MMX TOOLKIT V0.1.0 USERS GUIDE

John L. Schnase
Code 606.3 - Data Science Group
NASA Goddard Space Flight Center
john.l.schnase@nasa.gov

SYSTEM CONFIGURATION

MacBook Pro (Mac14,7) Apple M2
maxOS Monterey Version 12.6.3
Cores 8 (4 performance and 4 efficiency)
Memory 24 GB

R version 4.1.3 (2022-03-10) -- "One Push-Up"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)
Libraries: sys, rJava, ENMeval, raster, MASS, dismo, Hmisc, paletteer, colorRamps, usdm, ncdf4, rgdal, gganimate, gifski, png, animation, tidyverse, rtsVis, spatialEco, terra, pals, plyr, scales, rgbif, stringi

RStudio 2022.07.1+554 "Spotted Wakerobin" Release
(7872775ebddc40635780ca1ed238934c3345c5de, 2022-07-22) for macOS
Mozilla/5.0 (Macintosh; Intel Mac OS X 12_6_3) AppleWebKit/537.36 (KHTML, like Gecko)
QtWebEngine/5.12.10 Chrome/69.0.3497.128 Safari/537.36

Java OpenJDK Version "20.0.2" 2023-07-18
OpenJDK Runtime Environment (build 20.0.2+9-78)
OpenJDK 64-Bit Server VM (build 20.0.2+9-78, mixed mode, sharing)

Python 3.9.12 (main, Apr 5 2022, 01:53:17)
[Clang 12.0.0] :: Anaconda, Inc. on darwin
Libraries: os, sys, subprocess, time, shutil, random, csv, datetime, time, numpy

MaxEnt Maximum Entropy Species Distribution Modeling
Version 3.4.1

MMX TOOLKIT CONTENTS / DIRECTORY STRUCTURE

bash	Bash shell scripts, numbered for cross-reference with this guide.
data	Occurrence and predictor data sets, for use as examples and testing.
Guide.pdf	MMX Users Guide
Guide.R	MMX Users Guide R script
MMX	MERRA/Max variable selection programs and directories.
R	R scripts, numbered for cross-reference with this guide.
resources	Required program resources and other helpful background documents.

TIME SERIES PROCESSING

1) **SET UP SYSTEM DIRECTORIES (Manual step)**
   a) <u>Symbolic links</u> -- This is a manual step. Install the MMX_Toolkit and create an MMX_Experiment directory in locations of your choosing, then create symbolic links from your home directory to these two locations. For example, from my home directory ~/ I see:
   ~/MMX_Experiment -> /Users/myname/MMX-Project/MMX-Experiments/Exp000
   ~/MMX_Toolkit -> /Users/myname/MMX-Project/MMX-Development/MMX_Toolkit_V0.1.0
   b) <u>Configuration file</u> -- Copy the _mmx_config file from the ~/MMX_Toolkit/resources directory to the new ~/MMX_Experiment directory and edit the file as needed to configure for the current run.

2) **PREPARE DATA**
   (This is a semi-automatic step. It is works best to do data prep as a separate, preliminary step.)
   a) <u>Occurrence files</u> – Run ***~/MMX_Toolkit/R/02a_get_occurrences.R*** to assemble a collection of GBIF observations for each five-year interval of the time series. The script makes a connection to the GBIF server and can be flaky at time but keep trying. Fill in the _mmx_config configuration file's GBIF_TAXON_KEY and SPECIES_NAME fields to specify the data to be gathered. A list of potential species of interest is provided in the ***~/MMXToolkit/resources/ species_list.txt*** file. Results are delivered to the directory specified in _mmx_config's OF_DST_DIRECTORY field. The default name of the output directory is the GBIF species taxon key + species name. If years in the specified range lack data, the script crashes; try again with a narrowed range of TEMPORAL_EXTENT_START_YR and TEMPORAL_EXTENT_STOP_YR years. After download, you may want to thin the observations or apply other treatments before proceeding. The download creates a full collection of yearly files; the files that you want to use for further processing at the five-year aggregates with the name "OF-year.csv" ...
   b) <u>MERRA-2 Predictors</u> – Run ***~/MMX_Toolkit/R/02b_m2_ws_builder.R*** to build a working_set collection of MERRA-2 predictors for each five-year interval of the time series. The script draws from a base_collection of .nc files as specified in the _mmx_config file's BASE_COLLECTION field. Results are delivered to the directory specified in _mmx_config's WS_DST_DIRECTORY field. The default name of the output directory is the GBIF species taxon key + species name.
   c) <u>MERRAclim-2 Predictors</u> – Run ***~/MMX_Toolkit/R/02c_mc_ws_builder.R*** to build a working_set collection of MERRAclim-2 predictors for each five-year interval of the time series. The script draws from a base_collection of .tif files as specified in the _mmx_config file's BASE_COLLECTION field. Results are delivered to the directory specified in _mmx_config's WS_DST_DIRECTORY field. The default name of the output directory is the GBIF species taxon key + species name. (This version of MERRAclim-2 uses MERRA-2's t2m max and min and the prectotcorr observation-corrected precipitation estimate.)

3) **SET UP ENM TIME SERIES RUN**
   a) <u>Experiment directories</u> -- Run ***~/MMX_Toolkit/bash/03a_create_exp_directories.sh***. This will create the experiment's top-level directories.
   b) <u>Populate time series directories</u> -- Run **~/MMX_Toolkit/bash/03b_populate_ts_directories.sh**. This will create the time series second-level directories and populate the the occurrence_file and working-set directories prepared in Step 2.

4) **BUILD ENM TIME SERIES**
   a) <u>Variable selection</u> – This step uses MMX (MERRA/Max) to screen for the top ten most contributory variables in each of the time series five-year interval's working_set of predictors. It can take a lot of time depending on the number of available cores and the size of the working sets. Run ***~/MMX_Toolkit/bash/04a_mmx_select_vars_local.sh*** to process the entire time series on the local host; run with a year parameter from the command line to process a single five-year interval. Results are placed in each interval's selection_set directory. (The current version of the MMX Toolkit performs selection on the local host only.)

b) <u>Variable linting</u> – This step removes NA-rich variables from the selection_set. The step is optional and not generally used. If used, run **~/MMX_Toolkit/R/04b_linter.R** and repeat as needed until the collection is "clean." Results reside in the selection-set folder with deleted variables indicated by a .xxx file name extension.

c) <u>Covariance reduction</u> -- Run **~/MMX_Toolkit/R/04c_reducer.R** to remove colinear variables in the selection_set collection. Results are deposited in the model_set directory.

d) <u>Time series build</u> -- Run **~/MMX_Toolkit/R/04d_timeseries_builder.R** to find optimal MaxEnt tuning parameters and generate a final model for each time interval. Results are deposited in the model directory.

5) **SET UP TREND ANALYSIS RUN**

a) <u>Trend analysis directories</u> -- Run **~/MMX_Toolkit/bash/05a_setup_ta_directories.sh**. This will create ENM and VAR trend analysis directories and populate the ENM directory with predictors. (This version of the MMX Toolkit performs trend analysis on the suitability time series only.)

6) **RUN ENM TREND ANALYSIS**

a) <u>Theil-Sen trend</u> -- Run **~/MMX_Toolkit/R/06a_enm_theilsen_trend.R** to find the Theil-Sen trend in predicted environmental suitability over the ENM time series.

7) **SET UP VELOCITY ANALYSIS RUN**

a) <u>ENM velocity analysis directories</u> -- Run **~/MMX_Toolkit/bash/07a_setup_va_directories.sh** to create the ENM velocity analysis directory. (This version of the MMX Toolkit performs velocity analysis on the suitability time series only.)

8) **RUN ENM VELOCITY ANALYSIS**

a) <u>ENM (biotic) velocity</u> -- Run **~/MMX_Toolkit/R/08a_enm_velocity.R** to find the bioclimatic velocity over the ENM time series.

9) **COLLECT RESULTS**

a) <u>Gather run results</u> -- Run **~/MMX_Toolkit/bash/09a_collect_results.sh** to collect run results and place them in the Summaries directory.