Visual Analytics of the Impacts of Climate Change on Migratory Bird Habitats
Technical Document

Jacob Vogt, Mihika Krishna, Catherine Kang, Hangyul Yun

Professor Dongyu Liu

University of California, Davis

June 2024

Contents

Introduction

Our Senior Capstone project is comprised of two components:

1. A species distribution model (SDM) capable of predicting how climate change will affect future bird habitats up to year 2100
2. A web-app that visualizes SDM output and displays other relevant information such as bird migration patterns and climate trends.

This technical document will overview how each of these components work, as well as the required data and file structure for them to operate correctly.

1 Species Distribution Model

1.1 Datasets

The species distribution model (SDM) uses two datasets, one for climate data and one for species distribution data.

Climate Data

NEX-GDDP-CMIP6 datasets are granular by year and offered for four (future) Shared Socioeconomic Pathways (SSP). These dastasets are offered in NETCDF (.nc) format. Descriptions of each SSP can be found at this link.

Climate scenarios used were from the NEX-GDDP-CMIP6 dataset, prepared by the Climate Analytics Group and NASA Ames Research Center using the NASA Earth Exchange and distributed by the NASA Center for Climate Simulation (NCCS).

Thrasher, B., Wang, W., Michaelis, A. et al. NASA Global Daily Downscaled Projections, CMIP6. Sci Data 9, 262 (2022). https://doi.org/10.1038/s41597-022-01393-4

Thrasher, B., Wang, W., Michaelis, A. Nemani, R. (2021). NEX-GDDP-CMIP6. NASA Center for Climate Simulation. https://doi.org/10.7917/OFSG3345

Species Distribution Data

The occurrence datasets for our bird species come from the Global Biodiversity Information Facility (GBIF). The citations for each dataset is below.

- *Numenius americanus Bechstein, 1812 in GBIF Secretariat (2023). GBIF Backbone Taxonomy. Checklist dataset https://doi.org/10.15468/39omei accessed via GBIF.org on 2024-06-09.*

- *Setophaga striata (J.R. Forster, 1772) in GBIF Secretariat (2023). GBIF Backbone Taxonomy. Checklist dataset* `https://doi.org/10.15468/39omei` *accessed via GBIF.org on 2024-06-09.*

- *Anser albifrons (Scopoli, 1769) in GBIF Secretariat (2023). GBIF Backbone Taxonomy. Checklist dataset* `https://doi.org/10.15468/39omei` *accessed via GBIF.org on 2024-06-09.*

- *Haliaeetus leucocephalus (Linnaeus, 1766) in GBIF Secretariat (2023). GBIF Backbone Taxonomy. Checklist dataset* `https://doi.org/10.15468/39omei` *accessed via GBIF.org on 2024-06-09.*

- *Numenius phaeopus subsp. phaeopus in GBIF Secretariat (2023). GBIF Backbone Taxonomy. Checklist dataset* `https://doi.org/10.15468/39omei` *accessed via GBIF.org on 2024-06-09.*

GBIF datasets are downloaded in .csv format.

1.2 R Preprocessing

Pre-processing is necessary to transform raw datasets into a form factor that is usable by our SDM. This pre-processing step uses R due to its specialty in working with biological and spatial data.

There are two different R scripts, one for preprocessing training data and another for preprocessing prediction data. These scripts are named `preprocess_training.R` and `preprocess_prediction.R` respectively.

Packages

- *dismo*: Methods for species distribution modeling, that is, predicting the environmental similarity of any site to that of the locations of known occurrences of a species.

- *sf*: Support for simple features, a standardized way to encode spatial vector data. Binds to 'GDAL' for reading and writing data, to 'GEOS' for geometrical operations, and to 'PROJ' for projection conversions and datum transformations. Uses by default the 's2' package for spherical geometry operations on ellipsoidal (long/lat) coordinates.

- *raster*: Reading, writing, manipulating, analyzing and modeling of spatial data.

- *ncdf4*: Provides a high-level R interface to data files written using Unidata's netCDF library (version 4 or earlier), which are binary data files that are portable across platforms and include metadata information in addition to the data sets.

Prerequisites

The R scripts were developed using the latest version of R as of June 9, 2024 (v4.4.0). Different versions of R including v4.4.0 can be found at https://cran.r-project.org/bin/windows/base/old/. Versions older than v4.3.2 are not guaranteed to work.

RStudio is recommended to run the R scripts. Running the scripts from an independent R terminal could work in theory, but is not tested. RStudio can be installed from https://posit.co/download/rstudio-desktop/

The above packages should either be installed with RStudio's built-in package installer or by running the command `install.packages(c("sf", "raster", "dismo", "ncdf4"))` in an R terminal.

Both of these scripts require datasets that can be downloaded from the Box folder at https://ucdavis.box.com/s/l5iky1y6z526r6ewifvtr6c4ccp5jsjz. The `data` folder from the above link should be placed in the root folder of the repository.

Running the Scripts

If the above prerequisites are met, then the R scripts should run without issue. To do this in RStudio, open the desired script and click the "Source" button.

If a custom species is desired, the species CSV must be downloaded from GBIF and stored under `/data/GBIF/[species name].csv`. Furthermore, the `species_list` parameter in `training-preprocessing.R` must be altered to reflect the new species.

Script Descriptions

See the above section for instructions on obtaining the appropriate inputs.

`training-preprocessing.R`

```
Input:
  * Species datasets (.csv)
  * Historical climate datasets (.nc)
Parameters:
  * species_list: an array of strings describing the desired species
  * e: extent variable that describes longitude/latitude boundaries
Output:
  * Presence/absence files (.geojson)
  * Historical climate raster files (.tif)
```

Presence/absence files are created by using 1) occurrences from the species CSV file as presences and 2) an equal number of climate points from the climate

NetCDF file as pseudo-absences. Pseudo-absences are used due to lack of real absence data. GeoJSON files are stored under `/model/input/geojson`.

Climate raster files are converted from NetCDF to TIFF format for compatibility with python libraries. TIFF files are stored under `/model/input/historical`.

All outputs are cropped to the extent variable `e`. Default `e` is set to both Americas.

`prediction-preprocessing.R`

```
Input:
  * Future climate datasets (.nc)
Parameters:
  * e: extent variable that describes longitude/latitude boundaries
Output:
  * Future climate raster files (.tif)
```

Climate raster files are converted from NetCDF to TIFF format for compatibility with python libraries. Unlike the climate rasters for model training (of which there are only two), there exists a pair of future rasters for each year for each SSP model, making a total of 680 raster files. TIFF files are stored under `/model/input/future`, organized by SSP model and year. Raster files are cropped to the extent variable `e`. Default `e` is set to both Americas.

1.3 Species Distribution Modeling in Python

The SDM was prototyped in Jupyter Notebook (see `birdmig_sdm.ipynb`) and implemented in python script (see `model_training.py` and `model_prediciton.py`). Python scripts were used over Jupyter Notebook for implementation due to ease of use with loops and terminal.

Requirements

Python scripts were developed using Python 3.10.12, which can be found at this link. While you may use newer versions of Python, they are not guaranteed to work.

Required libraries include `os, tqdm, glob, pyimpute, pickle, rasterio, matplotlib, geopandas, sklearn, numpy,` and `pandas`. These libraries can be installed in a Python terminal using `pip`.

Ensure that preprocessed data from R exists under `/model/input/`. There should be three folders: `geojson, historical,` and `future`, each containing the appropriate data. `geojson` and `historical` are required for model training, while `future` and `/data/pickle` is required for model predicting. Example pickle files can be found at the Box link here.

1.3.1 SDM Training

The model is trained by associating presence/absence data with zonal averages for temperature and precipitation. In other words, the model learns what zonal environmental factors typically correlate to a presence or absence of a species. When given a file containing possible future climate readings, the model can then predict how the distribution of a species may react to that climate.

Training models for each bird species goes as follows:

1. Data is loaded into training vectors using the `load_data()` function
2. The random forest classifier is chosen for model training
3. Classifier's accuracy is measured by averaging K-fold evaluation (k=5)
4. Model is fit with training vectors and saved to a pickle file

Random forest was chosen for the model because it showed the highest and most consistent accuracy when compared to other classifiers (extra trees, SVM, xgboost, lightgbm) See `birdmig_sdm.ipynb`.

A description of the `load_data()` function is as follows:

```
load_data():
    Parameters:
        species: string representing species name
        n: optional int representing sample size (default=None)
    Return:
        train_xs: 2D array of floats where each inner array represents
                  zonal averages for temperature and precipitation
        train_y:  array whose values represent presence or absence &
                  are indexed to correlate to each array in train_xs

    load_data() loads appropriate files from /model/input/ into training
    vectors for SDM training. It takes an optional n parameter that
    randomly samples the presence/absence data in order to reduce the
    size of the training vector.

    load_data() primarily wraps the load_training_vectors() function
    from pyimpute, which is responsible for finding the zonal averages
    to begin with. load_data() is also responsible for postprocessing
    these vectors, removing NaN values to avoid training errors.
```

`model_training.py` is set up in such a way that it trains and saves models for each bird species in a loop, tracking the process using `tqdm`. If a custom species is desired, the `species_list` variable must be modified to reflect it (ensure that the custom bird species has already undergone R preprocessing). These models are saved to `/model/output/models` via the `pickle` library.

1.3.2 SDM Predicting

SDM prediction takes place in `model_prediciton.py`. Its prototyping can again be found in `birdmig_sdm.ipynb`.

Species distribution prediction uses the following process:

1. Future climate rasters are loaded from `/model/input/future` and processed into an appropriate form factor
2. Pickle files for each species is loaded from `/model/output/models`
3. `impute()` from `pyimpute` is used to create distribution predictions in the form of TIFF images for each climate raster.
4. TIFF images are converted to PNGs for use in the web application and stored in `/public/png-images`

There exists a future climate raster for each year (2015-2100) for each SSP (ssp126, ssp245, ssp370, ssp585). These rasters are used to predict distribution for five species, making a total of 1,720 images.

`model_prediciton.py` is comprised of a triple-nested for loop (species ¿ ssp ¿ year) and two functions, `make_predictions()` and `create_pngs()`. The former function is responsible for steps 1-3 listed above, while the latter function is responsible for step 4. TIFF images are converted to PNG through the use of `matplotlib`, and are saved in high resolution.

2 Web Application

2.1 APIs

2.1.1 React

Reactjs was used to created the frontend, and the main component is found under `/src/App.js`. In App.js, there are calls to the backend to gather data for the various components located in the `/src/components` folder. These components include:

- `BirdInfo.js`: returns the summary of the desired bird
- `ClimateChart.js`: returns temperature and precipitation graphs
- `HeatMap.js`: returns heatmap graph (see Leaflet section)
- `PolylineMap.js`: return bird trajectory graph (see Leaflet section)
- `SDMchart.js`: returns png of SDM output

- `PredictionControl.js`: contains code to display the year slider and SSP buttons that change the Climate and SDM Charts

In addition to these components, the sidebar and header are created in `App.js` which allows users to view different birds. All the styling is written up in `/src/App.css`.

2.1.2 Leaflet

Leaflet was used to map bird migration patterns in two ways under the "Trajectory" component.

The first map, "Individual Path", is created in `/src/components/PolylineMap.js`. The function fetches data from the csv files under `/backend/app/data`. The function first fetches all the tagged Bird IDs from the CSV for the desired bird. Then, all the trajectories for the first Bird ID are fetched from the backend and drawn on OpenStreetMap using Leaflet. Arrows are calculated between two points in `calculateBearings` function and added onto the map as well.

The second map, "Aggregated Path", is created in `/src/component/HeatMap.js`. The function fetches all the latitude and longitude coordinates from the CSV of the desired bird in the `/backend/app/data` folder. Using Leaflet and the Leaflet plugin, `L.heatLayer`, a heat layer is generated showing the combined trajectory path of all the tagged birds on top of OpenStreetMap layer.

2.1.3 FastAPI

The middleware and backend of the website was developed using python and FastAPI, in the file `base.py` located under `/backend/app`. The FastAPI backend primarily serves the purpose of quickly retrieving data and sending it to the front end as requested, allowing the front end user-interface to remain lightweight and easy to load quickly. FastAPI was used to build the backend due to its ease in creating end points for a frontend application to call, natively supporting features such as delayed requests and an intuitive way of passing arguments to the FastAPI backend.

The backend is defined predominantly by functions that handle various requests the front end makes in order to retrieve data to display. These functions and their functions are discussed in greater detail in the following section for RestFUL API.

2.1.4 RestFUL

RestFUL API was used to define how the end points were set up, partially due to built-in compatibility with FastAPI, but also due to its simple and intuitive interface. While RestFUL API primarily defines 4 primary methods for GET, PUT, POST, and DELETE, the final application was operational with a stateless backend, meaning only GET and PUT requests were used. In the file base.py, GET and PUT requests were used to define the following end points:

- `get_temperature_data()`: a GET endpoint that in turn performs a POST request to grid2.rcc-acis.org to obtain temperature data for the state of California throughout a year, which is specified per the user's request. Currently, this endpoint only utilizes the SSP8.5 scenario as other endpoints have not yet been developed by RCC-ACIS.

- `get_precipitation_data()`: a GET endpoint that also performs a POST request to grid2.rcc-acis.org to obtain precipitation data for the state of California throughout a year, specified per the user's request. Like the temperature endpoint, it also currently only utilizes the SSP8.5 scenario due to the limited development of other climate scenarios by RCC-ACIS.

- `get_predictions()`: a PUT end point that retrieves the pre-computed predictions the Species Distribution Model makes about a specified bird, for a specified year based on specified emission data. While a GET request would have sufficed for its current utility, a PUT request was preferred as this could be expanded to implement a machine learning model to generate predictions dynamically should such a feature be planned in the future.

- `get_trajectory_data()`: a GET end point that retrieves a csv file containing the migration trajectory of a specific bird, identified by its unique bird ID, that can be used to generate individual trajectory maps of an individual bird. The trajectory information is discussed later in the subsection for Leaflet.

- `get_bird_ids()`: a GET end point that returns all unique bird IDs per a given species. This is used to by the front end component PolylineMap.js to search for a specified bird ID.

- `get_heatmap_data()`: a GET end point that retrieves heatmap data of a bird species, to display the aggregated paths of all birds catalogued for that species on the front end with HeatMap.js, as a heat map. This is discussed in further detail in the Leaflet subsection.

2.2 Usage

The website is designed to serve as an interactive tool for ornithologists, researchers, and bird watching enthusiasts. It allows users to:

- Retrieve detailed information about various bird species, including their scientific names, general characteristics, and migration patterns.

- View and interact with climate data visualizations to understand how different climate variables such as temperature and precipitation affect bird migration.

- Predict future distributions of bird species using species distribution models (SDM) based on selected climate scenarios and years.

- Explore bird migration trajectories either as individual paths or aggregated heatmaps, providing a visual understanding of migration routes and densities.

Users can interact with various controls to select different birds, years, and emission scenarios, and the application dynamically updates the information and predictions displayed.

2.3 Design

The website is developed using a React frontend and a FastAPI backend, leveraging modern web technologies and frameworks to ensure a responsive and intuitive user experience. Key design features include:

- **Modular Design:** The application is structured into several independent components like BirdInfo, SDMChart, and PolylineMap. Each component encapsulates a particular feature of the web app, facilitating seperation of concerns, readability, and scalability.
- **Data-Driven Interactions:** By integrating with external APIs and processing data through Python-based FastAPI, the application provides real-time data interactions and updates.
- **User-Centric Interfaces:** Emphasis on usability and accessibility, with interactive charts, maps, and dynamic controls that cater to both novice users and experienced researchers.

The application's architecture is designed to handle high volumes of data efficiently, utilizing caching and asynchronous data fetching to optimize performance and user experience.