**Explanation of scripts**

Note: an actively updated version of this document is available at: <https://github.com/jcoliver/biodiversity-sdm-lesson/blob/master/docs/script-explanation.md>

This document provides a more in-depth explanation of what the four species distribution modeling scripts are doing. All four scripts are documented in the actual code, so if you want *even more* detail, feel free to look through the R code itself. If you are running into problems with installation or running the analyses, see the document titled “HelpDocumentforCommonErrorsandHelpfulWebsites”.

The four scripts prefaced with "run-" in the scripts folder are all doing largely the same thing, but they differ in (1) the species observation data being used and (2) the climate data being used. Considering each script in turn, we start with the simplest one, and describe the remaining three in terms of modifications from this base script.

**run-sdm-single.R**

* **Summary**: This is the template for the creation and evaluation of a species distribution model (SDM) for a single species based on current climate data. It plots the output of this model onto a map, which is saved as a pdf file.
* **Process**:
1. Check to make sure data files exist and the additional R packages are installed
2. Format the data so they are analyzed appropriately
3. Run the species distribution model
	1. Determine the geographic extent of our data and restrict analysis to that geographic scale. We do this to reduce computation time; it is a fairly safe assumption that if we are analyzing a species from the Sonoran desert of southwestern North America, we do not need to evaluate suitability of the climate of northern Eurasia.
	2. Create "pseudo-absence" points. In order to evaluate the performance of the model, we need both presence points (where the species of interest is known to occur) and absence points (where the species of interest is known to *not* occur). The challenge is that most biodiversity observation data are *presence only*. One common work around for coercing presence-only data for use with presence/absence approaches is to use pseudo-absence, or "background" points. While "pseudo-absence" sounds fancy, it really just means that one randomly samples points from a given geographic area and treats them like locations where the species of interest is absent. A great resource investigating the influence and best practices of pseudo-absence points is a study by Barbet-Massin et al. (2012) (see Additional Resources below for full details).
	3. Divide data into a "training" set and a "testing" set. Because we will need to evaluate the performance of the SDM, we reserve some portion of the observation data (and pseudo-absence data), in this case 20% of the observations; this is the "testing" data. The remaining 80% of the observations are used for "training" the SDM. These *independent* data sets afford an unbiased evaluation of the models.
	4. Using the bioclim approach and the **training** data, estimate the effects of the climate variables on the probability that our species of interest will occur in a given location. From an abstract perspective, we are building a model *y = b1x1 + b2x2 + b3x3 + ... + bkxk*, where *xi* is the value of the *ith* climate variable (e.g. annual rainfall) for a particular geographic location and *bi* is the slope for the effect of that climate variable on the probability of presence of the species of interest at a particular geographic location. In this case the values of *xi* are the current climate data that were downloaded from the WorldClim site. You can find out more about the bioclim algorithm in the [documentation for the dismo package](https://www.rdocumentation.org/packages/dismo/versions/1.1-4/topics/bioclim).
	5. Using the **testing** data, including presence points and the randomly-generated pseudo-absence points, evaluate the SDM. This compares the probability of presence estimated by the SDM at a particular geographic location to the actual presence and (psuedo-)absence of the species of interest. Note this evaluation *only* considers locations for which we actually have data (presence or pseudo-absence).
	6. Use the current climate data to predict the probability of presence at **every** point in the geographic area of interest.
4. Plot the results of the species distribution model on a map and save it to a PDF file.

**run-future-sdm-single.R**

* **Summary**: This is the template for the creation and evaluation of a species distribution model (SDM) for a single species based on **future** climate data. It plots the output of this model onto a map, which is saved as a pdf file.
* **Process**: Identical to that of run-sdm-single.R, except that in step iii.f, this script uses estimated climate data for the year 2070.

**run-sdm-pairwise.R**

* **Summary**: This script provides the template for running two separate SDMs: one based on data for an insect and one based on data for the insect's host plant. The resultant predicted ranges are based on current climate data.
* **Process**: This is effectively running the single SDM script (run-sdm-single.R) twice, once for the insect data and once for the plant data. The two models are estimated and evaluated independently of one another. After the climate data are used to predict the probability of presence for each species, a single map is created, showing the predicted ranges of the insect, the plant, and areas where the ranges overlap.

**run-future-sdm-pairwise.R**

* **Summary**: This script provides the template for running two separate SDMs: one based on data for an insect and one based on data for the insect's host plant. The resultant predicted ranges are based on future climate data (2070).
* **Process**: Differs only from run-sdm-pairwise.R in the data used for predicting the ranges of insect and plant: in this case climate data estimated for 2070 are used, rather than current climate data.

**Additional Resources**

* [A study on the effect of pseudo-absences in SDMs (Barbet-Massin et al. 2012)](https://dx.doi.org/10.1111/j.2041-210X.2011.00172.x)