* 1. Model Prep: Generating Model Inputs

***\*\*Run for each species\*\****

These models screen and reformat data from the master data tables to produce tables that can be used in statistical models (Maxent, GLM, RandomForest). The results from these models include a preliminary table used to run the statistical models, but additional screening of response variables - screening that cannot be automated - is required before actual execution of the models.

* 1. 01 Create Model Input File
     1. Creates a subfolder named with the species name in the stats analysis root folder
     2. Join species presence records with response variables
     3. Identify HUC8s in which species occurs
     4. Remove all catchments not in those HUC8s
     5. Identify and remove response variables with null values
     6. *OUTPUT: CSV file listing presence/absence and valid response variables for all catchments in HUC8s in which species occurs.*
  2. 01 Calculate SH (Species Habitat) Correlations
     1. Calculate p-values of response variable correlations with presence/absence
     2. Generate list of response variables significantly correlated with p/a (p < 0.05)
     3. *OUTPUT: CSV file listing response variables significantly correlated with presence/absence along with the correlation coefficient and p-value.*
  3. 03 Calculate RV (Response Variable) Correlations
     1. Calculate correlations among response variable pairs.
     2. Identify redundant response variable pairs (coef. > 0.7)
     3. *OUTPUT: List of redundant response variable pairs along with their correlation coefficient.*
  4. 04 Visualize Correlations
     1. Generate an HTML page displaying response variables as nodes and correlations as edges.
        1. Nodes are colored by a ranking based on the variable's importance to management hooks; rankings are listed in the **Data/VariableRankings.xlsx** document.
           1. 0 = Management activities will have no impact on variable
           2. 1 = Secondary impact from any management activity
           3. 2 = Direct impact from one or more management activity, but impact cannot be quantified
           4. 3 = Direct impact from one or more management activity; impact can be quantified
        2. Node edges are colored red if variable reflects primary data (i.e. direct from source); nodes have no edges if variable reflects manipulated data.
        3. Node size reflects the variables coefficient with presence/absence. Hovering over the node reveals the coefficient value.
        4. Edges occur between correlated variable pairs. The thickness of the line corresponds to the magnitude of the correlation. Hovering over the edge reveals the correlation value.
     2. *OUTPUT: HTML document in species folder, which is displayed in a web browser.*
  5. Convert CSV to MaxentSWD
     1. Reads in records generated from the 01 Create Model Input File and the list of response variables found to be correlated with presence/absence for the given species.
     2. Recodes the presence/absence values to the species name and "background", respectively.
     3. Removes response variable columns if the variable was not found to be significantly correlated with presence/absence (i.e. included in the SH Correlation list.
     4. *OUTPUT: A Maxent Species With Data (SWD) format file used in Maxent habitat model as well as GLM and RF models. This file is screened for catchments (rows) falling outside of HUC8s in which the species was found and response variables (columns) either with null values or found not to correlate significantly with presence/absence. It still, however, includes redundant variables.*

* 1. Running Models: Maxent

The final screening of response variables - removing redundant variables - is done in these models and is the final step required before running Maxent habitat models for each species. These models facilitate that as well as migrate Maxent results into files that can be used in subsequent analysis and modeling. Outputs include a listing of variable importance to each species as well as a prediction of habitat likelihood based on the data used to produce the model.

* 1. Create Maxent batch file

*(run for each species)*

* 1. Open the HTML file that displays redundant response variables.
  2. Identify redundant variables to eliminate (based on rank, ecology, etc.)
  3. Select eliminated redundant variables in the Create Maxent batch file tool.
  4. Creates an "Output" folder within the species folder to hold Maxent results.
  5. *OUTPUT is a* ***RunMaxent.bat*** *file in the species folder which can be double-clicked to execute the Maxent analysis. This file also includes a record of response variables tagged as redundant.*
  6. *Run Maxent ..*
     1. *For each species* double click the RunMaxent.bat file created above.
     2. *OUTPUTS are stored in the "Output" subfolder created when the RunMaxent.bat file was created.*
  7. Extract Variable Importance from Maxent results

*(iterate across species names)*

* 1. Locates the Maxent results file (maxentResults.csv) in the Maxent output folder. This file contains information on the model run, including variable importance and jackknifing results.
  2. Reads in the values from the above file and creates dictionaries of each variables contribution to gain, permutation importance, gain with only the variable, and gain without the variable.
  3. Writes these values to a CSV table for the given species.
  4. *OUTPUT is a CSV format table listing each modeled response variable and indicators of variable importance. These should be named "<species>\_MEVars.csv" stored in the habitat modeling root folder.*
  5. Display Maxent results (optional)

*(iterate across species names)*

* 1. Reads in the logistic probabilities from the "\_samplePredictions.csv" file in the Maxent output folder.
  2. Reads in the logistic threshold to use from the "maxentResults.csv" file in the Maxent output folder.
  3. Copies the NHD catchment feature class to a new output feature class.
  4. Joins the table of logistic probabilities to the feature class.
  5. Adds a field for binary habitat and calculates values > the threshold as 1, otherwise 0.
  6. *OUTPUT is a feature class of catchments labeled with Maxent modeled habitat likelihood and thresholded habitat classification. Only HUC8s modeled will have data.*

* 1. Running Models: GLM and Random Forest (RF)

With the final screening step completed in generating the Maxent model runs, running GLM and RF models is a simple process of running these models which call R statistical scripts to produce outputs similar to those of the Maxent model runs for each species, namely a listing of variable importance to each species as well as a prediction of habitat likelihood based on the data used to produce the model.

* 1. Run GLM on Species

*(iterate through each species)*

* 1. Locates the species subfolder within the stats root folder using the species' name.
  2. Locates the SWD file for the species within this folder; this file contains all the necessary input data for the GLM and RF models.
  3. Locates the RunMaxent.bat file within this folder; this file contains a listing of the redundant response variables that should be omitted from the GLM and RF models.
  4. Locates the SH\_Correlations.csv file within this folder; this file contains the list of response variables correlated with presence/absence and correlation values (used for sorting variables).
  5. Loads the PypeR module, used to execute R commands from Python. *(This must be installed on the machine prior to use.)*
  6. In R, reads all data in the SWD file into the sppAll data frame
  7. In R, reads the first column of the sppAll into the spp variable; this a list of presence absences
  8. In R, converts the values in the spp variable to a binary (1's and 0's) list named sppBin.
  9. Creates a dictionary of response variables correlated with presence/absence for the species, provided in the SH\_Correlations.csv file (key=variable name; value=correlation coefficient).
  10. Creates a list of response variables sorted on its correlation with presence/absence.
  11. Removes variables in this list that are identified as redundant in the RunMaxent.bat file.
  12. In R, creates a data frame, called habData, of response variables from the list created, sorted, and screened above.
  13. In R, creates a GLM model, named sppGLM, from the spp and habData objects.
  14. In R, creates a jackknifed table of variable importances using the jackGLM.R script
  15. In R, writes the jackknifed variable importance table to an output CSV file.
  16. In R, creates a table of predictions from the GLM model and writes this table to an output CSV file.
  17. In R, loads the RandomForest library.
  18. In R, creates a RF model, named sppForest, from the spp and habData objects.
  19. In R, calculates probabilities (rfProbs) and binary habitat (rfPedictions) from the RF model
  20. In R, writes a table of RF predictions and probabilities to an output CSV file.
  21. In R, saves the R session image to a file, .RData, in the species folder.
  22. Also saves a log of all R commands in a file called <species>.R, in the species folder. This file can be opened in R to repeat all R steps above.
  23. *OUTPUTS (stored in the habitat modeling folder) include CSV format tables for:*
      1. *GLM model variable importances (\*GLMVars.csv)*
      2. *GLM model predictions on the input response variables (\*GLMPred.csv)*
      3. *RF model variable importances (\*RFVars.csv)*
      4. *RF model predictions on the input response variables (\*RFPreds.csv)*

*Also (strore in the species subfolders):*

* 1. *Log of R commands*
  2. *R workspace file containing GLM and RF model that can be used later in modeling habitat likelihood under alternate scenarios.*

* 1. Compiling tables on variable importance

*Variable importance tables are used to indicate the habitat sensitivity to a given response variable. A high value for a given response variable indicates that modifying it will likely have a larger impact on the species habitat. Tables are compute for each species and then averaged across species and are generated for each modeling approach (Maxent, GLM, RF).*

* 1. Merge Maxent Variable Importance Tables
     1. Creates a copy a table listing all response variables from the VariableRankings worksheet in the ResponseVaraiables.xls file in the data folder.
     2. Extracts each species Maxent Variable Importance files (*\*MEvars.csv*)
     3. Joins the variable importance columns to the response variable table for the given species.
     4. After all species tables are joined, mean values are computed and added as additional columns to the table.
     5. OUTPUT is a CSV format table listing the mean variable importances determined by MaxEnt as well as importances calculated for each individual species.

* 1. Merge GLM Jackknife Tables
     1. Creates a copy a table listing all response variables from the VariableRankings worksheet in the ResponseVaraiables.xls file in the data folder.
     2. Extracts each species GLM Variable Importance files (*\*GLMvars.csv*)
     3. Joins the variable importance columns to the response variable table for the given species.
     4. After all species tables are joined, mean values are computed and added as additional columns to the table.
     5. OUTPUT is a CSV format table listing the mean variable importances determined by MaxEnt as well as importances calculated for each individual species.
  2. Merge RF Variable Importance Tables
     1. Creates a copy a table listing all response variables from the VariableRankings worksheet in the ResponseVaraiables.xls file in the data folder.
     2. Extracts each species RF Variable Importance files (*\*GLMvars.csv*)
     3. Joins the variable importance columns to the response variable table for the given species.
     4. After all species tables are joined, mean values are computed and added as additional columns to the table.
     5. OUTPUT is a CSV format table listing the mean variable importances determined by MaxEnt as well as importances calculated for each individual species.

* 1. Generating inputs for alternate scenarios

*The goal of each uplift model is to produce an updated response variable table with values reflecting the changes in each affected variable. These response variable tables are then used in subsequent models to model the change in habitat likelihood -- or uplift - associated with implementing them.*

* 1. Compute Forest Buffer Response Variables
     1. Copies the master response variable table
     2. Updates FLNLCD4 to be equal to sum of FLNCD4 + 5 + 7 + 8 + 9
     3. Zeros out values in FLNCD4 + 5 + 7 + 8 + 9

* 1. Computing uplift from alternate scenarios