megaSDM Setup Instructions

NOTE: An example run has been provided with the code on GitHub (https://github.com/brshipley/megaSDM/blob/master/EXAMPLE.zip). The scripts and configuration file for this run are in EXAMPLE/datadirectory/scripts.

Required Data:

To run this program, the following data and programs are required:

1. A ***.csv*** file listing the species to be used in the analysis with two columns:
    a. Column 1: The higher-order taxon of each species to be used: this can be any taxonomic level, but class or order is recommended
        i. If occurrence points need to be download from GBIF, the high-order taxon should follow the GBIF Backbone Taxonomy (e.g.: Soricomorpha instead of Eulipotyphla). https://www.gbif.org/species/search
    b. Column 2: The scientific name of each species to be used
    c. Column 3 (*OPTIONAL*): The subset of environmental variables used to construct the distribution models for each species
        i. This allows for species-specific model generation
        ii. Environmental variables should be the exact same as the file name for that variable and should be separated by commas (e.g., *Bio1,Bio12*)
    d. **NOTE:** The titles of the columns do not matter, but the order of the columns does!

	A	B	C
1	Taxon	Scientific.Name	EnvVars (optional)
2	Mammalia	Alces alces	Bio1,Bio12,Bio15
3	Mammalia	Procyon lotor	Bio1,Bio12
4	Mammalia	Canis lupus	Bio1,Bio6

2. Environmental data in raster form (for example ***.bil***, ***.asc***, or ***.tif***) for all desired time periods and climate scenarios
    a. These data must be projected, but any projection is valid (see steps 4 and 5).
    b. All climate variables must be named using the exact same file name across time-periods and scenarios (folder names will differ)
    c. If categorical environmental variables are used (e.g., soil type, landforms), mark those rasters with a distinguishing prefix defined in Step 5 of the config file:
        i. e.g., *categ_landforms.bil* has the distinguishing prefix “categ”
3. RStudio (version 1.1 or higher)
4. If MaxEnt modelling is required, the executable Java script ***maxent.jar*** (which can be found on GitHub at https://github.com/mrmaxent/Maxent)
    a. Ensure that ***maxent.jar*** loads when opened (see picture):
    b. NOTE: If ***maxent.jar*** does not load when opened (this is common when using MacOS), the Java Runtime Environment may need to be downloaded as well (https://www.oracle.com/java/technologies/javase-jre8-downloads.html).

5. The configuration file accompanying this package (***config.txt***)

a. ***Config.txt*** can be downloaded using the getConfig() function or by downloading it from https://github.com/brshipley/megaSDM/blob/master/config.txt.

Optional Data:

If available, additional data may be provided by the user:

1. Occurrence points for each species in a ***.csv*** file with three columns:
    a. The name of the species repeated
    b. The longitude of each occurrence
    c. The latitude of each occurrence
    d. Each file should be named with the scientific name of the species
        i. e.g., *Cervus_elaphus.csv*
    e. **NOTE:** The exact titles of the columns do not matter, but they should be descriptive (e.g.: "long", "Longitude" or "x" for the longitude column; "lat", "Latitude", or "y" for the latitude column)

	A	B	C
1	Species	Longitude	Latitude
2	Cervus elaphus	-83.118639	35.648375
3	Cervus elaphus	-105.935073	32.880342
4	Cervus elaphus	-110.846022	43.500302

2. A buffer around the occurrence points (in ***.shp*** or raster form) to preferentially generate background points close to the occurrence points
    a. Named with the scientific name of the species
        i. e.g., *Cervus_elaphus.shp*
3. Background points (in the same format and coordinate system as the occurrence points)
    a. Named "species*_background_X.csv*", where X is 1 → number of replicates wanted
        i. e.g., *Cervus_elaphus_background_2.csv*
4. The dispersal rates for each species in a ***.csv*** file with two columns (in order):
    a. List of species
        i. **NOTE:** the names of the species should be the same as in the species list (Required Data: 1), with proper capitalization.
    b. Dispersal Rate (km/year)
5. Protected areas shapefile polygon(s) (***.shp***) detailing protected areas within the study region and (if available) other time periods
    a. All protected areas at each time period examined must be in a single shapefile
    b. If multiple protected area files (for different time periods) are used, name all protected area files with the time period the layer should be applied to
        i. E.g., *PA_2070.shp*
6. Binary urban data raster(s) for the study area and (if available) other time periods
    a. 0 = non-urbanized
    b. 1 = urbanized
    c. If future/projected urbanized files are used, name all urban files with the time period the layer should be applied to
        i. e.g., *urb_2070.bil*

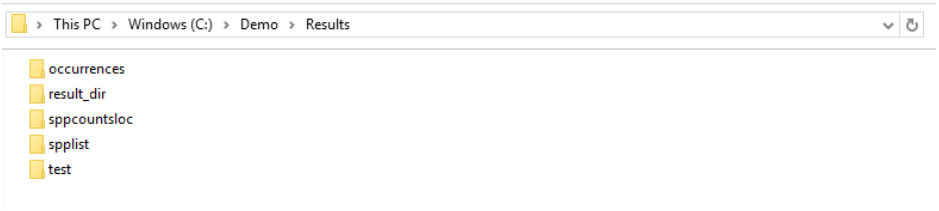
Setup, Configuration, and File Management:

To begin the setup process, download all megaSDM files. If you have previously run megaSDM, be sure that the "results" and "test" folders are empty. Fill out ***config.txt*** with the desired parameters:

NOTE: There should be no spaces in any folder names or directories

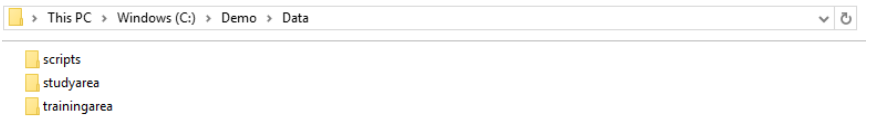
1. All occurrence files, background files, buffers, and results are included in ***TrialDirectory***. Fill out Step 1 of *config.txt* (Trial Directory) with directory paths and create each directory and subdirectory within ***TrialDirectory***
    a. e.g., TrialDirectory = "C:/Demo/Results"
        i. NOTE: for TrialDirectory and DataDirectory (Step 2), full path names are required!
    b. e.g., result_dir = "/result_dir"
    c. Move the list of species to be used in the analysis to the path given by ***spplist*.** See Required Data: 1 above for details about formatting.
    d. Copy ***maxent.jar*** into the ***occurrences*** subfolder. See Required Data: 4 for details on retrieving this file.
    e. If you are providing occurrence points for each species, place these in the "occurrences" folder. See Optional Data: 1 above for details about formatting.

    After Step 1, this is what ***TrialDirectory*** should look like:



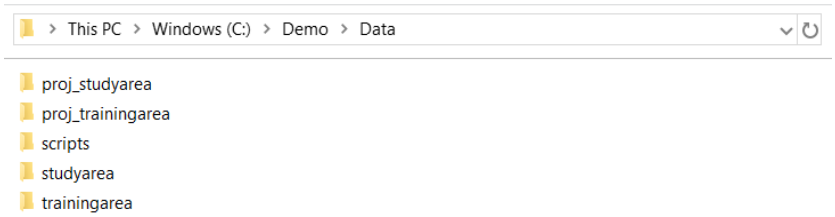
2. All environmental data and scripts required for running megaSDM should be included in ***DataDirectory***. Fill out Step 2 of *config.txt* (Input Data Directory) with directory paths and create each directory and subdirectory within ***DataDirectory***
    a. e.g., DataDirectory = "C:/Demo/Data"
    b. e.g., scripts = "/scripts"
    c. Place all megaSDM scripts plus the *config.txt* into the ***scripts*** subdirectory. These have been provided to you.
    d. Move all environmental data into the correct directories. See Required Data: 2 above for details about formatting.
        i. Training Area Data are environmental data for the region where the model coefficients will be generated (all occurrence and background points will be inside this region)
        ii. Study Area Data are environmental data for the region of interest (i.e. where the model will be applied to)

    After Step 2, this is what ***DataDirectory*** should look like:



3. Fill out Step 3 of *config.txt* (GIS Layer Projection) with the desired parameters**.**
    a. Set CoordinateProjectionStep = “N” if environmental rasters are already in the desired coordinate reference system and units.
    b. Create the required subdirectories within ***DataDirectory***
        i. e.g., proj_trainingarea = “/proj_trainingarea”
        ii. If CoordinateProjectionStep = “Y”, the rasters will be projected and copied into these subdirectories.
        iii. If CoordinateProjectionStep = “N”, the unprojected rasters will be copied into these subdirectories.

    After Step 3, this is what ***DataDirectory*** should look like:



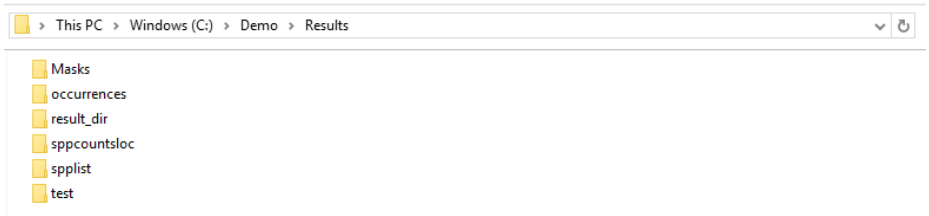
4. Fill out Step 4 of *config.txt* (Coordinate Systems) with the desired Coordinate Reference System (desiredCRS) and the Coordinate Reference System of the **occurrence data** (defaultCRS).
    a. CRS definitions should be in PROJ4 form; an overview of common coordinate systems may be found at https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/OverviewCoordinateReferenceSystems.pdf

5. Fill out Step 5 of *config.txt* (Raster Clipping and Resampling) with the desired parameters.
    a. Set ClipEnvDataStep = "N" if all rasters have the same resolution and both the training and the study rasters are the desired extents.
    b. Set Both ClipEnvDataStep and EnvTrainingAreaClip = “Y” if provided training rasters have a larger extent than the desired training area, and provide the desired extent:
        i. TrainClipLatitude = minlatitude, maxlatitude
        ii. TrainClipLongitude = minlongitude, maxlongitude
    c. Resolution should be provided in meters
    d. If you are using categorical environmental variables, set “Categorical” to whichever distinguishing prefix was used (see Required Data: 2 above).

6. Fill out Step 6 of *config.txt* (Downloading GBIF Occurrences) with the desired parameters
    a. Set gbifstep = “N” if occurrence files have already been generated for each species and provided in Step 1
    b. Geographic extent should be provided in min, max form
        i. e.g., decimalLatitude = 20,40
        ii. e.g., decimalLongitude = -57, -30

7. Fill out Step 7 of *config.txt* (Generating Background Points) with the desired parameters
    a. Set backgroundPointsStep = "N" if background points have already been generated for each species
        i. Create a directory within ***TrialDirectory/test*** called "backgrounds"
        ii. Place all background points within the "backgrounds" directory
        iii. See Optional Data: 3 above for formatting details.
    b. If backgroundPointsStep = "Y", background points will be generated
        i. MegaSDM can generate background points in 3 different ways:
            1. Sampled randomly throughout the training area ("random")
            2. Sampled from within a buffer around the occurrences ("constrained")
            3. A combination of the two ("combined")
        ii. For the "random" method, set speciesBufferStep = "N"
        iii. For the "constrained" method, set speciesBufferStep = "Y" and spatial_weights = 1
        iv. For the "combined" method:
            1. set speciesBufferStep = "Y" and spatial_weights < 1, > 0
            2. Spatial_weights describes the proportion of background points sampled from within the buffer.
        v. If speciesBufferStep = "Y", create ***buff_dir*** in ***TrialDirectory***
        vi. If you wish to provide buffers rather than creating them, place your buffer files in ***buff_dir***. See Optional Data: 2 above for formatting details

    After Step 7, this is what ***TrialDirectory*** should look like (buff_dir = "/Masks"):



8. Fill out Step 8 of *config.txt* (Environmental Subsampling) with the desired parameters. Recommended parameter values are provided in the configuration file

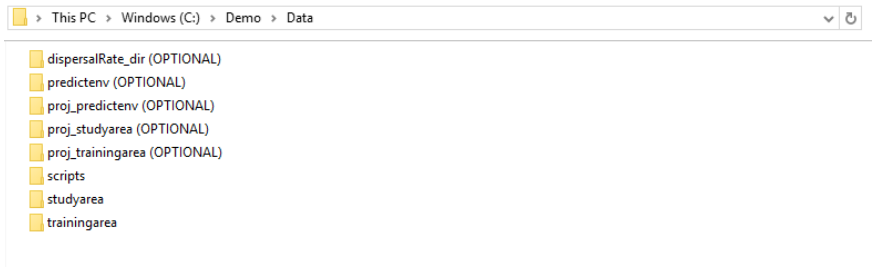
9. Fill out Step 9 of *config.txt* (MaxEnt) with the desired parameters
    a. If species distribution modelling has already been conducted (either from a previous run of megaSDM or from different modelling software) and habitat suitability maps have already been generated, set maxent Step to "N". Otherwise, set to "Y".
    b. If species-specific environmental variables are required, set variableEnvStep == "Y"
        i. This will use the optional 3rd column of the provided species list (***spplist***) to constrain the model to only the mentioned environmental variables. See Required Data: 1 for formatting details.
    c. Recommended parameter values area provided in the configuration file.

10. Fill out Step 10 of *config.txt* (Hindcasting/Forecasting) with the desired parameters
    a. If forecasting/hindcasting is desired, create subdirectories within ***DataDirectory.*** These directories must be formatted as:

i. ***DataDirectory/predictenv***/scenario folder/date folder/environmental files
ii. e.g., *C:/Demo/Data/predictenv/RCP8.5/2070/Bio1.bil*

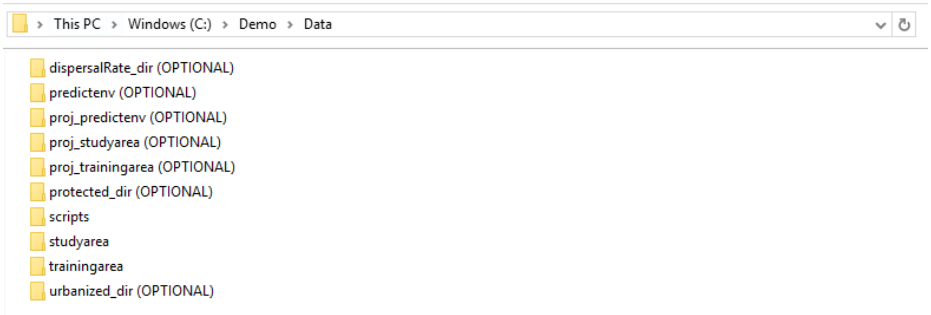
11. Fill out Step 11 of *config.txt* (Dispersal Rate) with the desired parameters
    a. If dispersal rate is desired (dispersalStep = "Y"), create subdirectory ***dispersalRate_dir*** within ***DataDirectory*** and copy the dispersal rate .csv file into ***dispersalRate_dir.*** See Optional Data: 4 above for formatting details.

    After Step 11, this is what ***DataDirectory*** should look like:



12. Fill out Step 12 of *config.txt* (Species Richness) with the desired parameters. Set RichnessStep = "Y" if richness maps for all modelled species are desired

13. Fill out Step 13 of *config.txt* (Urban Analysis & Protected Area Analysis)
    a. Create the subdirectories within ***DataDirectory***
    After Step 13, this is what ***DataDirectory*** should look like:



Running megaSDM:

1. Begin running ***megaSDM_run.R***

2. When prompted ((file.choose()) in ***megaSDM_run.R***), navigate to and select ***megaSDM_run.R***

3. Once the script has entirely run, delete or move all files located in the "results" and "test" folders if re-runs are necessary or desired.

4. The exact processing time varies widely, depending on factors such as the number of occurrence points, study region extent, and replicates run. However, the example run provided should take roughly 20-30 minutes to run to completion.