Chapter 5. Running a SDM in MaxEnt: from Start to Finish

Below is a brief overview of *my view* of the **best practices** of correlative species distribution modeling and how SDMtoolbox will facilitate achieving them. This overview focuses on modeling in MaxEnt, but many steps are applicable to all types of distribution modeling. For overview of major assumptions and other considerations, see table at the end of this document.

Species distribution modelling (SDM) occurs in two phases: 1) **Data compilation** and 2) **Model creation, calibration, and validation**

Data compilation

This step includes collecting occurrence records of the focal species and environment data for its habitats.

Occurrence Data

The single most important component of any SDM is the input occurrence records. Extra care should go into selecting, and then processing, these points. The quality, distribution and number of points are directly related to the accuracy of the model. Use as many high-quality locality points as possible (e.g. GPS data collected with confident taxonomic identification) and try to collect occurrence records that are evenly sampled throughout the species' range and avoid biases in the sampling method (e.g., sampling only from road transects). It is better to have only a limited number of points that satisfy the above conditions than many points of vague credence (e.g. be skeptical of points downloaded from internet databases, particularly those that are georeferenced from locality info) (Chan et al. 2011).

Environment Data

The environment data provide the landscape-level data to quantify the focal species' ecological tolerances. Include variables that are likely to be directly relevant to the species being modeled. However do not add all available climate data without regard to the redundancy of the data. Many environmental variables are tightly correlated making some redundant, this makes interpreting the influences of each variable in the model difficult. If not included in your model, consider the effects of the following items on the present distribution of your species: fire history, glaciations, contagious diseases, anthropogenic factors, recent geological changes, the species' movement potential through the landscape or biotic interactions.

1. Preparing Worldclim Climate Data for use in MaxEnt Analyses

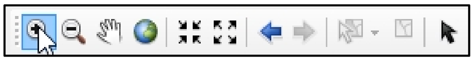
1A. Preparing Worldclim Climate Data: Clip the raster to area of species' extent

Tools: Extract by Mask (Folder) and Raster to ASCII

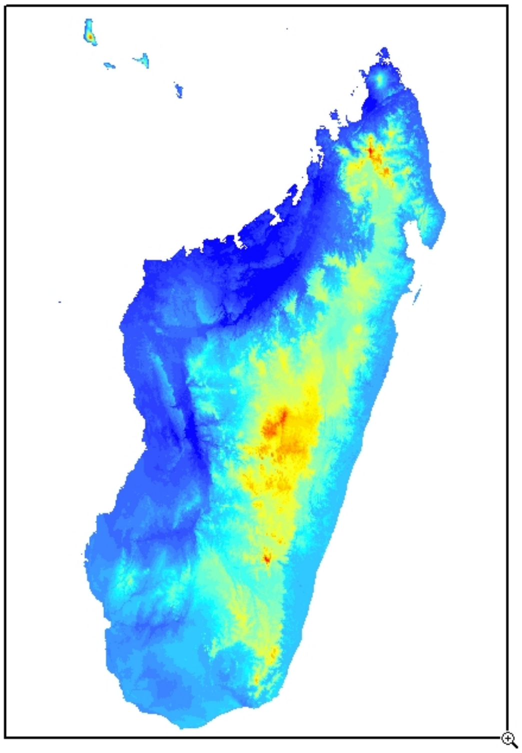
ARCGIS STEP-BY-STEP GUIDE:

1. Open a fresh ArcMap document
2. Download ESRI grid climate data (e.g. the 30 arc-second bioclim) from worldclim.org
3. Open one of the newly downloaded layers in ArcMap

4. There are several ways to define the area to clip the climate data into. One of the easiest ways is to simply zoom the display window the desired extent (Image below, where I wanted to reduce climate data to Madagascar). Other ways include defining the max-min XY coordinates (a bounding box) or using another GIS layer as a templates (such as, a country's boundary). Select one of these methods and continue. Note the area should encompass an area about 50-100 km (ca. 0.5-1 degree) greater than total distribution of **all** your focal species. We will then use bias files to limit background selection of each species to meaningful areas within this area.



Zoom tool



Zoom to desired area

5. Double-click the 'Basic Tools →Raster Tools → 1. Extract by Mask (Folder)
6. Continue to tool interface instructions (below)

SDMTOOLBOX STEP-BY-STEP GUIDE:

1. Input folder containing the full extent Worldclim data, input raster should be ESRI grid format
2. Check the box depicting that the rasters are ESRI grid format.
3. Select output folder location
4. Select either a 'mask' or select 'extent' and choose the appropriate method for defining the extent (here I used 'Same as Display').
5. Execute tool

1B. Preparing Worldclim Climate Data: Convert rasters to ASCII

Tool: Raster to ASCII

ArcGIS Step-by-Step Guide:

1. Double-click the 'Basic Tools →Raster Tools → 2a. Raster to ASCII (Folder)
2. Continue to tool interface instructions (below)

SDMtoolbox Step-by-Step Guide:

1. Input folder containing the clipped Worldclim data output from previous tool
2. Select raster type "Tiff(.tif)"
3. Select output folder location
4. The climate data are ready for use in MaxEnt

1C. Preparing Worldclim Climate Data: Define projection of ASCII climate data

Tool: 3d. Define Projection as WGS84 or ArcMaps's Define Projection

IF **input data are coordinates and WGS84 then:**

ArcGIS Step-by-Step Guide:

1. Double-click the 'Raster Tools → 3d. Define Projection as WGS84 (folder)'
2. Continue to tool interface instructions (below)

SDMtoolbox Step-by-Step Guide:

1. Input folder with clipped ASCII worldclim data.
2. Execute tool

IF **input data are another projection, use ArcMap's tool: "Define Projection".**

2. Optional Step. Which variables should I use? Testing Autocorrelations of Environmental Data

If you are interested interpreting how each input environmental variable contributes to your species distribution model, then you need to reduce autocorrelation of your input environmental data by removing highly correlated variables. It is widely known that many climate variables are highly correlated with each other. While including all these will not affect the predictive quality of your MaxEnt model, it does seriously limit any inference of the contribution of any correlated variables (*i.e.* the MaxEnt outputs from 'Analysis of variable contributions' and to some degree 'Jackknifing environmental variables'). This is mainly because when a model is built in MaxEnt, if a highly correlated variable is included in the model, this often excludes all other highly correlated variables from being incorporated. This is because these variables likely would contribute similarly to the models. Since they are not included, they will not be properly represented in the output 'Analysis of variable contributions'.

Tool: Explore Climate Data: Remove Highly Correlated Variables

ARCGIS STEP-BY-STEP GUIDE:

1. Double-click the 'Basic Tools →SDM Tools → Universal SDM Tools → Explore Climate Data → Remove Highly Correlated Variables '
2. Continue to tool interface instructions (below)

SDMTOOLBOX STEP-BY-STEP GUIDE:

1. Select all the clipped Worldclim data ('control+shift' will allow you to select all items in a folder). Layers that you wish to retain (vs. the other correlated layers) should be first in the list. All correlated layers that occur after will be excluded. For interpreting influence of environmental layers in the SDM, I prefer to place layers that depict metrics frequently used in non-SDM ecology and evolution studies [such as: BIO1 = Annual Mean Temperature, BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp)), BIO12 = Annual Precipitation]. Further for simplicity, these layers often best represent the original input climate data (as they directly reflect the actual measurements) and are not derived from several layers or a subset of the data.
2. Maximum correlation allowed. Multiple values can be input separated by semicolon (';'). Input a value between 0-1. The absolute value of the correlation coefficients range from 0 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly. A value of 0 implies that there is no linear correlation between the variables.
3. Input NoData Value. Note this must be the same for all values or else correlations will not be accurate. Since we used only Worldclim here, this should not be an issues (as all values are the same). To check NoData values, import layers into ArcGIS and right click the layer and select 'Properties' and then go to the 'Source' tab. Alternatively, you can simply open your ".asc" files in a text editor and at the top of header will be the NoData value.
4. Select output folder location. Output will be two tables with the correlation coefficients among all comparisons and a table with the final list of rasters to include in your model.

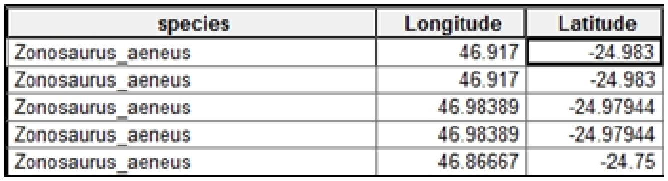
3. Preparing Occurrence Data for use in MaxEnt Analyses

3A. Preparing Occurrence Data: Import Species Occurrence Records

Tool: *CSV, TXT, XLS to shapefile*

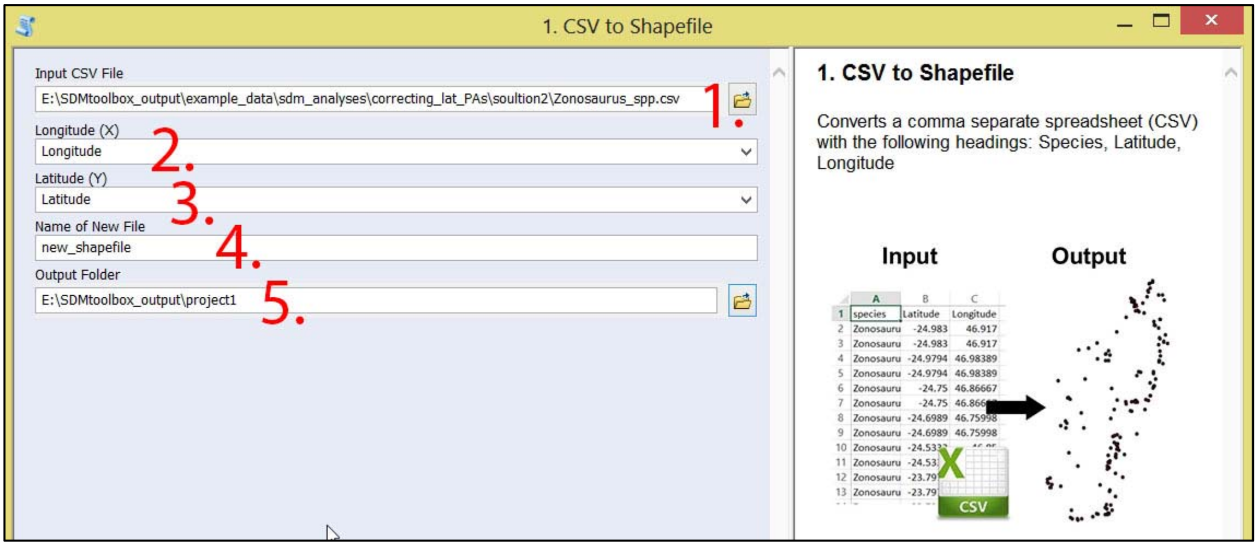
ARCGIS STEP-BY-STEP GUIDE:

1. Open a fresh ArcMap document
2. Import CSV, TXT or XLS file with occurrence records. Table of species occurrences containing ONLY: Species ID, Longitude and Latitude. The fields must be in that order. For species ID, do not use non-alphanumeric characters in names (e.g., * : \ / < > | " ? [ ] ; = + & £ $), replace spaces with "_", and remove periods.
3. Double-click the ‘CSV & Shapefile Tools → 1. CSV to Shapefile’
4. Continue to tool interface instructions (below)



species	Longitude	Latitude
Zonosaurus_aeneus	46.917	-24.983
Zonosaurus_aeneus	46.917	-24.983
Zonosaurus_aeneus	46.98389	-24.97944
Zonosaurus_aeneus	46.98389	-24.97944
Zonosaurus_aeneus	46.86667	-24.75

MaxEnt Input Species Data Format: three columns with species, longitude and latitude--- as ordered here.



1. CSV to Shapefile

Input CSV File
E:\SDMtoolbox_output\example_data\sdm_analyses\correcting_lat_PAs\soultion2\Zonosaurus_spp.csv

Longitude (X)
Longitude

Latitude (Y)
Latitude

Name of New File
new_shapefile

Output Folder
E:\SDMtoolbox_output\project1

1. CSV to Shapefile

Converts a comma separate spreadsheet (CSV) with the following headings: Species, Latitude, Longitude

	A	B	C
1	species	Latitude	Longitude
2	Zonosauru	-24.983	46.917
3	Zonosauru	-24.983	46.917
4	Zonosauru	-24.9794	46.98389
5	Zonosauru	-24.9794	46.98389
6	Zonosauru	-24.75	46.86667
7	Zonosauru	-24.75	46.866
8	Zonosauru	-24.6989	46.75998
9	Zonosauru	-24.6989	46.75998
10	Zonosauru	-24.53	[illegible]
11	Zonosauru	-24.53	[illegible]
12	Zonosauru	-23.79	[illegible]
13	Zonosauru	-23.79	[illegible]

CSV to Shapefile tool interface

SDMTOOLBOX STEP-BY-STEP GUIDE:

1. Input CSV file with columns with latitude and longitude
2. Field with longitude
3. Field with latitude
4. Name of new shapefile
5. Select output folder location

3B. Preparing Occurrence Data: Define projection of occurrence points shapefile

Tool: 6b. Define Projection as WGS84 or ArcMaps's Define Projection

ARCGIS STEP-BY-STEP GUIDE:

IF input data are coordinates and WGS84 (if not see below):

1. Double-click the 'CSV & Shapefile Tools → 6b. Define Projection as WGS84'
2. Continue to tool interface instructions (below)

SDMTOOLBOX STEP-BY-STEP GUIDE:

1. Input newly imported shapefile points
2. Execute tool

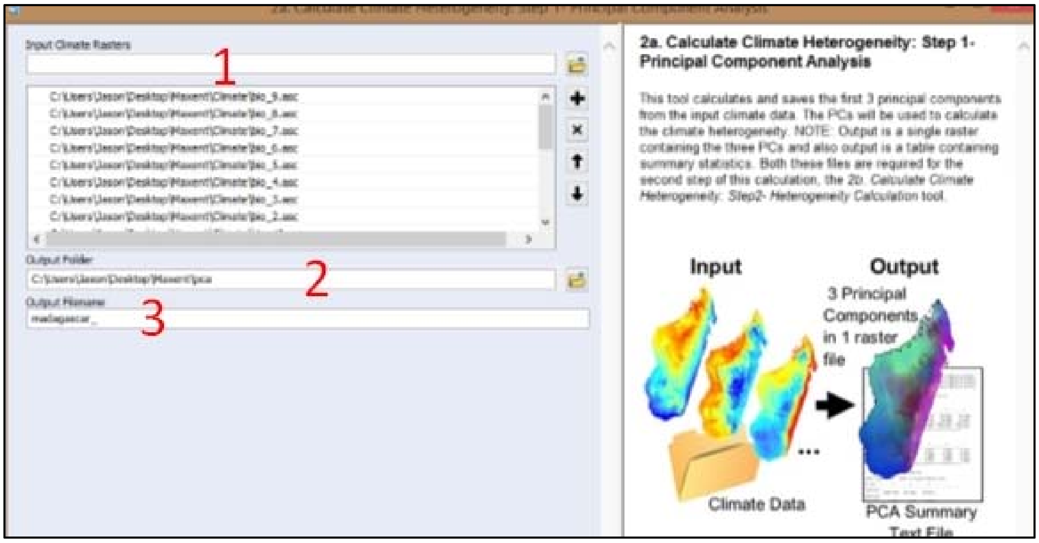
IF input data are another projection, use ArcMap's tool: "Define Projection".

3C. Preparing Occurrence Data: PCA of Climate Variables to Estimate Heterogeneity

Tool: 2a. Calculate Climate Heterogeneity: Step 1- Principal Component Analysis

ARCGIS STEP-BY-STEP GUIDE:

1. Double-click the 'SDM Tools → 1. Universal Tools → 2a. Calculate Climate Heterogeneity: Step 1- Principal Component Analysis' tool
2. Continue to tool interface instructions (below)



Input Climate Rasters
1
C:\Users\Jason\Desktop\Maxent\Climate\bio_9.asc
C:\Users\Jason\Desktop\Maxent\Climate\bio_8.asc
C:\Users\Jason\Desktop\Maxent\Climate\bio_7.asc
C:\Users\Jason\Desktop\Maxent\Climate\bio_6.asc
C:\Users\Jason\Desktop\Maxent\Climate\bio_5.asc
C:\Users\Jason\Desktop\Maxent\Climate\bio_4.asc
C:\Users\Jason\Desktop\Maxent\Climate\bio_3.asc
C:\Users\Jason\Desktop\Maxent\Climate\bio_2.asc
Output Folder
C:\Users\Jason\Desktop\Maxent\pca
2
Output Filename
madagascar_
3

2a. Calculate Climate Heterogeneity: Step 1- Principal Component Analysis

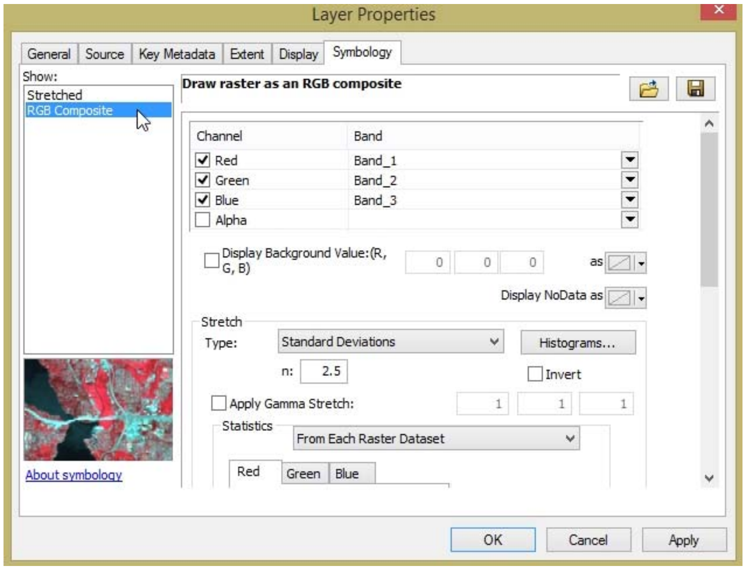
This tool calculates and saves the first 3 principal components from the input climate data. The PCs will be used to calculate the climate heterogeneity. NOTE: Output is a single raster containing the three PCs and also output is a table containing summary statistics. Both these files are required for the second step of this calculation, the 2b. Calculate Climate Heterogeneity: Step2- Heterogeneity Calculation tool.

2a. Calculate Climate Heterogeneity: Step 1- Principal Component Analysis tool interface

SDMTOOLBOX STEP-BY-STEP GUIDE:

1. Input all your climate rasters that depict continuous data. Note: hold 'shift' to select many rasters at once.

2. Select output folder location. This should be a new empty folder. If not empty this can cause the analysis to fail, particularly if temporary files from a previous analysis were not properly removed (e.g. this can happen if another SDMtoolbox analysis is terminated early).
3. Name of output file. Note that "_Climate_PCs" will be appended to file name.



Layer Properties

General | Source | Key Metadata | Extent | Display | Symbology

Show:
Stretched
RGB Composite

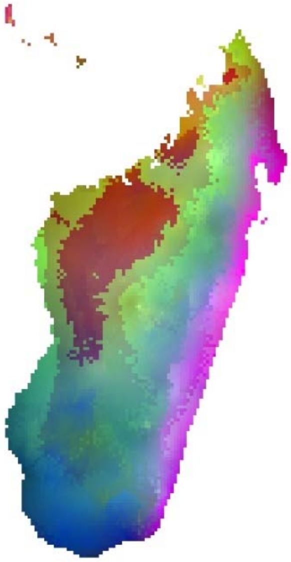
Draw raster as an RGB composite

Channel	Band
☑ Red	Band_1
☑ Green	Band_2
☑ Blue	Band_3
☐ Alpha	

☐ Display Background Value:(R, G, B) 0 0 0 as

Display NoData as

Stretch
Type: Standard Deviations
n: 2.5
☐ Invert
☐ Apply Gamma Stretch: 1 1 1
Statistics: From Each Raster Dataset



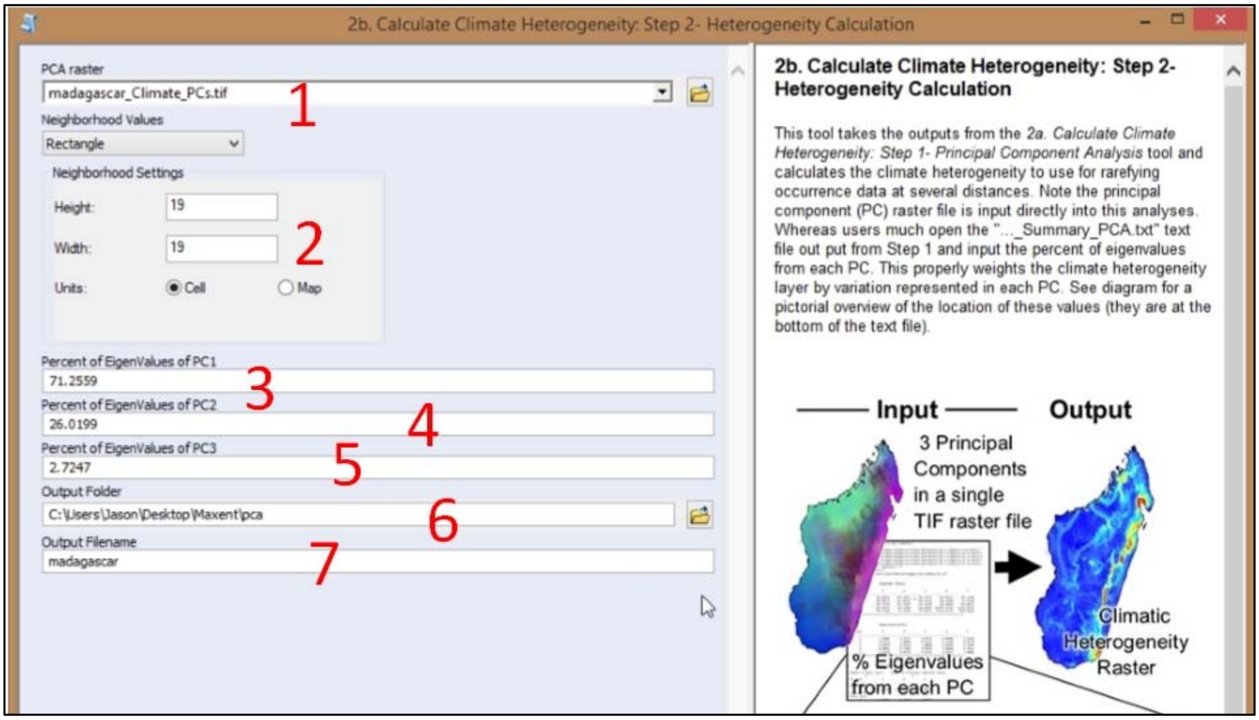
To visualize all three bands output from the PCA tool, right click layer and select ‘Properties’. Then select the ‘Symbology’ tab and select ‘RGB Composite’. The output raster depicts climate space: the more similar the colors the more similar values.

3D. Preparing Occurrence Data: Measure Spatial Heterogeneity of Climate PCs

Tool: 2b. Calculate Climate Heterogeneity: Step 2- Heterogeneity Calculation

ARCGIS STEP-BY-STEP GUIDE:

1. Double-click the 'SDM Tools → 1. Universal Tools →2b. Calculate Climate Heterogeneity: Step 2- Heterogeneity Calculation' tool
2. Continue to tool interface instructions (below)



2b. Calculate Climate Heterogeneity: Step 2- Heterogeneity Calculation

PCA raster: madagascar_Climate_PCs.tif

Neighborhood Values: Rectangle

Neighborhood Settings

Height: 19

Width: 19

Units: ● Cell ○ Map

Percent of EigenValues of PC1: 71.2559

Percent of EigenValues of PC2: 26.0199

Percent of EigenValues of PC3: 2.7247

Output Folder: C:\Users\Jason\Desktop\Maxent\pca

Output Filename: madagascar

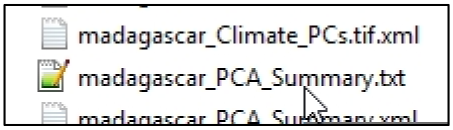
2b. Calculate Climate Heterogeneity: Step 2- Heterogeneity Calculation

This tool takes the outputs from the *2a. Calculate Climate Heterogeneity: Step 1- Principal Component Analysis* tool and calculates the climate heterogeneity to use for rarefying occurrence data at several distances. Note the principal component (PC) raster file is input directly into this analyses. Whereas users much open the "..._Summary_PCA.txt" text file out put from Step 1 and input the percent of eigenvalues from each PC. This properly weights the climate heterogeneity layer by variation represented in each PC. See diagram for a pictorial overview of the location of these values (they are at the bottom of the text file).

Input
Output
3 Principal Components in a single TIF raster file
% Eigenvalues from each PC
Climatic Heterogeneity Raster

2b. Calculate Climate Heterogeneity: Step 2- Heterogeneity Calculation tool interface

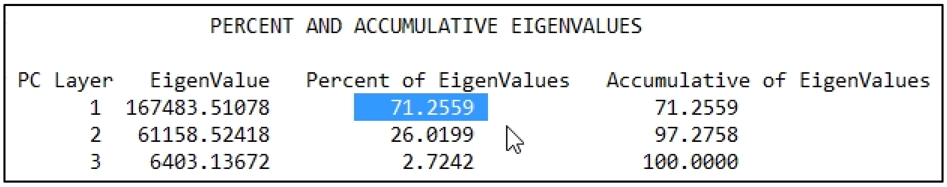
SDMTOOLBOX STEP-BY-STEP GUIDE:



madagascar_Climate_PCs.tif.xml
madagascar_PCA_Summary.txt
madagascar_PCA_Summary.xml

Output PCA Summary table

1. Input climates PCA raster output from previous step
2. This is the spatial scale used to calculate the heterogeneity of the landscape. E.g., if 3 x 3 rectangle and cell units are selected: heterogeneity values will be calculated from each raster pixel and the 8 cells neighboring the focal cell.
3. Percent of EigenValues of PC1. This value is input from the "..._PCA_summary.txt" file. Go to the bottom of the text file to the heading 'PRECENT AND ACCUMULATIVE



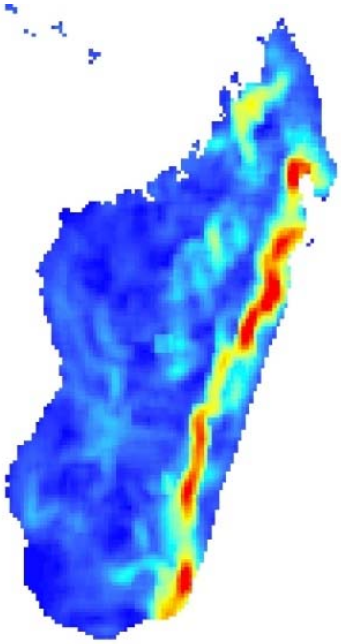
```
                    PERCENT AND ACCUMULATIVE EIGENVALUES

PC Layer    EigenValue    Percent of EigenValues    Accumulative of EigenValues
       1  167483.51078              71.2559                   71.2559
       2   61158.52418              26.0199                   97.2758
       3    6403.13672               2.7242                  100.0000
```

Section of the table that is necessary for input into this tool

EIGENVALUES' and go to the row corresponding to PC1 and input the value from the 'Percent of EigenValues'.

4. Percent of EigenValues of PC2.
5. Percent of EigenValues of PC3.
6. Output Folder
7. Output filename. Note: "_clim_hetero.tif" will automatically be appended.



Above. **Output climate heterogeneity raster. Here warm colors depict high areas of climatic heterogeneity.**

3E. Preparing Occurrence Data: Graduated Spatial Rarefying

An Unbiased Sample: a Need for Spatial Rarefying

Most SDM methods require input occurrence data to be spatially independent to perform well. However, it is common for researchers to introduce environmental biases into their SDMs from spatially auto-correlated occurrence points. The elimination of spatial clusters of localities is important for model calibrating and evaluation. When spatial clusters of localities exist, often models are over-fit towards environmental biases (reducing the model's ability to predict spatially independent data) and model performance values are inflated (Veloz 2009; Hijimans *et al.* 2012; Boria *et al.* 2014). The *spatially rarefy occurrence data* tool addresses this issue by spatially filtering locality data by a user input distance, reducing occurrence localities to a single point within the specified Euclidian distance. This tool also allows users to spatially rarefy their data at several distances according to habitat, topographic or climate heterogeneity (Table 1d). For example, occurrence localities could be spatially filtered at 5 km^2, 10 km^2 and 30 km^2 in areas of high, medium and low environmental heterogeneity, respectively. This graduated filtering method is particular useful for studies with limited occurrence points and can maximize the number of spatially independent localities.

-Veloz, S. D. (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography,* 36, 2290–2299.
-Hijmans, R.J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology,* 93, 679–688.
-Boria R. A., Olson L.E., Goodman S.M. & Anderson R.A. (2014) Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modeling,* 275, 73-77.

Tool: Spatially Rarefy Occurrence Data

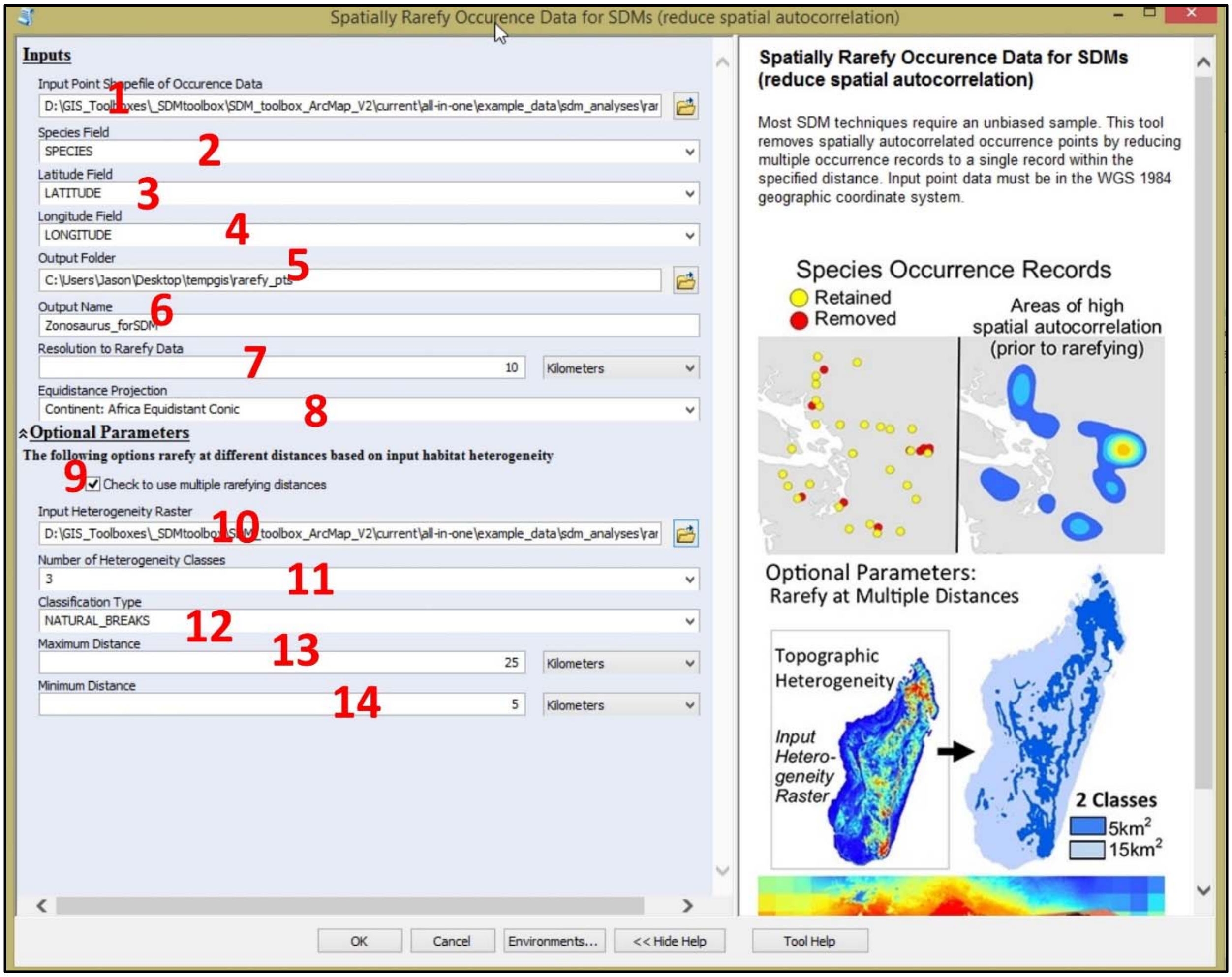
ArcGIS Step-by-Step Guide:

1. Double-click the 'SDM Tools → 1. Universal Tools → Spatially Rarefy Occurrence Data for SDMs (reduce spatial autocorrelation)' tool
2. Continue to tool interface instructions (below)

SDMtoolbox Step-by-Step Guide:

1. Input shapefile of species occurrence data, here select 'occurences.shp'
2. Shapefile field corresponding to species identity, here select 'SPECIES'
3. Shapefile field corresponding to latitude, here select 'LATITUDE'
4. Shapefile field corresponding to longitude, here select 'LONGITUDE'
5. Select output folder location. This should be a new empty folder. If not empty this can cause the analysis to fail, particularly if temporary files from a previous analysis were not properly removed (e.g. this can happen if another SDMtoolbox analysis is terminated early).
6. Name of output file. Note that "_rarefied_points" will be appended to file name.
7. The spatial resolution to rarefy the data. Here use the default settings. Note that this value will NOT actually be used here because you will be executing the multi-distance occurrence data rarefying.
8. Please select proper equidistance projection.
9. Placing a check mark in the box to execute the multi-distance data rarefying
10. Input heterogeneity raster.
11. The number of classes, here 5
12. Classification type, here 'NATURAL_BREAKS'

13. The maximum distance, here 25 Kilometers
14. The minimum distance, here 2 Kilometers



Spatially Rarefy Occurence Data for SDMs (reduce spatial autocorrelation)

Inputs

Input Point Shapefile of Occurence Data: D:\GIS_Toolboxes_SDMtoolbox\SDM_toolbox_ArcMap_V2\current\all-in-one\example_data\sdm_analyses\rar

Species Field: SPECIES

Latitude Field: LATITUDE

Longitude Field: LONGITUDE

Output Folder: C:\Users\Jason\Desktop\tempgis\rarefy_pts

Output Name: Zonosaurus_forSDM

Resolution to Rarefy Data: 10 Kilometers

Equidistance Projection: Continent: Africa Equidistant Conic

Optional Parameters

The following options rarefy at different distances based on input habitat heterogeneity

☑ Check to use multiple rarefying distances

Input Heterogeneity Raster: D:\GIS_Toolboxes_SDMtoolbox\SDM_toolbox_ArcMap_V2\current\all-in-one\example_data\sdm_analyses\rar

Number of Heterogeneity Classes: 3

Classification Type: NATURAL_BREAKS

Maximum Distance: 25 Kilometers

Minimum Distance: 5 Kilometers

Spatially Rarefy Occurence Data for SDMs (reduce spatial autocorrelation)

Most SDM techniques require an unbiased sample. This tool removes spatially autocorrelated occurrence points by reducing multiple occurrence records to a single record within the specified distance. Input point data must be in the WGS 1984 geographic coordinate system.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
Species Occurrence Records
Retained
Removed
Areas of high spatial autocorrelation (prior to rarefying)
Optional Parameters: Rarefy at Multiple Distances
Topographic Heterogeneity
Input Hetero-geneity Raster
2 Classes
5km²
15km²
OK
Cancel
Environments...
<< Hide Help
Tool Help

***Spatially Rarefy Occurrence Data* tool interface**

4. Creation of Bias Files

Background Selection via Bias Files

A subset of python scripts create bias files used to fine-tune background and occurrence point selection in Maxent. Bias files control where background points are selected and the density of background sampling. Proper use of bias files can avoid sampling habitat greatly outside of a species' known occurrence or can account for collection sampling biases with coordinate data.

Background points (and similar pseudo-absence points) are meant to be compared with the presence data and help differentiate the environmental conditions under which a species can potentially occur. Typically background points are selected within a large rectilinear area, within this area there often exist habitat that is environmentally suitable, but was never colonized. When background points are selected within these habitats, this increases commission errors (false-positives). As a result, the 'best' performing model tends to be over-fit because selection criterion favor a model that fail to predict the species in the un-colonized climatically suitable habitat (Anderson & Raza 2010, Barbet-Massin et al. 2012). The likelihood that suitable unoccupied habitats are included in background sampling increases with Euclidian distance from the species' realized range. Thus, a larger study spatial extent can lead to the selection of a higher proportion of less informative background points (Barbet-Massin et al. 2012). Researchers should *not* avoid studying species with broad distributions or those existing in regions that do not conform well to rectilinear map layouts, rather they simply need to be more selective in the choice of background points in Maxent (and pseudo-absences in other SDM methods)(Barve et al. 2011; Merow et al. 2013).

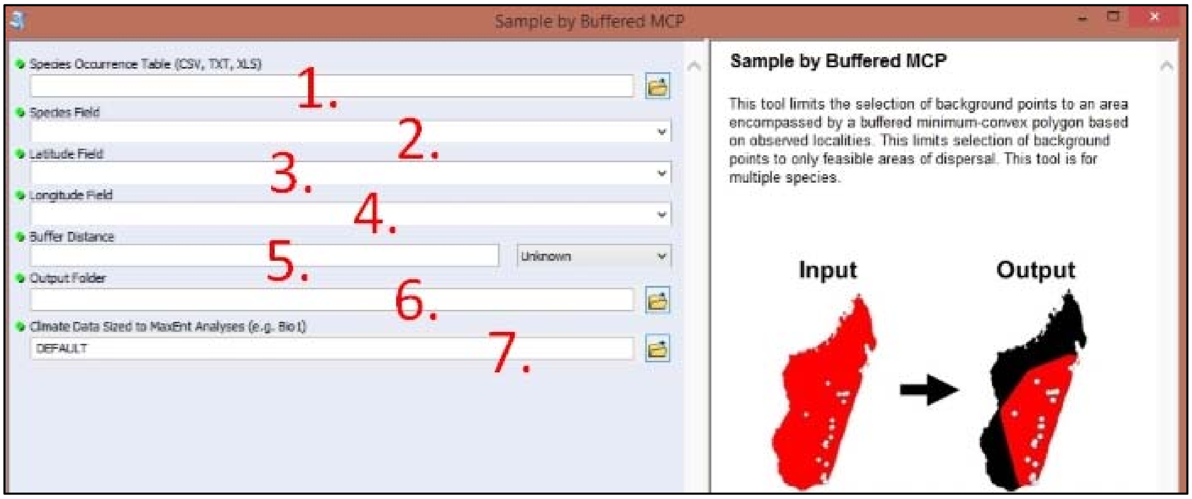
To circumvent this problem, many researchers have begun using background point and pseudo-absence selection methods that are more regional. SDMtoolbox contains three tools to facilitate more sophisticated background selection for use in Maxent. The *Sample by Distance from Obs. Pts.* tool (see: SDM Tools →2. MaxEnt Tools → Background Selection via Bias Files) uses a common method that samples backgrounds within a maximum radial distance of known occurrences (see Thuiller et al. 2009). The Sample by buffered MCP tool restricts background selection with a buffered minimum-convex polygons based on known occurrences (see following guide).

-Anderson, R. P. & Raza, A. (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography*, 37, 1378-1393.
-Barbet-Massin, M., Jiguet, F., Albert, C. H. & Thuiller, W. (2012) Selecting pseudoabsences for species distribution models: how, where and how many? *Methods in Ecology and Evolution,* 3, 327–338.
-Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudoabsence data. *Ecological Applications*, 19, 181-197.
-Thuiller, W., Lafourcade, B., Engler, R. & Araujo, M. B. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, 32, 369–373.

Tool: Background Selection: Sample by Buffered MCP, Sample by Distance from Obs. Pts. or Sample by Buffered Local Adaptive Convex-Hull

ARCGIS STEP-BY-STEP GUIDE:

1. Decide if you want to restrict background sampling to: i) a buffered minimum convex polygon based on observation localities, ii) a radial distance from all occurrence points or an intermediate between those two iii) using the tool: *Sample by Buffered Local Adaptive Convex-Hull*.
2. Double-click the corresponding tool 'SDM Tools →2. MaxEnt Tools → Background Selection via Bias Files→ Background Selection: Sample by Buffered MCP' or 'Background Selection: Sample by Distance from Obs. Pts.' or 'Background Selection: Sample by Buffered Local Adaptive Convex-Hull'
3. Continue to tool interface instructions (Buffered MCP below)



Sample by Buffered MCP

- Species Occurrence Table (CSV, TXT, XLS)
- Species Field
- Latitude Field
- Longitude Field
- Buffer Distance: Unknown
- Output Folder
- Climate Data Sized to MaxEnt Analyses (e.g. Bio1): DEFAULT

Sample by Buffered MCP

This tool limits the selection of background points to an area encompassed by a buffered minimum-convex polygon based on observed localities. This limits selection of background points to only feasible areas of dispersal. This tool is for multiple species.

Background Selection: Sample by Buffered MCP **tool interface**

SDMTOOLBOX STEP-BY-STEP GUIDE:

1. CSV file of species occurrences (from previous steps)
2. Field corresponding to species identity
3. File field corresponding to longitude
4. File field corresponding to latitude
5. The distance outside of minimum-convex-polygon included in background selection.
6. Select output folder location. This should be a new empty folder. If not empty this can cause the analysis to fail, particularly if temporary files from a previous analysis were not properly removed (e.g. this can happen if another SDMtoolbox analysis is terminated early).
7. Climate data sized to extent of MaxEnt Modeling. Here use the 'Bio_1.asc' layer. Select one of your climate files sized to your modeling extent (e.g. Bio1.asc). This file will be used to match the bias file to proper extent and resolution (no change will be made to this file).

Model calibration and validation

5. Spatial Jackknifing and independent tests of parameters

Why use SDMtoolbox for MaxEnt modeling:

I. Spatial Jackknifing

Spatial jackknifing (or geographically structured k-fold cross-validation) tests evaluation performance of spatially segregated spatially independent localities. SDMtoolbox automatically generates all the GIS files necessary to spatially jackknife your MaxEnt Models. The script splits the landscape into 3-5 regions based on spatial clustering of occurrence points (e.g. if 3: A,B,C). Models are calibrated with k-1 spatial groups and then evaluated with the withheld group. For example if k=3, models would be run with following three subgroups:

1. Model is calibrated with localities and background points from region AB and then evaluated with points from region C
2. Model is calibrated with localities and background points from region AC and then evaluated with points from region B
3. Model is calibrated with localities and background points from region BC and then evaluated with points from region A

II. Independent Tests of Model Feature Classes and Regularization Parameters

Equally important, this tool allows for testing different combinations of five model feature class types (FC) and regularization multiplier(s) (RM) to optimize your MaxEnt model performance. For example, if a RM was input (here 5), this tool kit would run MaxEnt models on the following parameters for each species:

1. RM: 5 & FC: Linear, 2. RM: 5 & FC: Linear and Quadratic, 3. RM: 5 & FC: Hinge, 4. RM: 5 & FC: Linear, Quadratic and Hinge, 5. RM: 5 & FC: Linear, Quadratic, Hinge, Product and Threshold

III. Automatic Model Selection

Finally, the script chooses the best model by evaluating each model's: 1. omission rates (OR)*,2. AUC**, and 3. model feature class complexity. It does this in order, choosing the model with the lowest omission rates on the test data. If many models have the identical low OR, then it selects the model with the highest AUC. Lastly if several models have the same low OR and high AUC, it will choose the model with simplest feature class parameters in the following order: 1. linear; 2. linear and quadratic; 3. hinge; 4. linear, quadratic, and hinge; and 5. linear, quadratic, hinge, product, and threshold.

Once the best model is selected, SDMtoolbox will run the final model using all the occurrence points. If desired, at this stage models will be projected into other climates, environmental variables will be jackknifed to measure importance, and response curves will be created.

*For each iteration, OR is weighted by the number of points in the evaluation subgroup. This is necessary because spatial groups may not have identical number of points. The weighing gives equal contribution to all points included in model evaluation.

**AUC is calculated from the total study area in the input bias file (if k=3, then all groups: ABC)

For info and justification for each of these methods see:
-Boria, R. A., L. E. Olson, S. M. Goodman, and R. P. Anderson. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecological Modelling, 275:73-77.
-Radosavljevic, A. and R. P. Anderson. 2014. Making better Maxent models of species distributions: complexity, overfitting, and evaluation. Journal of Biogeography. 41:629-643
-Shcheglovitova, M. and R. P. Anderson. 2013. Estimating optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. Ecological Modelling, 269:9-17.

Tool: Run MaxEnt: Spatially Jackknife

ArcGIS Step-by-Step Guide:

1. Double-click the 'SDM Tools →2. MaxEnt Tools → Modeling in MaxEnt → Run MaxEnt: Spatially Jackknife' tool
2. Continue to tool interface instructions (below)

SDMtoolbox Step-by-Step Guide:

IMPORTANT NOTE 1: none of the input and output file names/file paths can have spaces in them. If there are any spaces, the output batch scripts will likely fail to work properly.

IMPORTANT NOTE 2: Upon first use of this tool, due to its unique syntax, you need to specify the location of the menu file. If you do not due this, the menu presented by ArcGIS will not make complete sense. For a detailed overview of how to do this (it will take 30 seconds to do), go to: http://www.sdmtoolbox.org/menu-fix-spatial-jackknife

1. Folder with the MaxEnt program. Downloadable from: http://www.cs.princeton.edu/~schapire/maxent/
2. CSV file output from spatial rarefied occurrence data (output from step 3D)
3. CSV field corresponding to species identity
4. CSV field corresponding to longitude
5. CSV field corresponding to latitude
6. Folder with clipped ASCII climate data from step 3C.
7. Select environmental layers from this folder that depict categorical data.
8. Select environmental layers to exclude from analyses (not list in output from step 2).
9. Folder containing bias files from step 4.
10. Output folder for GIS, python scripts and MaxEnt batch files. Output will include all GIS files necessary to run models.
11. For spatial jackknifing this format output will be Logistic, regardless of selection
12. For spatial jackknifing this format output will be ASCII, regardless of selection

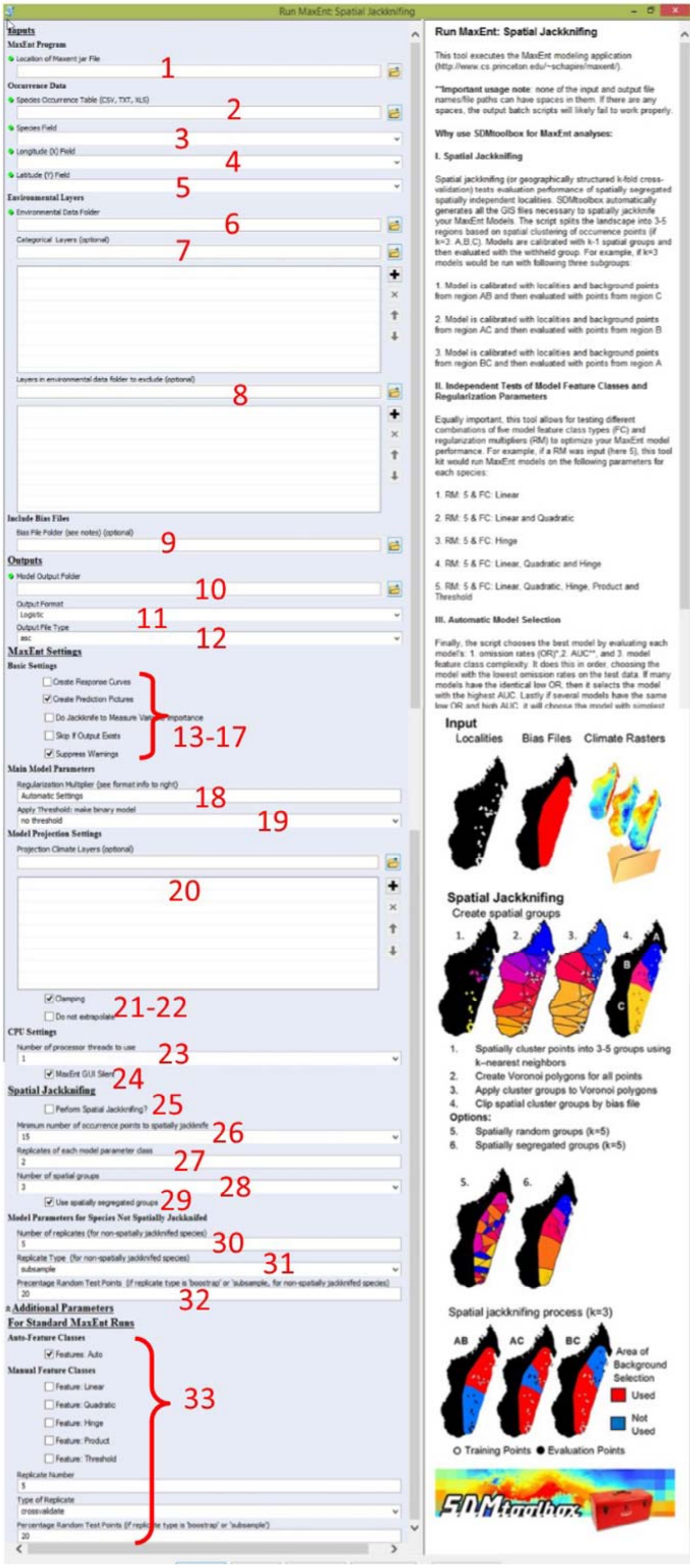
13. Create graphs showing how predicted relative probability of occurrence depends on the value of each environmental layer
14. Create an image of each output model
15. Measure variable importance by jackknifing the variables. Each variable is excluded in turn and a model created with the remaining variables. Then a model is created using each variable in isolation.
16. This will skip the model if an output exists.
17. This will suppress any warnings encountered during modeling. All warnings are always written to the log files.
18. This is the breadth of the model. A higher number gives a more spread-out distribution. Input many values separated by semi-colons.

For example:

0.5; 1; 1.5; 2; 2.5; 3; 4; 5

The default value is 1.

Remember the more values input will produce more SDMs created and require more computation time. For each regularization multiplier (RM), 15 models will be run (5 feature class groups and 3-5 spatial jackknife groups, 5x3=15 to 5x5=25). This number is multiplied for each replicate and each species


Run MaxEnt: Spatial Jackknifing

Inputs
MaxEnt Program
Location of Maxent jar File
Occurrence Data
Species Occurrence Table (CSV, TXT, XLS)
Species Field
Longitude (X) Field
Latitude (Y) Field
Environmental Layers
Environmental Data Folder
Categorical Layers (optional)
Layers in environmental data folder to exclude (optional)
Include Bias Files
Bias File Folder (see notes) (optional)
Outputs
Model Output Folder
Output Format: Logistic
Output File Type: asc
MaxEnt Settings
Basic Settings
□ Create Response Curves
☑ Create Prediction Pictures
□ Do Jackknife to Measure Variable Importance
□ Skip If Output Exists
☑ Suppress Warnings
Main Model Parameters
Regularization Multiplier (see format info to right): Automatic Settings
Apply Threshold: make binary model: no threshold
Model Projection Settings
Projection Climate Layers (optional)
☑ Clamping
□ Do not extrapolate
CPU Settings
Number of processor threads to use: 1
☑ MaxEnt GUI Silent
Spatial Jackknifing
□ Perform Spatial Jackknifing?
Minimum number of occurrence points to spatially jackknife: 15
Replicates of each model parameter class: 2
Number of spatial groups: 3
☑ Use spatially segregated groups
Model Parameters for Species Not Spatially Jackknifed
Number of replicates (for non-spatially jackknifed species): 5
Replicate Type (for non-spatially jackknifed species): subsample
Precentage Random Test Points (if replicate type is 'boostrap' or 'subsample, for non-spatially jackknifed species): 20
Additional Parameters
For Standard MaxEnt Runs
Auto-Feature Classes
☑ Features: Auto
Manual Feature Classes
□ Feature: Linear
□ Feature: Quadratic
□ Feature: Hinge
□ Feature: Product
□ Feature: Threshold
Replicate Number: 5
Type of Replicate: crossvalidate
Percentage Random Test Points (if replicate type is 'boostrap' or 'subsample'): 20

Run MaxEnt: Spatial Jackknifing

This tool executes the MaxEnt modeling application (http://www.cs.princeton.edu/~schapire/maxent/).

**Important usage note: none of the input and output file names/file paths can have spaces in them. If there are any spaces, the output batch scripts will likely fail to work properly.

Why use SDMtoolbox for MaxEnt analyses:

I. Spatial Jackknifing

Spatial jackknifing (or geographically structured k-fold cross-validation) tests evaluation performance of spatially segregated spatially independent localities. SDMtoolbox automatically generates all the GIS files necessary to spatially jackknife your MaxEnt Models. The script splits the landscape into 3-5 regions based on spatial clustering of occurence points (if k=3: A,B,C). Models are calibrated with k-1 spatial groups and then evaluated with the withheld group. For example, if k=3 models would be run with following three subgroups:

1. Model is calibrated with localities and background points from region AB and then evaluated with points from region C

2. Model is calibrated with localities and background points from region AC and then evaluated with points from region B

3. Model is calibrated with localities and background points from region BC and then evaluated with points from region A

II. Independent Tests of Model Feature Classes and Regularization Parameters

Equally important, this tool allows for testing different combinations of five model feature class types (FC) and regularization multipliers (RM) to optimize your MaxEnt model performance. For example, if a RM was input (here 5), this tool kit would run MaxEnt models on the following parameters for each species:

1. RM: 5 & FC: Linear

2. RM: 5 & FC: Linear and Quadratic

3. RM: 5 & FC: Hinge

4. RM: 5 & FC: Linear, Quadratic and Hinge

5. RM: 5 & FC: Linear, Quadratic, Hinge, Product and Threshold

III. Automatic Model Selection

Finally, the script chooses the best model by evaluating each model's: 1. omission rates (OR)*,2. AUC**, and 3. model feature class complexity. It does this in order, choosing the model with the lowest omission rates on the test data. If many models have the identical low OR, then it selects the model with the highest AUC. Lastly if several models have the same low OR and high AUC, it will choose the model with simplest

Spatial Jackknifing
Create spatial groups

1. Spatially cluster points into 3-5 groups using k-nearest neighbors
2. Create Voronoi polygons for all points
3. Apply cluster groups to Voronoi polygons
4. Clip spatial cluster groups by bias file

Options:
5. Spatially random groups (k=5)
6. Spatially segregated groups (k=5)

Spatial jackknifing process (k=3)

modeled. Thus, if you have 2 species, 5 RMs and 2 replicates; this would result in 300 or 500 models run for 3 and 5 spatial jackknife groups, respectively (2 species *2 replicates* 5 RM * 15-25 models per run)

19. Apply a threshold to make binary model. This will generated a binary model in addition to the continuous model. In none is supplied, SDMtoolbox will use '10 percent minimum training presence' to calculated omission rates. If you prefer another threshold, please select it here.
20. Projection Climate Layers. Folders containing environmental data for projecting the MaxEnt models (often are future are past climates). To select multiple folders at once hold to shift. The layers MUST match the input environment layers, e.g. if Bio27 is used to build the model then the projection folder must contain an analog variable and have the identical name. Here resolution and spatial extent do not need to match input environmental layers.
21. Apply clamping when projecting
22. If checked, will not predict areas of climate space outside of limits encountered during training
23. Number of CPUs to use for modeling
24. This will *not* display the MaxEnt GUI when running models- preferred
25. Check this box to perform all the analyses described into the information window. If not checked this will run the modeling as if executed from the MaxEnt GUI (no spatial jackknifing or independent evaluation of RV or feature classes).
26. This is the minimum number of points to execute spatial jackknifing. If *below* this value, the models will be trained and evaluated using either cross-validation, bootstrapping or sub-sampling (as specified below in steps: 30-32). Each non-spatially jackknifed group is optimized with independent tests of different combinations of the five model feature class types (FC) and input regularization multiplier (RM) values.
27. Replicates of each model parameter class in spatial jackknife runs
28. Number of groups to subdivide the landscape into. Higher the number the more models run, but also the more training points included in each model run.
29. If selected:

    Groups will be spatially segregated and numbers of occurrences within groups may not be equal. This analysis is more focused on natural spatial groups. This method is best if projecting models into other climates (i.e. current or past) and is particularly useful for training and evaluating model performance in potentially non-analogous climates.

    If not selected:

    Spatial jackknife groups will be spatially randomly and numbers of occurrences within groups will be equal (+/- 1, due to unequal group sizes for some combinations of occurrences records and group number). This method is best if not projecting models into other climates.
30. Replicates of each model parameter class for species with too few points to spatially jackknife
31. Replicate Type. If replicates are >1, then multiple runs are performed by this type:

    Crossvalidate: MaxEnt makes k number of folds of your occurrence data to train and test the data. Here you are not be able to tell MaxEnt how many replicates you would like to run or the percentage of occurrence data you would like withheld for model validation (test occurrences). Optimal if you have a large number of species occurrences.

Bootstrap: Replicates samples sets are chosen by sampling with replacement

Subsample: Replicate sample sets are chosen by removing the random test percentage (input in the following window) without replacement, the variables not included are then used for model evaluation

32. If replicate type is 'boostrap' or 'subsample, input the percentage of points used for subsampling.
33. Additional parameters not used in spatial jackknifing.
34. Run the tool.
35. After files are created, to execute models go to output folder. Click the batch file "Step1_Optimize_MaxEnt_Model_Parameters.bat" this will run all models and summary stats. Know this may take several hours, or even days (if you have many species and RV), to finish.
36. Once all the models from Step 1 are run, the "Step2_Run_Optimized_MaxEnt_Models.bat" will be populated with the best model parameters. Run this file to get final models.
37. To see model ranks for each species, open the corresponding folder and open the “species_name_SUMSTATS_RANKED_MODELS.csv.” Here the best model is the first row. Feature number corresponds to feature class group with: 1=linear; 2=linear & quadratic; 3= hinge; 4=linear, quadratic, and hinge; and 5=linear, quadratic, hinge, product, and threshold.

Final remarks

Great job—only a small proportion of all published papers using SDMs/ENMs address the following best practices of modeling:

1. Used species-specific regional background sampling
2. Spatially rarefied occurrence data
3. Spatially jackknifed SDMs to calibrate model parameters
4. Independently evaluated feature class parameters and regularization multiplier(s)
5. Reduced correlation of input climate variables for interpreting influence on model (optional here)

Now consider the following SDMtoolbox tools for further analyses:

If projecting models into future or past climates:

1. Limit Dispersal in Future SDMs
2. Overprediction Correction: Clip Models by Buffered MCPs
3. Distribution Changes Between Binary SDMs:
    a. Centroid Changes (Lines)
    b. Distribution Changes Between Binary SDMs

To create betters models:

4. Correcting Latitudinal Background Selection Biases

5. Gaussian kernel density of sampling localities

To assess landscape connectivity:

6. Among all sites or between shared haplotypes
7. Create friction layers

To measure spatial biodiversity patterns of many SDMs:

8. Calculate species richness and endemism (weighted endemism & corrected weighted endemism)

Now go publish you results!

Summary of some basic considerations when generating SDMs

Table modified from: Alvarado-Serrano, D. F. and Knowles, L. L. (2014), Ecological niche models in phylogeographic studies: applications, advances and precautions. Molecular Ecology Resources, 14: 233–248. Please see paper for more details.

Assumptions that may affect SDMs	Specific considerations
Data compilation **occurrence records:** Are species presences (and absence) records representative of the actual distribution?	▪effects of species' natural history ▪geographic/environmental bias ▪intraspecific variability ▪positional uncertainty ▪sample size ▪sampling bias (e.g. towards more accessible areas) ▪taxonomic accuracy (e.g. subspecies or races) ▪temporal coverage in relation to environmental data
Data compilation **Environmental variables:** Do environmental variables accurately capture the association between species subsistence and the environment at the relevant scale?	▪data quality and biases ▪effect on species distribution (direct vs. indirect) ▪resolution in space and time ▪spatial autocorrelation ▪spatial extent ▪temporal coverage and stability ▪type (categorical vs. continuous)
Model generation and calibration Is the modelling algorithm appropriate given the data available and research question?	▪algorithm assumptions ▪algorithm performance ▪under different scenarios ▪input data type (e.g. presences only vs. presence/absences) ▪output generated (e.g. presence/absence vs. continuous prediction) ▪sensitivity to model parameters
Model generation and calibration Is the model appropriately calibrated for the data available and research question?	▪model complexity ▪model selection procedure ▪setting of model parameters ▪variable selection strategy
Model validation Is validation performed on truly independent data and under appropriate settings?	▪assumptions/limitations of accuracy measurement ▪importance of use of multiple metrics ▪sensitivity to model parameters ▪threshold transformation of continuous predictions
Model projection Is the species environment relationship likely to be maintained in space and/or time?	▪availability of validation data in projected regions ▪likelihood of niche shifts ▪model uncertainty ▪model transferability ▪risks of interpolation and extrapolation