Schnase JL, Carroll ML. The MMX Toolkit: High performance, reanalysis-based climatic suitability modeling to advance avian conservation. In: Albani S, Lumnitz S, Soille P, editors. Proceedings of the 2023 conference on Big Data from Space – 6-9 November 2023. Vienna: European Commission, Joint Research Centre, Publications Office; 2023. (to appear)

THE MMX TOOLKIT: HIGH PERFORMANCE, REANALYSIS-BASED CLIMATIC SUITABILITY MODELING TO ADVANCE AVIAN CONSERVATION

John L. Schnase and Mark L. Carroll

Office of Computational and Information Sciences and Technology
NASA Goddard Space Flight Center, Greenbelt, Maryland USA

ABSTRACT

Ecological niche modeling (ENM) is an established approach to studying an organism's response to climate change. Earth observations and climate model outputs are among the most important environmental predictors employed in such studies. Unfortunately, significant big data challenges often limit the use of these data sets in ENM. The MMX Toolkit provides resources that can automatically combine data from NASA's Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) with field observations spanning the past 40 years to perform scalable, high-performance retrospective ENM. This innovative new capability enables historical time series studies of the changing climatic suitability of biological species, which can greatly improve the practical use of big data from space in a wide range of global species conservation efforts.

Index Terms — MERRA, MaxEnt, ecological niche modeling, Bendire's Thrasher

1. INTRODUCTION

The eminent conservation biologist Thomas Lovejoy often said, “If you take care of the birds, you'll take care of most of the world’s environmental problems” [1]. That is because birds are the only group of terrestrial vertebrates that covers the entire planet, essentially occupying all of the Earth's diverse ecoregions. In addition, birds are keen harbingers of environmental change, making them a particularly important class of indicator organism in conservation biology. Ecological niche modeling (ENM) is one of the primary ways of studying an organism's relationship to the environment. ENMs statistically correlate a species' occurrence records with the environmental conditions associated with those occurrences to predict the relative viability of local conditions for the species. ENM has particular value in conservation science, because it allows field biologists and policy makers insight into the current and future range, population health, and overall ecology of a species, information that is critical to defining our societal response to climate change.

Global climate model (GCM) outputs are gaining increased importance in these efforts. GCMs combine Earth observations from a wide array of sources to create global representations of the climate system, including historical simulations and future projections for hundreds of climate variables [2]. Despite their relevance, the use of GCM outputs in ENM-based conservation planning has been limited by a suite of big data challenges. GCMs produce complex, petabyte-scale data sets comprising hundreds of attributes, which complicates variable selection and limits their use in ENM. Further, it has been shown that climatic suitability studies can benefit from historical analyses, the use of ensemble models, and the incorporation of time specificity into traditional modeling frameworks, all of which add to the scale and complexity of modeling's computational and technological challenges. Coupled with the urgency imposed on conservation work by a rapidly changing climate, these challenges have raised calls for innovations in process automation that would enable quick and meaningful analyses.

NASA has engaged this challenge by addressing a need in conservation biology for better information about how life on this planet has responded in the past to environmental change. Conservation policy decisions are often based on what is known about a species' current circumstance and how it might be altered by projected climate change. However, there is much to be gained in formulating these projections when past responses are known. Longitudinal, retrospective ENM in support of climate change research, thus, became the science challenge that inspired the technology innovations described here.

2. THE MMX TOOLKIT

The MMX Toolkit bundles the capabilities of NASA's MERRA/Max system with utilities, data, documentation, and other resources to enable historical, ENM-based analysis of the changing climatic suitability of bird species. MERRA/Max enables the use of GCM outputs in ENM through a MaxEnt-enabled, Monte Carlo optimization that screens a large collection of variables for potential predictors [3]. Based on a machine learning approach to maximum entropy modeling, MaxEnt is one of the most popular

software packages in use today by the modeling community [4]. Among its many features, MaxEnt ranks the contribution of predictor variables in the formation of its models. MERRA/Max's Monte Carlo method exploits this capability in an ensemble strategy whereby many independent MaxEnt runs, each drawing on a random pair of variables in a large collection of variables stored in the underlying filesystem, converge on a global estimate of the top contributing variables in the collection being screened.

MERRA/Max currently operates on a dedicated set of ten 10-core Debian Linux 9 Stretch virtual machines (VMs) in the NASA Center for Climate Simulation's (NCCS's) Explore cluster computing environment (https://www.nccs.nasa.gov). It is important to note that MERRA/Max's independent MaxEnt runs are entirely parallelizable, and, because they operate on external files, the approach is infinitely scalable. MERRA/Max, the core enabling functionality of the MMX Toolkit, thereby directly addresses a key big data challenge that has limited use of GCM outputs in ENM: it automates variable selection when manual selection is impractical.

In developing the MMX Toolkit, we adopted birds as our first target group of organisms for the reasons described above. In addition, the great public and scientific interest in birds has resulted in a number of large, citizen-scientist observational data sets and long-standing surveys that make birds particularly amenable to historical studies.

For environmental variables, we turned to MERRA-2. NASA's Modern-Era Retrospective Analysis for Research and Applications, Version 2 is a product of the Goddard Earth Observing System, Version 5 (GEOS-5) modeling system [5]. With reanalysis, a data assimilation system is used to reprocess historical, meteorological observations from satellite, airborne, and in-situ sensors. The system typically employs an iterative, time-stepping procedure in which an existing modeled state of the atmosphere is compared with new observations, then updated to reflect those observations in a new modeled state of the atmosphere. The process relies on an underlying climate model to combine disparate observations in a consistent manner, enabling production of gridded data sets for a broad range of variables, including ones that are sparsely measured or not directly observed. MERRA-2 comprises a global, temporally and spatially uniform synthesis of over 600 climate-related variables. Its native spatial resolution is 1/2° latitude × 5/8° longitude (i.e., 55.5 × 69.4 km at the equator) × 72 vertical levels extending through the stratosphere. Its temporal resolution is hourly and extends from 1979 to the present. The MERRA-2 collection is about one petabyte in size and growing.

2.1. Toolkit organization and system requirements

The current version of the MMX Toolkit is designed to run on a MacBook Pro, which functions as a local host with connectivity to remote services as described below. Required software includes the most recent versions of R, RStudio, Java, Python, and MaxEnt. The compressed, downloadable MMX Toolkit directory has a self-documenting structure and provides a user's guide with configuration details. An interactive R version of the user's guide functions as an orchestrator where each step of the workflow can be invoked manually or as a batch to perform a fully automated run. Settings for a particular analysis run are conveyed to the system through a user-specified configuration file. As illustrated in in Fig. 1, the MMX Toolkit's core capabilities are grouped into three major workflow components.

2.2. Component architecture

2.2.1. Data assembly

The MMX Toolkit's *get_occurrences.R* and *ws_builder.R* scripts assemble the required input data for an analysis run (Fig. 1A). To create a set of longitudinal occurrence records, the MMX Toolkit uses R's rgbif library to obtain species observations from the Global Biodiversity Information Facility (GBIF) (https://www.gbif.org). Georeferenced bird records for the years 1980 through 2019 are downloaded based on GBIF's species identifiers. The *get_occurrences.R* script merges the observations into a time-series dataset comprising eight, five-year aggregated collections: 1980–84, 1985–89, 1990–94, 1995–99, 2000–2004, 2005–09, 2010–2014, and 2015-19. Optional filtering steps can be applied at this point to reduce sampling biases and record density.

The *ws_builder.R* script assembles a *base collection* of 30 MERRA-2 variables known to be potential environmental determinants of climate suitability. These are drawn from four, hourly, time-averaged, two-dimensional collections in which each variable represents one surface-level spatial grid across the landscape: (1) M2T1NXSLV, consisting of air temperatures, wind components, total precipitable water vapor, etc., (2) M2T1NXFLX, consisting of surface fluxes, such as observation-corrected total precipitation, surface air temperature, specific humidity, wind speed, and re-evaporation, (3) M2T1XRAD, consisting of radiation estimates, such as surface albedo, cloud area fraction, cloud optical thickness, solar radiation, and (4) M2T1NXLND, which is made up of an assortment of variables of interest to ENM applications, such as surface soil wetness, root zone soil wetness, soil temperatures at various layers, and important elements of the land energy and water balance equations.

The script obtains MERRA-2 data in its native, Network Common Data Form 4 (NetCDF4) format from research collections housed in the NCCS; however, the data are also available to the general public through the Goddard Earth Sciences Data and Information Services Center (GES DISC). The script draws on the subsetting capabilities of Python's xarray library to assemble global, five-year aggregate collections from the original downloads. These groupings corresponded to the eight intervals of the occurrence data time series. Each five-year collection contains maximum, minimum, and average values for the 30 selected variables.

From this base collection, the script uses R's rgdal library to extract eight collections of average values for each variable

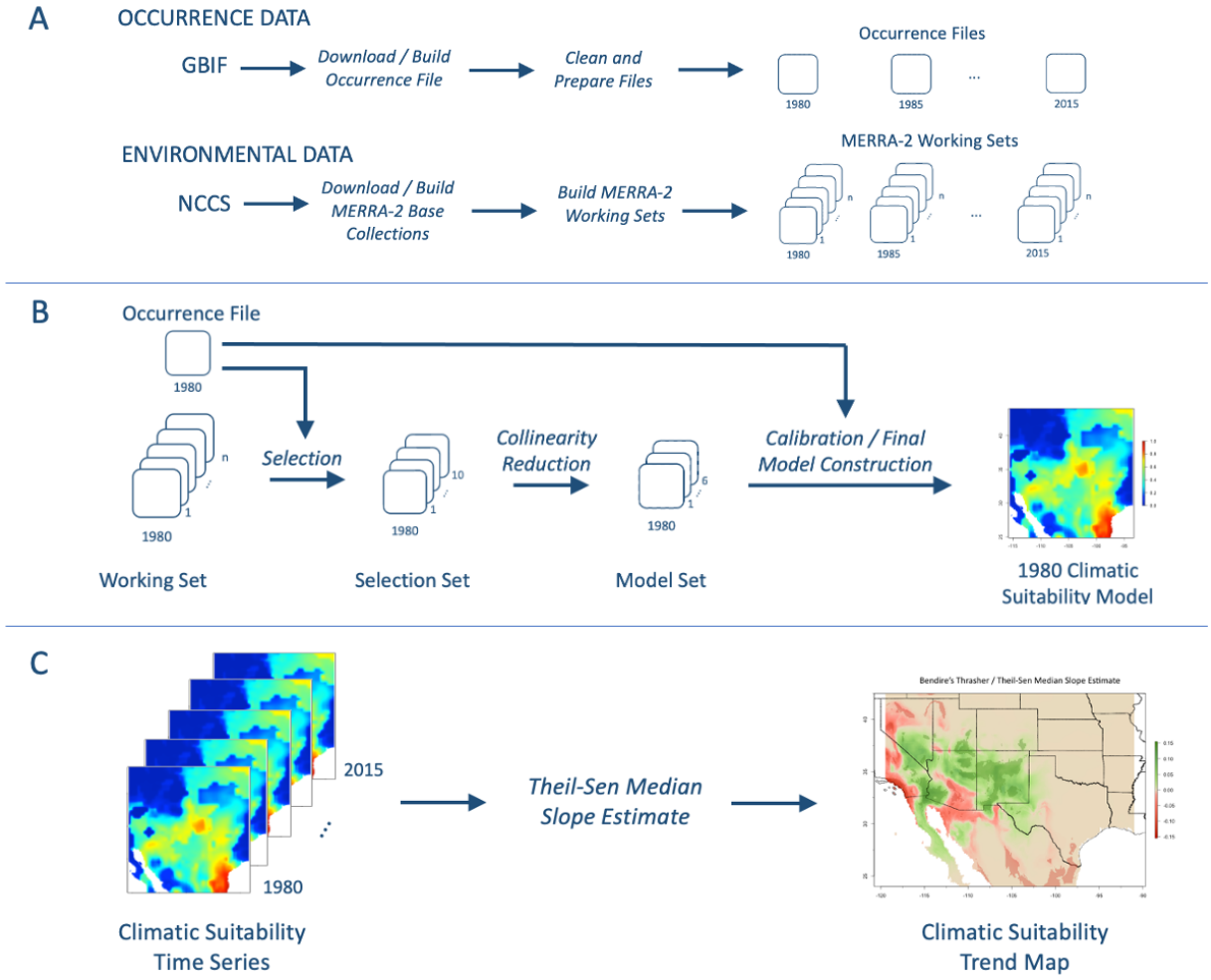

Fig. 1. Major MMX Toolkit workflow steps: (A) data assembly, (B) time series construction, and (C) time series analysis.

that are tailored to the specific requirements of the study. We refer to these study- and site-specific predictor datasets as *working set* collections. Each environmental layer is clipped to the coverage extent of the study area, re-projected, and formatted for further use by the system. The resulting working set collections correspond to the eight intervals of the occurrence data time series. To smooth the representation of local environmental conditions, the MERRA-2 layers are resampled from their native spatial resolution of 1/2° latitude × 5/8° longitude to 5.0 arc-minutes (1/12°) resolution using the R raster library's bilinear interpolation routine.

2.2.2. Time series construction

These data assembly products are then used by the MMX Toolkit's scripts to construct a time series of climatic suitability models spanning the years 1980 to 2019 in eight five-year intervals (Fig. 1B). The first step of this process calls on MERRA/Max to screen the working set variables in each of the eight five-year spans for the most contributory variables. The Toolkit's *select_vars_remote.sh* bash script coordinates interactions between local and remote hosts to do the selection; the *select_vars_local.sh* script can perform variable selection on the local host, but takes much longer. MERRA/Max returns the top ten most contributory variables for each time interval in a *selection set*, which can be further refined by the use of various utilities. For example, the Toolkit's *reducer.R* utility uses variance inflation factor (VIF) analysis to reduce collinearities in the selected predictors. VIF shows the degree to which standard errors are inflated due to the levels of multicollinearities. The script calculates the VIF for the selected variables and removes highly correlated variables using R's usdm library to create a *model set* of MERRA-2 predictors for each five-year interval.

At this point, the paired model set predictors and observations for each of the time series' eight five-year intervals are brought together in a model calibration and final model construction step performed by the MMX Toolkit's *timeseries_builder.R* script. The script uses R's ENMeval

package with default settings to scan for optimal MaxEnt feature class and regularization multiplier settings. The combination of settings resulting in the lowest value for Akaike's information criterion corrected for small sample size (AICc) is taken to be an optimal tuning configuration, which is then used to construct a final MaxEnt climatic suitability model for each five-year interval in the time series.

With a working set of 30 variables, at 100 random samples each, MERRA/Max selection for each of the eight five-year intervals in the time series would require 1,500 independents bivariate MaxEnt sampling runs and a total compute time of about one day on a single processor. But it takes only 15 minutes on our 100-core development cluster and would require *less than one minute* in a fully provisioned, 1,500-core environment. In theory, the MMX Toolkit is capable of reducing an intractable manual workflow and eight days of computing into an automated task of a few minutes.

2.2.3. Time series analysis

The resulting collection of time series models provides a platform for a wide range of analysis treatments. The current version of the MMX Toolkit includes a *theilsen_trend.R* script that can be used to assess trends in climatic suitability for the target species over the past 40 years (Fig. 1C). The script applies the Theil-Sen method, which is a non-parametric means of robustly fitting a line to points in a collection of raster planes by calculating the median of slopes through all the points in the planes. The script uses R's spatialEco library for trend analysis and applies a Mann-Kendall test to determine statistical significance of the trends.

3. BENDIRE'S THRASHER USE CASE

Bendire's Thrasher (*Toxostoma bendirei*) is a medium-sized species of thrasher native to the southwestern United States and northwestern Mexico. It is one of many bird species threatened by climate change and identified as a species of greatest conservation need. We used the MMX Toolkit to model changes in climatic suitability for Bendire's Thrasher from 1980 to 2019 (Fig. 2). The probabilities shown here range from 0.0 to 1.0, with warmer colors indicating more favorable conditions, which have varied over the years.

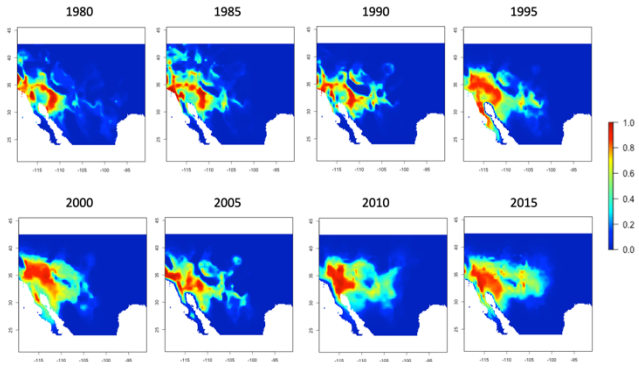

Fig. 2. Climatic suitability time series for Bendire's Thrasher.

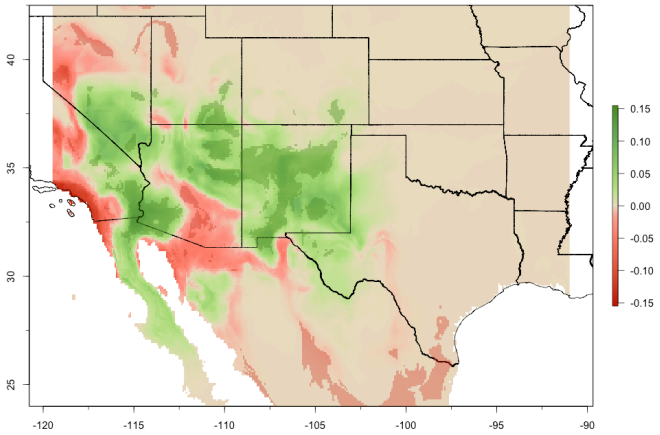

Fig. 3. 40-year climatic suitability trend for Bendire's Thrasher.

The resulting Theil-Sen trend analysis provides a dramatic example of how climate suitability for Bendire's Thrasher has been changing over the past 40 years (Fig. 3). Here we see positive trends indicated in green, negative trends in red. Units represent the change in probabilities per five-year interval. It is a unique snapshot, the story of one species' relationship to a changing climate throughout the satellite era, a view that before now would be nearly impossible to create.

4. CONCLUSIONS

The MMX Toolkit is in active development. Our long-term goal is to generalize the software to other species and make these capabilities freely available to the public. Together, with our colleagues in the conservation science community, we are identifying extensions, improvements, and a suite of science questions we hope to address in the coming months using these new capabilities. We believe this project provides an excellent example of how insights driven by big data from space can improve decisions addressing one of society's most important challenges. For a view into the ongoing project and access to the Toolkit, users guide, example data sets, and additional technical detail on the system, please go to: https://github.com/jschnase/MMX_Toolkit.

REFERENCES

[1] Gray E. The Birdsong Project: What the Birds Tell Us. 2022. https://www.audubon.org/news/introducing-birdsong-project-what-birds-tell-us. Audubon. 29 Jul 2022 [cited 9 Jun 2023].

[2] Edwards PN. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge: MIT Press; 2010.

[3] Schnase JL, Carroll ML. Automatic variable selection in ecological niche modeling. PLOS ONE. 2022;17: e0257502. doi:10.1371/journal.pone.0257502.

[4] Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME. Opening the black box: An open-source release of Maxent. Ecography. 2017;40: 887–893.

[5] Gelaro R, Mccarty W, Su MJ, Todling R, Molod A, et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). JOURNAL OF CLIMATE. 2017;30: 36.