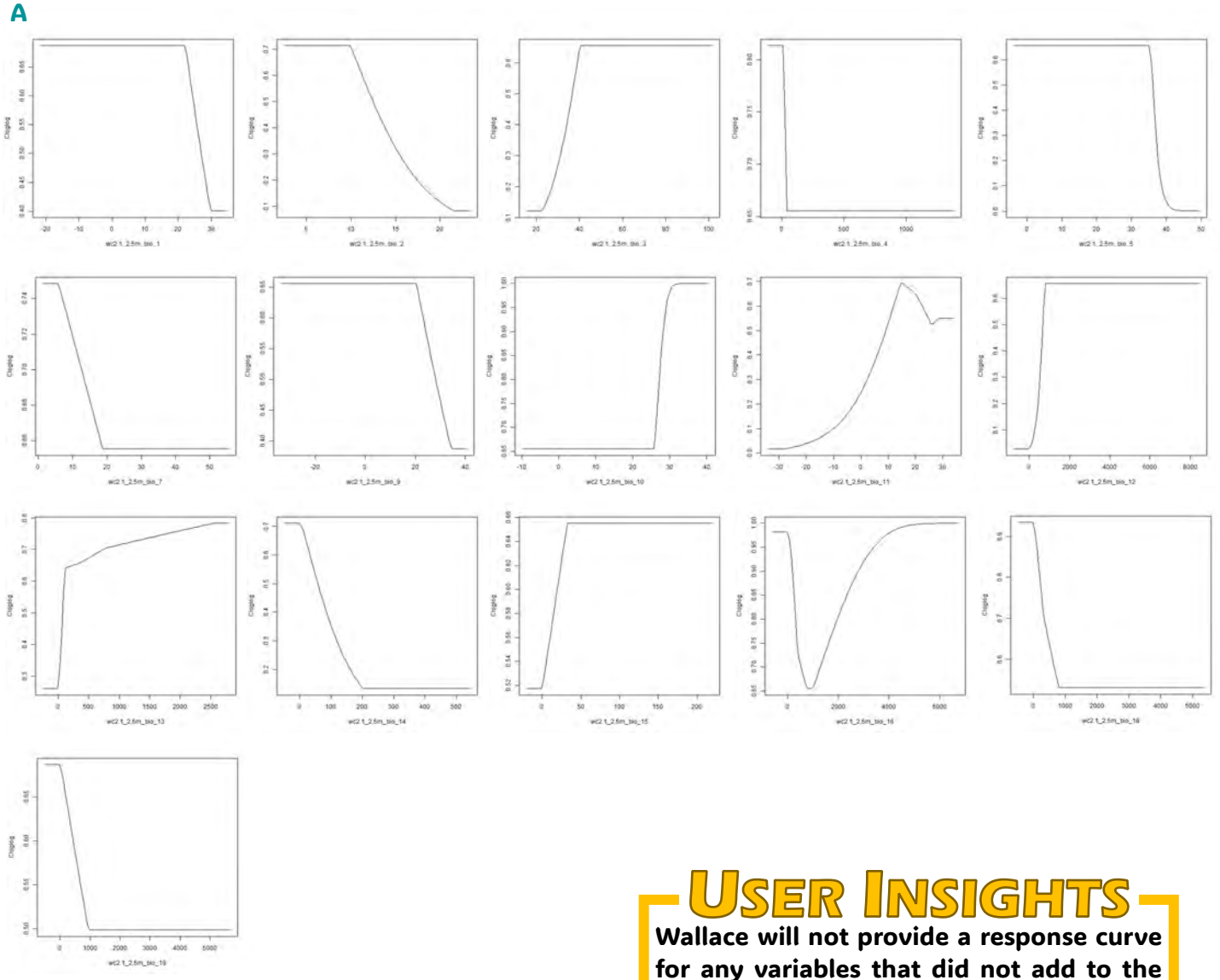
Variable	Percent contribution	Permutation importance
wc2.1_2.5m_bio_12	19.8	10.8
wc2.1_2.5m_bio_6	19.5	0
wc2.1_2.5m_bio_5	11.8	21.3
wc2.1_2.5m_bio_11	8	18.5
wc2.1_2.5m_bio_14	7.1	2
wc2.1_2.5m_bio_2	6.8	11
wc2.1_2.5m_bio_16	5.8	1.7
wc2.1_2.5m_bio_10	5.4	12.1
wc2.1_2.5m_bio_18	4	7.6
wc2.1_2.5m_bio_15	3.8	0.6
wc2.1_2.5m_bio_19	2.4	1.2
wc2.1_2.5m_bio_4	2	0.2
wc2.1_2.5m_bio_3	1.2	1.8
wc2.1_2.5m_bio_17	0.9	0.8
wc2.1_2.5m_bio_9	0.9	3.8
wc2.1_2.5m_bio_1	0.4	4.5
wc2.1_2.5m_bio_13	0.1	2.2
wc2.1_2.5m_bio_7	0.1	0
wc2.1_2.5m_bio_8	0	0

14. Look at the response variables from Wallace.

This gives us a better idea of its biological importance, rather than just its importance to the model. For example, a variable that greatly changes suitability is likely more important biologically than a variable that does not influence suitability. True, this is likely imparted in the Jackknife results, but this information can be useful when removing a variable and is important to think about for your results and discussion sections anyways.

A) The response curves for the top AUC model, H1. The response curves figures include an increase in suitability on the Y-axis (labeled Cloglog because we projected using a Cloglog algorithm) and an increase of the bioclimatic variable on the X-axis. Note that the axis scale is different dependent on the figure.

A

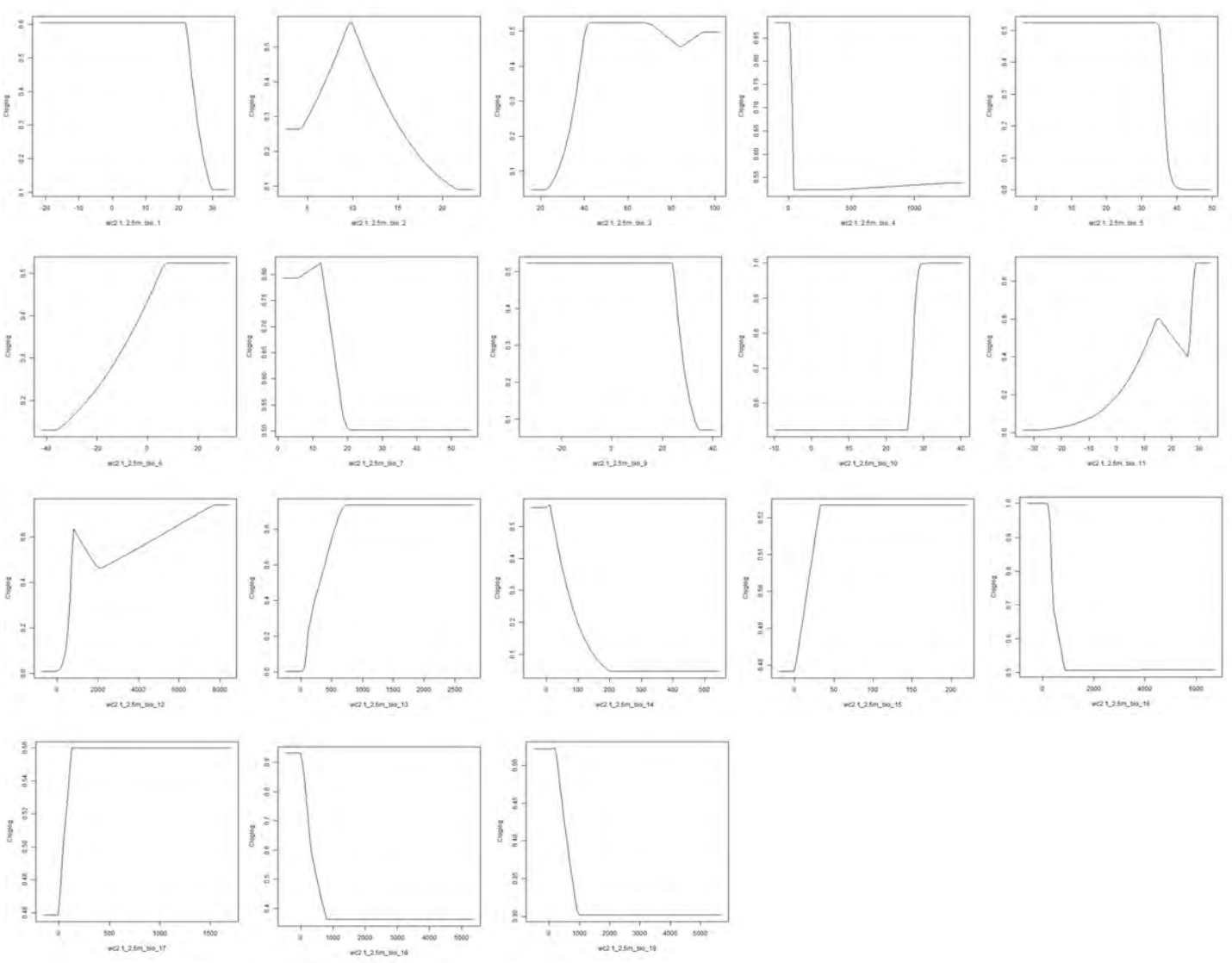


User Insights

Wallace will not provide a response curve for any variables that did not add to the model. This is why there are 16 curves instead of 19 for the H1 model.

B) The response curves for the top AIC model, H2. The response curves figures include an increase in suitability on the Y-axis (labeled Cloglog because we projected using a Cloglog algorithm) and an increase of the bioclimatic variable on the X-axis. Note that the axis scale is different dependent on the figure.

B



15. Let's work on the actual table. The first step is easy, mark the variables that are not highly correlated with any other variable. Fill the cells with a yellow color. Now we do not have to consider these with other comparisons.

J) Bioclimatic variables 02, 15, 18, and 19 are not highly correlated with any other variable. The cells are filled with yellow, so we will not look at them. The Bio labels are also yellow, indicating that we will keep these in the model.

J

	Bio01	Bio02	Bio03	Bio04	Bio05	Bio06	Bio07	Bio08	Bio09	Bio10	Bio11	Bio12	Bio13	Bio14	Bio15	Bio16	Bio17	Bio18	Bio19
Bio01	1																		
Bio02	-0.3797365	1																	
Bio03	0.4741836	-0.2499448	1																
Bio04	-0.6366495	0.46869613	-0.8918558	1															
Bio05	0.86147342	-0.0369545	0.08719365	-0.2002503	1														
Bio06	0.9500117	-0.5357776	0.66308564	-0.8199557	0.67813569	1													
Bio07	-0.5966092	0.69438016	-0.8263187	0.94811009	-0.1188495	-0.8103238	1												
Bio08	0.90885352	-0.2635731	0.26906355	-0.4020789	0.87537703	0.79207158	-0.3721437	1											
Bio09	0.94275944	-0.3768243	0.57294834	-0.7148757	0.7573655	0.94151176	-0.6681333	0.76417372	1										
Bio10	0.92424128	-0.2372738	0.15363032	-0.2970426	0.973294	0.77526495	-0.2713661	0.92263405	0.81818807	1									
Bio11	0.96731461	-0.4400894	0.65425822	-0.8104116	0.72194676	0.99068743	-0.7628107	0.82027695	0.95089693	0.79996259	1								
Bio12	0.5379045	-0.6504759	0.53273373	-0.6119873	0.23476758	0.6517886	-0.693392	0.39877698	0.55665327	0.36501698	0.60897563	1							
Bio13	0.49622347	-0.538955	0.28765269	-0.4732037	0.27986594	0.54312372	-0.5106259	0.40513167	0.48556147	0.37721906	0.52922582	0.83025853	1						
Bio14	0.22120005	-0.3805332	0.42279042	-0.2870833	0.04681926	0.32649555	-0.403771	0.1145828	0.27524336	0.13984719	0.26575871	0.62134926	0.2049566	1					
Bio15	-0.1113273	0.27700214	-0.2998232	0.1832219	0.01172168	-0.2199077	0.30644504	-0.0049024	-0.1502506	-0.0585901	-0.1518013	-0.344269	0.10671648	-0.6759874	1				
Bio16	0.51632056	-0.5655151	0.33417764	-0.5156838	0.27743346	0.57445303	-0.5548918	0.41695508	0.5079442	0.3808723	0.55832983	0.87594058	0.98747925	0.2470704	0.0523133	1			
Bio17	0.24716583	-0.4093666	0.44243113	-0.3130402	0.06087458	0.3564936	-0.4330923	0.13416982	0.30225137	0.15902521	0.29389031	0.65945328	0.23808822	0.99218855	-0.6876731	0.28195648	1		
Bio18	0.27366948	-0.4976746	0.20093746	-0.3155487	0.01960136	0.33009293	-0.4303327	0.27463251	0.22022941	0.16103829	0.30259016	0.68086329	0.61431488	0.36340102	-0.184775	0.65157704	0.38093401	1	
Bio19	0.30312131	-0.416934	0.50461008	-0.4218502	0.10638455	0.43268077	-0.4997363	0.15370258	0.38614007	0.18103716	0.37461395	0.67467399	0.41414169	0.62300181	-0.4360032	0.44890278	0.659212	0.17175208	1

16. Next, look at Bio01. It is highly correlated with 6 other bioclimatic variables. If it is not important to the model, we can remove it. Bio01 was ranked 18 out of 19 in the Jackknife test. Fill the cells with a grey color to show that we are removing it.

17. Let's look at the top ranked bioclimatic variables and make sure they are kept in the model. Bio12, Bio05, and Bio06 should be kept. Fill those with yellow.

18. An easy one to compare is Bio 17 vs 14. These are only highly variable with each other. Bio14 is ranked higher with the Jackknife test. Additionally, Bio17 was not included in the Wallace H1 model. So, we will keep Bio14. Fill the Bio14 cells with yellow and Bio17 with grey.

K) The table after steps 16, 17, and 19.

K

	Bio01	Bio02	Bio03	Bio04	Bio05	Bio06	Bio07	Bio08	Bio09	Bio10	Bio11	Bio12	Bio13	Bio14	Bio15	Bio16	Bio17	Bio18	Bio19
Bio01	1																		
Bio02	-0.3797365	1																	
Bio03	0.4741836	-0.2499448	1																
Bio04	-0.6366495	0.46869613	-0.8918558	1															
Bio05	0.86147342	-0.0369545	0.08719365	-0.2002503	1														
Bio06	0.9500117	-0.5357776	0.66308564	-0.8199557	0.67813569	1													
Bio07	-0.5966092	0.69438016	-0.8263187	0.94811009	-0.1188495	-0.8103238	1												
Bio08	0.90885352	-0.2635731	0.26906355	-0.4020789	0.87537703	0.79207158	-0.3721437	1											
Bio09	0.94275944	-0.3768243	0.57294834	-0.7148757	0.7573655	0.94151176	-0.6681333	0.76417372	1										
Bio10	0.92424128	-0.2372738	0.15363032	-0.2970426	0.973294	0.77526495	-0.2713661	0.92263405	0.81818807	1									
Bio11	0.96731461	-0.4400894	0.65425822	-0.8104116	0.72194676	0.99068743	-0.7628107	0.82027695	0.95089693	0.79996259	1								
Bio12	0.5379045	-0.6504759	0.53273373	-0.6119873	0.23476758	0.6517886	-0.693392	0.39877698	0.55665327	0.36501698	0.60897563	1							
Bio13	0.49622347	-0.538955	0.28765269	-0.4732037	0.27986594	0.54312372	-0.5106259	0.40513167	0.48556147	0.37721906	0.52922582	0.83025853	1						
Bio14	0.22120005	-0.3805332	0.42279042	-0.2870833	0.04681926	0.32649555	-0.403771	0.1145828	0.27524336	0.13984719	0.26575871	0.62134926	0.2049566	1					
Bio15	-0.1113273	0.27700214	-0.2998232	0.1832219	0.01172168	-0.2199077	0.30644504	-0.0049024	-0.1502506	-0.0585901	-0.1518013	-0.344269	0.10671648	-0.6759874	1				
Bio16	0.51632056	-0.5655151	0.33417764	-0.5156838	0.27743346	0.57445303	-0.5548918	0.41695508	0.5079442	0.3808723	0.55832983	0.87594058	0.98747925	0.2470704	0.0523133	1			
Bio17	0.24716583	-0.4093666	0.44243113	-0.3130402	0.06087458	0.3564936	-0.4330923	0.13416982	0.30225137	0.15902521	0.29389031	0.65945328	0.23808822	0.99218855	-0.6876731	0.28195648	1		
Bio18	0.27366948	-0.4976746	0.20093746	-0.3155487	0.01960136	0.33009293	-0.4303327	0.27463251	0.22022941	0.16103829	0.30259016	0.68086329	0.61431488	0.36340102	-0.184775	0.65157704	0.38093401	1	
Bio19	0.30312131	-0.416934	0.50461008	-0.4218502	0.10638455	0.43268077	-0.4997363	0.15370258	0.38614007	0.18103716	0.37461395	0.67467399	0.41414169	0.62300181	-0.4360032	0.44890278	0.659212	0.17175208	1

19. Look at what is highly correlated with the top variables. Which ones should be removed?

Bio16, precipitation of wettest quarter, and Bio13, precipitation of wettest month, are highly correlated with Bio12 and with each other. Should we remove both or just one? High amounts of precipitation is biologically important. Is there another variable that represents just high precipitation that we can keep? Thinking of our tropical and subtropical ranges, Bio18, precipitation of warmest quarter, likely represents the wet season and is already maintained in the model. Thus, high precipitation is covered with Bio18 as well as Bio12, annual precipitation. We will remove both Bio16 and Bio13.

Bio05 is the max temp. of warmest month and Bio06 is the minimum temp. of coldest month. They are highly correlated with several variables. Bio08 was not included in either model's response curves and is ranked last with the Jackknife test. Remove Bio08, mean temperature of wettest quarter.

Bio10 and Bio11 are > 97% correlated with Bio05 and Bio06. Remove Bio10 and Bio11.

The last three variables are all measurements of temperature variability. We need to keep one of these as a variability measurement has not been maintained yet, but it is biologically important. We can easily remove Bio07. This is highly correlated with Bio03, Bio04, Bio06.

The last variables are Bio03 and Bio04. These are highly correlated with each other. Bio4 is also correlated with Bio06. Looking at the response curves, an increase in Bio03 creates greater variability in suitability than Bio04. Due to both of these factors, we will remove Bio04.

L) Thus, the new optimized model will contain the variables 02, 03, 05, 06, 12, 14, 15, 18, and 19.

L

	Bio01	Bio02	Bio03	Bio04	Bio05	Bio06	Bio07	Bio08	Bio09	Bio10	Bio11	Bio12	Bio13	Bio14	Bio15	Bio16	Bio17	Bio18	Bio19
Bio01	1																		
Bio02	-0.3797365	1																	
Bio03	0.4741836	-0.2499448	1																
Bio04	-0.6366495	0.46869613	-0.8918558	1															
Bio05	0.86147342	-0.0369545	0.08719365	-0.2002503	1														
Bio06	0.9500117	-0.5357776	0.66308564	-0.8199557	0.67813569	1													
Bio07	-0.5966092	0.69438016	-0.8763187	0.94811009	-0.1188495	-0.8705238	1												
Bio08	0.90885352	-0.2635731	0.26906355	-0.4020789	0.87537703	0.79207158	-0.3721437	1											
Bio09	0.94275944	-0.3768243	0.57294834	-0.7148757	0.7573655	0.94151176	-0.6681333	0.76417372	1										
Bio10	0.92424128	-0.2372738	0.15363032	-0.2970426	0.973294	0.77526495	-0.2713661	0.92253405	0.81818807	1									
Bio11	0.96731461	-0.4400894	0.65425822	-0.8104116	0.72194676	0.99068743	-0.7628107	0.82027695	0.95089693	0.79996259	1								
Bio12	0.5379045	-0.6504759	0.53273373	-0.6119873	0.23476758	0.6517886	-0.693392	0.39877698	0.55665327	0.36501698	0.60897563	1							
Bio13	0.49622347	-0.538955	0.28765269	-0.4732037	0.27986594	0.54312372	-0.5106259	0.40513167	0.48556147	0.37721906	0.52922582	0.83025853	1						
Bio14	0.22120005	-0.3805332	0.42279042	-0.2870833	0.04681926	0.32649555	-0.403771	0.1145828	0.27524336	0.13984719	0.26575871	0.62134926	0.2049566	1					
Bio15	-0.1113273	0.27700214	-0.2998232	0.1832219	0.01172168	-0.2199077	0.30644504	-0.0049024	-0.1502506	-0.0585901	-0.1518013	-0.344269	0.10671648	-0.6759874	1				
Bio16	0.51632056	-0.5655151	0.33417764	-0.5156838	0.27743346	0.57445303	-0.5548918	0.41695508	0.5079442	0.3808723	0.55832983	0.87594058	0.98747925	0.2470704	0.0523133	1			
Bio17	0.24716583	-0.4093666	0.44243113	-0.3130402	0.06087458	0.3564936	-0.4330923	0.13416982	0.30225137	0.15902521	0.29389031	0.65945328	0.23808822	0.99218855	-0.6876731	0.28195648	1		
Bio18	0.27366948	-0.4976746	0.20093746	-0.3155487	0.01960136	0.33009293	-0.4303327	0.27463251	0.22022941	0.16103829	0.30259016	0.68086329	0.61431488	0.36340102	-0.184775	0.65157704	0.38093401	1	
Bio19	0.30312131	-0.416934	0.50461008	-0.4218502	0.10638455	0.43268077	-0.4997363	0.15370258	0.38614007	0.18103716	0.37461395	0.67467399	0.41414169	0.62300181	-0.4360032	0.44890278	0.659212	0.17175208	1

20. Remove colinear variables for *L. draconnis*. Your results may vary from our own due to your decisions. Our finished Pearson table will be on the next page.

M) The *L. draconnis* Pearson table pre-formatting.

M

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Bio01	Bio02	Bio03	Bio04	Bio05	Bio06	Bio07	Bio08	Bio09	Bio10	Bio11	Bio12	Bio13	Bio14	Bio15	Bio16	Bio17	Bio18	Bio19
2	1	-0.401722	0.3823168	-0.5415969	0.6925624	0.8943119	-0.5186799	0.8171049	0.9197801	0.829501	0.9333973	0.432248	0.3792983	0.1216903	-0.0010077	0.4060251	0.1406965	-0.0286409	0.3375436
3	-0.401722	1	-0.2213496	0.4486312	0.1821397	-0.6363058	0.7611291	-0.1802249	-0.3766443	-0.1754414	-0.4681421	-0.5831949	-0.4589496	-0.3615024	0.3055539	-0.4814879	-0.3868311	-0.3240709	-0.4098332
4	0.3823168	-0.2213496	1	-0.8638649	-0.1781253	0.6472519	-0.77007	0.0200461	0.5808769	-0.1052835	0.640733	0.5068541	0.230739	0.4354943	-0.3288003	0.2791153	0.4543258	0.046465	0.5191593
5	-0.5415969	0.4486312	-0.8638649	1	0.1248374	-0.7997882	0.8962051	-0.0819389	-0.692725	0.0143109	-0.8052486	-0.5713304	-0.4086604	-0.2557938	0.1557067	-0.4570234	-0.2795914	-0.0982455	-0.4368043
6	0.6925624	0.1821397	-0.1781253	0.1248374	1	0.3327264	0.2382892	0.7970513	0.5417674	0.9180801	0.4410173	-0.0592983	0.0402237	-0.1810024	0.2428054	0.0314562	-0.1815051	-0.3208258	0.0090216
7	0.8943119	-0.6363058	0.6472519	-0.7997882	0.3327264	1	-0.8365738	0.5599469	0.9206714	0.5405841	0.9709791	0.6155494	0.4635904	0.3079865	-0.1853609	0.5014857	0.3354802	0.0512352	0.5240898
8	-0.5186799	0.7611291	-0.77007	0.8962051	0.2382892	-0.8365738	1	-0.1136231	-0.633432	-0.0233696	-0.7437738	-0.6683873	-0.4540711	-0.4223412	0.3319572	-0.4981919	-0.4509483	-0.2391511	-0.5345047
9	0.8171049	-0.1802249	0.0200461	-0.0819389	0.7970513	0.5599469	-0.1136231	1	0.6229017	0.9049547	0.6065673	0.1849763	0.1967598	-0.0093997	0.0943452	0.2100104	0.0005593	0.0186124	0.1107588
10	0.9197801	-0.3766443	0.5808769	-0.692725	0.5417674	0.9206714	-0.633432	0.6229017	1	0.648031	0.9478866	0.4966839	0.3914388	0.2089822	-0.0711626	0.4221051	0.2317063	-0.0926183	0.465599
11	0.829501	-0.1754414	-0.1052835	0.0143109	0.9180801	0.5405841	-0.0233696	0.9049547	0.648031	1	0.5808956	0.1354725	0.1772779	-0.0161475	0.0965844	0.1748192	-0.0088519	-0.1422399	0.1348494
12	0.9333973	-0.4681421	0.640733	-0.8052486	0.4410173	0.9709791	-0.7437738	0.6065673	0.9478866	0.5808956	1	0.5460158	0.4384619	0.1985075	-0.0705004	0.4766705	0.2220905	0.0039244	0.4332538
13	0.432248	-0.5831949	0.5068541	-0.5713304	-0.0592983	0.6155494	-0.6683873	0.1849763	0.4966839	0.1354725	0.5460158	1	0.7894417	0.6252046	-0.3705139	0.8448806	0.6601921	0.5195595	0.6773055
14	0.3792983	-0.4589496	0.230739	-0.4086604	0.0402237	0.4635904	-0.4540711	0.1967598	0.3914388	0.1772779	0.4384619	0.7894417	1	0.1613859	0.1557171	0.9823359	0.1855011	0.3928805	0.4345288
15	0.1216903	-0.3615024	0.4354943	-0.2557938	-0.1810024	0.3079865	-0.4223412	-0.0093997	0.2089822	-0.0161475	0.1985075	0.6252046	0.1613859	1	-0.7703029	0.2037239	0.9916708	0.3825197	0.5470026
16	-0.0010077	0.3055539	-0.3288003	0.1557067	0.2428054	-0.1853609	0.3319572	0.0943452	-0.0711626	0.0965844	-0.0705004	-0.3705139	0.1557171	-0.7703029	1	0.1032535	-0.7873479	-0.1891415	-0.4238955
17	0.4060251	-0.4814879	0.2791153	-0.4570234	0.0314562	0.5014857	-0.4981919	0.2100104	0.4221051	0.1748192	0.4766705	0.8448806	0.9823359	0.2037239	0.1032535	1	0.2301675	0.440328	0.4714673
18	0.1406965	-0.3868311	0.4543258	-0.2795914	-0.1815051	0.3354802	-0.4509483	0.0005593	0.2317063	-0.0088519	0.2220905	0.6601921	0.1855011	0.9916708	-0.7873479	0.2301675	1	0.3876845	0.5830957
19	-0.0286409	-0.3240709	0.046465	-0.0982455	-0.3208258	0.0512352	-0.2391511	0.0186124	-0.0926183	-0.1422399	0.0039244	0.5195595	0.3928805	0.3825197	-0.1891415	0.440328	0.3876845	1	0.0204369
20	0.3375436	-0.4098332	0.5191593	-0.4368043	0.0090216	0.5240898	-0.5345047	0.1107588	0.465599	0.1348494	0.4332538	0.6773055	0.4345288	0.5470026	-0.4238955	0.4714673	0.5830957	0.0204369	1

N) The *L. draconnis* Pearson table after removing collinear variables. We will retain variables Bio02, Bio04, Bio05, Bio06, Bio08, Bio12, Bio13, Bio14, Bio15, Bio18, and Bio19.

Look to see which variables, if any, are different from the ones that we maintained. Why did you retain the ones that you did? Why do you think we retained the ones that you did? Do you think the differences will change the distribution plots?

N

	Bio01	Bio02	Bio03	Bio04	Bio05	Bio06	Bio07	Bio08	Bio09	Bio10	Bio11	Bio12	Bio13	Bio14	Bio15	Bio16	Bio17	Bio18	Bio19
Bio01	1																		
Bio02	-0.401722	1																	
Bio03	0.38231681	-0.2213496	1																
Bio04	-0.5415969	0.44863117	-0.8638649	1															
Bio05	0.69256245	0.18213972	-0.1781253	0.1248374	1														
Bio06	0.89431187	-0.6363058	0.64725192	-0.7997882	0.33272639	1													
Bio07	-0.5186799	0.76112907	-0.77007	0.89620511	0.23828917	-0.8365738	1												
Bio08	0.81710488	-0.1802249	0.02004615	-0.0819389	0.79705129	0.55994687	-0.1136231	1											
Bio09	0.91978014	-0.3766443	0.58087691	-0.692725	0.54176741	0.92067139	-0.633432	0.62290165	1										
Bio10	0.82950098	-0.1754414	-0.1052835	0.01431093	0.9180801	0.54058406	-0.0233696	0.90495472	0.64803099	1									
Bio11	0.93339733	-0.4681421	0.64073304	-0.8052486	0.44101734	0.97097906	-0.7437738	0.60656729	0.94788662	0.58089557	1								
Bio12	0.43224799	-0.5831949	0.50685409	-0.5713304	-0.0592983	0.61554936	-0.6683873	0.18497633	0.49668394	0.13547253	0.54601576	1							
Bio13	0.37929833	-0.4589496	0.230739	-0.4086604	0.0402237	0.46359045	-0.4540711	0.19675978	0.39143884	0.17727791	0.43846194	0.78944173	1						
Bio14	0.12169034	-0.3615024	0.43549426	-0.2557938	-0.1810024	0.30798651	-0.4223412	-0.0093997	0.20898219	-0.0161475	0.19850752	0.62520456	0.16138591	1					
Bio15	-0.0010077	0.30555385	-0.3288003	0.15570674	0.24280541	-0.1853609	0.33195717	0.09434517	-0.0711626	0.09658443	-0.0705004	-0.3705139	0.15571715	-0.7703029	1				
Bio16	0.40602506	-0.4814879	0.27911525	-0.4570234	0.03145618	0.50148569	-0.4981919	0.21001037	0.42210509	0.17481922	0.47667047	0.84468057	0.98233597	0.20372386	0.10325347	1			
Bio17	0.14069646	-0.3868311	0.45432583	-0.2795914	-0.1815051	0.33548024	-0.4509483	0.00055929	0.23170632	-0.0088519	0.22209052	0.66019207	0.18550114	0.99167078	-0.7873479	0.23016747	1		
Bio18	-0.0286409	-0.3240709	0.046465	-0.0982455	-0.3208258	0.05123519	-0.2391511	0.01861244	-0.0926183	-0.1422399	0.00392437	0.51955949	0.39288053	0.38251971	-0.1891415	0.44032801	0.38768447	1	
Bio19	0.33754356	-0.4098332	0.51915931	-0.4368043	0.00902164	0.52408977	-0.5345047	0.1107588	0.46559902	0.1348494	0.43325385	0.67730551	0.43452882	0.54700263	-0.4238955	0.47146727	0.58309572	0.02043692	1

The Wallace Platform, Back Again

Wallace settings, Components 1-8

We now go back to Wallace to create the final distribution maps. We will use the same settings as before, but the main difference is which bioclimatic variables are uploaded. We will only upload the variables that were retained after the Pearson method.

Here is a short reminder of the methods used.

Component 1:

→ **Upload the .csv files for one of the species. We suggest doing each species in separate Wallace sessions.**

Component 2:

→ **Upload the global rasters that were downloaded from WorldClim. Select just the rasters of the bioclimatic variables that were retained after the Pearson test.**

Component 3:

→ **Remove the N. American data point for the *L. draconnis* data.**

→ **Thin the data points by 10 degrees.**

Component 4:

→ **Create a 10 degree point buffer around the data points.**

→ **Collect 50,000 background points.**

Component 5:

→ **Environmental space can be skipped. We are modeling only one species.**

Component 6:

→ **Partition the data spatially with Block = 4 method.**

Component 7:

→ **Model using MaxEnt, feature classes L, H, LQ, LQH, RM 1-5, step size 1, Clamping = TRUE, Parallel = FALSE**

Component 8:

→ **Plot the top model for your species.**

→ **Save the response curves and MaxEnt plots**

After running the model with only the retained variables, the top models for *P. chromicus* were H1 and LQH1.

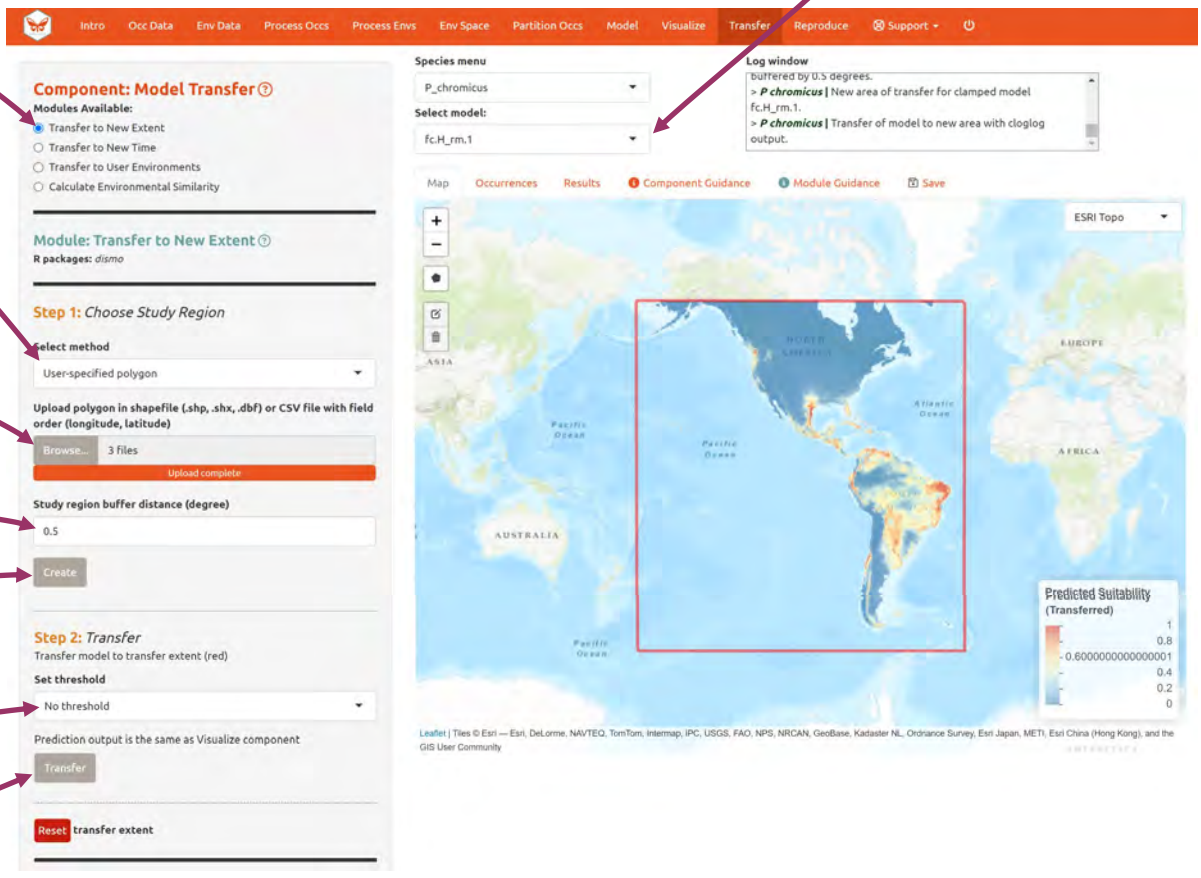
Component 9:

We now can view the model by transferring it to other places and times. We will do this with the top two models for both species. The models that we are working with are the models that only include the bioclimatic variables retained after the Pearson tests.

A. First, we need to transfer to a new extent.

Although we are interested in the global potential distribution, we need to break up our transfers into at least three pieces. If the extent is too large, the Wallace platform will freeze. We will first transfer to the America continents.

1. Click "Transfer to New Extent".
2. Select the H1 model.
3. Select "User-specific polygon".
   (Or you can draw the polygon yourself and save the polygon).
4. Upload the three files associated with the Americas.
5. Add a 0.5 buffer.
6. Click "Create".



Wallace provides a few transfer extents. For instance, results can be mapped by quantile. We will use the no threshold option, which produces a map on a continuous scale.

7. Select "No threshold".
8. Click "Transfer".

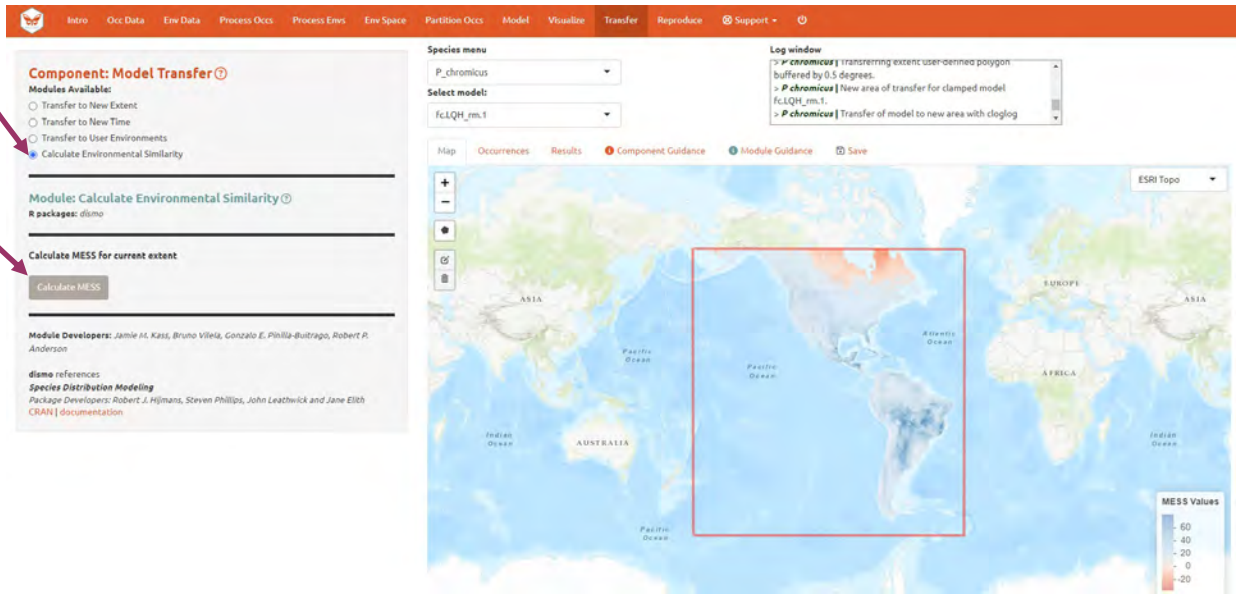
We also need to obtain the multivariate environmental similarity surface (MESS). This is a calculation of the similarity of the environments between the sampled and projected areas. The MESS map provided shows pixels that are similar as positive values and pixels that are dissimilar as negative values. We will eventually use this map to cover the areas that are dissimilar, removing them from analysis and any visual interpretation.

9. Click "Calculate Environmental Safety".

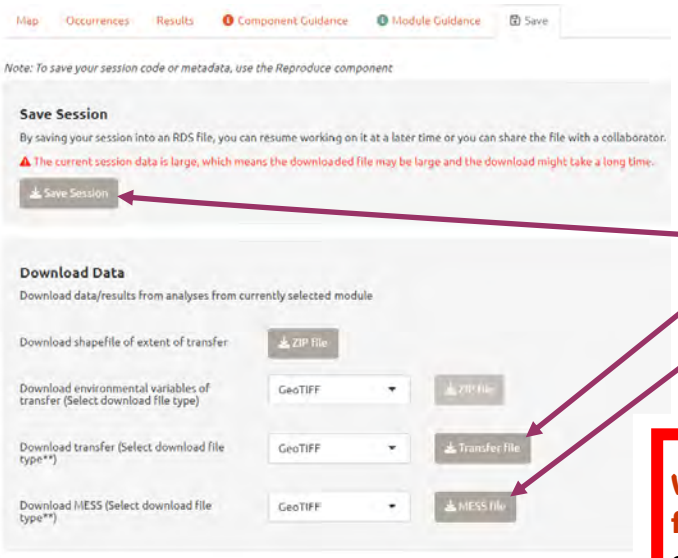
10. Click "Calculate MESS".

B. The Americas study region reaches 60°, -60°, 175°, -30° to the N, S, W, and E, respectively.

B



C



C. Make sure to save the files of the transfers. If you make another model, the previous ones will be removed.

11. Click the "Save" tab.

12. Download the Transfer file as a GeoTIFF.

13. Download the MESS as a GeoTIFF.

User Insights

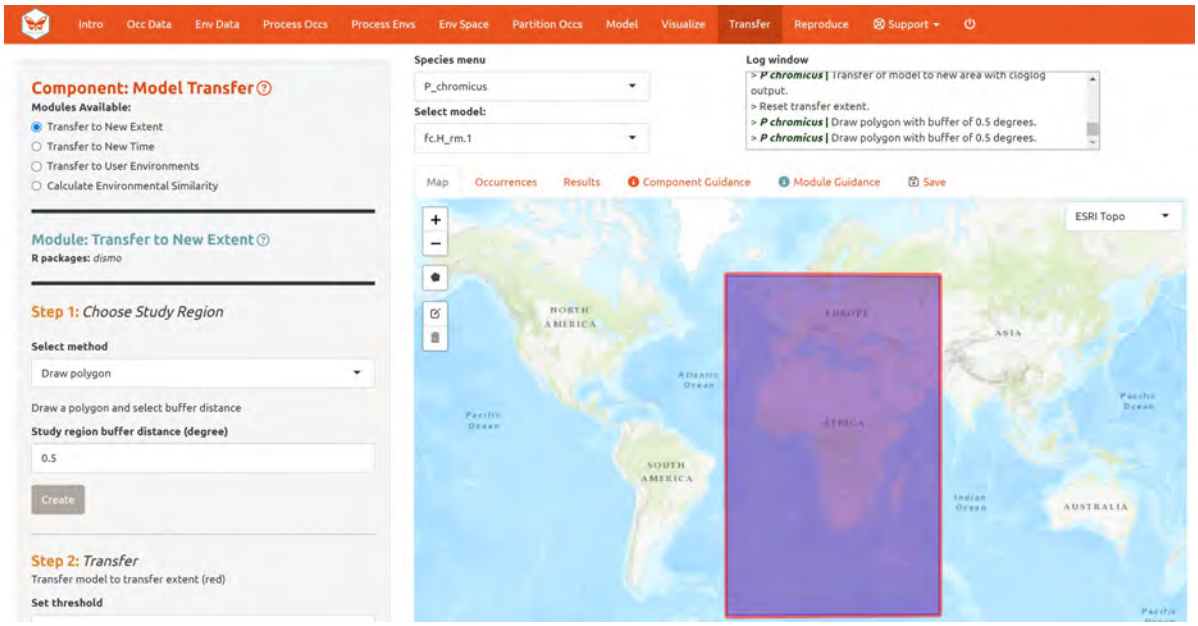
Go to your downloads folder and rename your files and place them in appropriate folders as soon as you download them. Files will automatically be named the same with the download, so they will be easy to confuse if you do not organize. Save each type of map in its own file (e.g., the 3 H1 distribution maps in one file).

Stop and Think!

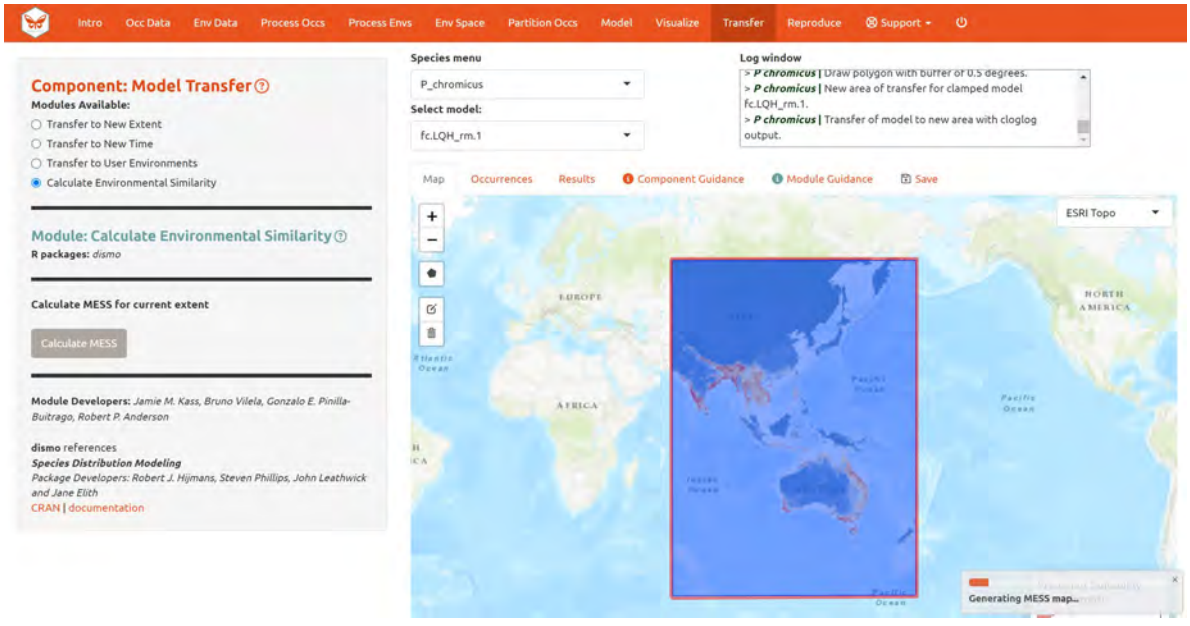
Why are we creating a MESS map? Why not use the full distribution map? Although we have used methods to help limit bias in sampling, e.g., thinning, and extrapolation, e.g., clamping, the model is still creating projections that can be over-fitted and over-generalized. This is particularly the case in novel areas that are not like those from the background extent taken in Component 4. The MESS lets the user know which areas are outside of the known conditions and, thus, what the model is uncertain of. If we assess these uncertain areas, we would over-reach our conclusions and provide biologically and statistically inaccurate statements.

14. Repeat with the other top models LQH1 and H2.
15. Reset the transfer extent by pressing the red "Reset".
16. Download the three files to create the African study region or draw your own polygon. If you draw, save the polygon.
17. Transfer to the new extent and create the MESS layer for H1, LQH1, and H2. Make sure to save between each model.
18. Repeat steps 16 and 17 with the Asia/Oceania study region.

D



E



RESOURCES

Simões et al. 2020. https://doi.com/10.17161/bi.v15i2.13376

Elith et al. 2010. https://doi.org/10.1111/j.2041-210X.2010.00036.x

D. The African/Europe study region reaches 60°, -60°, -30°, and 65° to the N, S, W, and E, respectively.

E. The Asian/Oceania study region reaches 60°, -60°, 65°, and 175° to the N, S, W, and E, respectively.

Initial maps in R:

We need to visually assess each of the top models to see if the top models are biologically sound. For instance, if a map has a good potential range in a more arctic region, then the model is likely not a best fit.

1. We will first load the libraries and set a working directory.

```
# Load the libraries
library(ggplot2)      ## Go to for plotting
library(dplyr)        ## Manipulates data structures
library(sp)           ## Needed to change raster file types
library(terra)        ## Works with rasters
library(sf)           ## Needed to change raster file types

# Set working directory
setwd("C:/WORKINGDIRECTORY")  ## This is an individualized file path to the folders where you have saved everything.
```

2. Next, we will merge the three different rasters from Wallace (Americas, Africa/Europe, Asia/Oceania) into one global raster. To do this, we make a mosaic. Any pixels that overlap will be averaged together.

```
# Import all files in a single folder as a list.
LIST = list.files(path = "H1/2020 Raster/",
                  pattern='tif$',                          # This will call all of the .tif or .tiff files from the folder. Make sure
                  full.names = TRUE, all.files = TRUE)     # That the only rasters in this folder are for the raster you are making.

# Merge rasters into a mosaic

MERGED = sprc(LIST)
M_RASTER = mosaic(MERGED)

# Save the mosaic as a tiff file in your working directory

writeRaster(M_RASTER, "H1 2020 P chomicus.tiff")
```

3. Repeat for the H1 MESS layer. Save the file as "H1 2020 MESS P chromicus.tiff".

4. Repeat for the *P. chromicus* LQH1 distribution. Save the file as "LQH1 2020 P chromicus.tiff".

5. Repeat for the *P. chromicus* LQH1 MESS layer. Save the file as "LQH1 2020 MESS P chromicus.tiff".

6. Repeat for the *P. chromicus* H2 distribution. Save the file as "H2 2020 P chromicus.tiff".

7. Repeat for the *P. chromicus* H2 MESS layer. Save the file as "H2 2020 MESS P chromicus.tiff".

6. Now that we have the four layers saved, we will make a quick map to visually assess the potential distribution of the top two models.

```
# Upload the distribution layer.

DISTMAP = rast("H1/H1 2020 P chomicus.tiff")          # File name and (if needed) directory to the tif map file
DISTMAP_df = as.data.frame(DISTMAP, xy = TRUE)        # Convert to a data frame for making a map in R

# Make the map

FILL_LAYER = DISTMAP_df[,3]   # This states that the potential distribution values are in the third column of the dataframe

ggplot(DISTMAP_df, aes(x = x, y = y)) +
  geom_raster(aes(fill=FILL_LAYER), show.legend = TRUE) +    # We can remove the legend by changing TRUE to FALSE
  scale_fill_viridis_c(na.value="white") +                   # The na.values are pixels without a distribution, such as over water
  theme_void() +                                             # The theme changes the formatting of the map, such as borders
  coord_quickmap()
```

7. Repeat step 6 for the LQH1 2020 P chromicus.tiff.

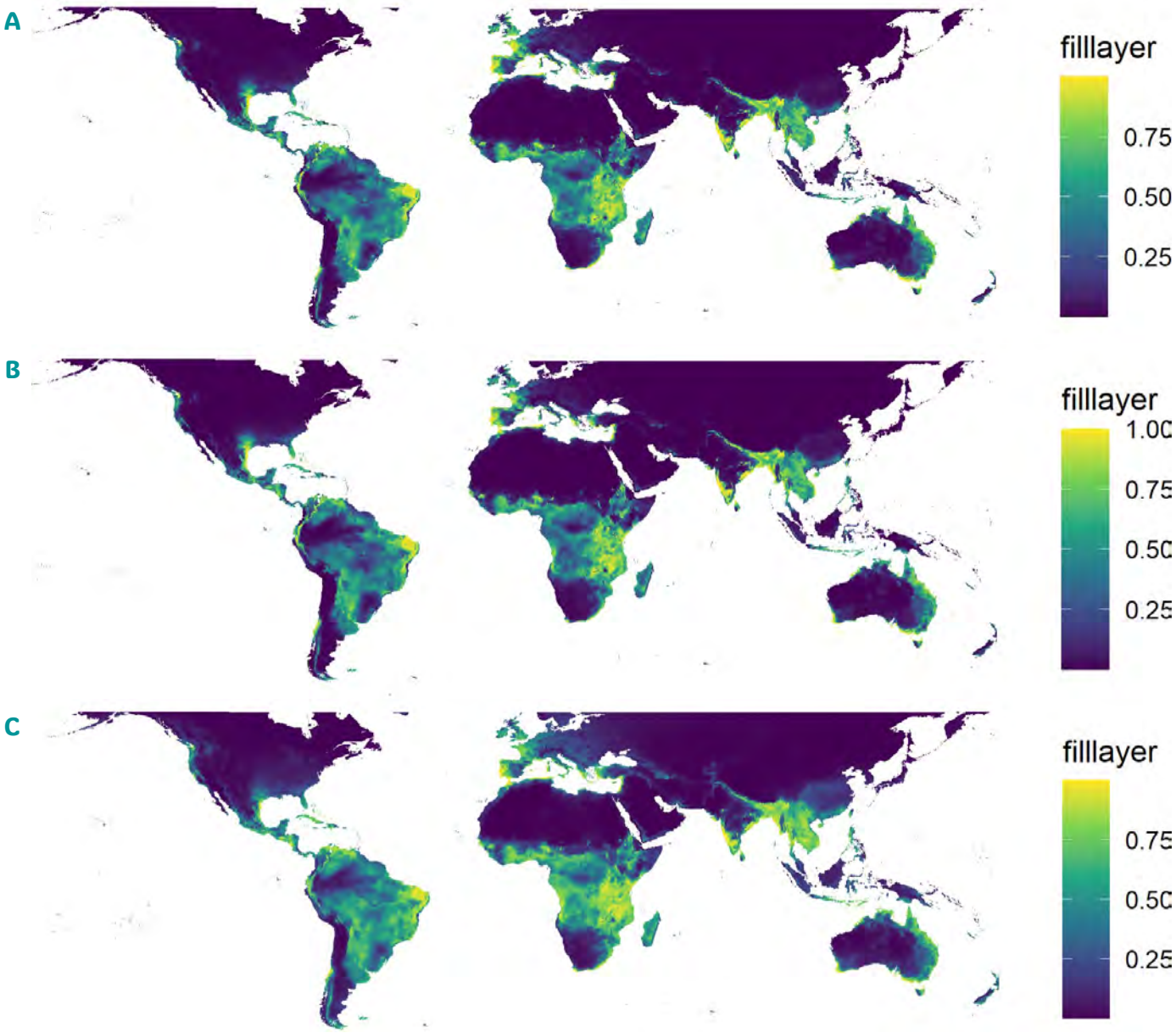
8. Repeat step 6 for the H2 2020 P chromicus.tiff.

9. Compare the three maps.

A) H1 global distribution map for *P. chromicus*.

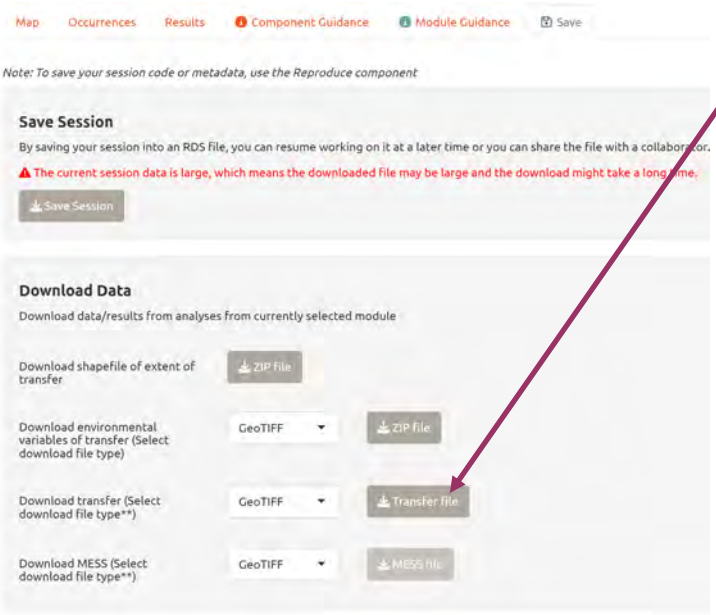
B) LQH1 global distribution map for *P. chromicus*.

C) H2 global distribution map for *P. chromicus*.



We need to choose just one of the models to project into the future and publish. What differences do you see between the maps? There are very few differences. This shows that our models are robust.

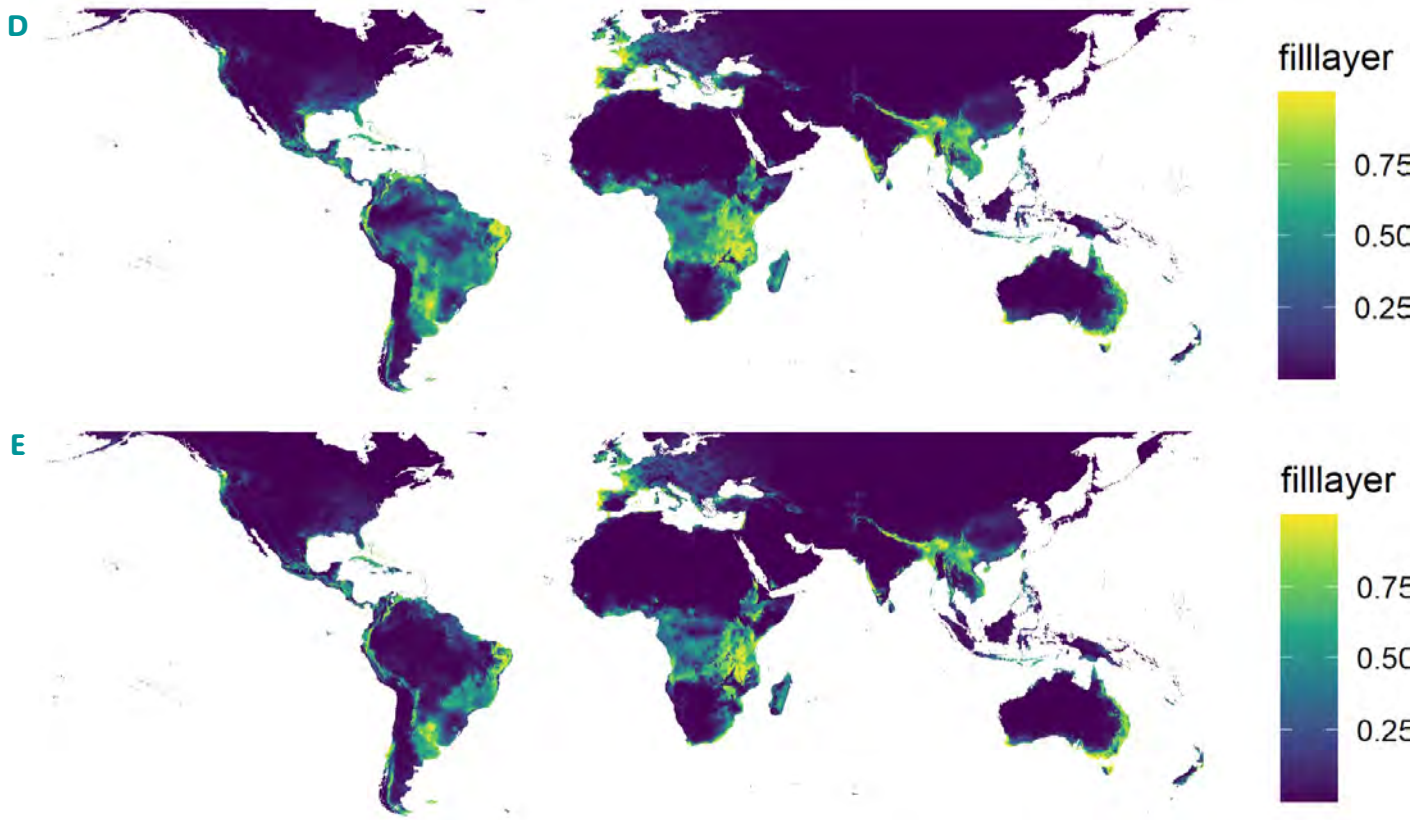
We will proceed with the H1 model. It was selected as the top AUC model and the second AICc model.



11. **Save the transfer tiff file using the "Save" tab. Make sure to rename the downloaded file.**
12. **Repeat steps 1-11 with the SSP 126.**
13. **Press the red "Reset" button to change the transfer extent.**
14. **Upload the files for the Africa and Europe transfer extent.**
15. **Repeat steps 1-11 with the SSP of 585 and then SSP 126.**
16. **Press the red "Reset" button to change the transfer extent.**
17. **Upload the files for the Asia and Oceania transfer extent.**
18. **Repeat steps 1-11 with the SSP of 585 and then SSP 126.**

D) *P. chromicus* H1 model projected to SSP126 (low climate change) 2061-2080 using the global circulation model CNRM-ESM2-1.

E) *P. chromicus* H1 model projected to SSP585 (high climate change) 2061-2080 using the global circulation model CNRM-ESM2-1.

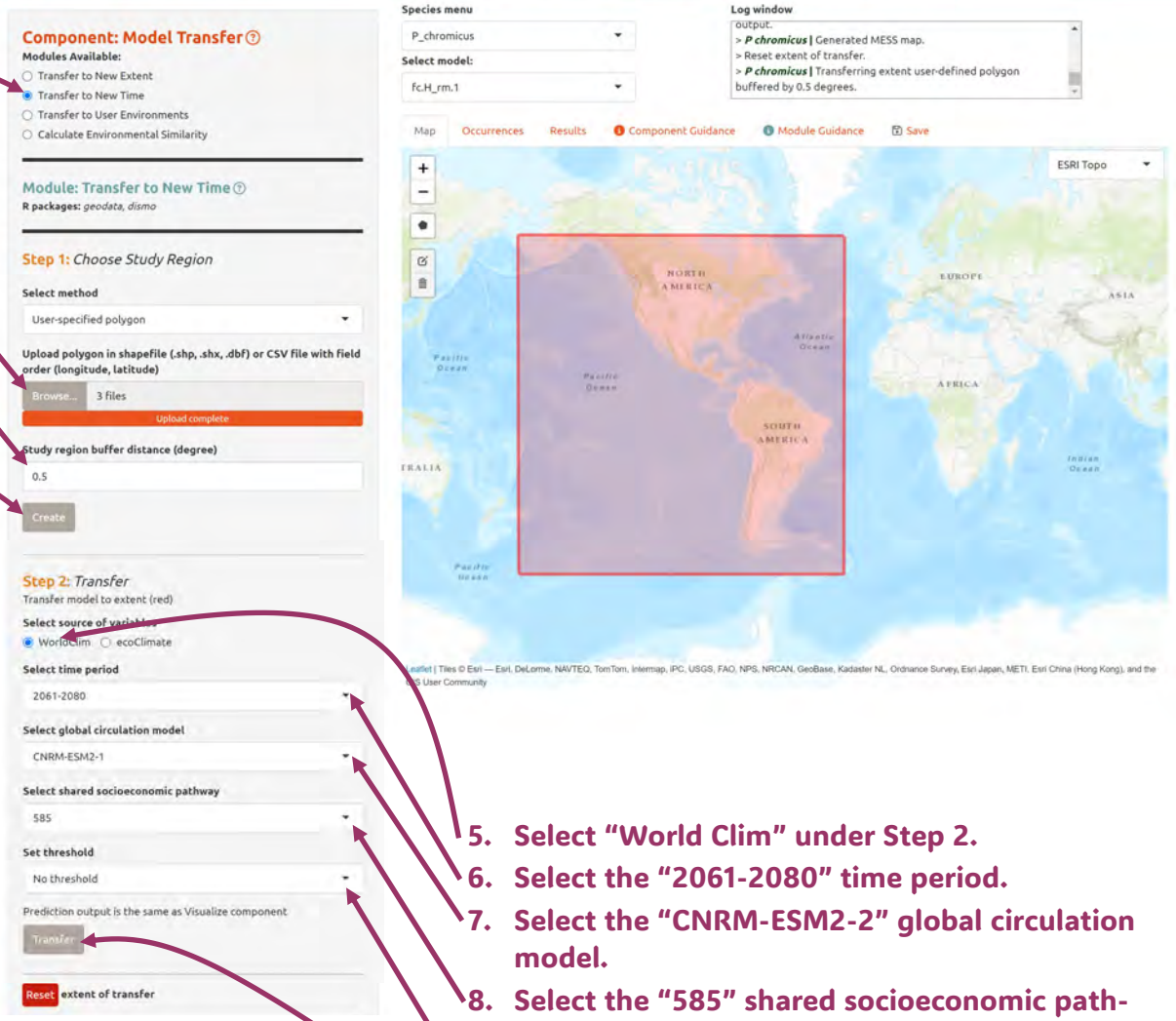


Component 9: Model Transfer

We will now model the future potential distributions under two climate change scenarios. There are many different global circulation models to select from within the Module "Transfer to New Time". We will use The CNRM-ESM2-1 model. This model is the second generation of the CNRM-CM6-1 model. The model incorporates several earth system components, interaction of atmospheric chemistry, aerosols, and land and carbon cycles.

We will project to high (SSP 585) and low (SSP 126) climate scenarios for the 2061-2080 time period.

1. **Click "Transfer to New Time".**
2. **Select the three files associated with the study region for the Americas region.**
3. **Create a study region buffer of 0.5.**
4. **Click "Create".**



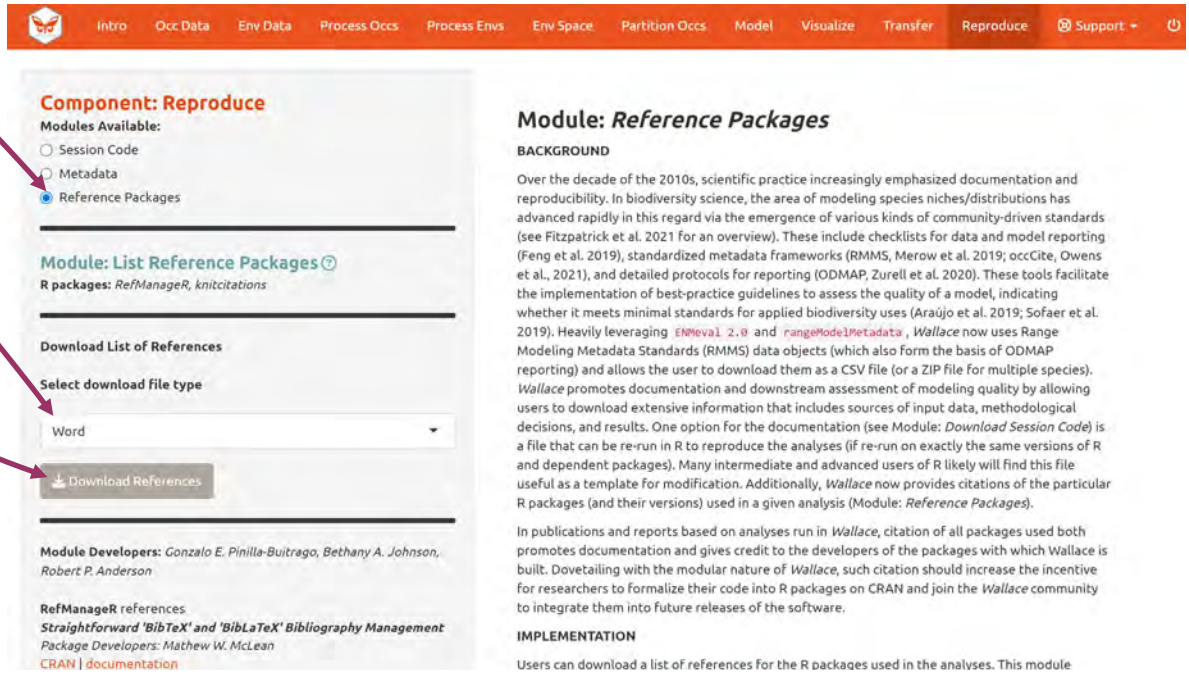
5. **Select "World Clim" under Step 2.**
6. **Select the "2061-2080" time period.**
7. **Select the "CNRM-ESM2-2" global circulation model.**
8. **Select the "585" shared socioeconomic pathway (SSP).**
9. **Select "No threshold".**
10. **Click "Transfer".**

Component 10: Reproduce

The last component provides information that you need for any published article. Wallace provides the R session Code that is nicely annotated. It also provides the references for each of the packages that were used during the session. For publications, you should provide the code as an online supplemental material and cite the references in the text and in the reference section.

1. Click the "Session Code" module.
2. Select the Rmd file type.
3. Click "Download the Session Code".

1. Select the "Reference Packages".
2. Select the "Word" file type (or other if you prefer).
3. Click "Download References".



Publication Quality Maps in R

Code for distribution map

Next, we will upload the global rasters to R to create publication quality maps. We will use color schemes that are accessible to those that are color blind.

1. We will first load the libraries and set a working directory. If your R session is still open from before, this can be skipped.

```
# Open Libraries if Needed

# Load the libraries
library(ggplot2)      ## Go to library for plotting
library(dplyr)        ## Manipulates data structures
library(sp)           ## Changes raster file types
library(terra)        ## Works with rasters
library(sf)           ## Changes raster file types

# Set working directory
setwd("C:/WORKINGDIRECTORY")  ## This is an individualized file path to the folders where you have saved everything.
```

2. Use the code from the previous module to combine the rasters downloaded from the future projections. This code uses the global rasters.

3. Upload the MESS layer and one of the distribution maps. We will then change the MESS layer from continuous to a binary layer that separates the similar and dissimilar environments (value = 0).

We show the code for the current distribution maps.

```
# Obtain a continuous mess layer and distribution layer from computer

CONTMESS = rast("H1/H1 MESS P chomicus.tiff")  # File name and (if needed) directory to the tif map file
CONTMESS_df = as.data.frame(CONTMESS, xy = TRUE)   # Convert to data frames for making a map in R

DISTMAP = rast("H1/H1 2020 P chomicus.tiff")
DISTMAP_df = as.data.frame(DISTMAP, xy = TRUE)

# Change continuous mess layer to binary

## This one will be used in code. Creates a column with TRUE or FALSE per pixel
MESSOV = (CONTMESS < 0)                    # Overlay areas from MESS map (not similar)
MESSOV_df = as.data.frame(MESSOV, xy = TRUE)

MESSREM = (CONTMESS > 0)
MESSREM_df = as.data.frame(MESSREM, xy = TRUE)
```

4. Set the colors and transparency values for the MESS map.

```
# Set color scheme for overlay map
MESSOV_df$color = ifelse(MESSOV_df[,3] == "TRUE", "grey", "black")
col = as.character(MESSOV_df$color)
names(col) = as.character(MESSOV_df$color)

# Set alpha level or transparency level for overlay
MESSOV_df$c_alpha = ifelse(MESSOV_df[,3] == "TRUE", 1, 0)
```

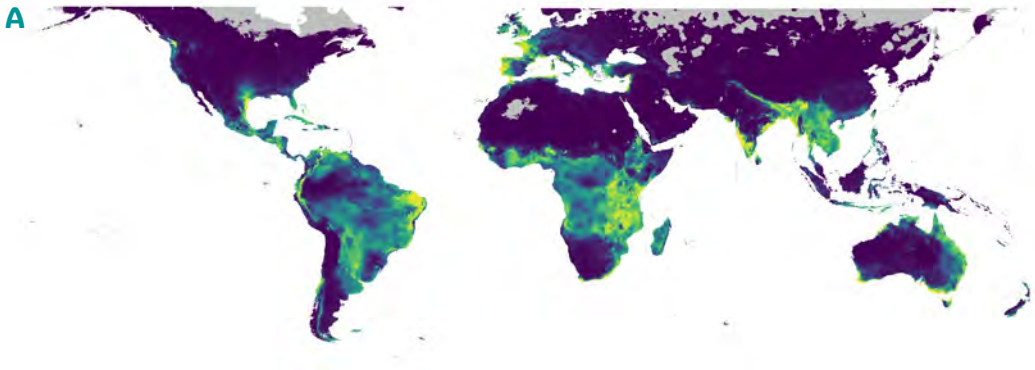
5. We will make the map. This code includes more specifics than the previous code that we used to make maps.

```
# Make the map

filllayer = DISTMAP_df[,3]     # This designates the third column of the data frame as the suitability levels

PLOT = ggplot(DISTMAP_df, aes(x = x, y = y)) +
    geom_raster(aes(fill=filllayer), show.legend = FALSE) +
    scale_fill_viridis_c(na.value="white") +                    # Change here for ocean NA values
    annotate(geom="raster", x=MESSOV_df$x, y=MESSOV_df$y, alpha=(MESSOV_df$c_alpha), fill = (col))+
    theme_void() +
    coord_quickmap()

# Show the map plot in the Plot tab
PLOT
```

6. Next, save the plot as a high quality tif file.

```
# Save file as high quality

tiff('P chromicus 2020 H1 map.tif', units = "in", width = 6, height = 2,
     res = 600, compression = "lzw")
PLOT                # This is calling all of the code written above to make the plot
dev.off()
```

A) *P. chromicus* H1 current 2020 potential distribution map.



We now save raster files to use in QGIS. These include files of just the MESS, the remaining area without the MESS, and areas of high distribution. To use these rasters in QGIS, we need file types that only include the pixels associated for each of the categories. At the moment, the rasters include NA values associated with water and FALSE values that represent the rest of the landmass.

7. Create rasters of just the subset of the data for the MESS layer and the remaining pixels.

```
## These will be saved and used in QGIS. It includes only the TRUE pixels, the other had FALSE and NA.

MESSONLY_df = subset(MESSOV_df, MESSOV_df[,3] == "TRUE")
MESSONLY_dfr = rast(MESSONLY_df)  # Transforms a data frame back to a raster by designating x and y coordinates

MESSREMONLY_df = subset(MESSREM_df, MESSOV_df[,3] == "FALSE")
MESSREMONLY_dfr = rast(MESSREMONLY_df)
```

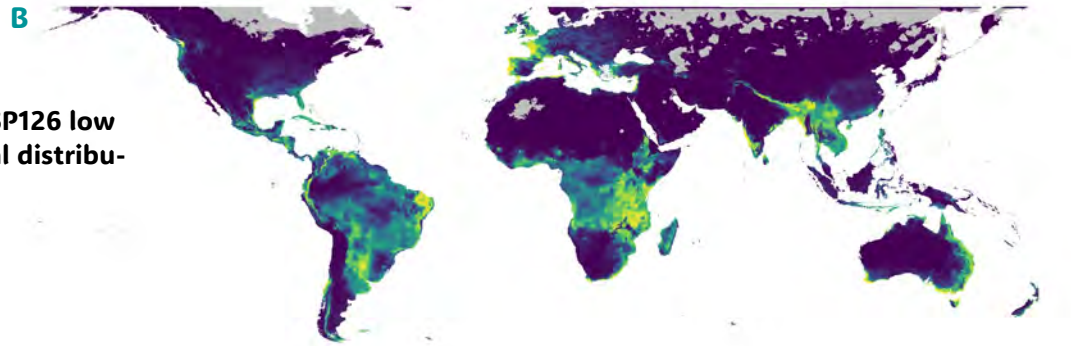
8. Select the pixels associated with high distribution (>75% suitability) and then create a subset raster of just those pixels.

```
HIGHDIST = (DISTMAP > 0.75)
HIGHDIST_df = as.data.frame(HIGHDIST, xy = TRUE)

## Includes only the TRUE pixels, the other had FALSE and NA

HIGHDIST_ONLY_df = subset(HIGHDIST_df, HIGHDIST_df[,3] == "TRUE")
HIGHDIST_ONLY_dfr = rast(HIGHDIST_ONLY_df)
```

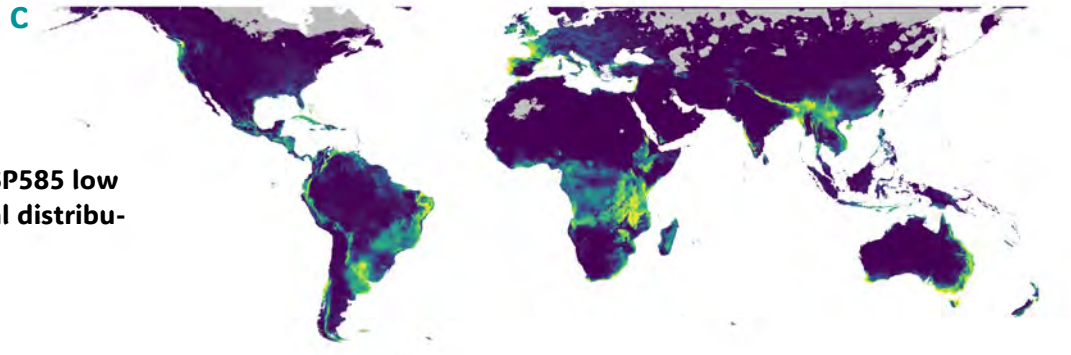
9. Save the rasters of just the subset data as .tiff files.

```
# Save the different layers as a tiff map each. These will be placed in the directory folder

writeRaster(MESSONLY_dfr, "H1 MESS Only.tiff")
writeRaster(MESSREMONLY_dfr, "H1 MESS Remaining Only.tiff")
writeRaster(HIGHDIST_ONLY_dfr, "2020 high dist raster.tiff")
```

10. Repeat steps 3 - 10 with the rasters representing the future projections.



B) ***P. chromicus*** **H1 2070 SSP126 low climate change potential distribution map.**



C) ***P. chromicus*** **H1 2070 SSP585 low climate change potential distribution map.**

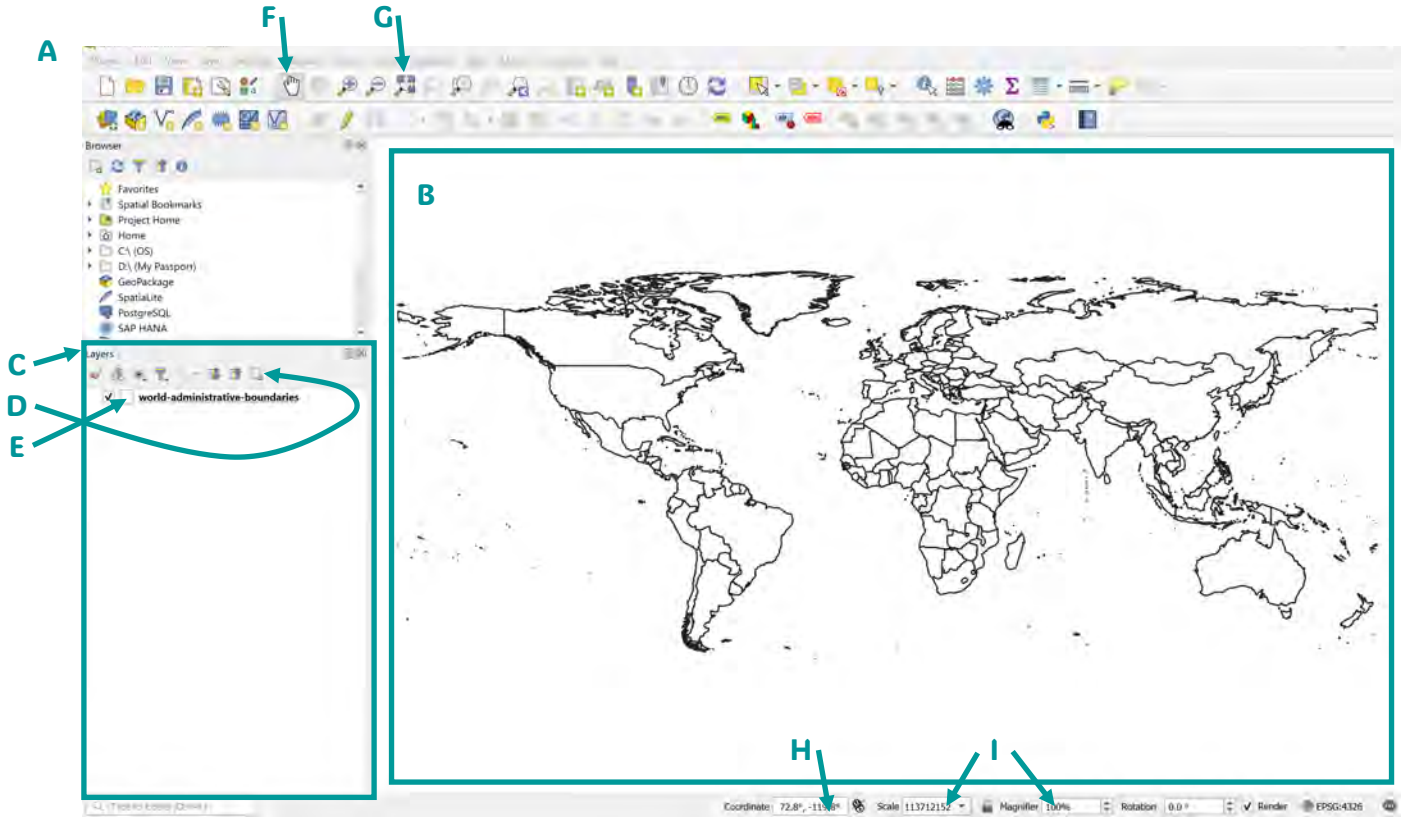
Publication Quality Maps in QGIS

QGIS layout

QGIS is an excellent, free platform that allows users to manipulate rasters. It is a lot easier to do some functions in QGIS than it is in R. During this module we will upload the raster layers that we created in the last module, transform them from rasters to polygons, and perform some raster maps by removing pixels and creating centroids.

The QGIS platform has several buttons and drop down menus. We will cover some of the basics here to help with general use of the program.

A) The QGIS display after uploading a layer file.
B) The main window is where layers of maps will appear when they are uploaded. This one happens to have a boundary layout of the countries already uploaded. The website for this download is https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/export/?flg=en-us. The shapefile is used in the window.
C) The Layers window is where rasters and other files can be uploaded and appear after they are created. Layers can be added from saved files by just dropping them into this window.
D) Files can be deleted using this icon.
E) The square icon beside the file name shows the color of the layer. Double-clicking on this provides a settings window to manipulate the layer. This includes changing the colors and/or shape.
F) This icon allows the user to move the map around freely with the mouse.
G) This is a useful icon that will center the map in the main window.
H) Coordinates of the mouse pointer are noted below the main window.
I) Changes in magnification and rotation can be altered below the main window.

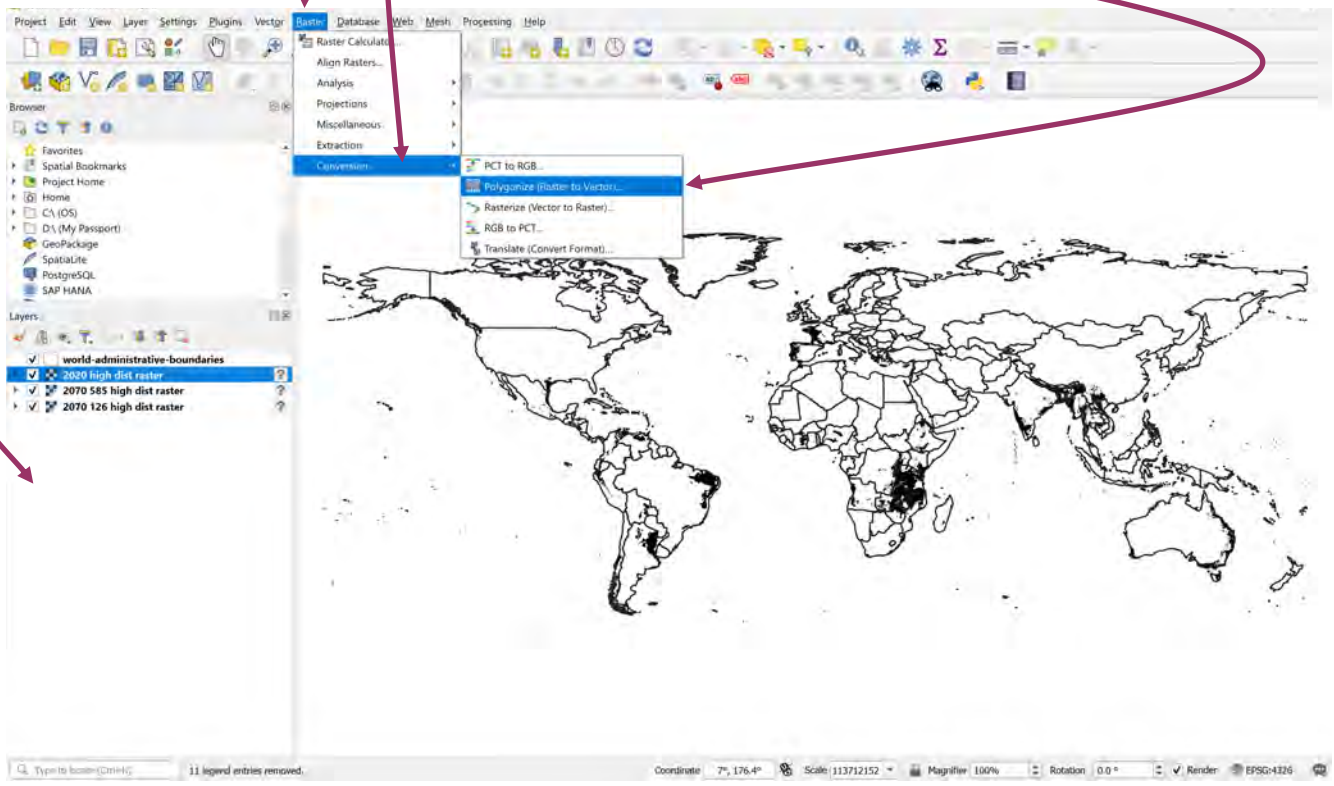


Making high distribution maps with centroids

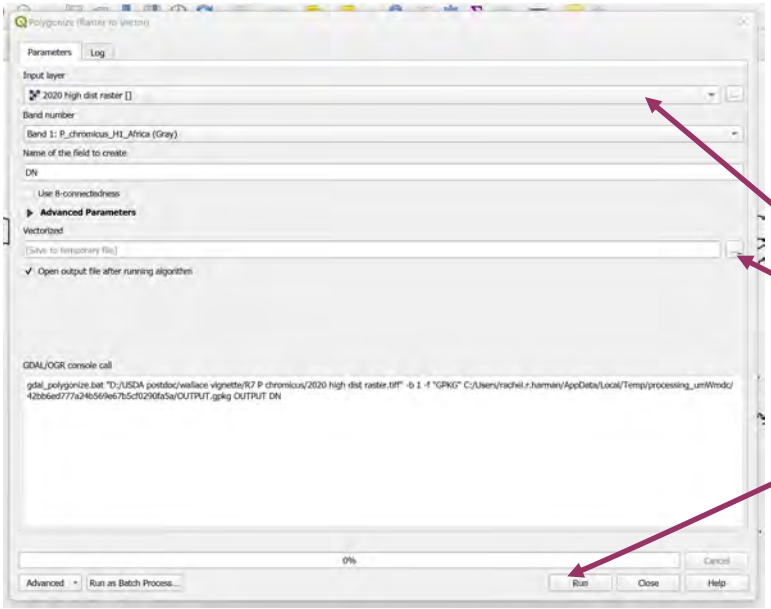
First, we will make maps of just the potential high distribution areas. This map will allow us to easily see if and where the potential distribution is predicted to stay the same, expand, or contract. First, we need to upload the rasters then transform them into shape files.

1. **Drag and drop the three .tiff files into the layers window associated with high distribution for current (2020) and future (2070 SSP585 and SSP126). Make sure you grab the raster .tiff files, not the plot .tif files.**
2. **To transform the raster into a polygon shape file click "Raster" on the top menu.**
3. **Select "Conversion" in the dropdown.**
4. **Select "Polygonize (Raster to Vector)".**

A



B



A) **The QGIS display after uploading the high distribution layers.**

B) **The Polygonize window will appear.**

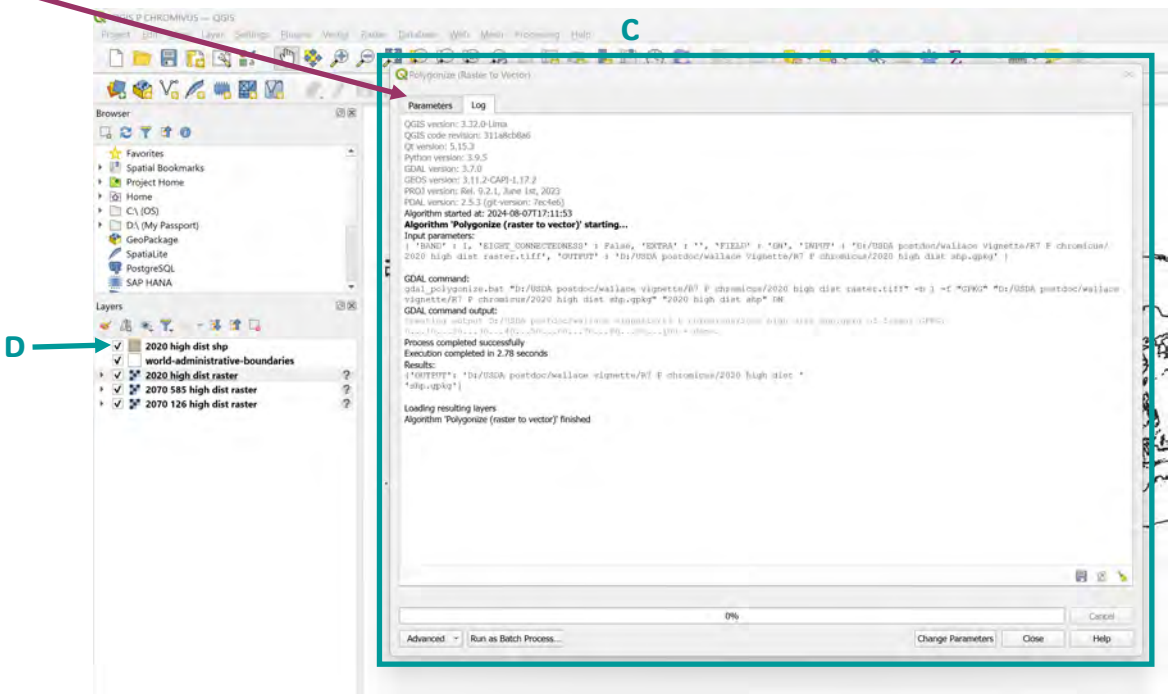
5. **Select the 2020 high distribution raster.**
6. **Select a folder and create a name for the polygon that is being created by clicking the button. We are naming this "2020 high dist shp".**
7. **Click "Run".**

C) The window will change over to the "Log" tab. Any errors will be revealed here in red.

D) The new shapefile will appear in the Layer window.

8. Click the "Parameters" tab.

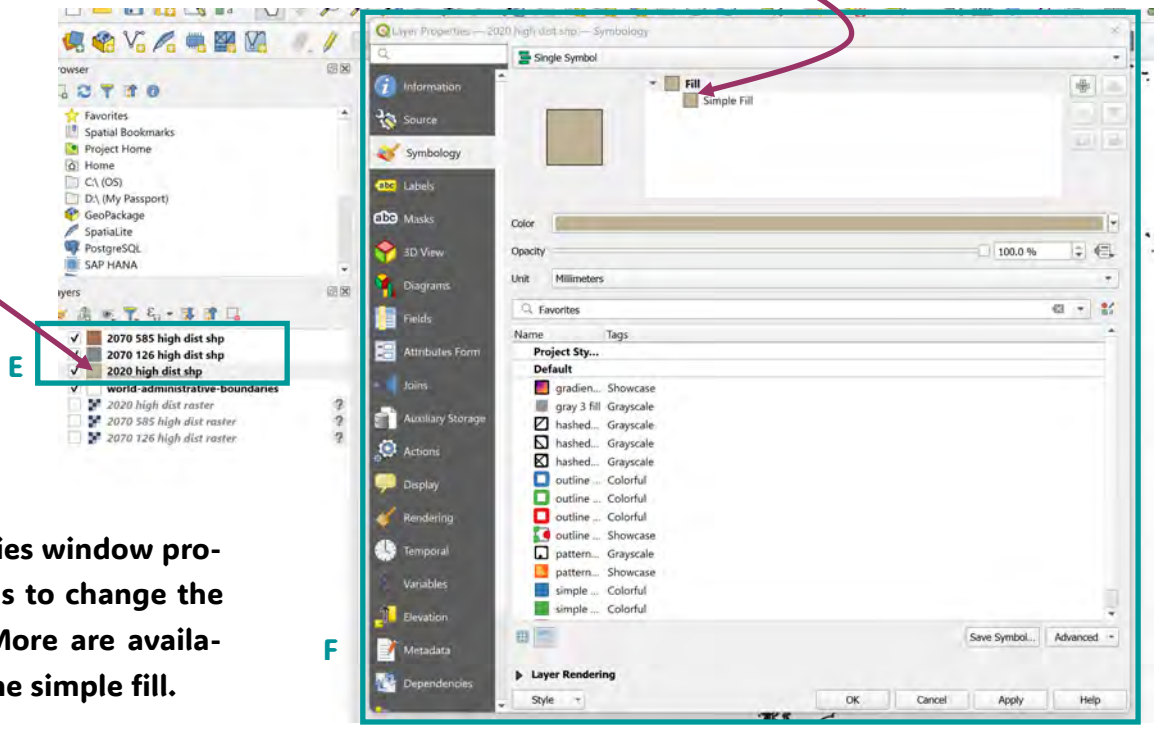
9. Repeat steps 5-7 with the future SSP126 and SSP585 high distribution rasters.



E) The three new shapefiles will upload as a random color. These need to be changed to clearer colors. It is also easier to tell the difference between the layers if the shapes do not have a border.

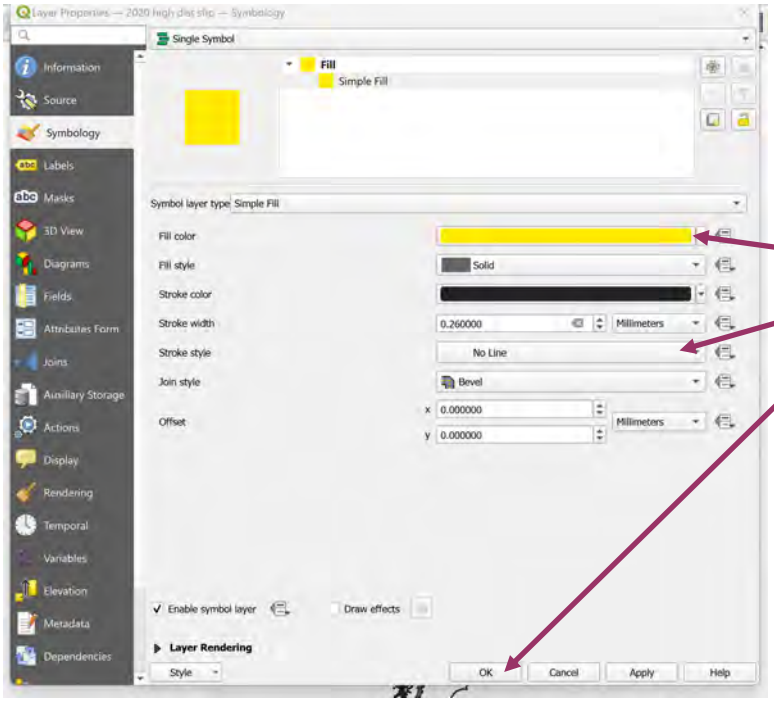
10. Double click on the box associated to the left of the 2020 high dist shp to open the layer properties window.

11. Click on the "Simple Fill" box icon under the "Fill" window.



F) The Layer Properties window provides basic options to change the layer's features. More are available by changing the simple fill.

G



G) The Simple Fill provides more options to format the shape.

12. Change the fill color to yellow.

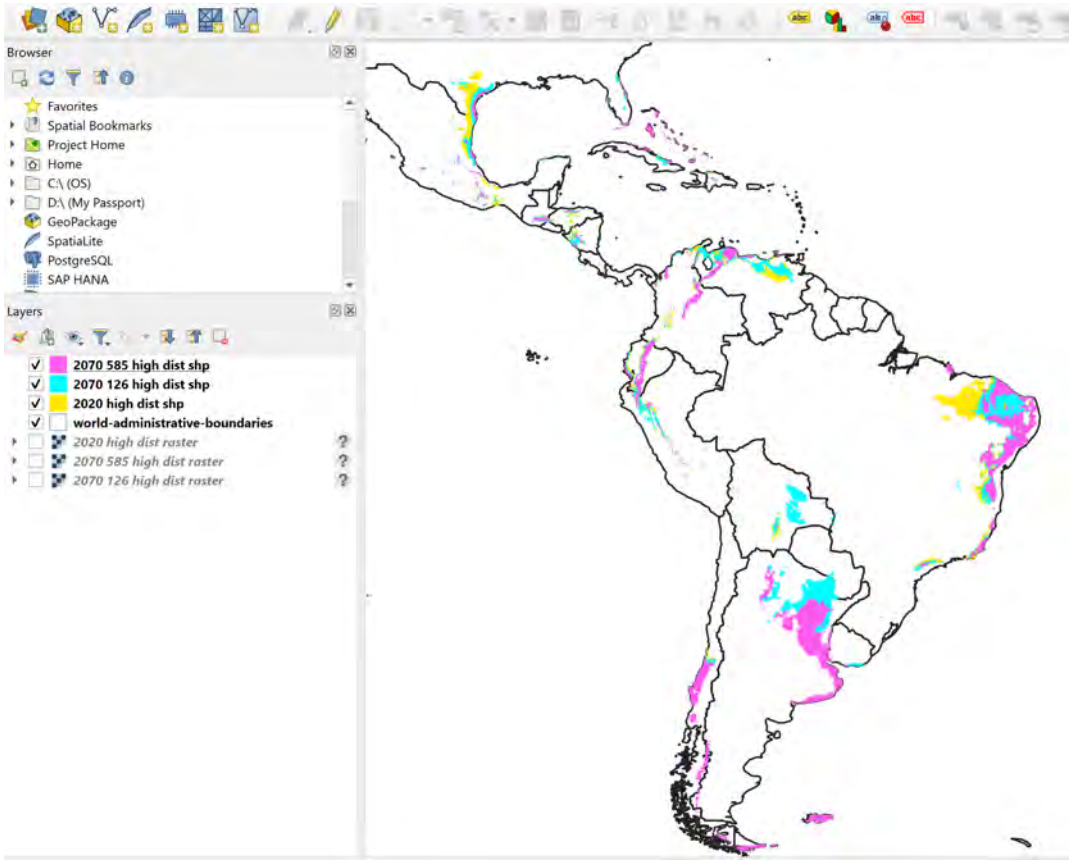
13. Change the Stroke Style to "No Line".

14. Press "OK".

15. Change the colors for the SSP126 and SSP585 shapefiles using steps 10-14. Change the SSP126 to blue and SSP585 to pink.

H) The colors now stand out. However, not all of each color is visible. Whichever shape file is higher on the list appears on top of those below. Thus, the 2070 585 high dist shp layer is fully seen whereas the 2020 high dist shp layer is covered by the future scenarios.

H



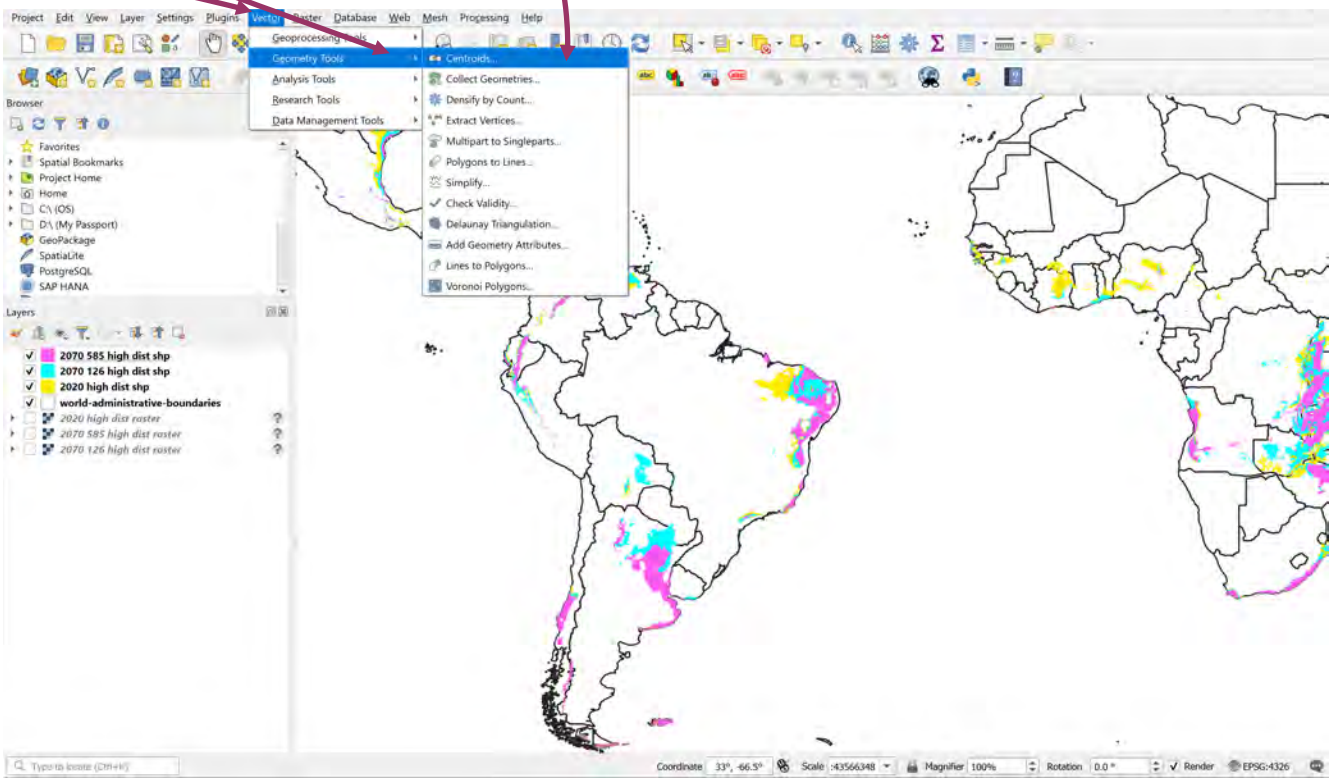
Adding centroids is a great way to show any possible shifts in the distribution. Centroids are placed at the center of a polygon. Thus, any changes in this center can be visually observable by adding a dot. We can also obtain the GPS coordinates of each centroid to measure the distance between centroids. Statistics can be run on this numeric data to test for significant differences between distances, such as comparing changes between current and SSP126 vs current and SSP585.

I) Now that we are working with shapefiles, we need to use the "Vector" tab, not the "Raster" tab.

16. Click the "Vector" tab.

17. Select "Geometry Tools" from the drop down menu.

18. Click "Centroids…"

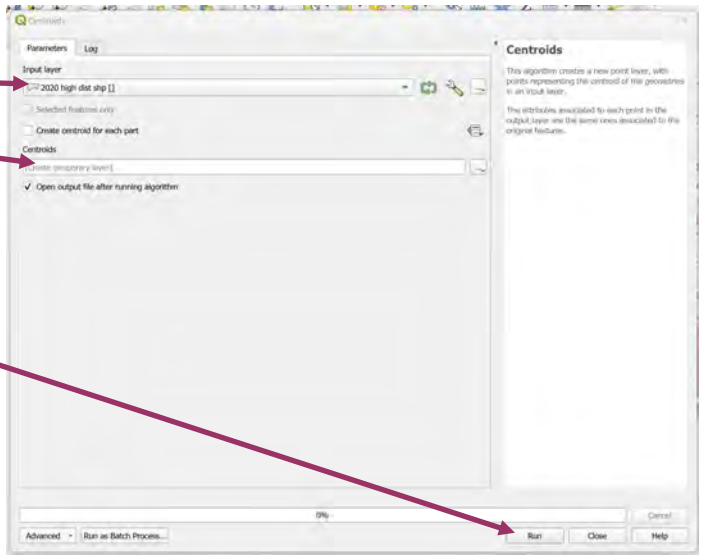


J) The Centroids window will appear.

19. Select the 2020 high dist shp.

20. Name the file and save it to a folder.

21. Click "Run".

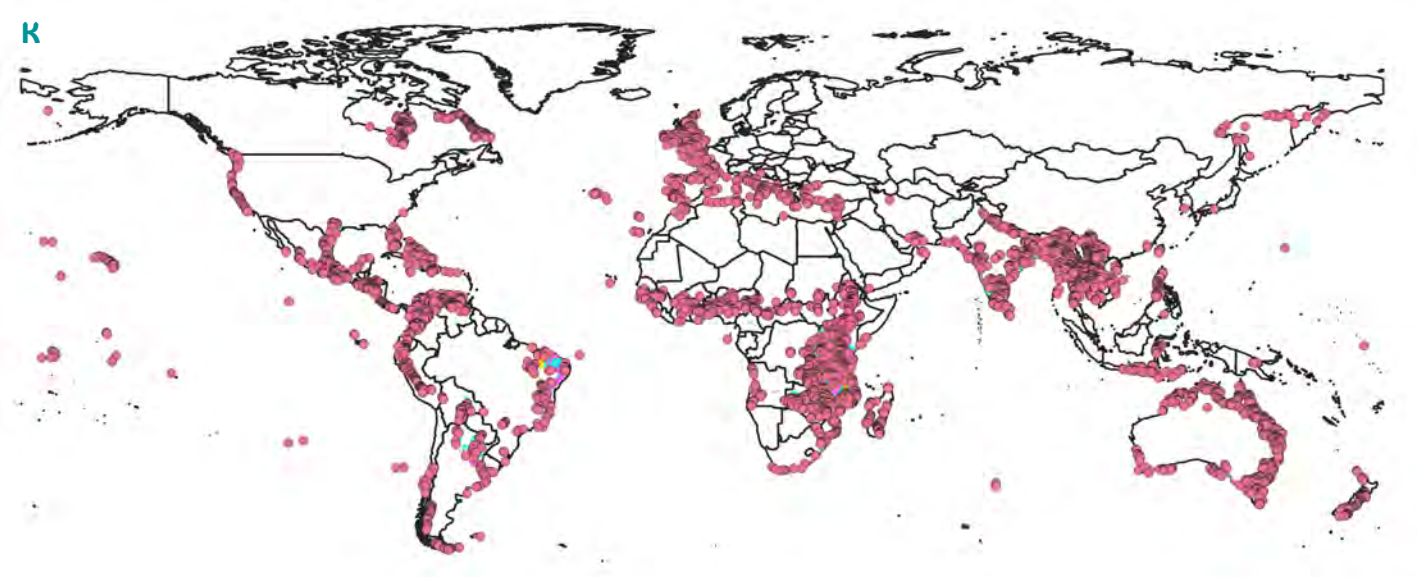


K) The centroids for the 2020 high dist shp file are everywhere. Does this tell us anything about the movement or allow us to compare it to others? Not really.

What happened? Why are there so many dots when we selected just one shape file? It may be one shape file, but it is made of several polygons. Each pixel has the potential to be an individual polygon if it is not adjacent to another pixel. This leads to many, many centroids.

L) Since this map is not interpretable, we need to provide QGIS with polygons that are statically and biologically meaningful.

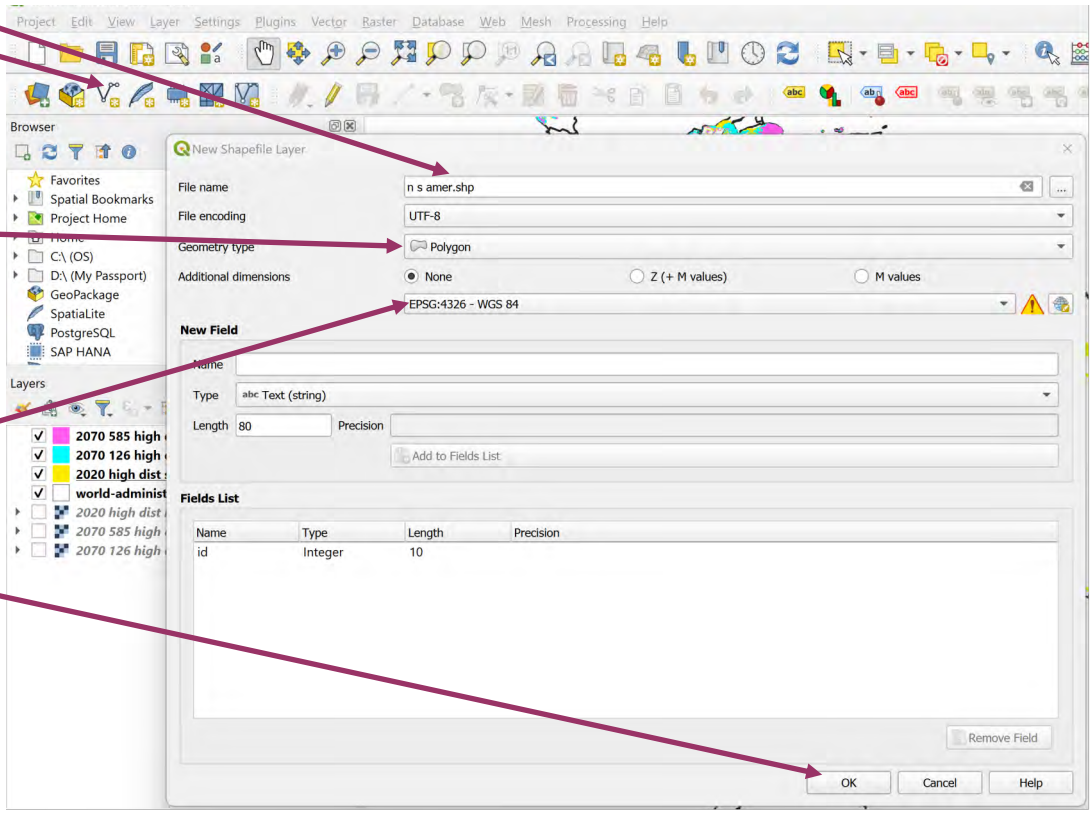
K



22. To create a polygon, click on the "Create New Shapefile…" icon. A window will appear.

23. The first shape we will make is around northern South America, So we will name the shape n s amer. Save this to a file.

L



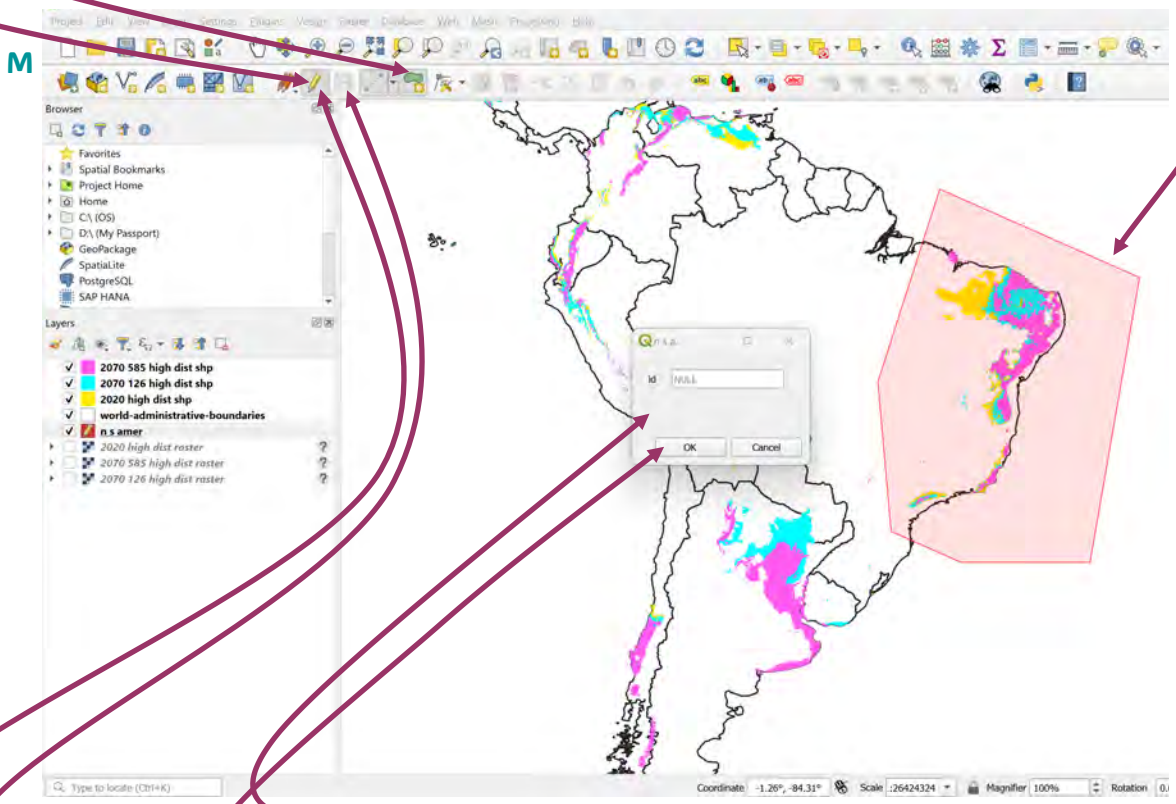
24. Select the "Polygon" geometry type.

25. Make sure the GPS dimension are selected for EPSG:4326 -WGS 84.

26. Click "OK".

M) When the new polygon is selected in the Layers window, several new icons are now visible.

27. Click the "Toggle Editing" pencil icon.
28. Click the "Add Polygon Feature" green icon.
29. Draw a polygon around the north eastern area of Brazil.
30. Right click to stop drawing the polygon.



31. A window will appear asking for an id. Start with 1 and make each new polygon a number greater than the previous.
32. Press "OK".
33. Click the "Save Layer Edits" save button
34. Turn off editing mode by clicking the "Toggle Editing" pencil icon again.

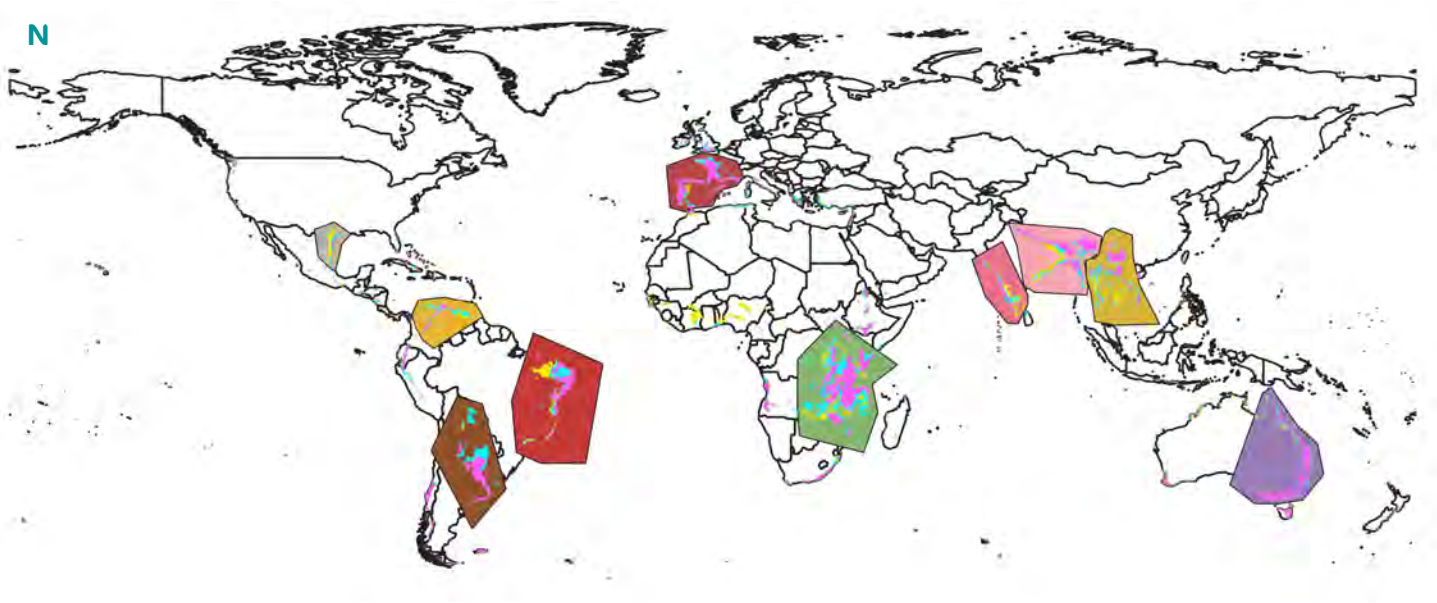
Stop and Think!

Why are we creating a polygon around just this area of Brazil? Think about what changes in potential highly suitable distribution represents. It is the areas that, if a population of the species arrives there, the individuals are >75% likely to survive, reproduce, and establish. Over time, any changes in the actual distribution of the species will occur because the population (or meta-populations) are moving, reproducing, or dying. Thus, a polygon around what visually looks like one population or connected meta-populations can provide us with centroids that show any changes for that range. If we instead made a polygon around the entire continent or a singular one for the globe, it would not tell us what populations are projected to experience in our model. We do not need to worry about every potential population area, just large ones that represent different areas around the globe and that we can use as replicates for data analysis later.

35. Repeat steps 24-34 with other populations around the globe.

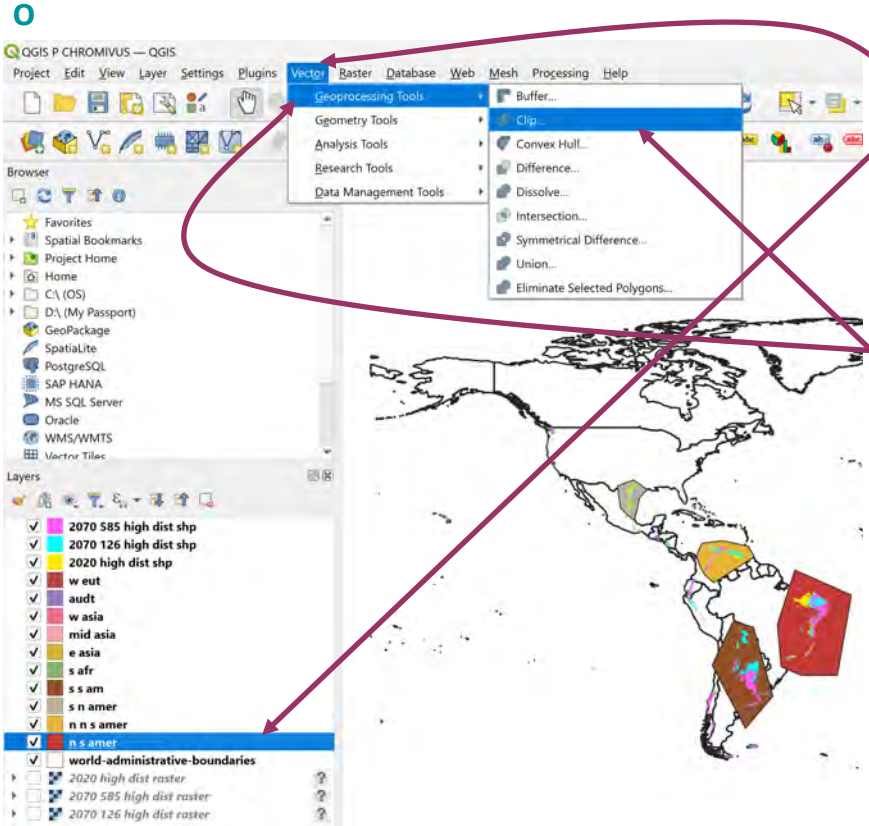
N) The 10 populations that we drew polygons around to create centroids.

N



O) We now need to clip the polygon to the potential distributions. This will make each potential distribution population one polygon. This needs to be done for each time frame. By the end, we will have 30 clipped polygons to create 30 centroids (10 polygons, three time periods).

O



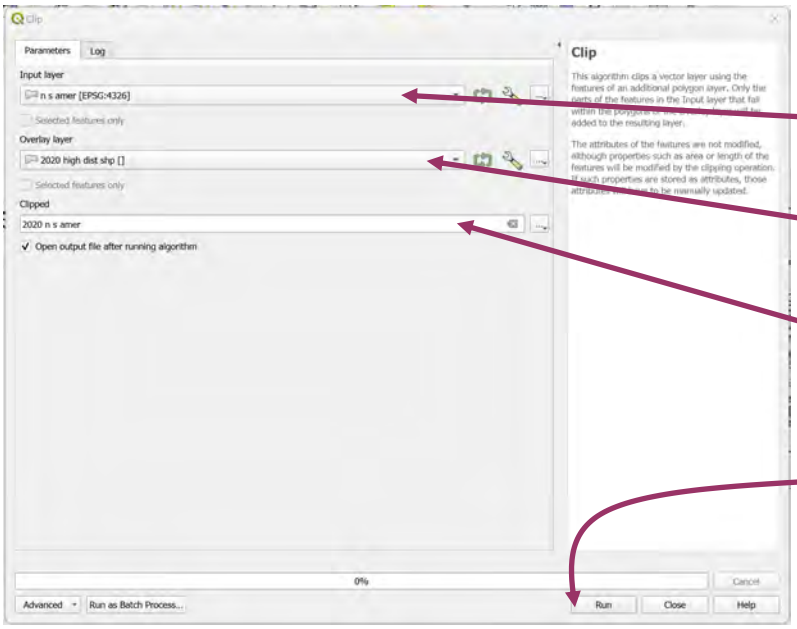
35. Click on the n s amer polygon (this will automatically select it in the drop down menu if its selected here).

36. Click on "Vector" tab.

37. Select "Geoprocessing Tools".

38. Click "Clip...".

P



P) The "Clip" window will appear.

39. The "Input layer" needs to be one of the polygons. Select the n s amer layer

40. The "Overlay layer" is the distribution shape. Select the 2020 high dist shp.

41. Name the new "Clipped" file something unique that you can recognize. We will name it 2020 n s amer shp.

42. Click "Run".

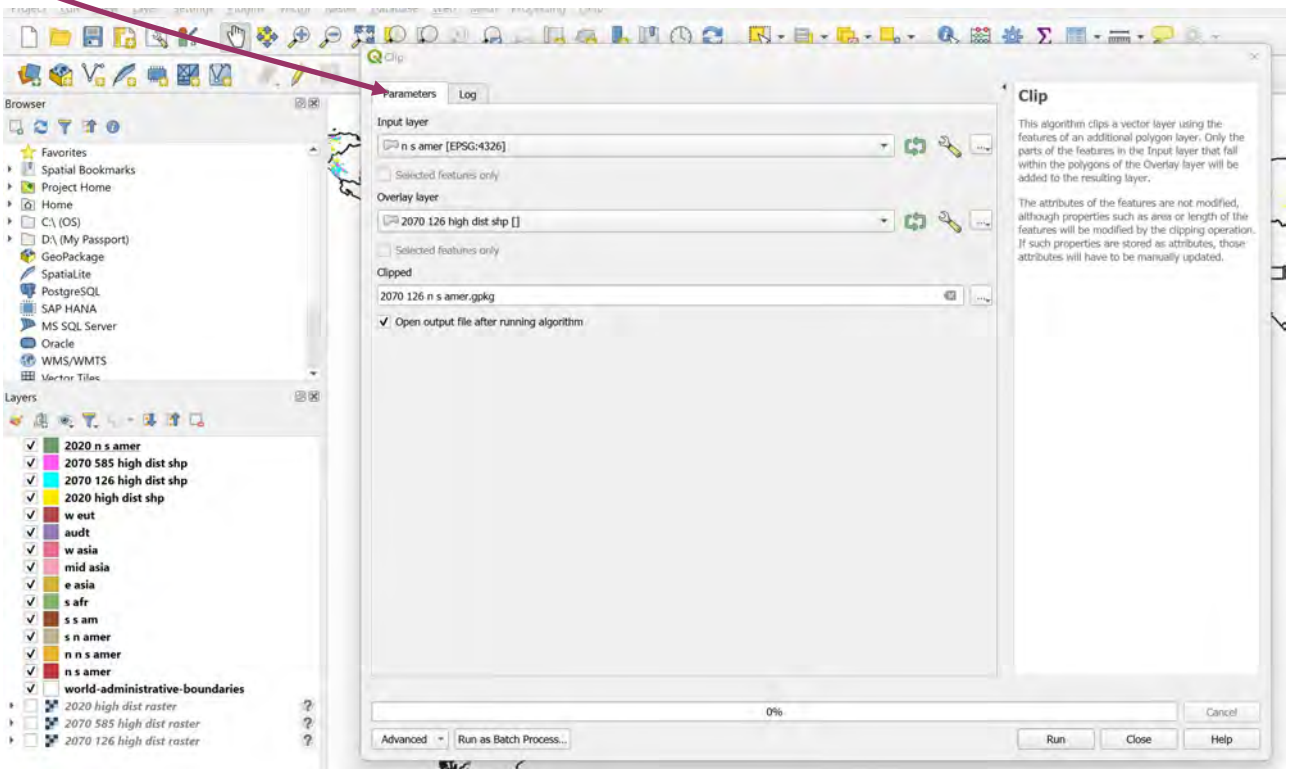
Q) As before, the new polygon will appear in the Layers window.

43. Return back to the "Parameters" tab.

44. Switch the Input or Overlay layers until all 30 combinations are created.

45. Name each layer with a name that is a combination of the input and overlay layers.

Q



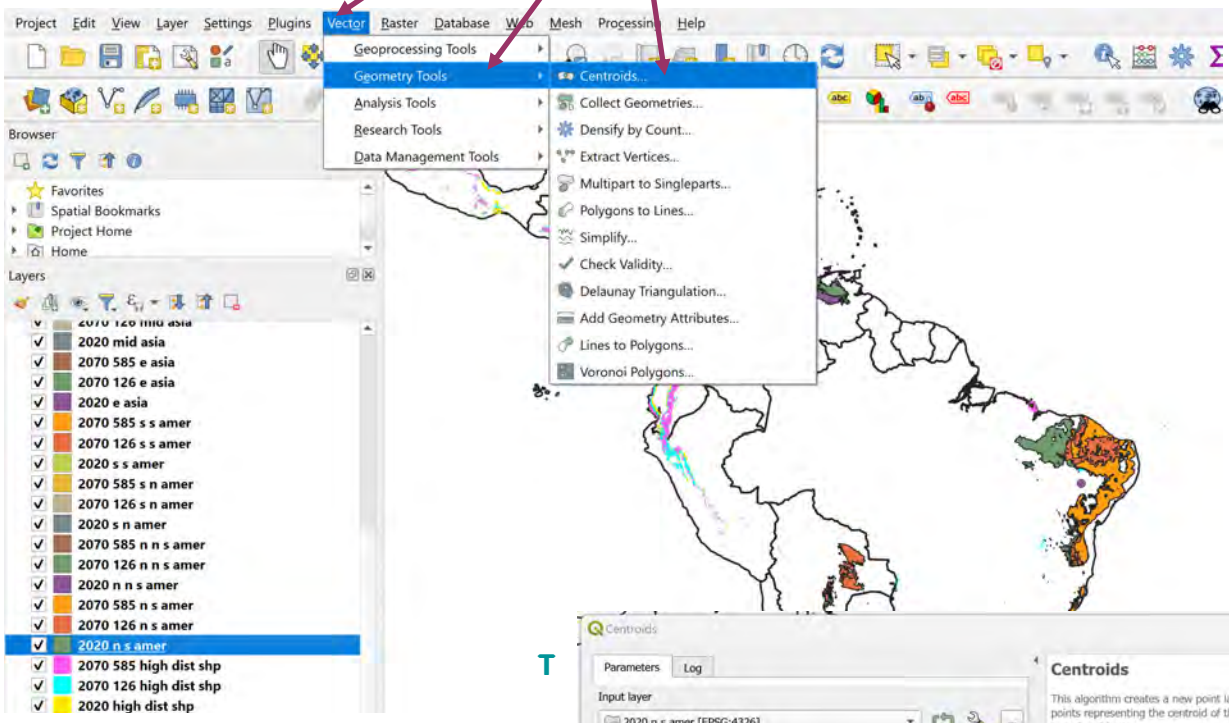
R

R) The 30 polygon layers that we created. Each of the three time periods are represented for each of the 10 population regions.

Now that we have just 10 polygons per time period instead of thousands, we can create centroids that are biologically and statistically sound.

S) We will again use the Centroids tool.

S

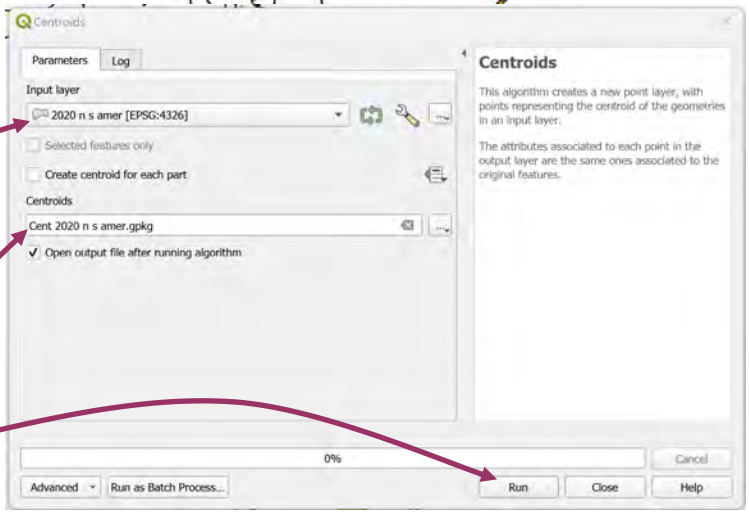


46. Click the "Vector" tab.

47. Select "Geometry Tools".

48. Click "Centroids..."

T



T) The Centroids window will appear.

49. Select the 2020 n s amer layer.

50. Name the centroid shape file. We are naming it "Cent 2020 n s amer". You will likely need to save with a file extension.

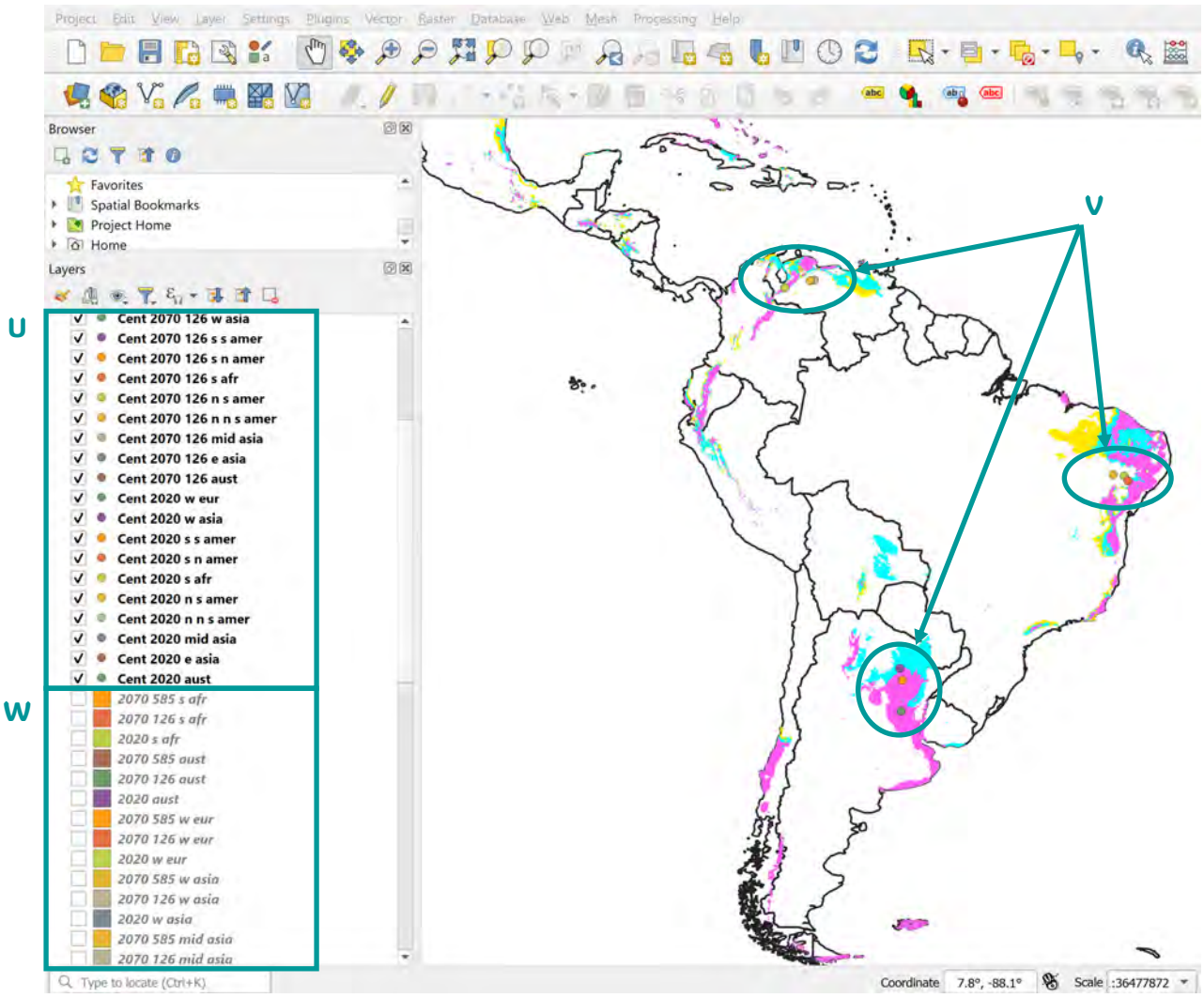
51. Click "Run". The new point will appear in the Layers window.

52. Repeat steps 49-51 with the other layers, naming them according to the layer.

U) Each of the centroids has been given a random color and is listed in the Layers window.

V) Each of the population areas now have three centroid points, one for 2020, 2070 SSP126, and 2070 SSP585.

W) We un-checked each of the population polygon layers so the global high distribution layers that were recolored are visible.



Next, we will change the colors of the centroids and the map to improve the visuals.

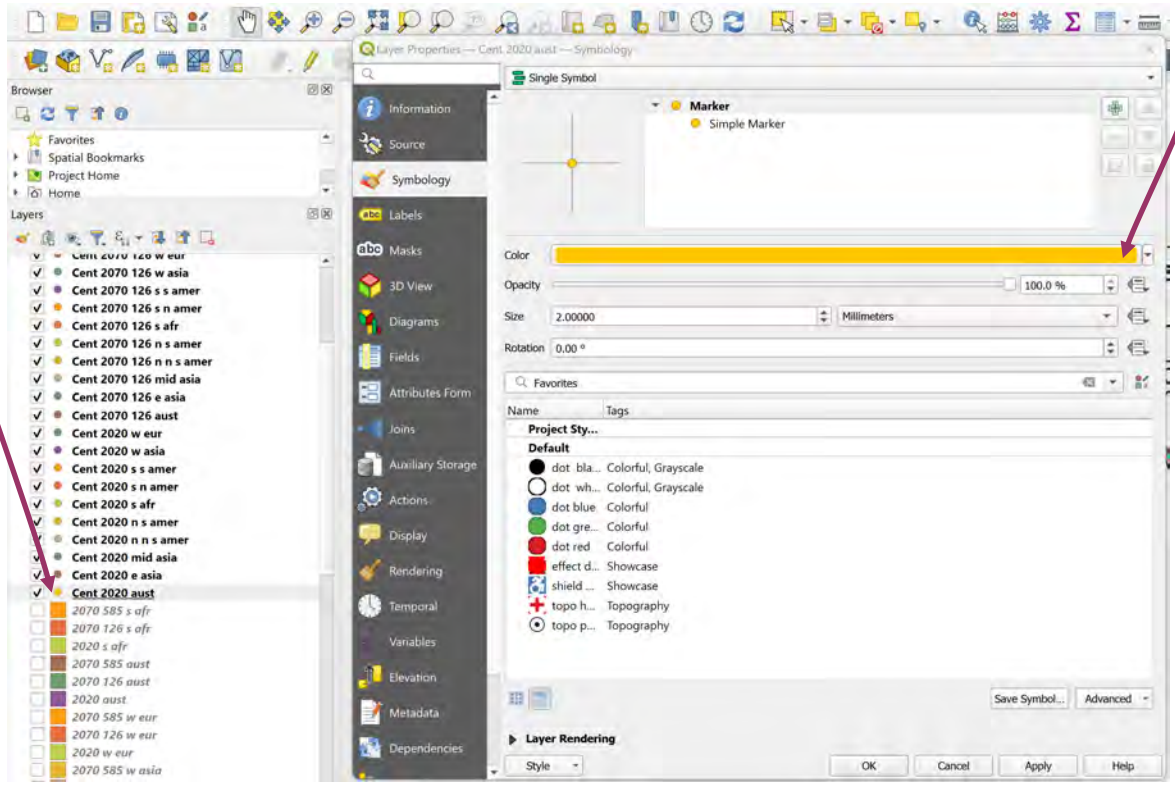
X) We will first color the centroids to match the distributions. 2020 centroids will be changed to orange, 2070 SSP126 to dark blue, and 2070 SSP585 to red.

53. Double click on one of the centroid points to open the Layer Properties window.

54. Change the color with the "Color" drop down.

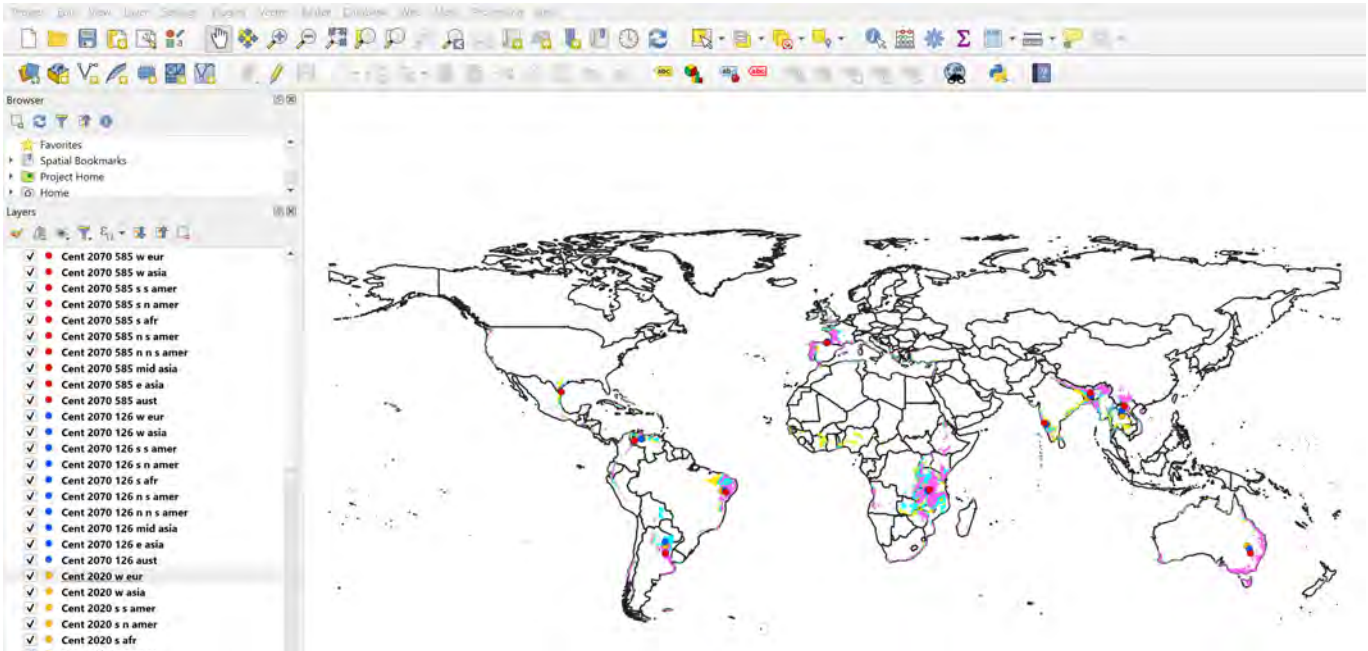
As before, more formatting options are available if you click on "Simple Marker" under the Marker layer.

X



Y) The map after changing the color for each of the centroids.

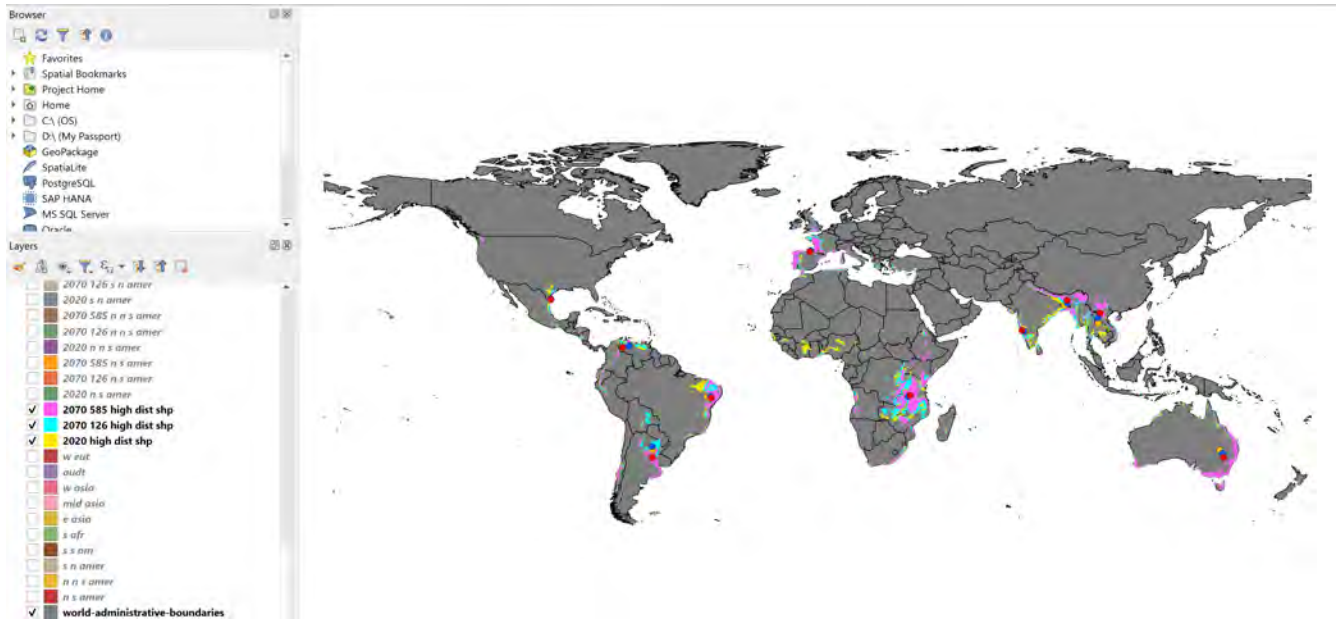
Y



Z) Next, change the fill of the world administrative boundaries to a dark grey. This will allow the other colors to pop. You may wish to change the line thickness as well.

55. Follow the steps 53-54 to change the world administrative boundaries polygon fill.

Z



AA) The last layer that needs to be added is the MESS layer.

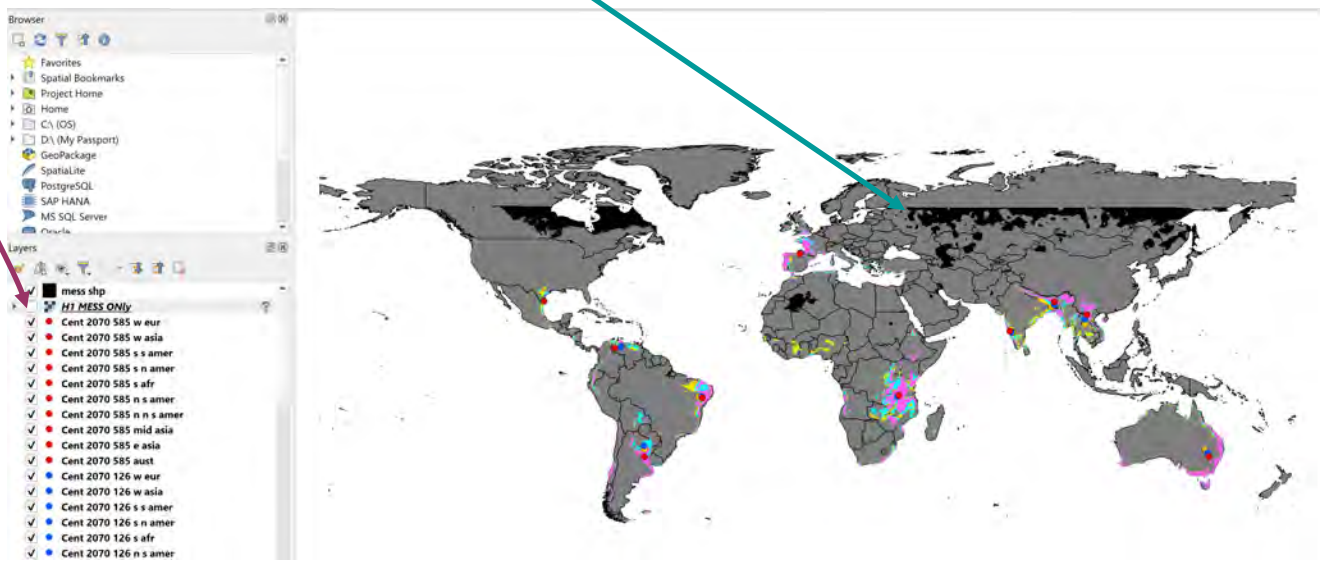
56. Drop the MESS only file into the Layers window.

57. Change the raster to a polygon using steps 2-7.

58. Un-click the raster layer.

The MESS layer has a line that stops at 60°N because that is as far as we created the extent in Wallace. We will save the image to only this extent, so this will not be evident.

AA



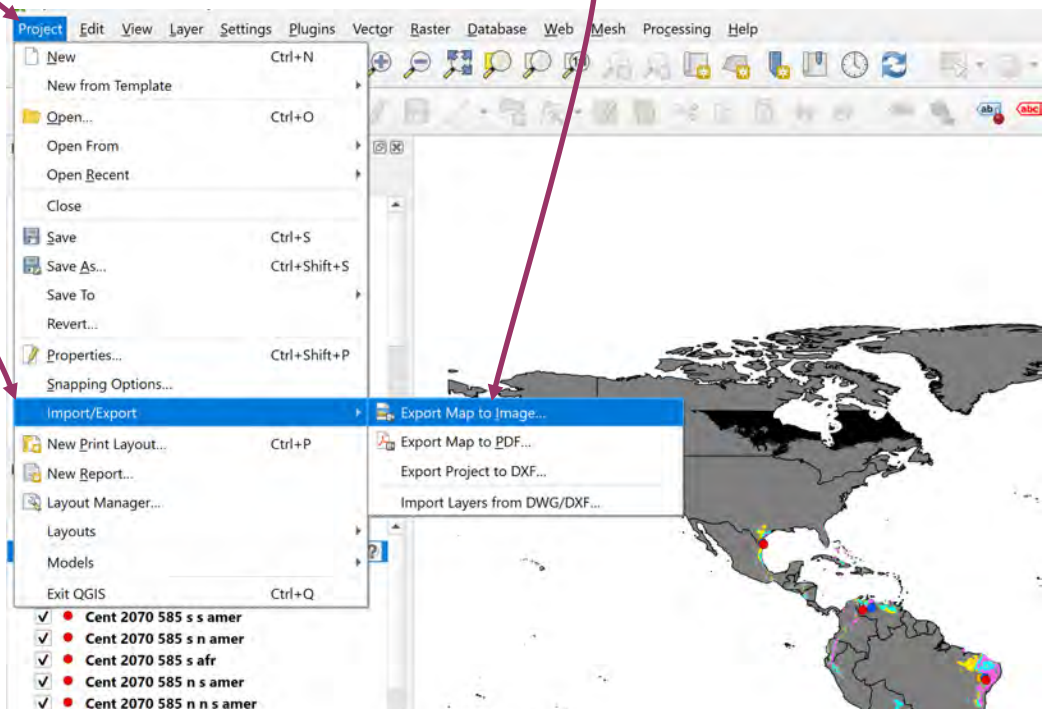
AB) Next we will export the map to an image.

59. Click the "Project" tab.

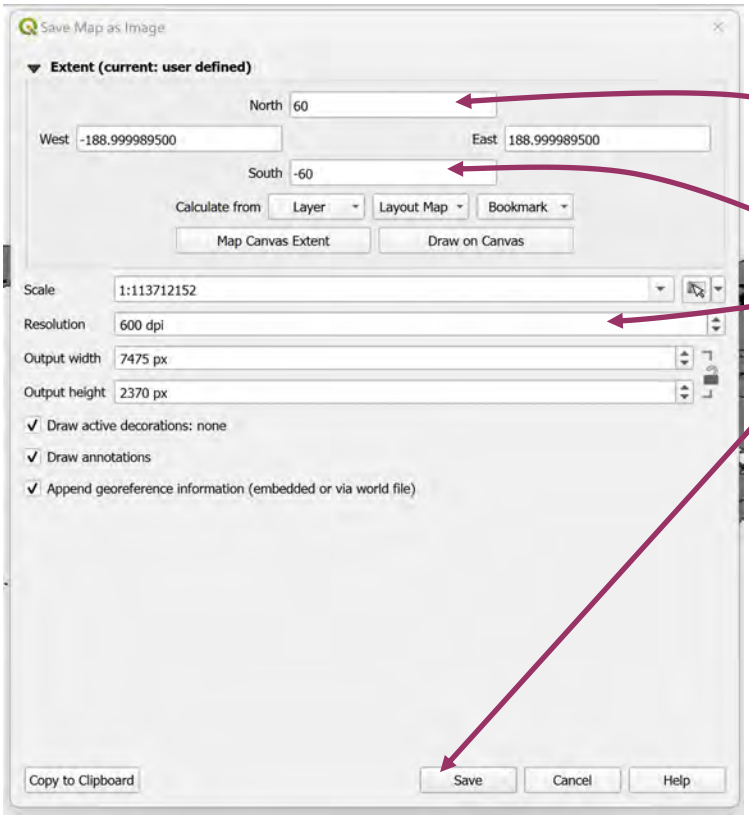
60. Select "Import/Export".

61. Click "Export Map to Image…".

AB



AC



AC) The Save Map as Image window will appear.

62. Change the North extent to 60.

63. Change the South extent to –60.

64. Change the Resolution to 600 dpi.

65. Click "Save".

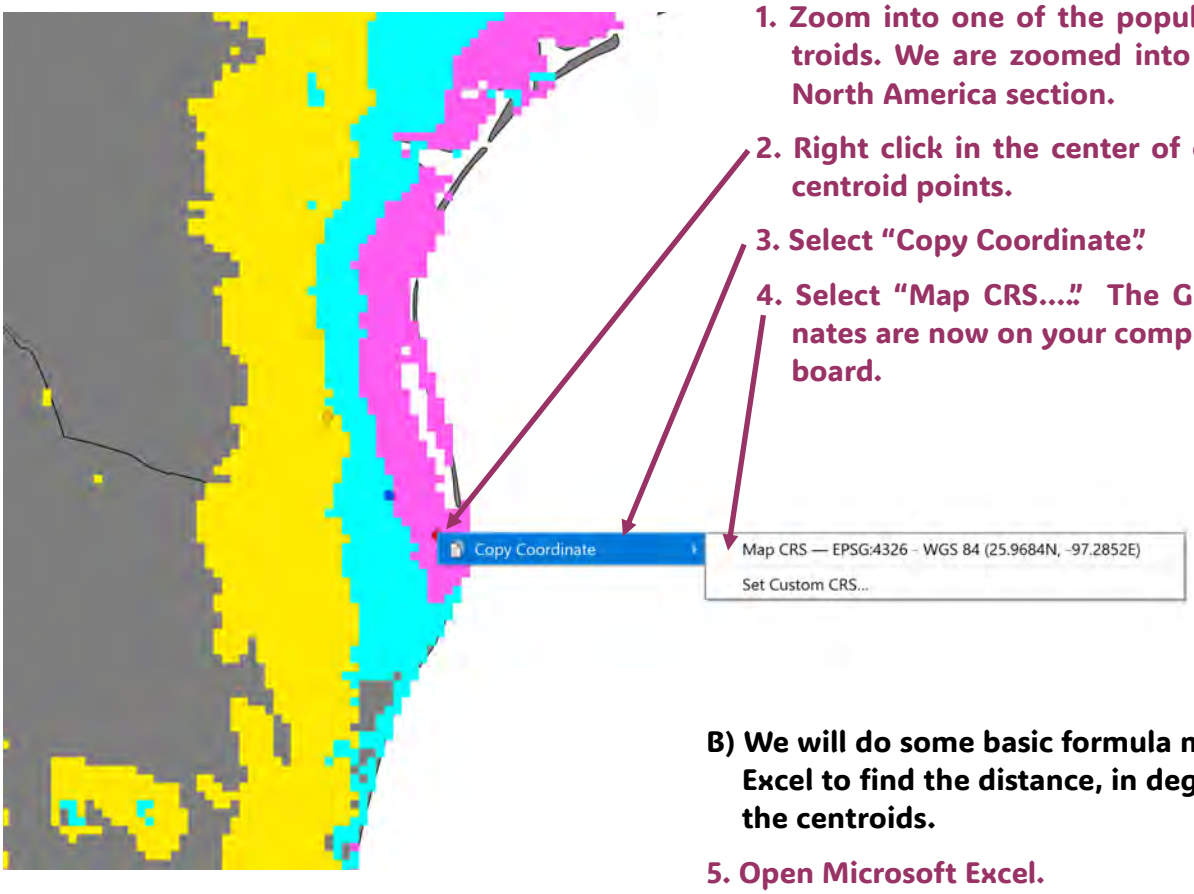
66. Save the file as an appropriate name and to an appropriate file using the next window.

Analyzing the Maps

Centroid distance

A) We can use the GPS locations of each of the centroids to calculate the distance between the shifts of the distributions. This will inform us if the shift is different between SSP levels. The direction of shift is also visually apparent.

A



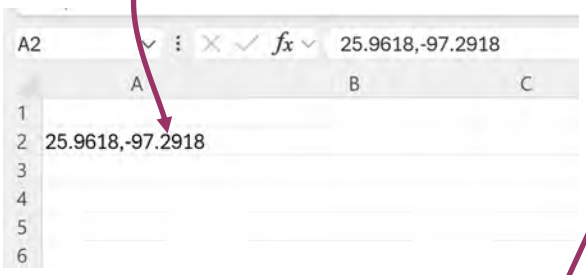
1. Zoom into one of the population centroids. We are zoomed into the south North America section.
2. Right click in the center of one of the centroid points.
3. Select "Copy Coordinate".
4. Select "Map CRS…." The GPS coordinates are now on your computer's clipboard.

B) We will do some basic formula map in Microsoft Excel to find the distance, in degrees, between the centroids.

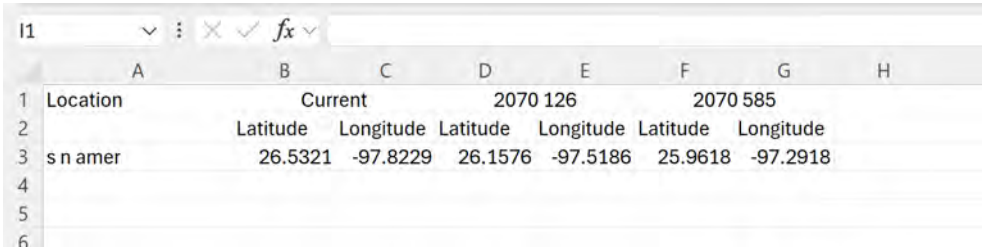
5. Open Microsoft Excel.
6. Paste the coordinates into a cell. The longitude and latitude are pasted into one cell.
7. Separate the longitude from the latitude.
8. Create a table with clear labels for the population location and the longitude and latitude for current, 2070 SSP126, and 2070 SSP585.

C) The layout for the Excel table.

B



C



9. Repeat steps 1-4 for the other 29 centroids.

D) The latitude and longitude for all 30 centroids.

D

	A	B	C	D	E	F	G
1	Location	Current		2070 126		2070 585	
2		Latitude	Longitude	Latitude	Longitude	Latitude	Longitude
3	s n amer	26.5321	-97.8229	26.1576	-97.5186	25.9618	-97.2918
4	n n s amer	9.262	-68.552	9.219	-68.889	8.559	-71.277
5	n s amer	-9.098	-40.233	-9.308	-39.152	-9.8	-38.829
6	s s amer	-28.63	-60.166	-27.366	-60.419	-31.516	-60.334
7	s afr	-9.013	33.281	-9.161	33.112	-8.782	33.723
8	w eur	43.452	-3.15	43.694	-2.612	43.42	-2.844
9	w asia	14.417	75.653	14.543	75.147	15.007	74.536
10	mid asia	23.793	89.981	24.678	90.95	25.689	90.866
11	e asia	17.493	102.349	19.495	102.244	21.075	103.023
12	aust	-28.84	146.702	-29.999	147.566	-31.516	148.008

E) We will now do a little basic math using the Pythagorean theorem to find the distance between the centroids. We need to calculate the difference between the current point and each of the 2070 points.

First, we need to calculate the difference between the latitudes and longitudes.

10. Create a table section for the calculation. We need two columns, one for latitude and one for longitude, for each calculation. A column for the locations is also needed.

11. Use the formula =ABS($B3-D3) for the first cell (highlighted). This will provide the absolute difference between the two latitudes.

12. Use a similar code for the other cells to calculate the difference for each particular cell.

E

	A	B	C	D	E	F	G	H
1	Location	Current		2070 126		2070 585		
2		Latitude	Longitude	Latitude	Longitude	Latitude	Longitude	
3	s n amer	26.5321	-97.8229	26.1576	-97.5186	25.9618	-97.2918	
4	n n s amer	9.262	-68.552	9.219	-68.889	8.559	-71.277	
5	n s amer	-9.098	-40.233	-9.308	-39.152	-9.8	-38.829	
6	s s amer	-28.63	-60.166	-27.366	-60.419	-31.516	-60.334	
7	s afr	-9.013	33.281	-9.161	33.112	-8.782	33.723	
8	w eur	43.452	-3.15	43.694	-2.612	43.42	-2.844	
9	w asia	14.417	75.653	14.543	75.147	15.007	74.536	
10	mid asia	23.793	89.981	24.678	90.95	25.689	90.866	
11	e asia	17.493	102.349	19.495	102.244	21.075	103.023	
12	aust	-28.84	146.702	-29.999	147.566	-31.516	148.008	
13								
14								
15	Difference between the Lats and Longs							
16		current - 126				current - 585		
17		Latitude	Longitude	Latitude	Longitude			
18	s n amer	=ABS($B3-D3)		0.3043	0.5703	0.5311		
19	n n s amer	ABS(number)	043	0.337	0.703	2.725		
20	n s amer	0.21		1.081	0.702	1.404		

User Insights

You can copy the code to other cells in Excel by clicking on and dragging the little box in the lower right corner of a selected cell. This will change the code so that the cells listed in the formula are dragged in a similar fashion. Double click in any cell to double check that the cells you want are used.

F) Now that we have the distance between the latitudes and longitudes, we can use the Pythagorean theorem to find the difference in distance (in degrees) between the points.

13. Create a table section for the new hypotenuse calculation. We need three columns, one for the locations, one for the difference for SSP126 and the last for the SSP585.
14. Use the formula =SQRT(B17^2 + C17^2) to calculate the distance for the first cell.
15. Use a similar formula for the other cells and incorporate the appropriate cells from the difference table. This is already filled out in the table F.

F

SUM =SQRT(B17^2 + C17^2)

	A	B	C	D	E
15	Difference between the Lats and Longs				
16		current - 126		current - 585	
17	s n amer	0.3745	0.3043	0.5703	0.5311
18	n n s amer	0.043	0.337	0.703	2.725
19	n s amer	0.21	1.081	0.702	1.404
20	s s amer	1.264	0.253	2.886	0.168
21	s afr	0.148	0.169	0.231	0.442
22	w eur	0.242	0.538	0.032	0.306
23	w asia	0.126	0.506	0.59	1.117
24	mid asia	0.885	0.969	1.896	0.885
25	e asia	2.002	0.105	3.582	0.674
26	aust	1.159	0.864	2.676	1.306
27					
28					
29	Hypotenuse = the distance between the two dots				
30		current - 126	current - 585		
31	s n amer	0.482544029	0.77930052		
32	n n s amer	0.339732248	2.81421996		
33	n s amer	1.101208881	1.56971972		
34	s s amer	1.289071371	2.89088568		
35	s afr	0.224644163	0.49872337		
36	w eur	0.589922029	0.30766865		
37	w asia	0.521451819	1.26324542		
38	mid asia	1.312320845	2.09237688		
39	e asia	2.004751606	3.64485939		
40	aust	1.445606101	2.97768568		
41					

Stats Chat

What can we do with this distance data?

There are some simple statistics that can be done. For instance, analyzing if the distances between the two SSP categories are statistically different. There are 10 replicates representing each of the population locations. If you also create rasters that represent the 2050 future scenarios, you can assess if there are differences between the current to 2050 versus 2050 to 2070 time periods. From an applied perspective, this information can let you know when it is critical to heighten conservation or maintenance practices, such as surveillance.

Resources

Excel tips: https://www.microsoft.com/en-us/microsoft-365-life-hacks/organization/excel-tips-spreadsheets-dont-have-to-be-scary

26 of the Best Excel Tips To Optimize Your Use: https://www.indeed.com/career-advice/career-development/excel-tips

Proportion of are per suitability level

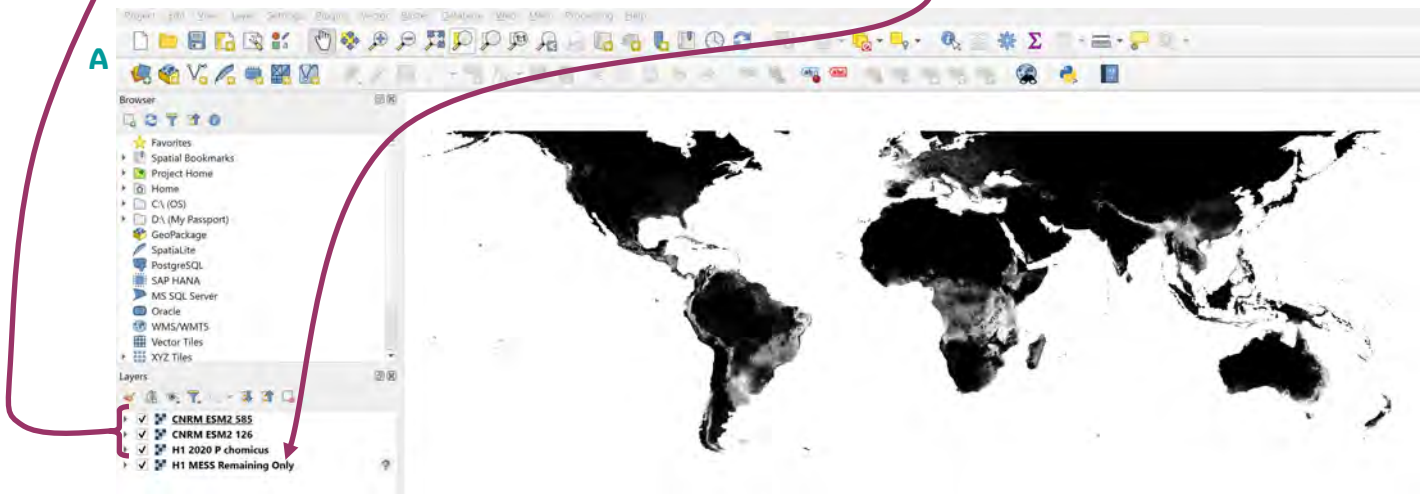
One more measurement that can be calculated is the number of pixels in different suitability levels. We will divide suitability into four categories including, low (0-25), poor (25-50), moderate (50-75), and high (75-100).

First, we need to remove the pixels that are covered by the MESS layer from the analysis in QGIS. Then, we will upload the new rasters into R to obtain the pixel count and create a histogram of the number of pixels per suitability level. We can compare this information between time periods and SSP levels.

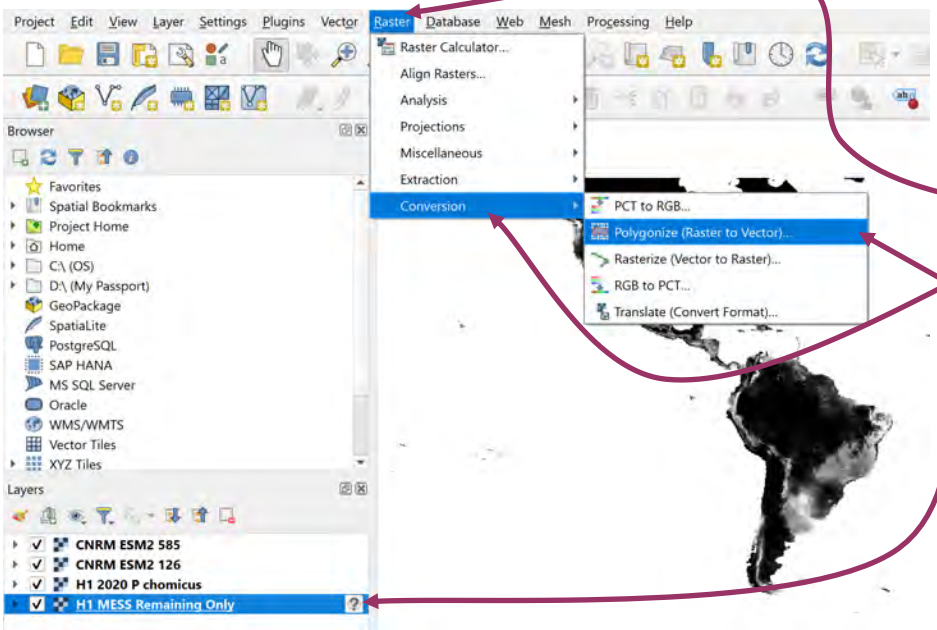
A) We will work in a new QGIS project.

1. **Open a new QGIS project.**
2. **Drag and drop the three rasters representing the 2020 (current), and 2070 SSP585 and SSP126 projections.**
3. **Drag and drop the MESS remaining Only layer.**

A



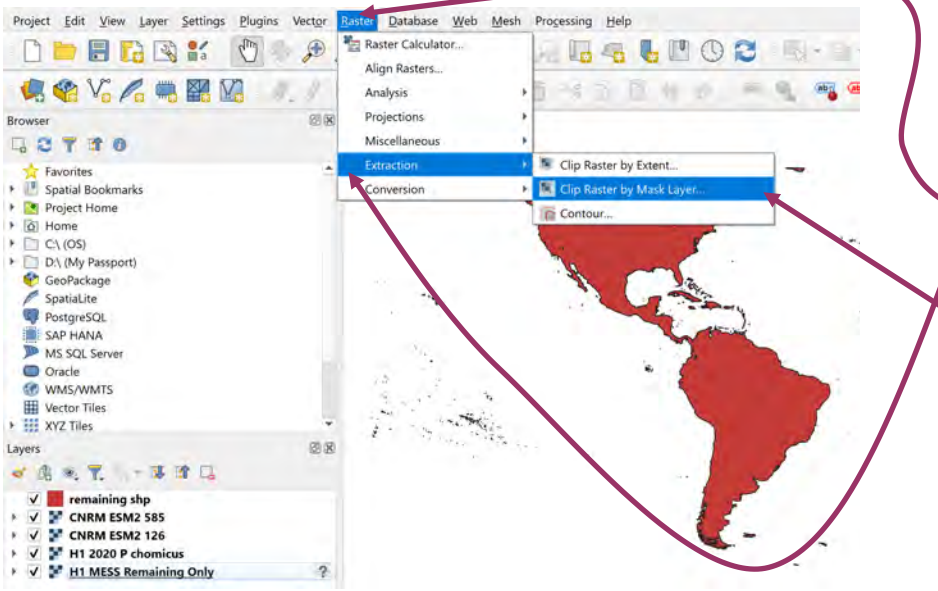
B



B) The MESS Remaining Only layer needs to be converted to a raster. The steps are the same as when we did this previously.

4. **Select the MESS Remaining Only layer.**
5. **Select the "Raster" tab.**
6. **Select "Conversion".**
7. **Click "Polygonize (Raster to Vector)....**
8. **In the popup window, save the new polygon to a file and run the conversion. We anmed it remaining shp.**

C



C) Now that we have a shape file, we can extract just the pixels associated with the remaining layer.

9. Click the "Raster" tab.

10. Select "Extraction".

11. Click "Clip Raster by Mask Layer...".

D

D) The Clip Raster by Mask Layer window will open.

12. Select the current 2020 layer as the "Input layer".

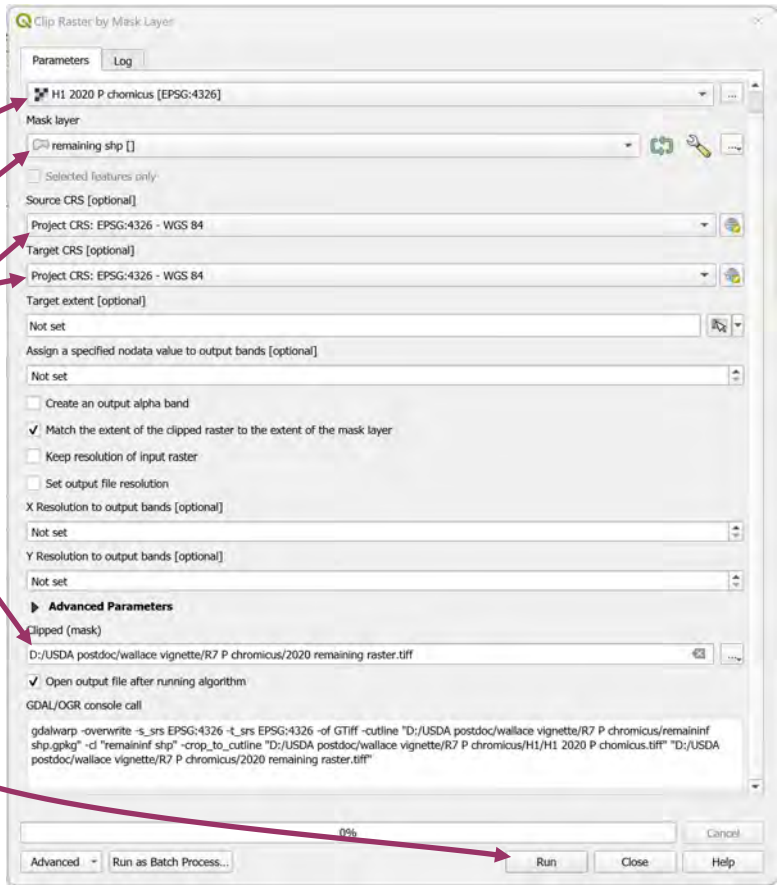
13. Select the new shape file, remaining shp as the "Mask layer".

14. Select the "Project CRS" for both the "Source CRS" and "Target CRS".

15. Name the new clipped raster and save it to a file. We are naming the file "2020 remaining raster.tiff".

16. Click "Run".

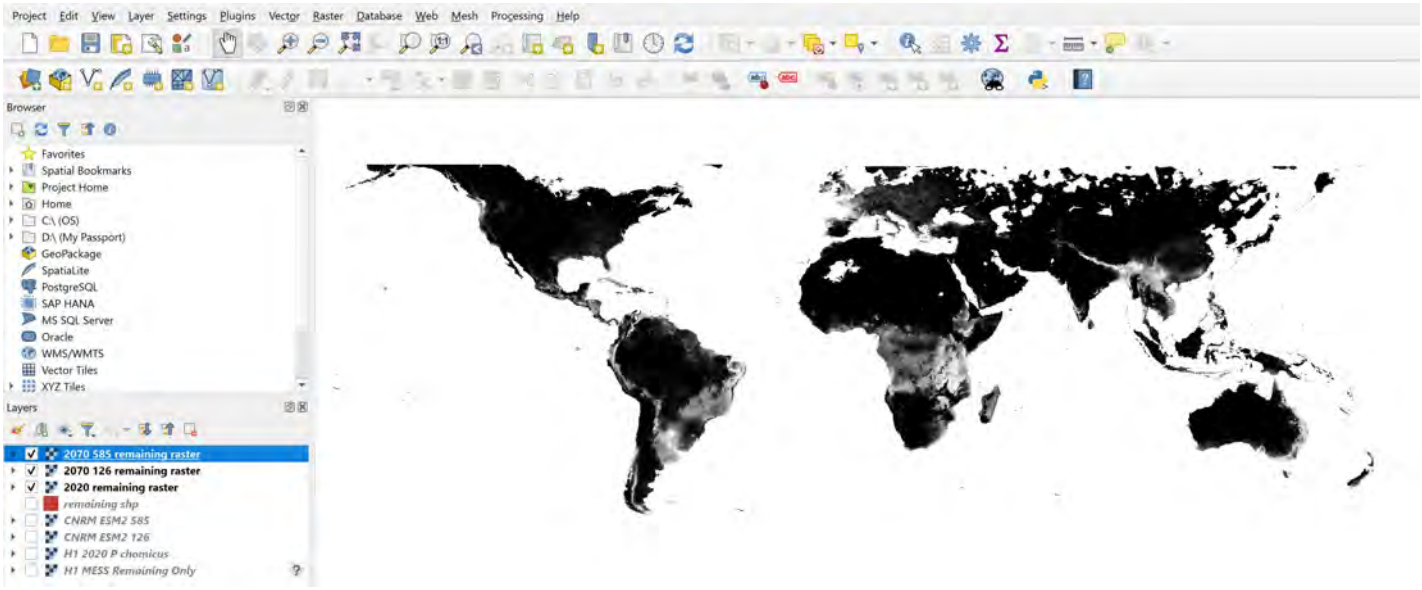
17. Repeat steps 12-16 with the SSP126 and SSP585 raster files. Save them under a relevant name.



E) The three rasters that are clipped to remaining pixels do not include the MESS layer.

We are now done with QGIS. These rasters can be used in R to obtain a pixel count for each of the histograms and make a histogram of the counts.

E



18. Open R Studio.

19. Upload the 2020 current raster and save it as a dataframe. Label the third column to use easily.

```
WHOLERAST <- ras t("2020 remaining ras ter.tiff")
WHOLERAST_df = as.data.frame(WHOLERAST, xy = TRUE)
LAYER = WHOLERAST_df[,3]   # Jus t the column needed for his ts, always saved as the third column in WAL-
    LACE
```

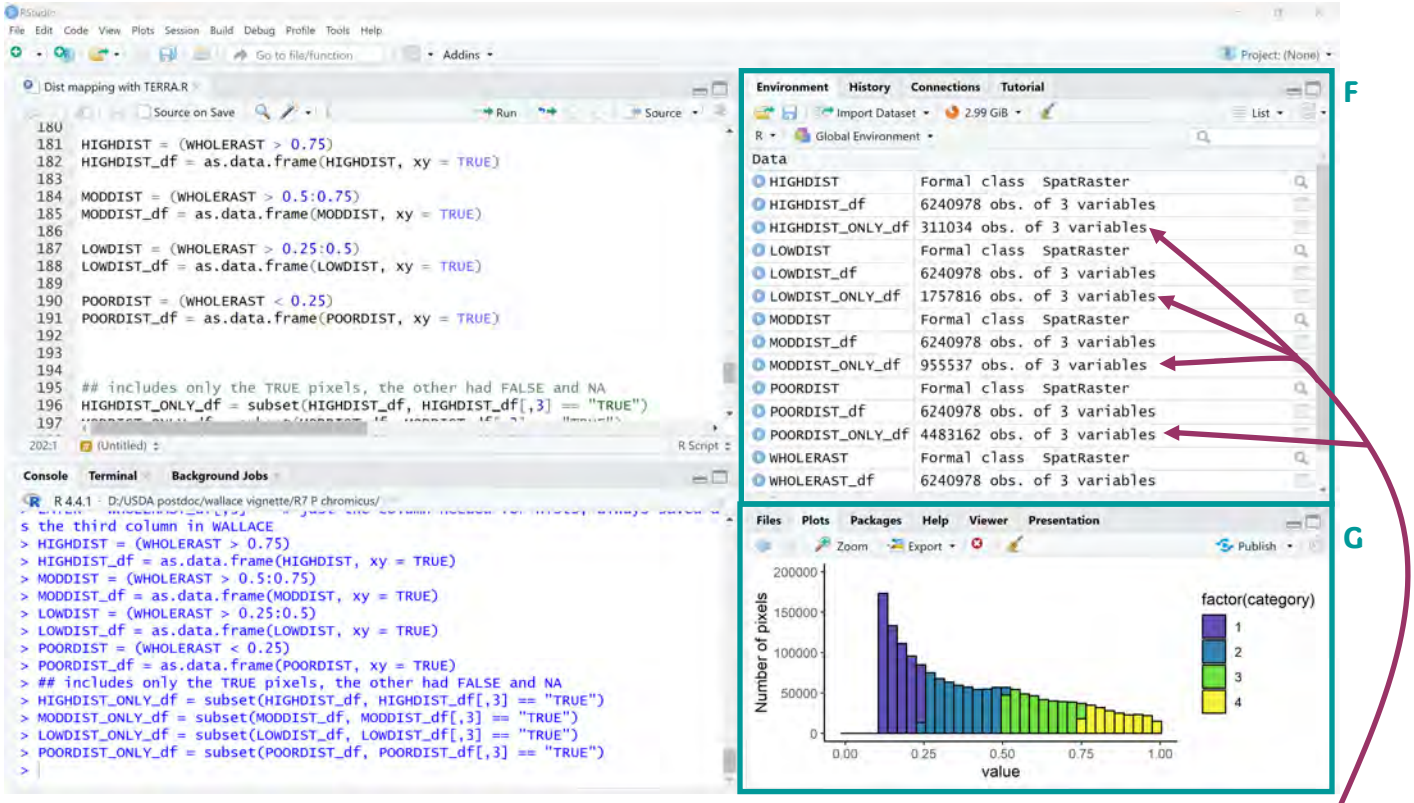
20. Create a histogram of the number of pixels per suitability level.

```
## Create a data frame with categories based on quantiles

data_df <- data.frame(value = LAYER)
data_df$category <- cut(data_df$value, breaks = c(0, 0.25, 0.5, 0.75, 1), # the breaks in color
                        labels = FALSE, include.lowes t = TRUE)

## Create the his togram plot with color-coded categories

HIST = ggplot(data_df, aes(x = value, fill = factor(category))) +
  geom_his togram(binwidth = 0.03, color = "black", alpha = 0.8) +  # binwidth changes the number of columns
scale_fill_manual(values = c("#3925a0", "#006699", "#51d715", "#f2ee0e" ))+ # The numbers are color codes
  scale_y_continuous(name = "Number of pixels", limits = c(0, 200000))+   # change names and scale of axis
  theme_classic()

HIST  # Will open the his togram in the plot window

## Save the his togram as a .tif file

tiff('2020 Dis tribution His togram.tif', units = "in", width = 6, height = 5,
         res = 600, compression = "lzw")
     HIST
dev.off()
```

20. An easy way to obtain the number of pixels in each suitability level is to divide the raster into each level and look at the number of observations (pixels) that are noted in the Data window.

```
## Create a separate raster for each suitability level

HIGHDIST = (WHOLERAST > 0.75)
HIGHDIST_df = as.data.frame(HIGHDIST, xy = TRUE)

MODDIST = (WHOLERAST > 0.5:0.75)
MODDIST_df = as.data.frame(MODDIST, xy = TRUE)

LOWDIST = (WHOLERAST > 0.25:0.5)
LOWDIST_df = as.data.frame(LOWDIST, xy = TRUE)

POORDIST = (WHOLERAST < 0.25)
POORDIST_df = as.data.frame(POORDIST, xy = TRUE)

## Includes only the TRUE pixels, removing FALSE and NA pixels

HIGHDIST_ONLY_df = subset(HIGHDIST_df, HIGHDIST_df[,3] == "TRUE")
MODDIST_ONLY_df = subset(MODDIST_df, MODDIST_df[,3] == "TRUE")
LOWDIST_ONLY_df = subset(LOWDIST_df, LOWDIST_df[,3] == "TRUE")
POORDIST_ONLY_df = subset(POORDIST_df, POORDIST_df[,3] == "TRUE")
```

F) The Data window states the number of observations that are within the dataframe.
G) The histogram will appear under the plot tabs.



21. Make an Excel table to type the number of observations, which represent the pixels, for each of the rasters labeled as ONLY.

22. Repeat steps 19-21 for the SSP126 and SSP585 rasters that were clipped to the remaining pixels.

H) The histogram (pixel count) of the current time period.

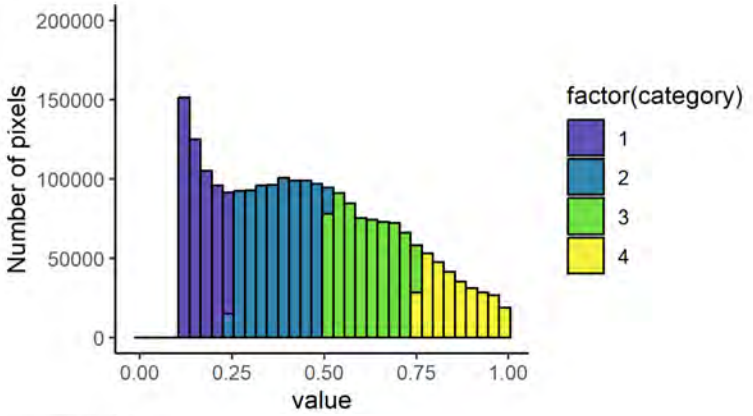
I) The histogram (pixel count) of 2070 SSP126.

J) The histogram (pixel count) of 2070 SSP585.

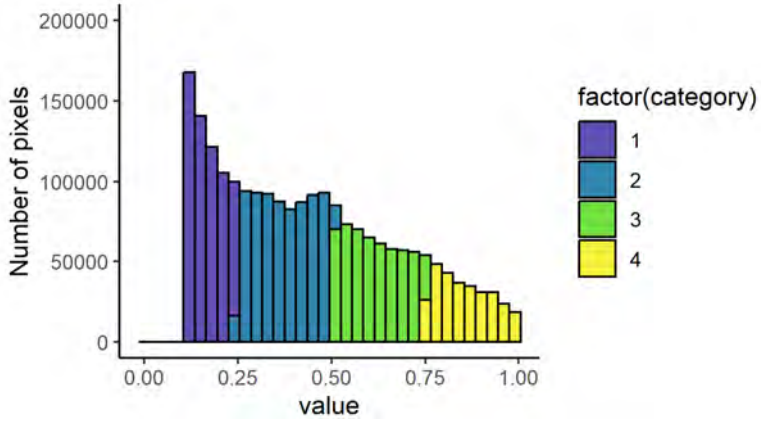
K) Table of the number of pixels in each distribution category by projection.

L) Table of the number of pixels in each distribution category by projection.

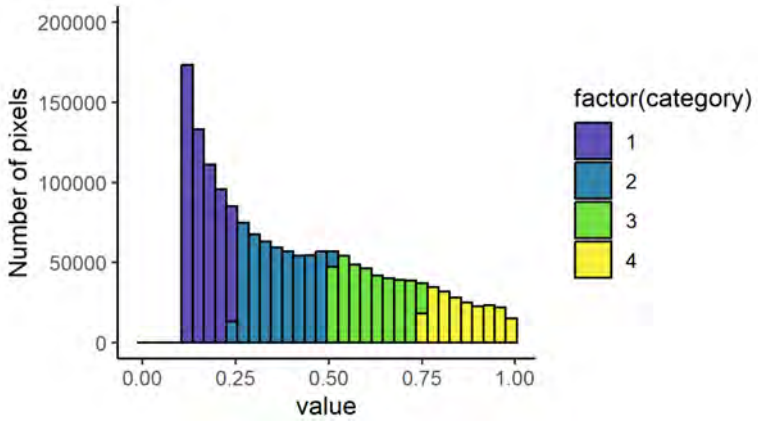
H



I



J



Stats Chat

What can I do with the histogram data? **The histogram is a great visual to show the change in pixels. Also, you can do a Chi Square test with the proportion of pixels to see if the different time periods, SSP scenarios, or distribution categories are significantly different from each other.**

	Poor	Low	Moderate	High
Current	4483162	1757816	955537	311034
SSP126	4659677	1581301	829202	292301
SSP585	5135527	1105451	595201	219997

M

	Poor	Low	Moderate	High
Current	59.71539	23.41398	12.72768	4.14295
SSP126	63.28949	21.47783	11.26254	3.970143
SSP585	72.7806	15.66643	8.435178	3.117794

Future Steps

1. **If you would like to work with more data or finish this vignette project, repeat everything from removing variables with the Pearson method and onwards with the *L. draconnis* data.**
2. **Make interesting comparison graphs with the predator and prey in QGIS.**
    - **Where do the species spatially overlap?**
    - **Does the overlap change over time or climate change scenario?**
    - **Does one of the species have a greater distance between centroids than the other?**
    - **Do the species differ in response to climate change (e.g., centroids show that one shifts away from the equator whereas the other has a contracted range.)**
3. **Look at the maps, response curves, and jackknife tests. Connect the species biology with your results without making assumptions.**
    - **What does the distribution mean for management?**
    - **Are there differences in distribution between the native and invaded ranges?**
    - **Do the response curves match with the known biology of the species?**
    - **How do the response curves of the two species compare? Do the responses to the bioclimatic variables explain differences in the distributions or what is known about the effectiveness of the predator (or how competitive they are, ect., if you are doing competing species, rather than a predator, prey community).**
4. **Write the manuscript. There are some tip and reminders in the next section.**

Resources

Araújo et al. 2005. https://doi.org/10.1111/j.1365-2486.2005.01000.x

Turbek et al. 2016. https://doi.org/10.1002/bes2.1258

General writing insights: https://besjournals.onlinelibrary.wiley.com/hub/journal/13652745/journal-resources/guide-to-scientific-writing

General writing insights: https://conservationbytes.com/2012/10/22/how-to-write-a-scientific-paper/

Writing the Manuscript

There are many peer-reviewed published manuscripts that include species distribution models. If you are writing to publish your own data and models, please keep reproducibility, accessibility, and your conclusions in mind. Unfortunately, many articles are particularly weak in these three areas. Make sure to strengthen your own writing so that you provide a clear message that is biologically sound. Use the check lists below to assess your manuscript as you are writing.

Reproducibility

- ☐ Can another individual repeat your methods using the same program?
  - ☐ Data is provided in supplemental materials.
  - ☐ Explain settings that were optimized.
  - ☐ List settings were done with the default settings.
  - ☐ State the version of the program that was used
- ☐ State reasoning for the different optimization methods.
- ☐ Mention the selected model, including the feature class and regularization multiplier, in the methods or results.
- ☐ Include all of the R packages and programs used with proper citations that will also be in the reference section.

Accessibility

- ☐ Are the figures formatted to be accessible?
  - ☐ Colors are discernible by the vast majority of the seeing population.
  - ☐ Font sizes are large enough.
  - ☐ Figure captions include descriptions that go beyond just stating colors and axis labels.

Conclusions

- ☐ Is your discussion more than just a repeat of your results?
  - ☐ Connect the known biology to the maps.
  - ☐ Make comparisons between your species and describe likely reasons for differences and similarities. Examples: Why and how do the distributions differ. Why and how do the responses to bioclimatic variables differ?
  - ☐ Incorporate the response curves and jackknife tests. Examples: Does it make biological sense that the top ranking bioclimatic variables most impact distribution? Are these values within the known life history of the species?
  - ☐ State the implications of your work. Examples: Are there areas that should be focused on for management? Is there a time line for when management should start? Will the predator-prey relationship become spatially decoupled?
- ☐ Check for conclusions that make overreaching arguments. Do they make connections that are beyond the results? If so, reword, find different evidence, or delete.