User Guide for Graph Neural Network-based Species Distribution Model (GNN-SDM)

1. Introduction

Species Distribution Models (SDMs) serve as indispensable instruments in ecological studies and conservation efforts by projecting where species might inhabit and evaluating habitat suitability. They function by examining how species occurrence records relate to environmental factors such as climate conditions, precipitation patterns, and land use. Through this analysis, SDMs help predict potential habitats in areas that have not been thoroughly surveyed. They are key to locating suitable environments, anticipating species' adaptive responses to environmental shifts, and assessing risks posed by phenomena like climate change and habitat degradation. Accordingly, SDMs have found extensive use in biodiversity conservation, management planning, and guiding species reintroduction and relocation programs.

However, for many threatened species, the limited availability and spatial bias of occurrence data often prevent effective species distribution modeling. This issue is particularly pronounced in approaches that rely solely on presence-only points, such as MaxEnt, Random Forest, and various ensemble models. Such reliance increases the difficulty of conducting large-scale assessments for threatened species and poses challenges to biodiversity conservation efforts. Although the growth of open-access databases has improved data availability, occurrence records collected by different organizations often lack standardized sampling protocols, as they are gathered at different times, locations, and under varying conditions. These inconsistencies introduce uncertainties that can cause biases when aggregating data for conservation planning across multiple species.

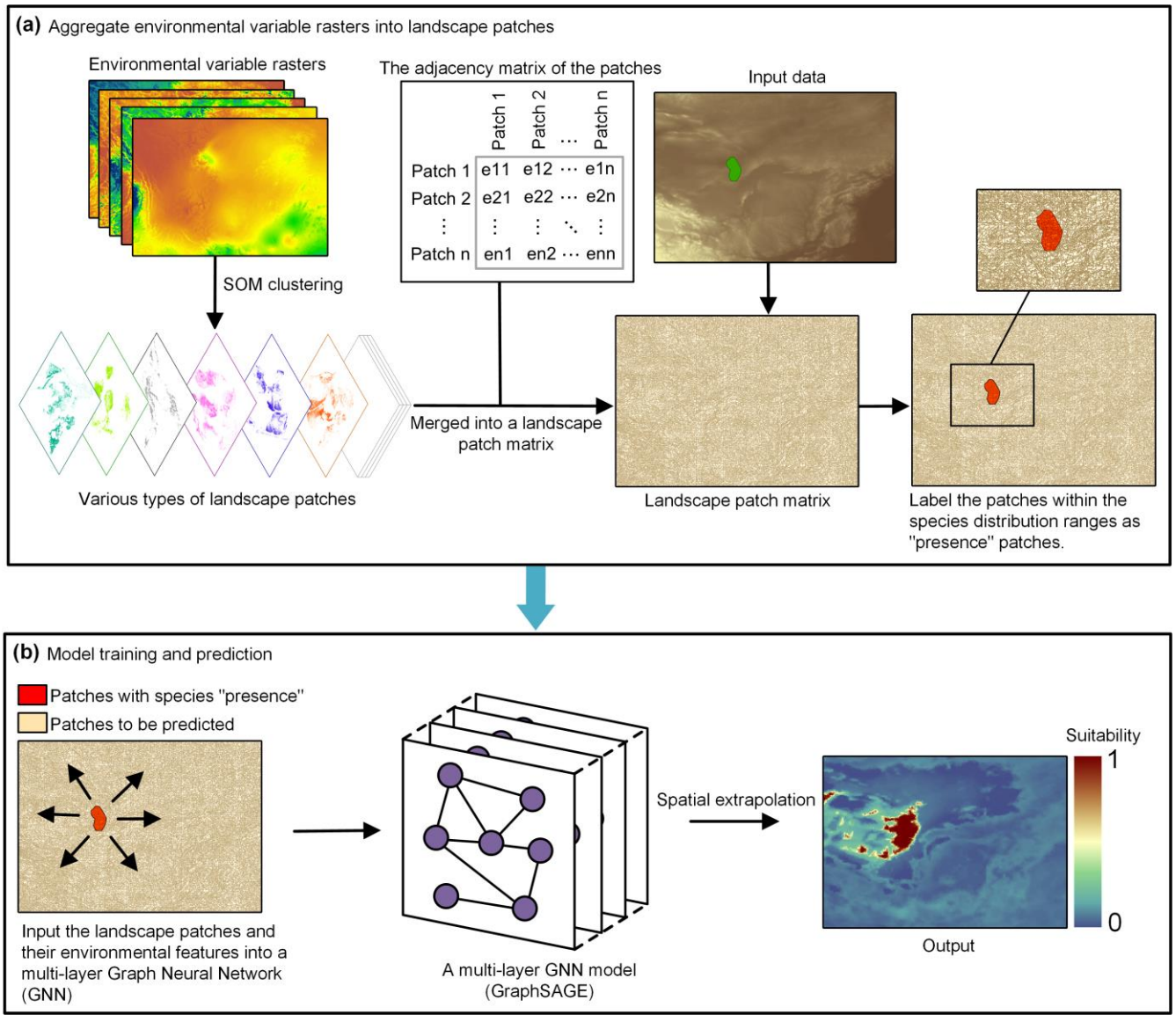
The IUCN Red List (https://www.iucnredlist.org/) provides spatial datasets that map species' geographic ranges, distinguishing between historical and current distributions across native and introduced regions. Offered in polygon (.shp) format, these datasets allow for detailed spatial analyses based on precise distribution boundaries. As products of rigorous scientific assessments, they deliver not only reliable distribution data but also critical information on species' threat levels, serving as important resources for biodiversity research and conservation planning.

GNN-SDM (Graph Neural Network-based Species Distribution Model) is a species distribution modeling framework that uses spatial polygon data representing species distribution ranges as input and incorporates landscape patches to account for complex environmental features and landscape patterns. This approach supports the use of high-quality, standardized species distribution data available from the IUCN Red List website (https://www.iucnredlist.org/). This document provides a detailed guide for implementing the method, including step-by-step instructions, practical tutorials, and case study examples to facilitate its application.

2. Tutorial: GNN-SDM Application Guide

In this tutorial, we demonstrate how to perform species distribution modeling using GNN-SDM with the provided scripts and example data. To facilitate first-time use, all data required for this tutorial have been organized into the \test folder. Running GNN-SDM involves two main steps: (1) generating landscape patch files and (2) training and predicting with the GNN model. The figure below provides an overview of the GNN-SDM workflow:

GNN-SDM Framework



2.1. Step 1 — Create Landscape Patches

First, a method called Self-Organizing Map (SOM) is used to cluster all pixels based on environmental raster variables into landscape patches, resulting in the output of a patch file named \patch.shp. This landscape patch file consists of numerous mosaic polygons of varying shapes and sizes, capturing the complex landscape structure and environmental gradients within the study area. Each patch represents a homogeneous landscape unit characterized by similar environmental features. The steps described in this section are implemented in R, and a simple example is provided for users to follow.

To create the landscape patch file, you need to run the provided R script \step_1 Generate landscape patches.R.

The code in this script demonstrates the process of generating a landscape patch matrix based on surface environmental conditions. We use multiple environmental variables covering climate, topography, vegetation cover, and human activities as predictors for clustering. These data can be sourced from remote sensing imagery, meteorological station observations, or publicly available environmental datasets. To ensure the script runs correctly, all raster datasets must undergo necessary preprocessing to maintain consistent spatial resolution, coordinate reference system, and spatial extent. In this tutorial, all environmental raster data used in the example are provided in the \test folder. Users may replace or modify these datasets according to their specific research needs.

Environmental raster data

Environmental variables and corresponding raster files

Environmental variables	Raster file
Annual mean temperature	\bio1.tif
Mean diurnal range	\bio2.tif
Annual precipitation	\bio12.tif
Precipitation seasonality	\bio15.tif
Wind speed	\ws.tif
Aridity Index	\ai.tif
Elevation	\ele.tif

Slope	\sp.tif
Drainage density	\dd.tif
Shannon index of habitat heterogeneity	\si.tif
Normalized difference vegetation index	\ndvi.tif
Canopy height	\ch.tif
Solar radiation	\sr.tif
Human footprint index	\hfp.tif

Note: All environmental variable data must be in GeoTIFF (.tif) format. In this tutorial, we use South America as the study area for demonstration purposes.

Study area file

The study area for this demonstration is defined by a polygon file representing the boundary of South America (\ecoregions.shp) and serves as the spatial extent for all analyses. If the polygon file used to define the study area contains multiple polygon features, with each feature representing a specific geographic unit such as an ecoregion, country, state, administrative unit, or subregion, the script will automatically generate a separate landscape patch file for each polygon feature. These individual patch files are then merged to create a global landscape patch matrix. This approach accommodates differences in extent, resolution, and ecological characteristics among different spatial units. Since this is a tutorial example, we use a regular grid covering South America instead of actual ecoregions or administrative boundaries.

Path setting

In the R script \step_1 Generate landscape patches.R, users are required to specify the file paths for all input data needed for this step. Specifically, users must set a working directory that contains all the environmental raster data as well as the boundary file defining the study area.

Before running the script, please ensure that all required file paths are correctly set：

work_path

The root directory of the project. The default is the example folder \work_path provided in this tutorial.

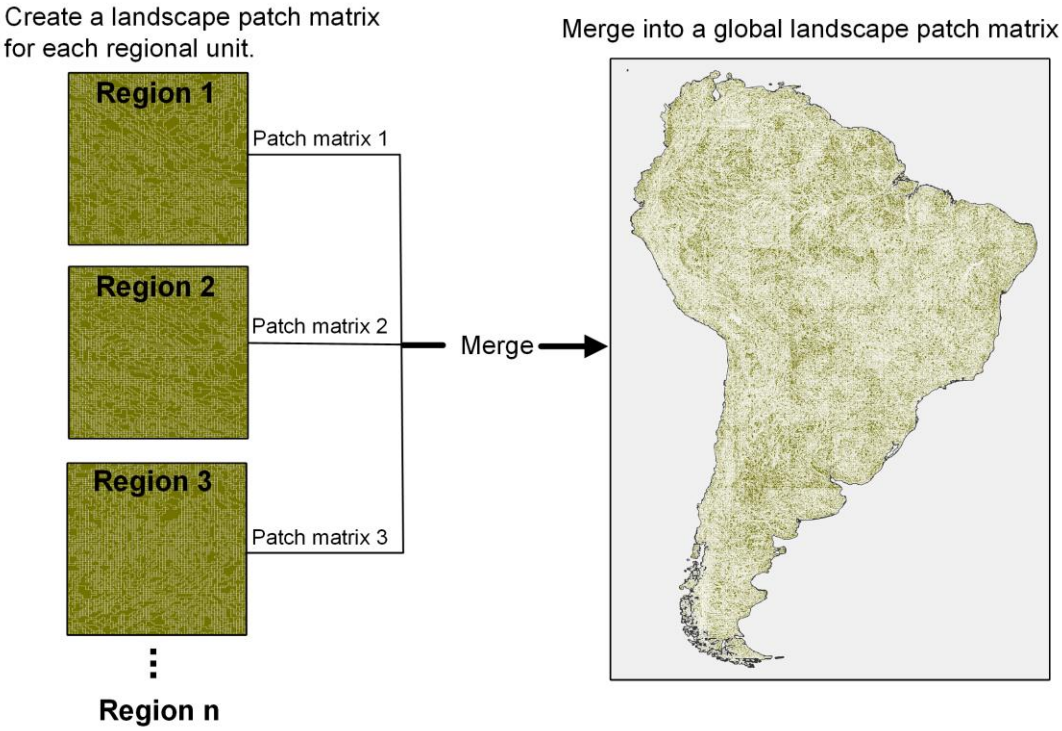
factors_path

The directory containing environmental raster data. The default is the example folder \work_path\factors provided in this tutorial.

study_area_path

The file path to the spatial polygon (.shp) that defines the study area extent. By default, it is set to the example file provided in this tutorial: \work_path\database\ecoregions.shp.

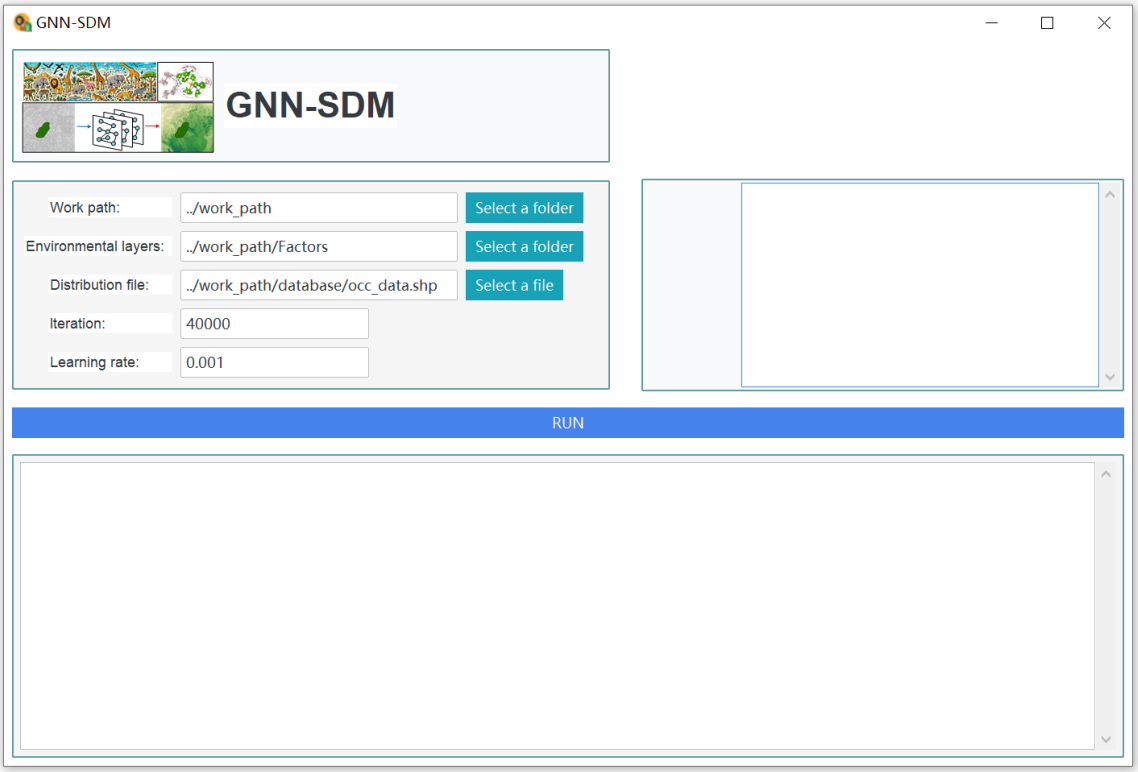
After running the R script \step_1 Generate landscape patches.R, the program will automatically generate landscape patches for each subregion within the study area and merge them into a single global landscape patch file, which is output as a spatial polygon file named patch.shp. This landscape patch matrix consists of multiple adjacent polygons, with each polygon representing a small unit characterized by similar environmental features.


Create a landscape patch matrix
for each regional unit.
Region 1
Patch matrix 1
Region 2
Patch matrix 2
Region 3
Patch matrix 3
Region n
Merge
Merge into a global landscape patch matrix

2.2. Step 2 — Training and Prediction of GNN-SDM

After generating the landscape patch file, users can proceed with species distribution modeling using GNN-SDM. We provide a graphical user interface (GUI) executable program to facilitate ease of use. The program is developed in Python and incorporates the GNN components of GNN-SDM using the PyTorch Geometric (PyG) (https://pytorch-geometric.readthedocs.io/en/latest/index.html) Python package.

To launch the program, users simply need to double-click the executable file \GNN_SDM.exe located in the \dist folder. The graphical user interface offers a clear and intuitive way to interact with the software, allowing users to manually set input file paths and parameters.



Note: Before running the program, users must first complete *Section 2.1: Step 1 - Create Landscape Patches* to generate the landscape patch file required for GNN-SDM training and prediction.

Parameter Description

Work Path

To run GNN-SDM, you need to specify the project working directory. In this tutorial, the default working directory is \work_path. This directory must contain the environmental raster folder \work_path\factors, the species distribution data \work_path\database\occ_data.shp, the study area boundary file \work_path\database\ecoregions.shp, and the landscape patch file \patch_x.shp with environmental features generated from *Section 2.1: Step 1 - Create Landscape Patches*.

Please ensure that all file and folder paths contain only English characters, as the use of other characters may cause encoding errors. It is recommended to manually select the file and folder paths.

Distribution File

Select the file path to the spatial polygon data that describes the species distribution range. The file must be in *.shp* format. Species distribution data can be obtained from the IUCN Red List website (https://www.iucnredlist.org/). You should directly select the file path, for example, \work_path\database\occ_data.shp.

In this tutorial, we use the small South American mammal *Aegialomys xanthaeolus* as an example. This species occupies distinct environmental conditions and is primarily distributed along the continental margins of South America. The input species distribution file must include an attribute column named "*binomial*" to identify the species name (or another unique species identifier), which can be assigned based on the scientific name or a custom identifier.

If the species distribution data contains multiple species, the binomial attribute column is used to identify the unique species name represented by each polygon feature. The program will iteratively predict habitat suitability for each species based on its distribution range and output the results for each species into the folder \work_path\predictions.

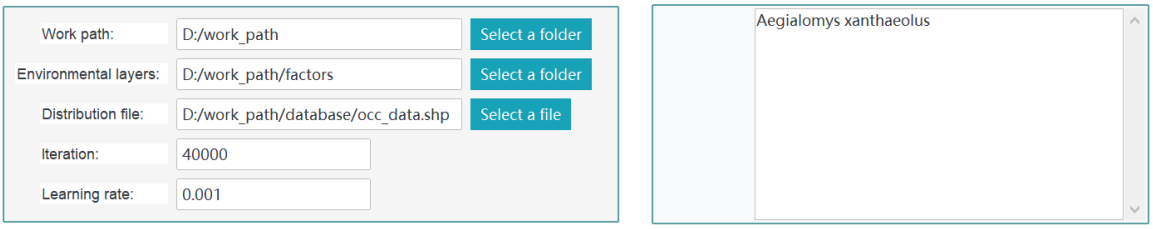
Table

occ_data

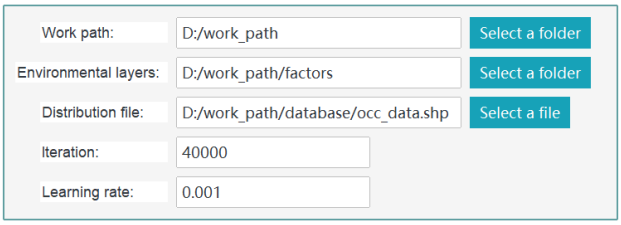
FID	Shape	binomial
0	Polygon	Adenomera araucaria
1	Polygon	Allobates masniger
2	Polygon	Allobates myersi
3	Polygon	Allobates nidicola
4	Polygon	Allobates olfersioides
5	Polygon	Amazophrynella javierbustamantei
6	Polygon	Amazophrynella moisesii
7	Polygon	Ameerega bilinguis
8	Polygon	Ameerega flavopicta
9	Polygon	Ameerega macero
10	Polygon	Ameerega petersi
11	Polygon	Adenomera bokermanni
12	Polygon	Aplastodiscus perviridis

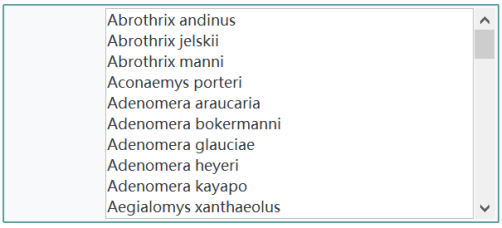
Note: Each species' distribution must be represented by a single feature class within the input file. If a species has multiple distribution areas represented by multiple features, the program will treat them collectively as the distribution range of the same species. If a species' distribution is composed of multiple features, it is recommended to use the "*Dissolve*" tool in ArcGIS (based on the "*binomial*" attribute) to merge them into a single feature. Once the spatial polygon file for species distribution is selected, all species names contained in the file will be displayed in the panel on the right.

Single species



Multiple species





Environmental Layers

The selected directory should contain all environmental raster files in .tif format. These raster files must have consistent spatial resolution and extent. For first-time users, it is recommended to directly select the example environmental raster folder provided in this tutorial: \work_path\factors.

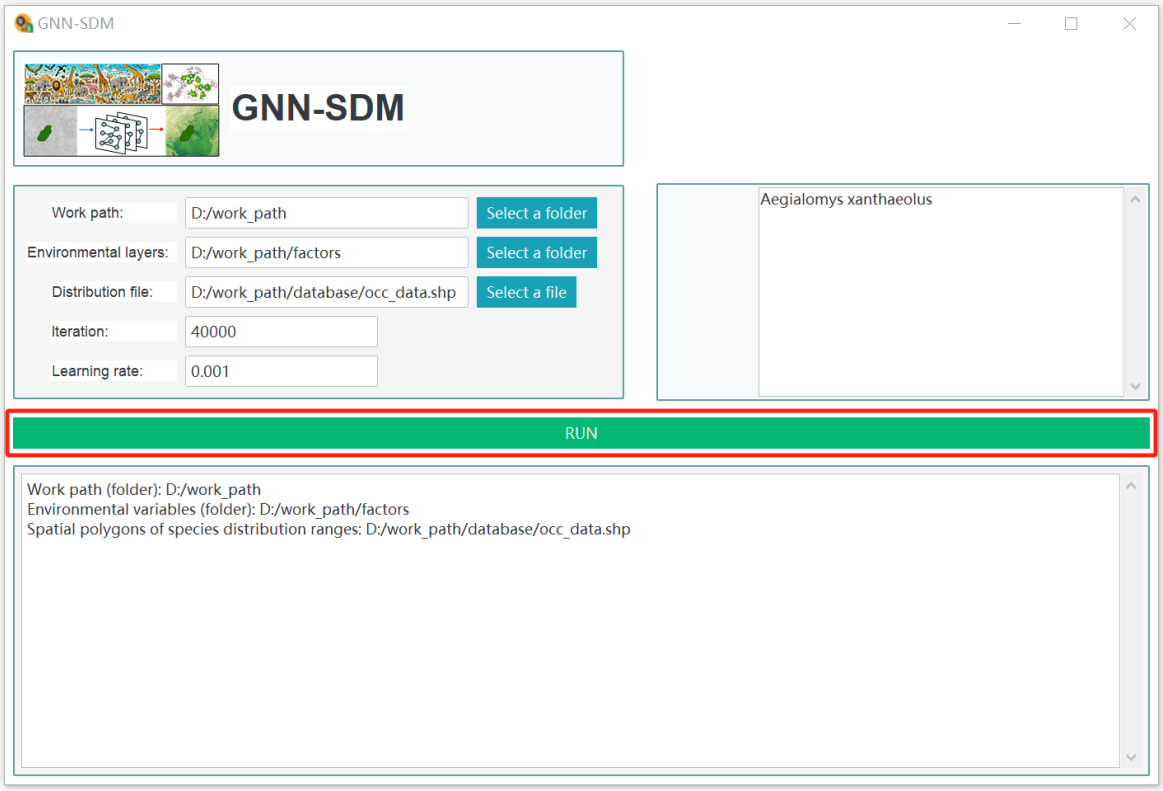
Iteration

The number of training iterations for the model, with a default value of 40,000. Users can adjust this value based on their specific needs.

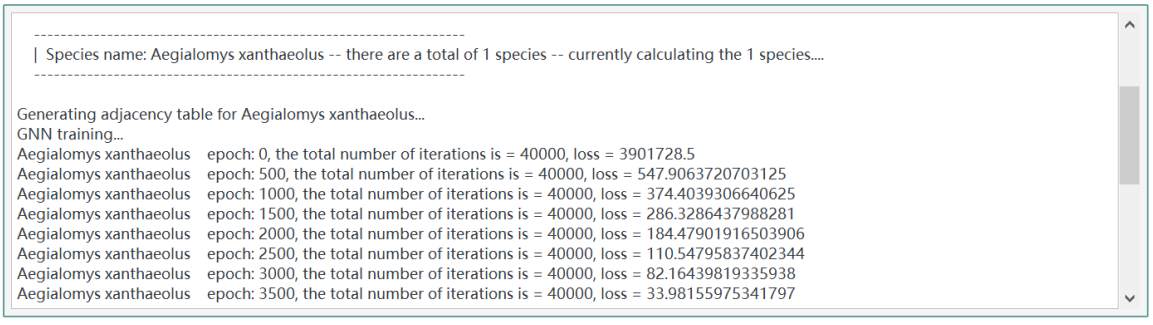
Learning Rate

The learning rate parameter used during the training of the GNN-SDM model. A higher learning rate can speed up convergence but may also cause training instability or prevent the model from reaching the optimal solution.

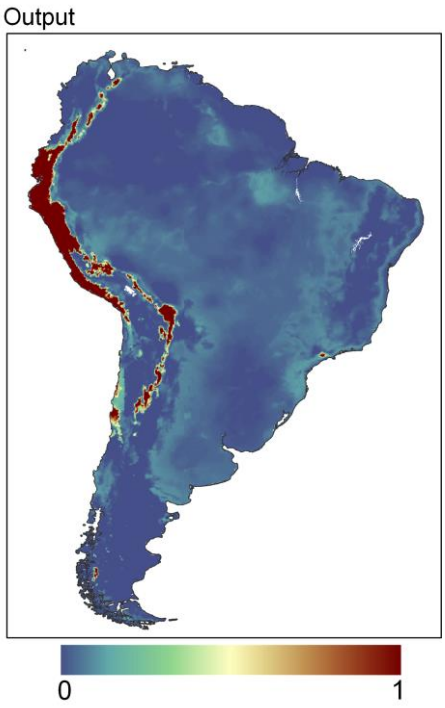
After configuring all parameters, click the "*RUN*" button to execute the program and initiate the model training and prediction process. The program will automatically train the model based on the input data and output habitat suitability predictions for each species.



During execution, the program generates a log text that provides detailed information about the GNN-SDM training process. If any errors occur, they will also be displayed in the log window, helping users monitor progress and quickly identify any issues encountered during execution for easier debugging and correction.



Upon completing the run, the prediction output file named out.tif will be saved in the \predictions folder within the working directory. This file is a GeoTIFF raster in which each pixel value represents the predicted habitat suitability probability for the corresponding location. The values range from 0 to 1, with higher values indicating areas that are more suitable for the target species.



Note: Each time the program is run, a \predictions folder is created within the working directory. This folder contains an output directory for each species, named according to its unique identifier. If you wish to re-predict the habitat suitability for a specific species, you must first delete the corresponding result folder; otherwise, the program may not overwrite the previous prediction results.