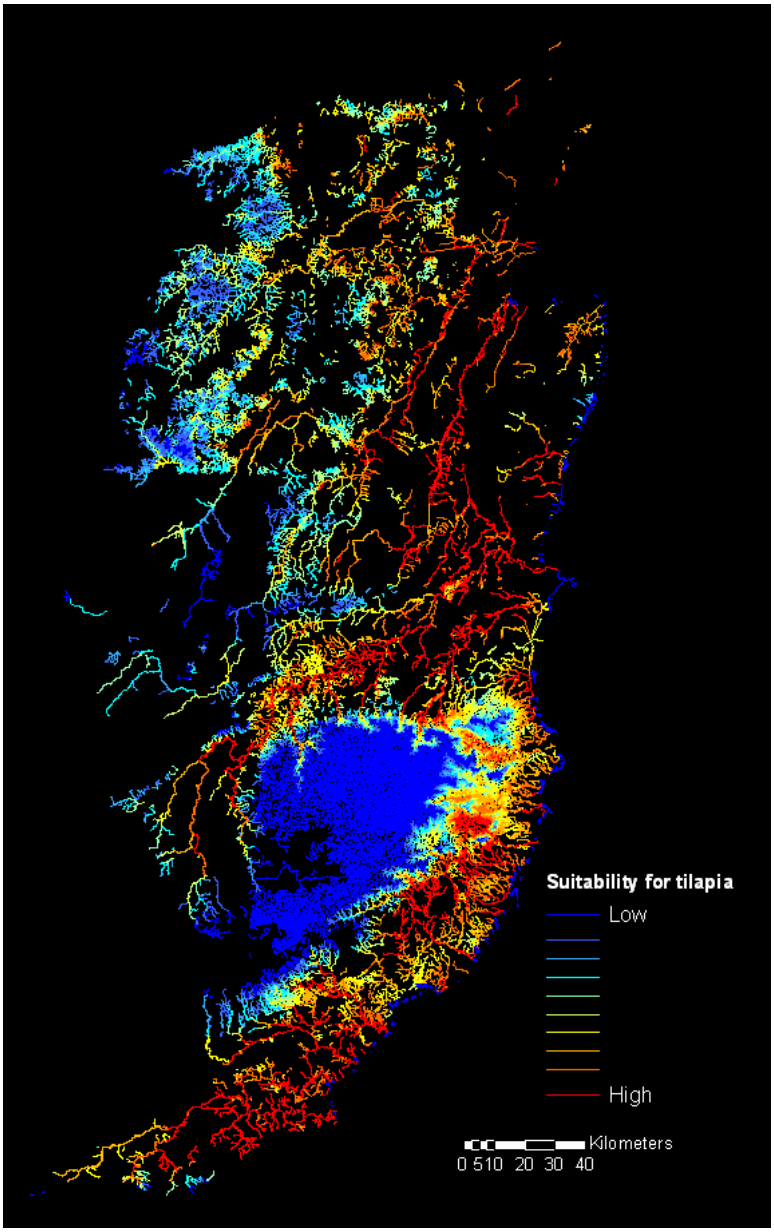
A Tutorial for the

Prediction of Aquatic Species Distributions



17 April 2007

Peter C. Esselman
School of Natural Resources and Environment
University of Michigan
Ann Arbor, MI 48109 USA
esselman@umich.edu

iabin
Red Interamericana de Información sobre Biodiversidad

Table of Contents

Introduction

Without accurate information about the biodiversity that exists in a place, it is difficult to make good decisions about how human societies can maintain a healthy environment to live in. In most places around the world, even in well-researched countries, there are many gaps in our knowledge of biodiversity simply because there are many places where scientists have not done species inventories. Information limitation is particularly acute in developing countries where the capacity to fund and carry out scientific research is limited. For this reason, it is necessary to leverage the utility of the information that *does* exist to work to the greatest extent possible to inform the decisions we make.

With increased access to computer processing power, freely available software, and increasingly available online datasets, it is now feasible for anyone with a good computer, a fast internet connection, and some patience to make their own quantitative predictions about where given species are likely to occur in a given landscape. These predictions can be made by creating empirical models—mathematical formulations that describe the relationships between two or more variables. Though modeling of species distributions involves mathematics, it does not require that the person who creates the models knows a lot about math; powerful modeling software exists that does the math for us. You give the software the data and the software gives you results and maps to interpret and utilize.

This guide is intended to be a basic "how-to" manual that will allow you to generate your own models and maps of species distributions from point occurrence data about where a species has been documented, and geographic information about environmental and habitat characteristics relevant to the species. In particular, the manual focuses on organisms that live in river and stream environments. Aquatic ecosystems like rivers and streams differ from terrestrial and marine ecosystems in that (1) they have hard boundaries at the water's edge and thus are very linear and spatially discrete; and (2) any given point in a river network is influenced not only by local conditions but by the integrated conditions in the upstream watershed. For these reasons, aquatic ecosystems can be considered a special case in the implementation of species distribution models. In practice, much of the logic that goes into species modeling is the same between aquatic ecosystems and terrestrial or marine systems—the main differences lie only in the preparation of the data. Thus, while this tutorial gives specific guidance on how to think about modeling aquatic species, much of the content is applicable to species in other ecosystem types.

Much of the bulk of this tutorial describes the process of preparing datasets in geographic information system software. Users with data already prepared for them, or with an interest in gaining a broader perspective on the modeling process are encouraged to skip over substantial portions of Step 3, as there is valuable conceptual and practical advice both before and after the detailed section covering GIS methodologies, including a brief overview of potential applications of species distribution models in the final section of the document.

Overview of the modeling process

The modeling process presented here consists of 5 steps that begin with the creation of an ecological conceptual model of a species, move through the development and implementation of empirical models, and conclude with the interpretation of maps and other model outputs (Figure 1). The user, upon finish the processes described in the tutorial, will not only have a wealth of valuable datasets, but should be equipped with maps and predictive models that can be used to visualize patterns of species distributions throughout river and stream networks.

Step 1 in the process is to develop a detailed ecological conceptual model that identifies factors thought to drive patterns in species presence within a well-defined geographic area. This step establishes both the geographic scale of the analysis, and defines in as much detail as possible, a list of potential predictor variables for consideration in the model. This is where the intellectual heavy-lifting occurs and where consultation with ecological experts and the scientific literature is very useful.

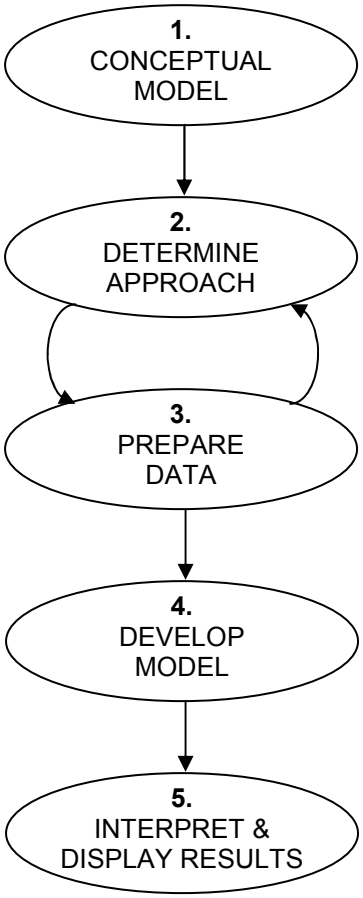
In Step 2 you must select the specific modeling approach to be used based on the amount and types of data available. Though there are many approaches to choose from, one specific technique is advocated here—an approach that is user friendly, quantitatively rigorous, requires only data about where a species is known to be present (not where it is absent), and works well with low sample sizes.

In Step 3, you take the idealized list of predictors developed in the conceptual model, seek appropriate data from your geographic area to satisfy this list, and prepare the data for entry into the model. It is in this step that the modelers must face the realities of data limitation and make decisions about which available data best satisfy the predictor variable list. Also in this step, species occurrence records must be compiled and appropriately formatted. Because the end product of the entire process is a map generated in a geographic information system (GIS), this is the stage where GIS data must be prepared to feed into the models. This step is likely to require the most effort and time investment.

Step 4 is the model development process. This step involves model training, validation, and evaluation, all of which are automated by the modeling software introduced here.

In Step 5, the final step, you will map the model predictions and interpret the performance of the model.

The entire process is computer driven and dependent on several different software programs. If you will be preparing many datasets, you will want to find the best workstation possible (fast processor, high RAM, plenty of hard disk space). You will also want to calibrate your expectations—species modeling can be an involved process. However, once you have done the leg work to prepare your datasets, you will be able to easily model many species, species invasion, scenarios for future landscape and climate change, and more.



	Goal	Main Activities
1. CONCEPTUAL MODEL	To formulate a clear idea of the factors that influence the ecology and distribution of your target species.	Hypothesize and diagram assumed relationships between the species and its local and regional environment.
2. DETERMINE APPROACH	To decide on a modeling approach that matches your needs and data limitations.	Establish the type of model that you will develop and the software you will use to do the work.
3. PREPARE DATA	To assemble and properly format all necessary data layers for entry into the modeling software.	Create GIS datasets that represent the variables hypothesized as important in your conceptual model.
4. DEVELOP MODEL	To use your data to train and evaluate a model that accurately predicts distribution of your species of interest.	Feed data in to the modeling software; parameterize and run the model.
5. INTERPRET & DISPLAY RESULTS	To judge the performance of your model and display results for interpretation.	Evaluate model outputs and create maps to aid with communication of results.

Figure 1. The five steps of modeling species distributions, and the goals and main activities involved in each step (simplified from Guisan and Zimmermann 2000).

Step 1. Develop ecological conceptual model

In Step 1 you must conceptualize the ecological situation you are trying to model, but first you should have a clear picture of why you want the model outputs. Are you interested in a particular species or all species? Do you want to predict species invasion or native species? How do you hope to use the information when you have maps of species distributions? The answers to these questions will be highly context dependent, but you will want to enter the modeling process with a clear idea of what you want to accomplish.

Once you have a clear picture of the species you want to model, you must develop a conceptual model for this species. This conceptual model will be the road map for the rest of the modeling process that follows. At the end of Step 1 you should have a concept diagram with a basic structure like the one in Figure 2. In this figure, the species is located at the base of the diagram as a triangle. Above the species is a layer of **key ecological attributes (KEAs)**: the natural conditions that must be met for a species to survive in a location. Above this, in the third layer, are the **landscape factors** that help determine the states of the KEAs at the local scale. Depending on your species mapping problem, you may or may not decide to include a final group of factors: **human activities**, which can strongly influence species and ecosystems. A

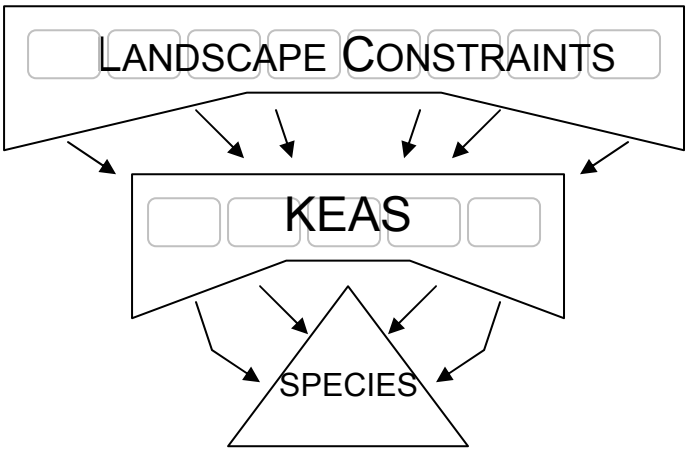


FIGURE 2. A generalized concept map showing relationships between factors that might constrain the occurrence of a species. KEAs are Key Ecological Attributes (defined in text).

more detailed description of each layer will help guide you in the development of your conceptual model.

KEAs.
Key ecological attributes are the ecological characteristics that shape a species' natural variation over time and space and ensure its long term persistence[1]. These are characteristics that influence the species at the local scale in the very specific space it occupies within the larger aquatic habitat. In aquatic ecosystems there are at least 5 categories of KEAs[2]:

1. Hydrologic regime – Factors that influence the timing, magnitude, duration, rate of change, and frequency of water flow in a river or wetland[3].
2. Chemical regime – The state of chemical parameters (e.g., pH, dissolved oxygen, turbidity) within the aquatic environment.
3. Physical habitat – The physical template of the ecosystem (habitat volume, substrate availability, channel structure and stability, etc.).
4. Energy regime – Basic food availability and rates of movement and cycling of organic nutrients through a food web.
5. Biotic interactions – Interactions between the species and other living things in the ecosystem; could include predation, competition, parasitism, disease, etc.

You should try to identify which specific KEAs are most important for your species. This may require consultation with an expert ecologist or some library research. A suggested pool of possible variables to consider can be seen in Figure 3. You should try to identify a subset of these that you can justify as likely to be important to the species.

Landscape factors.
Whereas KEAs are generally measured at the local scale, they are very often constrained by factors at much larger scales—the landscape scale. Large scale factors like geology, climate, and land cover can all influence the state of KEAs at the local scale, and acting through the KEAs can indirectly affect the presence and abundance of species. As with KEAs, it is a good

[1] Parrish et al. 2003
[2] Karr et al. 1986
[3] See Richter et al. 1996

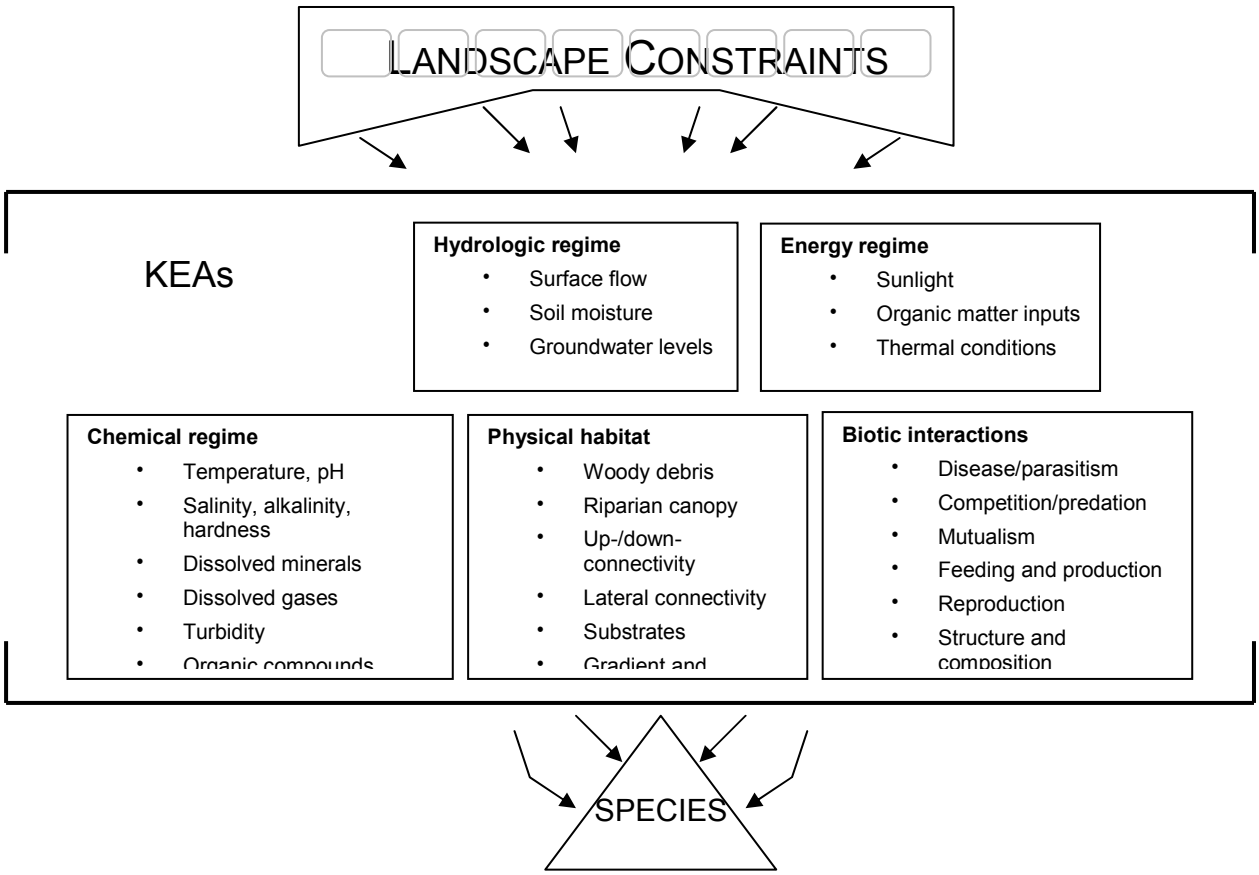


FIGURE 3. A detailed list of the types of ecological factors that might be important to species in each key ecological attribute category. A subset of these should be chosen based on scientific knowledge about the species in question.

idea to consult with experts on the species you are trying to model to determine what large scale factors may be important determinants of local scale KEAs. Generally speaking, climatic variables are important because they determine temperature, light, and flow conditions; geologic and soil conditions are important because they determine channel shape, substrates, and water chemistry; topography is important in determining flow conditions and physical habitat; and land cover can be important to determine chemical regime and energy regime (Figure 3). Several other landscape variables also deserve consideration. These include distance from sea, a variable that often correlates with species occurrence, and latitude and longitude which may correlate with many of the landscape and local scale variables. Adding these, our conceptual model becomes slightly more complex (Figure 4).

Human activities.

Human activities can have a very strong influence on the distributions of species. Dams, urban and agricultural land use, aquaculture facilities, and roads can all exert influences on the species by changing the status of KEAs or landscape factors that correlate to a species' presence in a location. Therefore, human activities are also good to consider in our conceptual model. In some cases, human activities can be represented as landscape factors, such as with agriculture and urban land uses, which can be represented as land cover classes. In other cases, human influences occur at the level of KEAs, such as with industrial point sources of pollution that influence chemistry regime. Finally, there are some cases where human activities

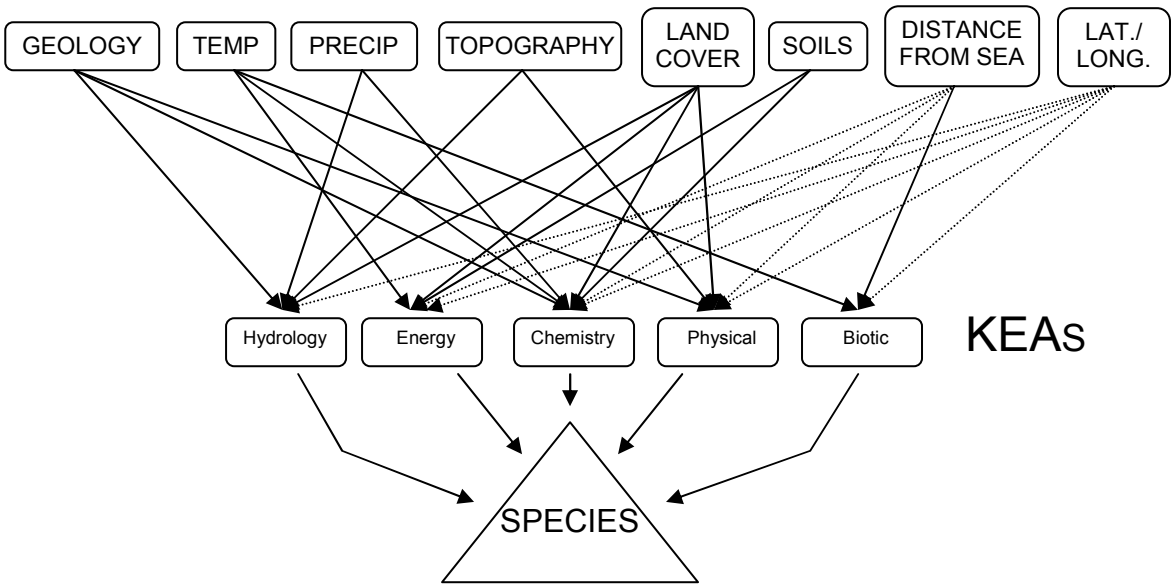


FIGURE 4. Common landscape characteristics known in the general scientific literature to influence KEAs. The variables listed are commonly used in aquatic species distribution models[4]. Dotted lines represent indirect relationships.

will influence a species directly, as in the case of commercial fish harvest. Because human activities can affect the species at multiple layers of the conceptual model, they can be placed at any level in the conceptual diagram (Figure 5).

In the special case of modeling non-indigenous species, human activities directly cause the non-indigenous species to be present in a watershed or at a site. Non-indigenous species are influenced by a similar set of factors as native species, but additional information that specifically pertains to these species can also be important. For instance, time since introduction, the locations of aquaculture facilities or fish stocking centers, the connectivity of habitats, and the level of impairment of local waters can be important because they relate to the success of non-indigenous species. These special considerations can be represented in our diagram as a special group of factors (Figure 5, bottom).

At the culmination of Step 1, you should have a well-supported conceptual model of how species success is influenced by ecological and anthropogenic factors. You should be as explicit as possible in documenting the logic contained in your conceptual model. The factors in the conceptual model represent an idealized variable list that, if quantified and used to develop predictive models, are likely to lead to accurate predictions of your species of interest.

[4] See Joy and Death 2004 for one example of a thorough variable set.

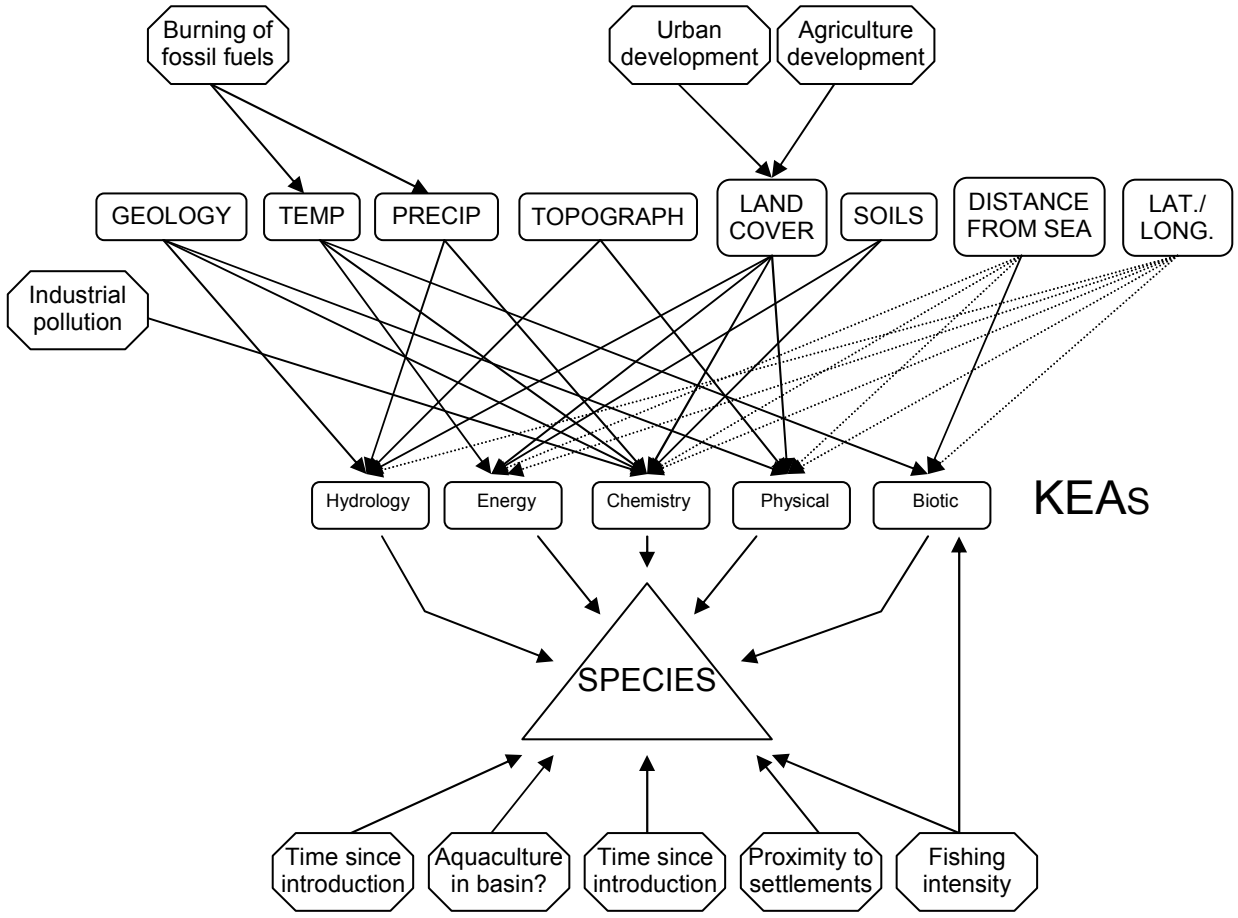


Figure 5. Conceptual diagram showing human considerations (hexagonal boxes) and their influences on landscape factors, KEAs, and on the species directly. The five hexagons at the bottom represent factors that could be considered in a conceptual diagram for a non-indigenous fish species. Dashed lines represent indirect relationships.

Step 2. Determine approach

There are many potential modeling approaches to choose from and the selection of an approach can be very involved[5]. The approach you select has important implications for how you prepare your data and run the models. Here, only one approach is discussed called maximum entropy or MaxEnt. As you will see, this approach is highly appropriate for creating models from limited data, and for use with the most commonly available type of species occurrence data; presence-only records.

Maximum entropy modeling

In any section of river, there is a greater probability that a species will occur in some habitats than others. If you assign a probability of occurrence to each location and add all the possibilities together, this is called a probability distribution. Maximum entropy (MaxEnt) is a mathematical approach that predicts an unknown probability distribution based on the principle that the estimated distribution must agree with everything that is known about its occurrence and be subject to no unfounded constraints. The approach estimates the most uniform distribution (e.g., the distribution with maximum entropy) across a defined area subject to

[5] See Guisan and Zimmermann 2000 and Elith et al. 2006 for academic treatments of this topic.

constraints imposed by information available about environmental conditions at the locations being modeled. In the scientific literature, such models are called "ecological niche models" because the combination of factors (including biotic factors) that limit a species' success in an ecosystem are called that species' ecological niche.

In mathematical terms the MaxEnt distribution maximizes the product of the probabilities of the sample locations, and takes the form:

$$P(x) = \exp(c1 * f1(x) + c2 * f2(x) + c3 * f3(x) \ldots) / Z$$

Here *x* represents a given habitat unit, *P(x)* represents the probability assigned to that unit, *c1*, *c2*, etc. are weights given to the *f*'s (*f1*, *f2*, ...), which are the numeric representations of the KEAs or landscape factors from the conceptual model, and *Z* is a scaling constant that ensures that *P* sums to 1 over all cells in the study area. The mathematical process that is implemented by MaxEnt iteratively adjusts the weights (*c1*, *c2*, ...) associated with each environmental variable to maximize the likelihood that the occurrence data used to train the model are correctly predicted. The weights are adjusted many times (e.g., 500) to cause the model prediction to get closer and closer to the "optimum" probability distribution—the best solution to the problem given the data available. The output of a MaxEnt model is a continuous surface of values ranging between 0 and 100, with higher values indicating a higher suitability of that area for the target species. The MaxEnt process is guaranteed to converge on the single set of values of *c1, c2,...*, and Z, that give the (unique) optimum distribution *P*[6]. In other words there is only one output for each given modeling problem.

Advantages and disadvantages of MaxEnt

There are many advantages to MaxEnt, and several disadvantages[7]. Some of the advantages are that: (1) It requires only presence data together with environmental data for the whole study area. This is extremely important because reliable absence data is exceedingly difficult to find and trust, and this allows you to mix datasets that were collected with different field methods. Most museum data are presence-only data. (2) It can use both categorical and continuous data, and can incorporate interactions between variables. (3) The mathematical process is guaranteed to converge on the optimal probability distribution—it always finds the best solution (and the same solution give the same data). (4) The MaxEnt probability distribution has a concise mathematical definition, and is easier to analyze and understand than some "black box" approaches. (5) The output is continuous, allowing for fine distinctions to be made between different areas and flexibility in choosing thresholds for deciding at what value you will judge the species to be present. Some disadvantages of MaxEnt include: (1) It is a relatively new statistical method in prediction of species distributions, so there are fewer guidelines for its use than some "more mature" statistical techniques. (2) Avoidance of overfitting, while addressed in the mathematical process, still requires more research. Overfitting occurs when the model does such a good job at estimating the data you give it, that it loses the ability to generalize to new or independent data. (3) Because of the math used, it can give very large predicted values for environmental conditions that occur outside of the range of values that you develop the model from.

Several recent studies have compared MaxEnt to other modeling techniques, including another popular presence-only technique (also with free software[8]) called Genetic Algorithm for Ruleset

[6] For a detailed overview of the mathematical process involved in the modeling see Phillips et al. 2004 and 2006.
[7] Phillips et al. 2006
[8] Available at http://www.lifemapper.org/desktopgarp/

Production (GARP). MaxEnt has consistently performed highly in modeling tests, often outperforming GARP[9], and displaying good predictive power even with extremely low species occurrence sample sizes (less than 10)[10]. This makes it a well-suited method for use in data poor areas, or for prediction of rare species.

A final advantage of using MaxEnt to model species distributions is that a software package is available for free on the internet to be run from the desktop of your computer.

Data required to run a MaxEnt model
Inputs to the MaxEnt software include species occurrence points with latitude and longitude information, and raster data layers (data held as a grid of equally sized pixels) of the different environmental attributes hypothesized as important in your conceptual model. Raster data are formatted as an evenly-spaced grid (like a piece of graph paper) with a value in each cell. MaxEnt requires that all layers have the same extent (boundaries), cell size, and coordinate system[11].

Ways to use MaxEnt to model aquatic species distributions
The types of models that you generate can take many forms depending on data availability, the spatial resolution of your input data, which GIS software package you have and how much skill you have using it, etc. What works for one person's purpose may not work for another's, but even very simple models can have utility to decision makers. Here are some options.

Continuous vs. constrained distribution models. Several published aquatic species prediction models project the data across the entire landscape as if the aquatic species were distributed continuously on land, and not actually constrained to lakes, streams and rivers[12] (Figure 6a). This type of presentation is useful for interpretation, but ultimately inaccurate because aquatic species do not live on land. Constraining predictions to water requires more GIS work, but predictions are more accurate, though potentially more difficult to present (Figure 6b). Depending on the resolution you select for your environmental data, constrained analysis may not be appropriate, because rivers are best represented as small grid cells in raster space. Representing a river line with pixel sizes of 500 m on a side would be clumsy, whereas representing them with 30 m pixels is realistic and linear. Making the latter selection has implications for the amount of memory and processing time it takes to create raster grids with many cells (e.g., more information to store), but you end up with higher resolution maps.

Local vs. watershed variables. As mentioned in the introduction, the habitat at given point in a river system is influenced not only by the conditions locally, but by the integrated conditions of the watershed upstream of a habitat. For example, geology is a landscape variable that has a great influence on many KEAs (e.g., substrates and water chemistry) and is an influence that is transmitted downstream. Thus, instead of representing geology as a variable with only local influence, it is more logical to represent geology as a proportion of the watershed upstream of a point. On the other hand, variables such as elevation may be best represented as local scale variables (e.g., elevation of the stream at each point) rather than as a watershed variable. Creating watershed variables adds to the GIS processing time necessary to produce your predictor datasets, but it adds realism to your models.

[9] Phillips et al. 2004, Hernandez et al. 2006, Phillips et al. 2006, Pearson et al. 2007.
[10] Hernandez et al. 2006, Pearson et al. 2007
[11] See http://training.esri.com/acb2000/showdetl.cfm?DID=6&Product_ID=826 for a free online tutorial about coordinate systems.
[12] Drake and Bossenbroek 2004, Iguchi et al. 2004, Dominguez-Dominguez et al. 2006

High vs. low resolution predictions. The resolution of your modeling effort relates directly to the size of the cells (pixels) within your raster grid and to how many cells the grid contains. The

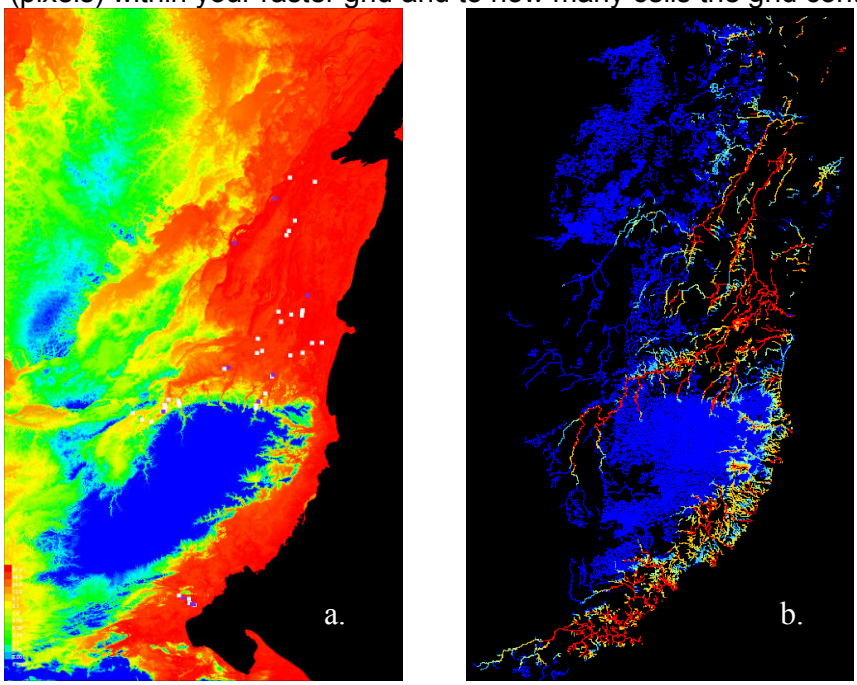


Figure 6. (a.) A simple model of African tilapia (*Oreochromis niloticus*) in Belize developed using only one variable (30 m resolution elevation; resampled from 60 m). Warmer colors (yellow, orange, red) represent increasingly greater habitat suitability for tilapias. Even though this is a model of a fish that lives only in rivers and wetlands, the model made a prediction for every possible location in the study area. This type of model is easy to implement (less steps in GIS), and valuable for display. (b.) A second model of tilapia that includes 20 input variables with a prediction constrained to streams and rivers only. This type of map, though not as visually striking as the continuous prediction, is still easy to interpret and far more accurate in that the species is predicted only in its true habitat. However, to accurately represent the locations of river, a small grid size is necessary, meaning more memory required and processing time for GIS applications, but greater ability to distinguish fine scale patterns.

smaller the cells, the higher the resolution, and also the higher the file sizes you will have to accommodate. A file with identical boundaries but a smaller cells size will require more memory, because it contains more information over the same area. The implications of this are (1) more free space will be necessary on your hard drive to store the data layers; (2) more processing time will be required for the computer to generate outputs; and (3) several extra steps may be necessary to feed the data into the MaxEnt software (see “Samples with data” in Step 4 below). So by working at a higher resolution, you gain the ability to make fine distinctions between habitats, and you gain the ability to realistically produce constrained models, but the process becomes more cumbersome and involved.

In the next step, the process necessary to prepare datasets for each of the above possibilities (continuous, constrained, local, watershed) are described in detail.

Step 3. Data Preparation

In this step, you will select and prepare datasets to use in the modeling process. It is here that you must face the realities of data limitation and make decisions about which of the available environmental data sets best satisfy your list of predictor variables. It is also here that you must locate and prepare your species occurrence records. The environmental and species data will be combined in the Next step to allow you to generate your models.

Figure 7 shows a workflow diagram of the entire modeling process. Each of these tasks are described in detail below. The main challenges of this step are to: (1) match the KEA and landscape variables from the conceptual model to existing or easy-to-create spatial datasets; (2) convert all acquired data to a consistent format (in GIS); and (3) locate, evaluate, and format species occurrence data.

Words of advice before starting

It will be necessary that you have a good computer that has a fact processor, GIS software, and ample hard drive space (between 10 – 100 Gb depending on the resolution you choose and the boundaries of your study area). There are several GIS packages on the market: the most common are probably ESRI's ArcView 3, ArcGIS 8, and ArcGIS 9 packages. An additional GIS package is available online for free download called DIVA-GIS[13]. DIVA-GIS is more limited than ArcGIS 9, but still has ample functionality to develop some of the model types presented here. Here procedures are presented for implementation in ArcGIS 9. This part of the modeling process is best implemented by individuals with some prior GIS experience. The beginner GIS user who wishes to undertake the process is encouraged to take a GIS tutorial to learn about basic GIS concepts and operations, and to establish contact with an experienced local GIS technician who can offer assistance through the process.

When handling many data sets in different projections and from different sources, it is easy to get confused, so it is very important that you stay organized throughout the process. Keep notes about what you do during any given work day. In particular, keep track of the names and important information of new files and folders that you create as you work, including the coordinate systems that different data are in, and the name of the source data that were used to create new data sets. It is also a good idea to come up with naming conventions to indicate the types of files you create. For example 2- or 3-letter suffixes at the end of your file name can describe the coordinate system a file is in (e.g., '[filename]_laz' to denote that the file is in Lambert Azimithal Equal Area projection).

Finally, because you want to end up with uniform extent and cell sizes for every dataset that you generate, it is a good idea to create or designate a raster "Reference Grid" that has the exact extent and cell size that you desire. Throughout the modeling process you will repeatedly reference this dataset to define the extent and grid size of the new raster datasets that you create. If you are using the IABIN-DGF hydrological derivatives, a good choice for a Reference Grid might be the Flow Direction grid.

Identify data sources

Your conceptual model from Step 1 represents your hypotheses about which factors influence the success, and therefore the distribution of your species. The variables defined in the conceptual model—KEAs, landscape factors, and human activities—represent the ideal list of variables that you will want to consider to create your species distribution model. Your task is to match this list of ideal variables with geographic datasets (GIS data) for your study area.

[13] http://www.diva-gis.org/

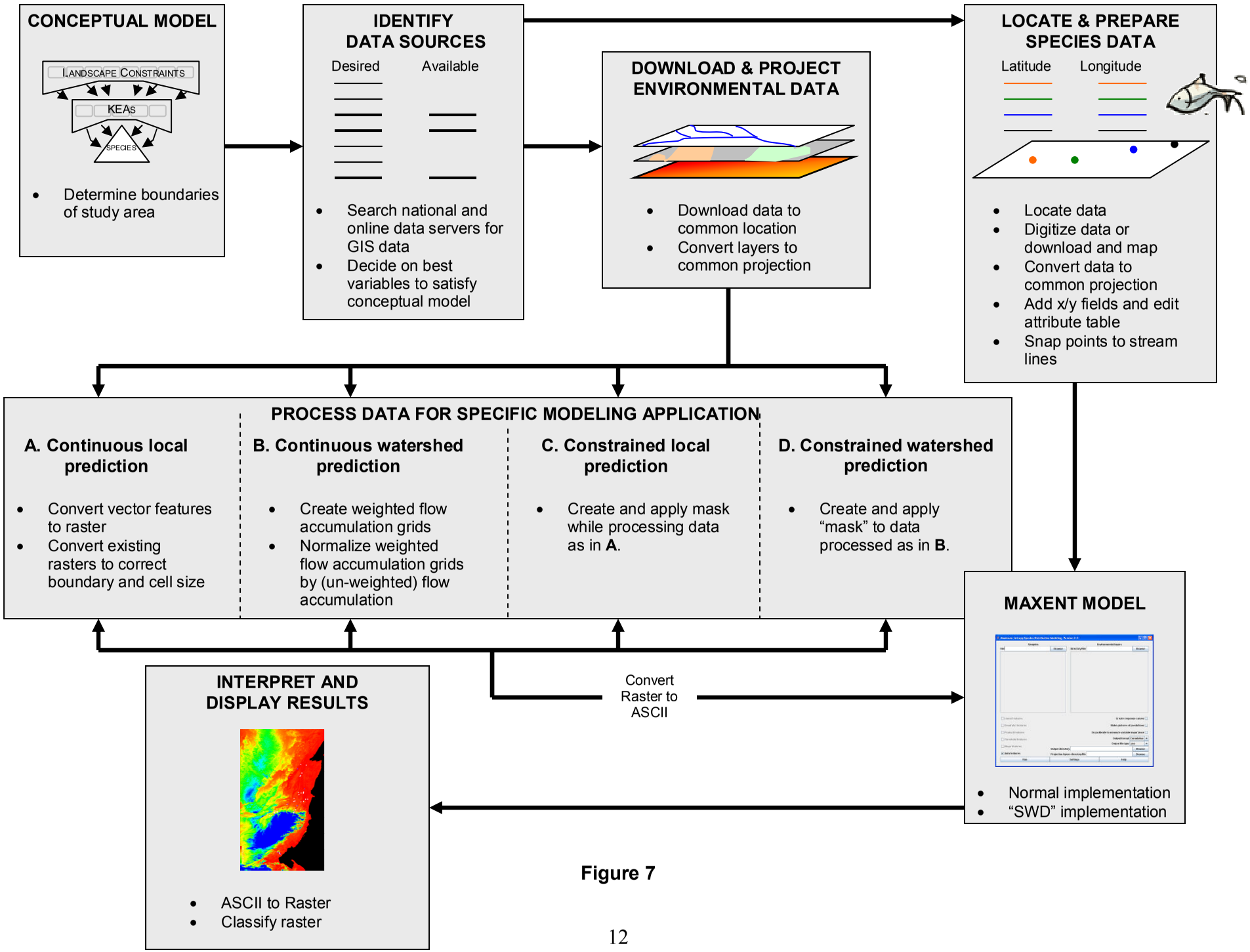

CONCEPTUAL MODEL
LANDSCAPE CONSTRAINTS
KEAs
SPECIES
Determine boundaries of study area
IDENTIFY DATA SOURCES
Desired
Available
Search national and online data servers for GIS data
Decide on best variables to satisfy conceptual model
DOWNLOAD & PROJECT ENVIRONMENTAL DATA
Download data to common location
Convert layers to common projection
LOCATE & PREPARE SPECIES DATA
Latitude
Longitude
Locate data
Digitize data or download and map
Convert data to common projection
Add x/y fields and edit attribute table
Snap points to stream lines
PROCESS DATA FOR SPECIFIC MODELING APPLICATION
A. Continuous local prediction
Convert vector features to raster
Convert existing rasters to correct boundary and cell size
B. Continuous watershed prediction
Create weighted flow accumulation grids
Normalize weighted flow accumulation grids by (un-weighted) flow accumulation
C. Constrained local prediction
Create and apply mask while processing data as in A.
D. Constrained watershed prediction
Create and apply "mask" to data processed as in B.
MAXENT MODEL
Normal implementation
"SWD" implementation
Convert Raster to ASCII
INTERPRET AND DISPLAY RESULTS
ASCII to Raster
Classify raster

Figure 7

There are many potential sources of geospatial information that you can draw on to assemble the necessary datasets. For instance, most natural resource agencies within national governments have GIS technicians on staff and centralized databases, and many universities are good sources of data. Many online resources can also help you find available GIS data. For modeling aquatic species in Mesoamerica, the IABIN-DGF hydrologic derivatives are a great data source with many of the important data layers you will want in your models. Another good resource is Mesostor—a free online data clearinghouse housed by the Mesoamerican Regional Visualization and Monitoring System (SERVIR). These and other sources are publicly available at the web sites listed in Table 1.

In some cases you will find data that directly represent the factors you want to include in your model; for instance, many places on the planet have GIS layers for geology. In other cases you may find that you must rely on "proxy variables"—variables known to correlate to your variable of interest. For example terrestrial vegetation type is often a good proxy for ecosystem type,

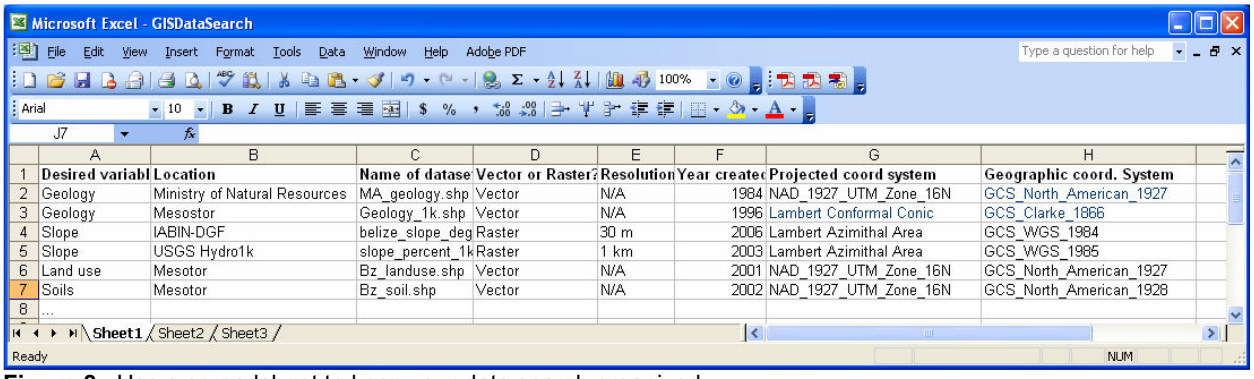
Table 1. Online sources of free geospatial data for potential use in species modeling.

Source (Region)	Worldwide Web link	Available Data
Geographic information		
Mesostor (Mesoamerica)	http://servir.nsstc.nasa.gov/MesoStor/	Aster, Landsat, photographic images; topographic, socioeconomic information, ecological, watersheds, rivers, political boundaries, protected areas
IABIN-DGF hydrologic derivatives (Central America and Yucatan, Mexico)	http://edcintl.cr.usgs.gov/iabin_datadownload.html	Hill shade, slope, watersheds, aspect, streams, flow accumulation, flow direction for all countries in Central America and part of Mexico; 30 m resolution
USGS Hydro1k Elevation derivatives (Global)	http://edc.usgs.gov/products/elevation/gtopo30/hydro/index.html	Elevation, slope, aspect, flow direction, flow accumulation, drainage basins, streams
Global Land Cover Facility (Global)	http://glcf.umiacs.umd.edu/index.shtml	Vegetation, land cover, elevation, protected areas
DIVA-GIS	http://www.diva-gis.org/Data.htm	Altitude, land cover, population density, boundaries, climate data, elevation, satellite data
Species data		
Global Biodiversity Information Facility (Global)	http://www.gbif.org/	Biodiversity data; many taxa (point locality data)
FishNet2 (Global)	http://www.fishnet2.net/index.html	Fish specimen collections from Museums (point locality data)

Some things that you should consider for any data that you locate are:

- *Do the data match the spatial resolution that you are interested in?* Data don't have to have the exact resolution that you want to do your analysis in, but many data are available at 1 km^2, which may be too coarse for the analysis you want to do. While it is true that 1 km data can be resampled to any size, you should make efforts to identify data that are close to the resolution you want. It is always better to resample data from a higher resolution (e.g., 30 m) to a lower resolution (e.g., 1 km), than vice versa.
- *Are the data raster or vector?* You will be able to use both types of data, but you will have to know which data are vector, so that you can convert these to raster data sources (the necessary format for use MaxEnt).
- *What is the coordinate system of the data?* You should know the names of the "projected coordinate system" and the "geographic coordinate system" in order to reproject data to the coordinate system of your choice.
- *Do the data match the time frame during which the specimens were collected?* Predicting species collected in 1932 with land use data from 2005 is obviously not appropriate. Some data types (e.g., geology) change over very long time scales and can be used with data from any year. Other data types (e.g., current land cover) may have a strong human footprint and therefore may be inappropriate for your purpose if all you have are older specimen data.

As you research potential data sources, be sure to keep track of what data sources are available, the details of each data set, and the location of the data. A spreadsheet may be a good way of keeping track of your data (Figure 8).



	A	B	C	D	E	F	G	H
1	Desired variabl	Location	Name of datase	Vector or Raster?	Resolution	Year create	Projected coord system	Geographic coord. System
2	Geology	Ministry of Natural Resources	MA_geology.shp	Vector	N/A	1984	NAD_1927_UTM_Zone_16N	GCS_North_American_1927
3	Geology	Mesostor	Geology_1k.shp	Vector	N/A	1996	Lambert Conformal Conic	GCS_Clarke_1866
4	Slope	IABIN-DGF	belize_slope_deg	Raster	30 m	2006	Lambert Azimithal Area	GCS_WGS_1984
5	Slope	USGS Hydro1k	slope_percent_1k	Raster	1 km	2003	Lambert Azimithal Area	GCS_WGS_1985
6	Land use	Mesotor	Bz_landuse.shp	Vector	N/A	2001	NAD_1927_UTM_Zone_16N	GCS_North_American_1927
7	Soils	Mesotor	Bz_soil.shp	Vector	N/A	2002	NAD_1927_UTM_Zone_16N	GCS_North_American_1928
8	...							

Figure 8. Use a spreadsheet to keep your data search organized.

Download and project environmental data layers

Once you have identified the best data layers to satisfy your variable list, you should establish working directories on your computer and begin consolidating the data in one place. This will be the location on your hard drive from which you do all of your data preparation.

When your environmental data are in one place, you must make sure all of your data share a common coordinate system. There are two types of coordinate systems: geographic and projected. A geographic coordinate system is used to locate objects on the curved surface of the earth. A projected coordinate system is used to locate objects on a flat surface—a paper

map or a digital GIS map displayed on a flat computer screen[14]. Each projection type attempts to model the earth and feature locations accurately. A map projection is used to convert data from a geographic coordinate system to a projected (planar) coordinate system. Just like there are many geographic coordinate systems, there are many different map projections as well—each preserves the spatial properties of data (shape, area, distance, and direction) differently.

Selecting a coordinate system can be tricky because there are many to choose from. Most countries have a common coordinate system that national data are held in. For studies at the national level, it may be a good idea to choose a coordinate system that matches the local standard so that data can be easily assimilated with other data sets. Consult a local GIS expert for advice on what coordinate system you should choose to do the modeling for your specific application.

GIS: What coordinate system is your data in?

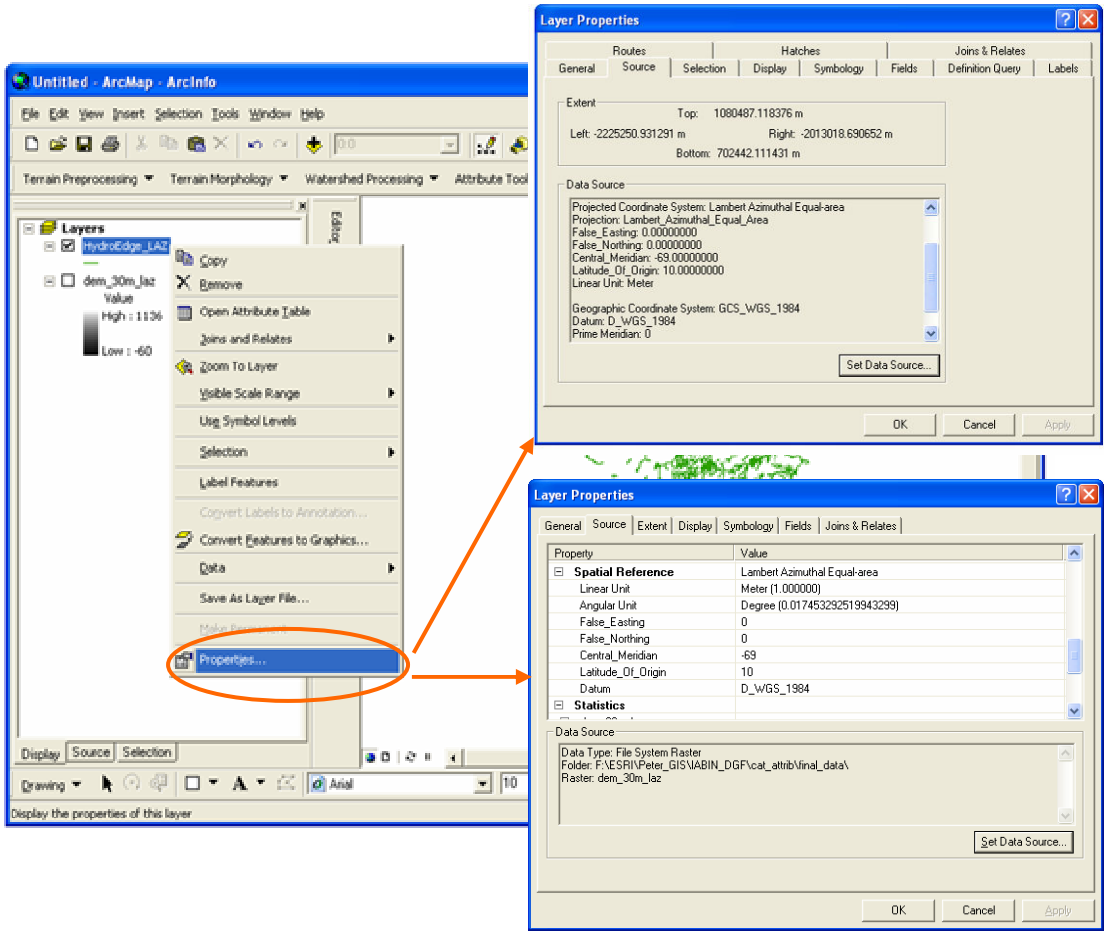
In ArcMap, load your data using the "Add data" button () on the Standard toolbar. When your data are loaded, right click on the name of your data set, select Properties, and the Layer Properties box will open. Select the Source tab in the Layers Properties box to display the Coordinate system information in the Data source window. Vector data will appear differently from raster data. For vector data scroll down until you see the Projected Coordinate System and the Geographic Coordinate System listed. For raster data scroll until you see Spatial Reference information. The coordinate system is listed as the Spatial reference.

If you see the words "<Undefined>" in these locations for either vector or raster data then you will need to define the coordinate system for the data layer before you can convert to a new one. If a coordinate system already exists and you want to change it to another, you will need to re-project the data.

GIS: Defining a coordinate system

If a dataset has no projection defined, you will need to figure out what the intended projection was for the data. You can accomplish this by either looking at the metadata, looking at the website where you got the information, or by comparing the data to other datasets with known projection information. To define a coordinate system for the data, you will need to use the ArcToolbox in either ArcMap or ArcCatalog. Clicking on the ArcToolbox button () will open the ArcToolbox menu. In ArcToolbox, select Data Management Tools→Projections and Transformations, and double left-click on Define Projection. The Define Projection window will open. For a Vector dataset, you will want to use the select file button () to the right of the "Input Dataset or Feature Class" box to select the dataset you want to define a projection for (you can also drag and drop the dataset into the Input box). Next, open the Spatial Reference Properties window by clicking on the button to the right of the Coordinate system box (). Within the Spatial Reference Properties box you have several options for selecting the desired Coordinate system: (1) you can click the "Select" button, which will open the Browse for Coordinate Systems window, from which you should select the appropriate system; (2) you can click the "Import…" button, which will allow you to specify the coordinate system from a data set with the coordinate system you want to use already defined; or (3) you can click the "New…" button which will open a window that allows you to define the parameters of

[14] For more information on managing projections in ArcGIS readers are directed to a free (English language) tutorial offered by ESRI: "Working with Map Projections and Coordinate Systems in ArcGIS" at http://training.esri.com/acb2000/showdetl.cfm?DID=6&Product_ID=826

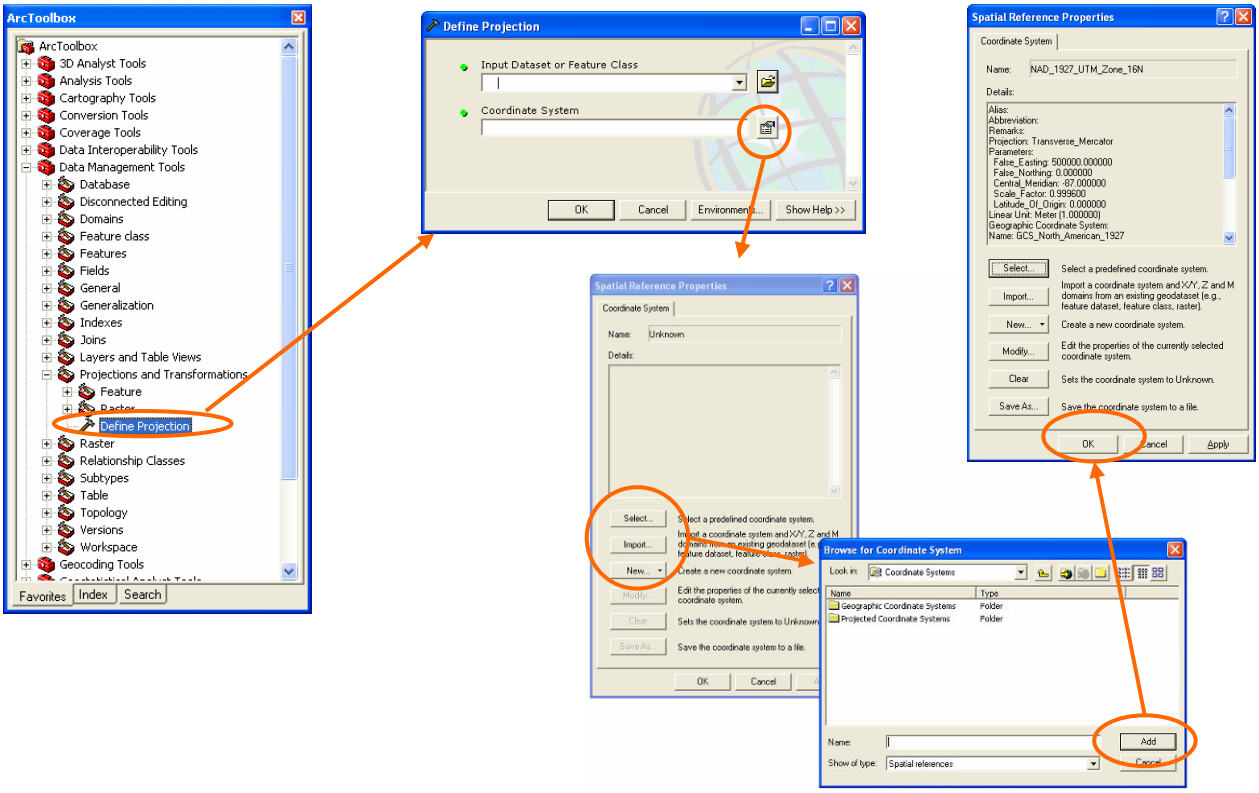


a new projection (use this only as a last resort). Select the most appropriate option for your situation consulting a local GIS expert if necessary. Click OK to leave the Spatial Reference Properties Window, and OK on the Define Projection Window, after which the projection should be defined successfully.

<u>GIS: Re-projecting data to a new coordinate system</u>

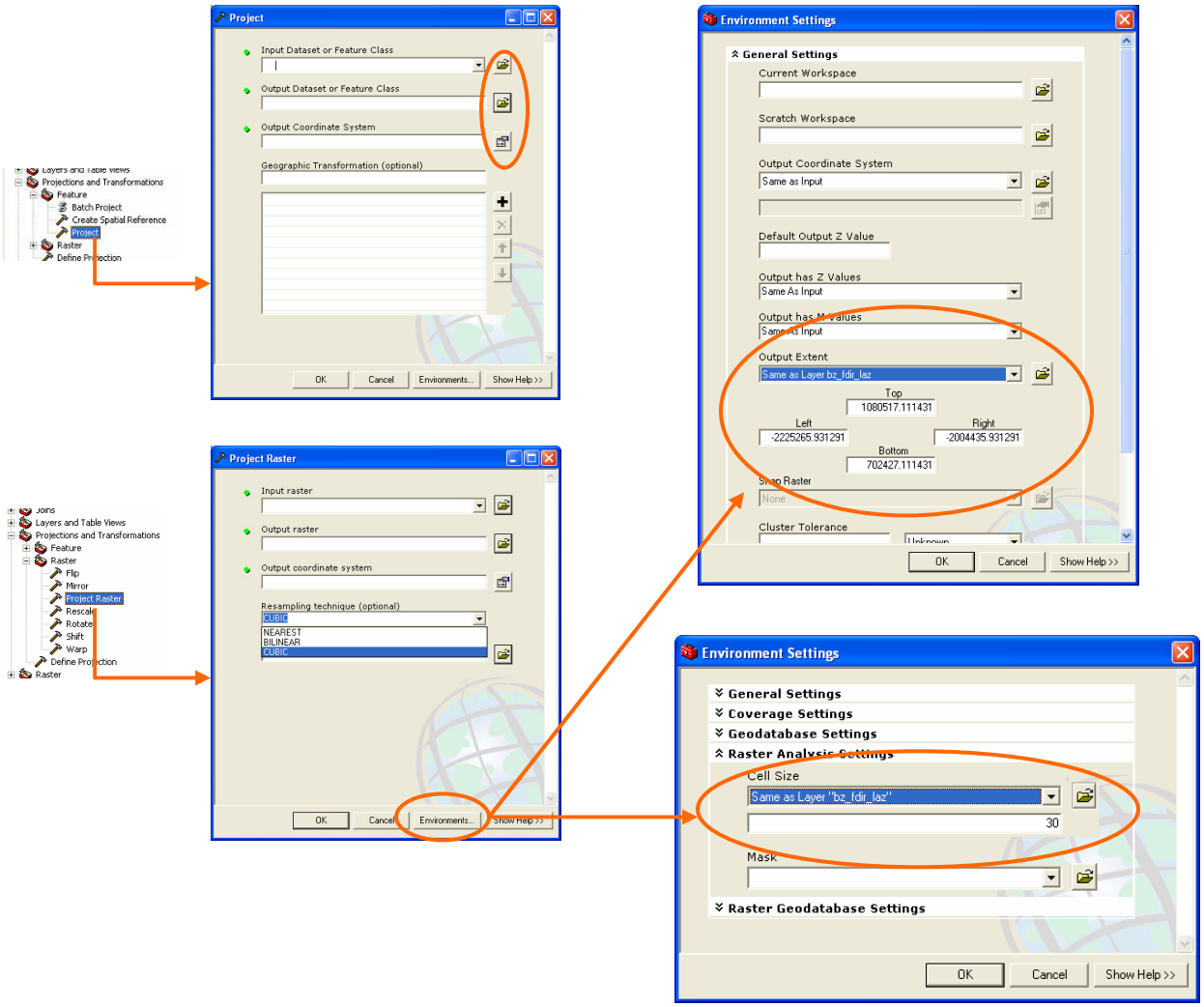
If your acquired data already have a projection, but are not in the desired coordinate system, you need to re-project your data using the ArcToolbox. Go to ArcToolbox→ Data Management Tools → Projections and Transformations, and then either to the Feature or Raster submenus. Use the Feature submenu for vector data (lines, points, or polygons), or the Raster selection for data in grid format.

For vector data, select the "Project" tool under the Features menu. When the new window opens, select the appropriate Input Dataset, then define the location and name for the new Output Dataset using the folder button ()(see diagram on following page). Next, select the Output Coordinate System. Finally, you may have the choice to select a Geographic Transformation. Each combination of projections will have different options for Geographic Transformations, so you may want to consult a local expert for advice on making a good selection if the correct choice is not obvious to you.



Reprojecting raster data involves creating a new grid and assigning values to each pixel in that grid (called “resampling”). There are three possible mathematical approaches to resampling raster data. It is recommended that you use the Cubic resampling option under the “Resampling technique” pulldown menu (see diagram on following page).

Any time you resample raster data, you have the opportunity to redefine the fundamental properties of the output raster grid. Of special relevance are (a) the boundaries of the grid, and (b) the cell size. To adjust the boundaries and the cell size you must adjust the “Environments” for the resampling process by pressing the Environments button. Doing so will open the Environmental Settings window. There are many settings that can be changed here, but the one you are interested in right now is “Output Extent” under the General Settings options. Constrain the extent of your resampled raster to match the Reference Grid, using the folder button ( ) next to the Output Extent dropdown menu. You will see the top, bottom, left, and right values automatically populated to match the reference data when you select the file. Now, when you reproject the data, the resulting grid will have the same exact boundaries as your Reference Grid. By a similar process you should expand the Raster Analysis Settings options in the Environmental Settings window and select your Reference Grid to populate the Cell Size dropdown menu. The cell size will be automatically set to equal your Reference Grid. Now you have adjusted the Environments and you can click OK to return to the Project Raster window, where you should hit OK again after you ensure that the Input Raster (the dataset you will change) and Output Raster (the name and location of the new dataset) fields have been properly set. ArcGIS will run the reprojection process and tell you when it was successful.



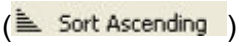
Data processing for specific modeling applications
As described in Step 2 above (page 9), there are at least 4 types of models that you can produce: (A) continuous local prediction, (B) continuous watershed prediction, (C) constrained local prediction, and (D) constrained watershed prediction. Each of these options requires slightly different GIS preparation for the environmental data layers that you will feed into your model.

A. Continuous local prediction. A continuous local prediction model estimates the suitability of each pixel for the species regardless of whether that pixel is on land or water, and uses only data representing the local environment (vs. combined watershed conditions). Continuous local prediction models are the easiest models to implement because they require the least amount of data processing. To prepare these models you must:

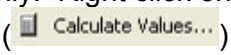
1. Convert all data to raster grids with the *exact same* extent and cell size
2. Save each grid to the format required to be imported by MaxEnt (ASCII file)

GIS: Preparing vector feature classes for conversion to raster

Any vector datasets that you want to use must be converted into raster datasets to use them with MaxEnt. Because each individual raster dataset can only represent the values of one variable, each vector attribute that you want to use must be converted to an individual raster dataset, and each of these fields must be made numeric before making the conversion. Thus, categorical variables with text names (e.g., igneous rock, sedimentary rock, etc.) must be modified so that each category has a numeric representation (e.g., 1, 2, …). This requires manipulation of the attribute table of the vector dataset to create the numeric values for any text classes, and then conversion of the vector feature to raster data.

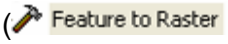
To create the numeric codes for the text categories in your vector dataset, you should load the dataset into ArcMap, and then open the Attribute table of the dataset by right clicking on it and selecting "Open Attribute Table". You must create a new "short integer" field by left clicking the Options pulldown menu (bottom right side of attribute table, see graphic on next page). Give the new field a meaningful name with 10 or less characters. Next, right click the field with the text categories that you want to reclassify, and select Sort Ascending (

) to reorder the classes in alphabetical order. Manually select the first class by clicking and dragging along the far left side of the attribute table, or by right clicking the column and choosing "Select by Attributes". Once you have selected all of the features in the first class, limit the view to "Show: Selected" (

). This limits the entries in the attribute table to the selected class only. Right click on the newly created short integer field and select Calculate Values (

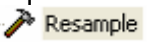
). The Field Calculator will open, which allows you to populate each selected column with a number. Type a unique integer into the text box where it says "[Column name]=" and press the OK button. That number will be placed in each row representing the text class variable. Be sure to make note of what class is represented by what number. Repeat this process for each class that you want to represent in the raster layer, and then move on to the next step.

GIS: "Rasterizing" vector feature classes

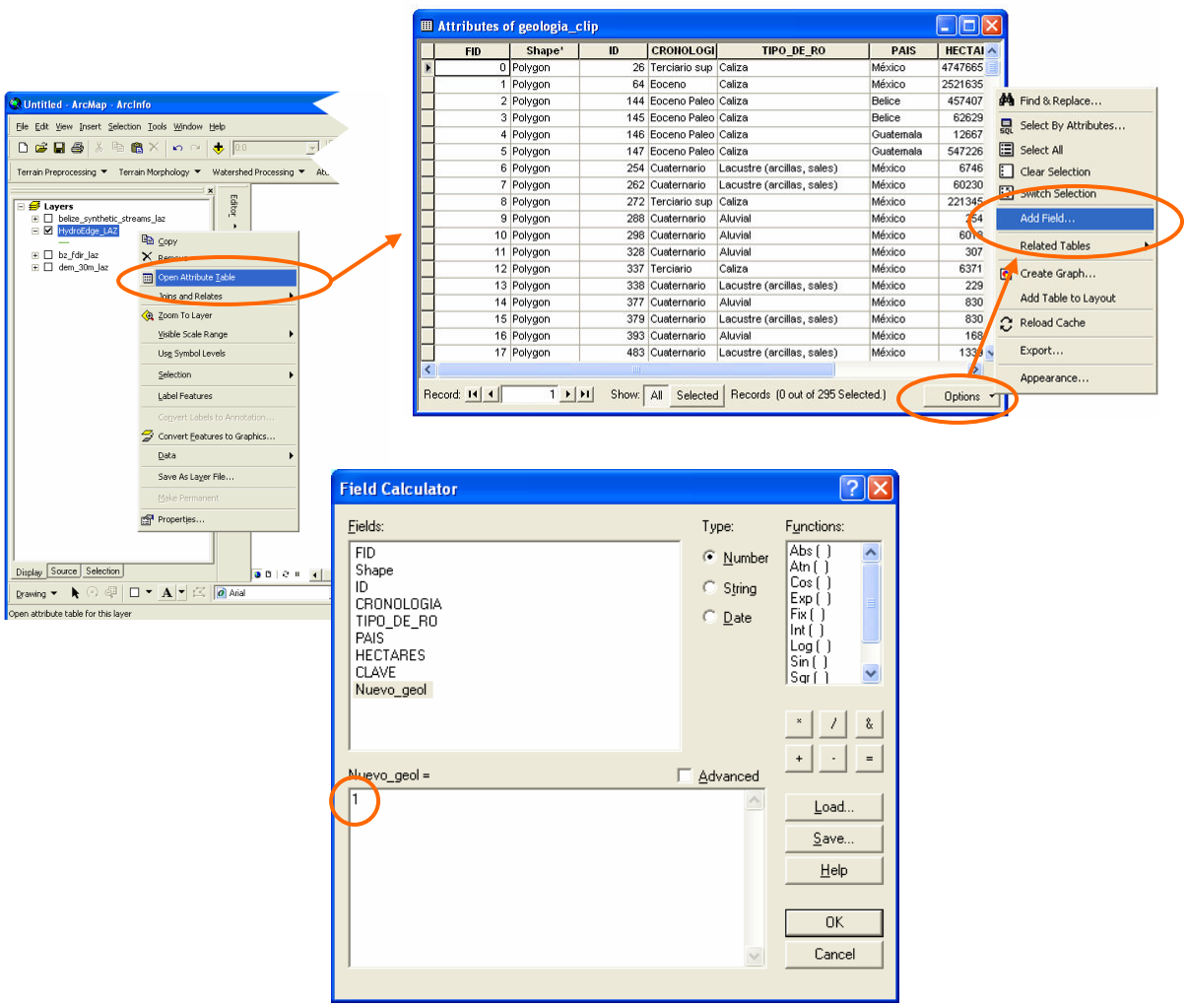
In the ArcToolbox, select Conversion Tools→To Raster and double click the Feature to Raster tool (

). This will bring up the Feature to Raster window. Select the vector layer that you want to convert in the "Input features" text box, then select the newly created field with numeric classes from the "Field" dropdown menu. Next, name the raster and determine its location in "Output raster", and finally set the "Environments" exactly as described above on page 17 (e.g., set the Extent and Grid Cell Size to equal your Reference Grid). Then click OK in the "Environments" window and OK in the Feature to Raster window. The feature will now be converted into a raster dataset with a numeric value for each category.

GIS: Resampling grid extent and cell size

You must adjust the size of the cells and the extent of all the grids that have not been made equivalent to your Reference Grid before you can use MaxEnt. Use the Resample tool (

) under Data Management Tools → Raster to do this. Put the grid that you want to adjust in the Input raster text box, define the Output raster name and location, select "Cubic" as your Resampling technique, and then set the Environments to equal your Reference Grid. Click OK several time until the process runs.



GIS: Convert grids to ASCII (.asc) format

The final step to prepare your environmental data for continuous local prediction is to convert all of your grids to the format required for use by the MaxEnt software. This format is a text file called an ASCII file with the file extension ".asc". By now all of your grids should have exactly the same extent and grid size. Before you convert the data from grid to ASCII, you should create a new folder on your hard drive to store the ASCII files in. It will save you time later to have all of the ASCII files in one place. To convert the files, go to ArcToolbox→Conversion Tools→From Raster, and select the Raster to ASCII tool (Raster to ASCII). In the Raster to ASCII window, put the grid dataset in the Input raster text box, and then click the File button ( ) next to the Output ASCII raster file text box. Find the new folder you created to receive your ASCII files, type an appropriate file name, and be sure to select "File (.ASC)" from the "Save as type:" pulldown menu. Click Save to leave the Save As window then click OK in the Raster to ASCII window to begin the conversion process.

Once you have converted all of your raster grids to ASCII files, you can proceed to Step 4: Modeling.

B. Continuous watershed prediction. A continuous watershed prediction model estimates the suitability of each pixel in the study area for the species, regardless of whether that pixel is on land or water, using environmental data representing the cumulative watershed influence on each pixel. Local variables can also be included into these models. To prepare these models you must:

1. Convert all data to raster grids with the *same exact* extent and cell size
2. Calculate flow accumulation for watershed variables
3. Normalize watershed variables (for proportion and average data)
4. Save each grid to the format required for use by MaxEnt (ASCII text file)

Flow accumulation is a process whereby a flow direction grid derived from a digital elevation model is used to calculate the number of pixels upstream of each pixel in the directionality grid (Figure 9). One of the most useful outputs of the IABIN-DGF project is a flow direction grid that was derived from a restricted 30 m digital elevation model (DEM) that USGS could access to create hydrologic derivatives. The flow direction grid gives you the ability to make high resolution calculations about watershed influences. **Weighted flow accumulation** is a variation of the flow accumulation process that uses a flow direction grid, but also uses a "weight grid" to calculate the values that flow into the next downstream cell. The values in all of the upstream weight grid cells are added together to calculate the value in each cell of a weighted flow accumulation grid (Figure 9). In this way, the state of different variables in the upstream watershed can be calculated and incorporated into your species distribution models.

There are several ways to use weighted flow accumulation. One way is to calculate how many cells above a given pixel in the watershed contain some class state, such as a type of geology. The weight grid to use for this purpose is a binary grid (1 or 0 values only) that has a 1 in cells representing locations with the class state present, and a 0 where it is absent. The weighted flow accumulation grid will calculate how many cells with a 1 flow into each pixel. This in turn can then be divided ("normalized") by an unweighted flow accumulation grid that represents only the number of pixels upstream of every pixel. Normalizing a weighted flow accumulation grid by the unweighted grid results in a grid with proportions (between 0-1) of the watershed in a given class state. Another way to use weighted flow accumulation is to calculate the sum total value (of elevation, precipitation, temperature, etc.) of all cells above each cell, which can also be divided by a normal flow accumulation grid to yield the average value (rather than the proportion) of the variable in the watershed above each pixel. Using weighted flow accumulation and normalizing it gives you the ability to represent proportional and average values for characteristics in the watershed above each pixel.

GIS: Preparing vector feature classes for conversion to raster

Any vector datasets that you want to use must be converted to raster datasets to be used with MaxEnt. Because each individual raster dataset can only represent the values of one variable, each vector attribute that you want to use must be converted to an individual raster dataset, and each of these fields must be made numeric before making the conversion. Unlike the previous model type where we classified each text category into a number and saved these into the same raster dataset (see "Preparing vector feature classes for conversion to raster" on page 19), we must now create a new binary field (with 1s and 0s only) for each class in our vector data sets and save each of these as an individual raster. Thus, where in the continuous local models we only had one raster containing all classes, here each binary class will have its own raster.

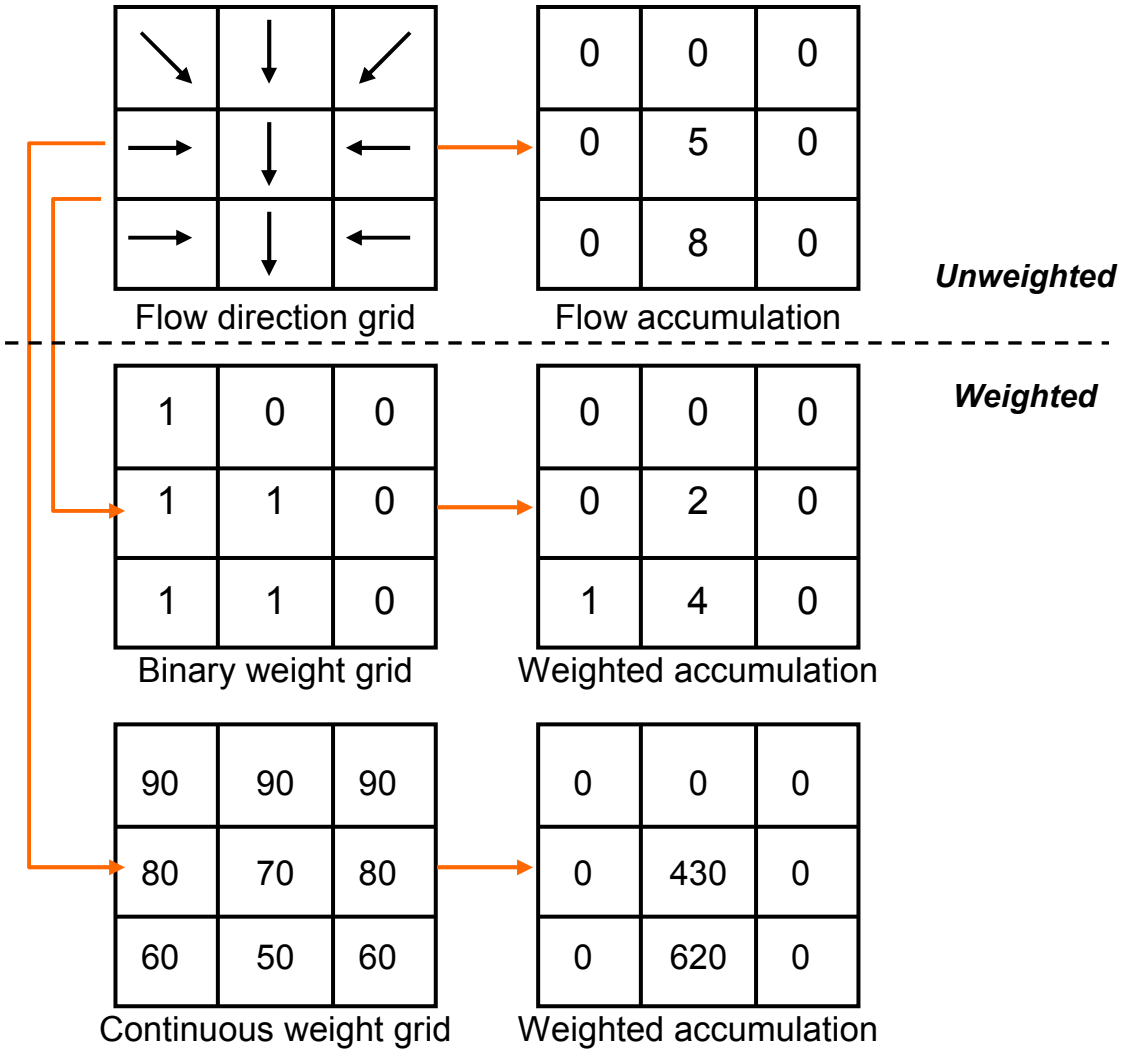


Figure 9. Flow accumulation is a process whereby a flow direction grid (upper left) is used to calculate the number of pixels upstream of each pixel (top right). Weighted flow accumulation also draws on a flow direction grid, but uses a weight grid to calculate the sum of the values that flow into each cell.

To create a binary code for each text class for your vector dataset, you should load the dataset into ArcMap, and then open the Attribute table of the dataset by right clicking and selecting "Open Attribute Table". You must create new short integer fields for *each* class in the dataset by left clicking the Options pull down menu (bottom right side of attribute table) and select "Add field". Each new field should be given a meaningful name with 10 or less characters. Next, right click the field with the class data in it and select Sort Ascending (Sort Ascending) to reorder the classes in alphabetical order. Manually select all the rows in the first class by clicking and dragging along the far left side of the attribute table, or by right clicking and choosing "Select by Attributes". Once you have selected all of the features in the first class, limit the view to "Show: Selected" (button at bottom of attribute table; Show: All Selected). This limits the entries in the attribute table to one class only. Now right click on the newly created short integer field that represents the class that you selected, and right click and choose Calculate Values (Calculate Values...). The Field Calculator will open, which allows you to populate the

selected rows with a number. Type the number 1 into the "[Column name]=" text box and press OK. Each row will be classified as present (1), and the other (unselected) rows will be designated 0 (absent) by default. Be sure to make note of what class is represented with what field name. Repeat this process for each class that you want to represent in a raster layer, and then move on.

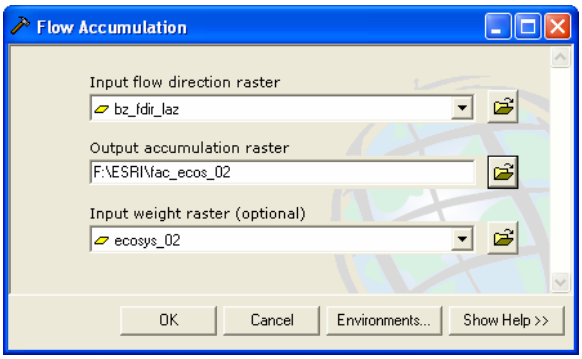
GIS: "Rasterizing" vector feature classes

Follow the instructions by the same title on page 19 to convert each of your binary (1/0) fields to a raster dataset representing the pixels where that class (e.g., geology, soil, etc.) is present. Be sure to set the extent and cell size of the raster in the Environments window to match your Reference Grid.

GIS: Flow accumulation and weighted flow accumulation

To calculate flow accumulation in ArcGIS 9, you will need a flow direction grid[15], and the spatial analyst extension for ArcGIS 9. Make sure that the Spatial Analyst extension is installed and turned on. To do this, go to the Tools dropdown menu in the ArcMap Main menu at the top of the page, and select the "Extensions" option. If you have Spatial Analyst it will show up in the list of extensions in the Extensions window that opens,. Place a check box next to Spatial Analyst in the extensions list and hit the Close button. Then right click in the empty part of the Main Menu toolbar (to the right of Help); this opens a popup window that should have Spatial Analyst listed and checked. If it is not checked activate it by clicking on it.

Once you have done this, you are ready to use the flow accumulation tool (Flow Accumulation), which is located in ArcToolbox→Spatial Analyst Tools→Hydrology. Double click the Flow Accumulation tool and the Flow Accumulation window will open. Load your flow direction grid into the "Input flow direction raster" text box, assign a name and location to the output flow accumulation raster, and then select a binary or continuous raster file for the "Input weight raster" (your weight grid). Be sure to set the Environments so that your output extent and cell size match your Reference Grid. Clicking OK will initiate the flow accumulation process. After doing this for each dataset that you want to use in your watershed analysis, you are ready for your next step.



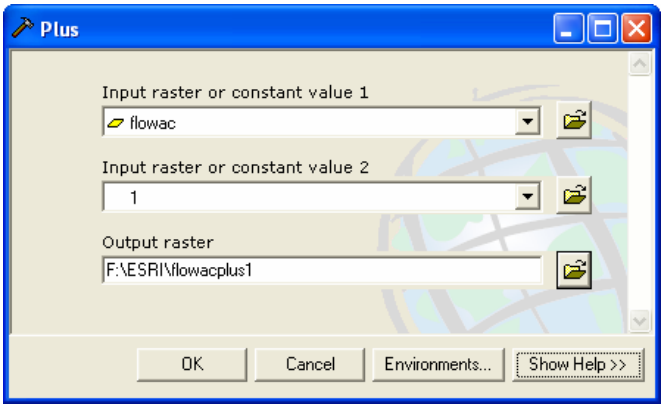
GIS: Normalize watershed variables

Normalizing a variable means that you divide it by a common denominator used across datasets so that your data become scaled or averaged and can be compared easily to one another. For instance, geology is best described as the proportion of the upstream watershed in a certain geology, because it allows you to compare the *relative* influence of that geology on a certain point in your landscape. This becomes obvious when you think about comparing two watersheds, one very large and one very small. If 100% of the small watershed and only 10% of the big watershed are occupied by geology type A, but the absolute area covered is greater in the larger watershed, then the weighted flow

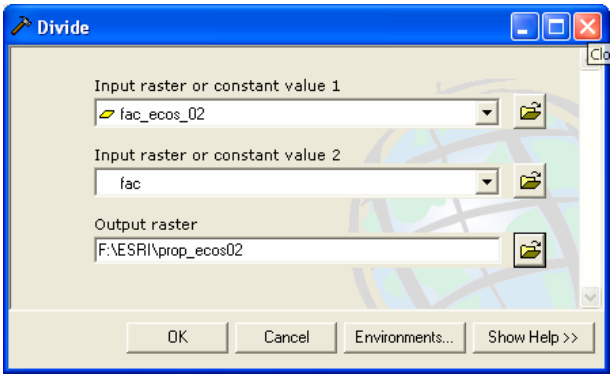
[15] Available at the IABIN-DGF website for all countries in Central America: http://edcintl.cr.usgs.gov/iabin_datadownload.html

accumulation will show a greater value for that geology in the large watershed, even though it is relatively less important there than in the small watershed. By normalizing the weighted flow accumulation by the unweighted flow accumulation value you will have a more meaningful measure of *relative* influence.

You want to normalize a weighted flow accumulation grid by an unweighted flow accumulation grid that only represents the number of pixels above a point. Before we can do the division however, we need to do a slight manipulation on our flow accumulation grid to eliminate 0 values that yield null values when divided and end up causing errors in the MaxEnt software. The manipulation we will perform is to add 1 to each of the cells in the flow accumulation grid to make a grid that represents the pixels flowing into each cell plus the cell itself. To add 1 to each grid cell, we will use the Plus tool (Plus) located in ArcToolbox→Spatial Analyst Tools→Math. Open the Plus window by double clicking the Plus tool, enter your flow accumulation grid in the "input raster or constant value 1" text box, enter the number 1 in the "Input raster or constant value 2" text box, name the Output raster and click OK. A value of 1 will be added to each pixel.



Next we will normalize the weighted flow accumulation grid by the modified flow accumulation grid. To do this division, you will use the Divide tool (Divide) under ArcToolbox→Spatial Analyst Tools→Math. Double click the Divide tool to open the Divide window. In the top text box, "Input raster or constant value 1", you want to put the weighted flow accumulation grid to be normalized (the numerator). In the "Input raster or constant value 2" text box you should place your "normal plus 1" flow accumulation grid. Select the name and location of your output raster file, set the Environments extent and cell size to equal the Reference Grid, and hit OK to run the divide.



Normalizing a weighted flow accumulation grid calculated from a binary base layer will yield an output grid of "proportion in watershed" values. Normalizing a weighted flow accumulation grid calculated from a base layer with continuous values will yield an "average in watershed" value.

GIS: Convert grids to ASCII (.asc) format

The final step to prepare your environmental data for use in continuous watershed prediction models is to convert all of your grids to the format required for import to the MaxEnt software. This process is exactly the same as the one presented in the directions in the previous model type on page 20.

C. Constrained local prediction. A constrained local prediction model estimates the suitability of each pixel for the species in your study area for only those pixels that represent a stream or river channel, using local environmental information only. To prepare data for constrained local prediction, you must work with continuous rasters to extract only those pixels that overlay water leaving pixels over terrestrial space blank. To prepare data for these models you must:

1. Convert all data to raster grids with the *exact same* extent and cell size
2. Create and apply a "water mask" to the grids to exclude terrestrial cells
3. Save each grid to the format required to be imported by MaxEnt (ASCII text file)

Development of constrained local prediction maps utilize all of the steps described in the directions for Continuous local prediction above (page 18), except for one, the application of a "mask" that will constrain the environmental grids to only the pixels that represent water. As mentioned earlier, low resolution analysis with large cell sizes (e.g., 1 km^2) is not very appropriate for constrained analysis, because streams and rivers are narrow and linear. Higher resolution analyses are more appropriate. There is no specific maximum grid size after which constrained analysis is unreasonable; you should use your judgment.

GIS: Convert all data to raster grids with the same extent and cell size

This topic is covered above (page 19), except that as you convert data from vector to raster, or resample rasters you will want to "Set a mask" in the Environments settings, as described in the next steps.

GIS: Create a river grid mask

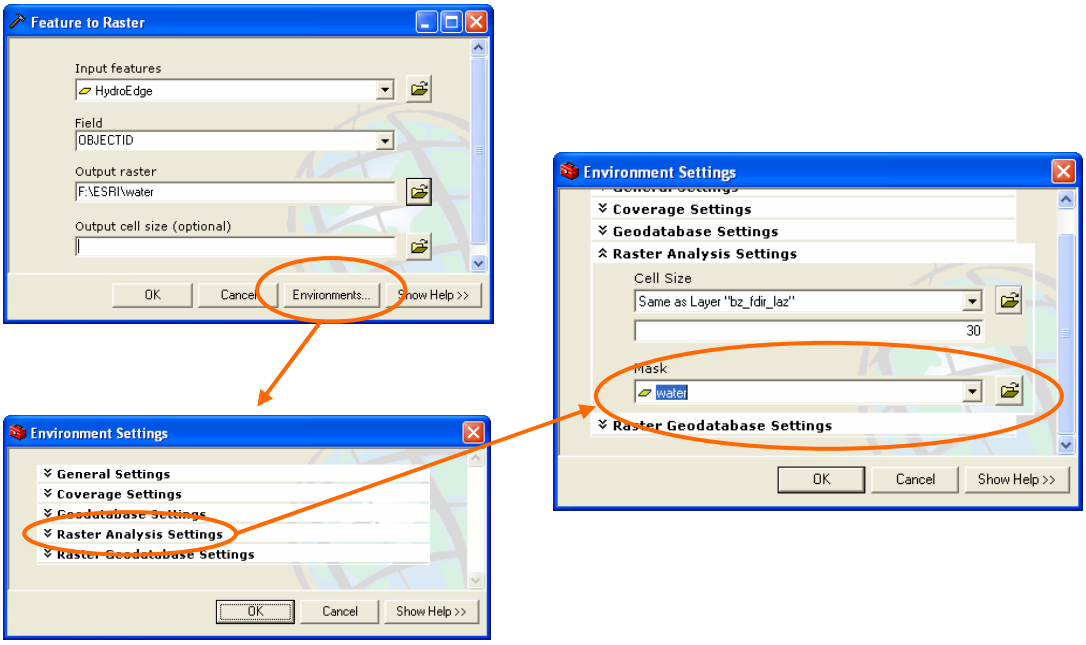
A mask identifies those cells within the analysis extent that will be considered when performing an operation or a function. Cells not covered by the mask layer will be excluded from the resulting raster grid. Because a constrained analysis only considers values where water is located, we need to create a mask that represents water and apply that mask. To create the mask you must convert a vector stream line feature to a raster using the instructions above in the "'Rasterizing vector feature classes" section (on page 19). You will need to select the Field to use as the values for your stream raster; you can select any numeric field with no missing values for this purpose.

One word of caution before you do this: if you use a stream line dataset that does not match the IABIN-DGF flow accumulation grid, then it is possible that the water mask will not line up exactly with the areas of maximum flow accumulation. This means that the pixels you extract with your river mask may not represent the areas where the flow direction grid suggests the river to be. For this reason it is important that your stream

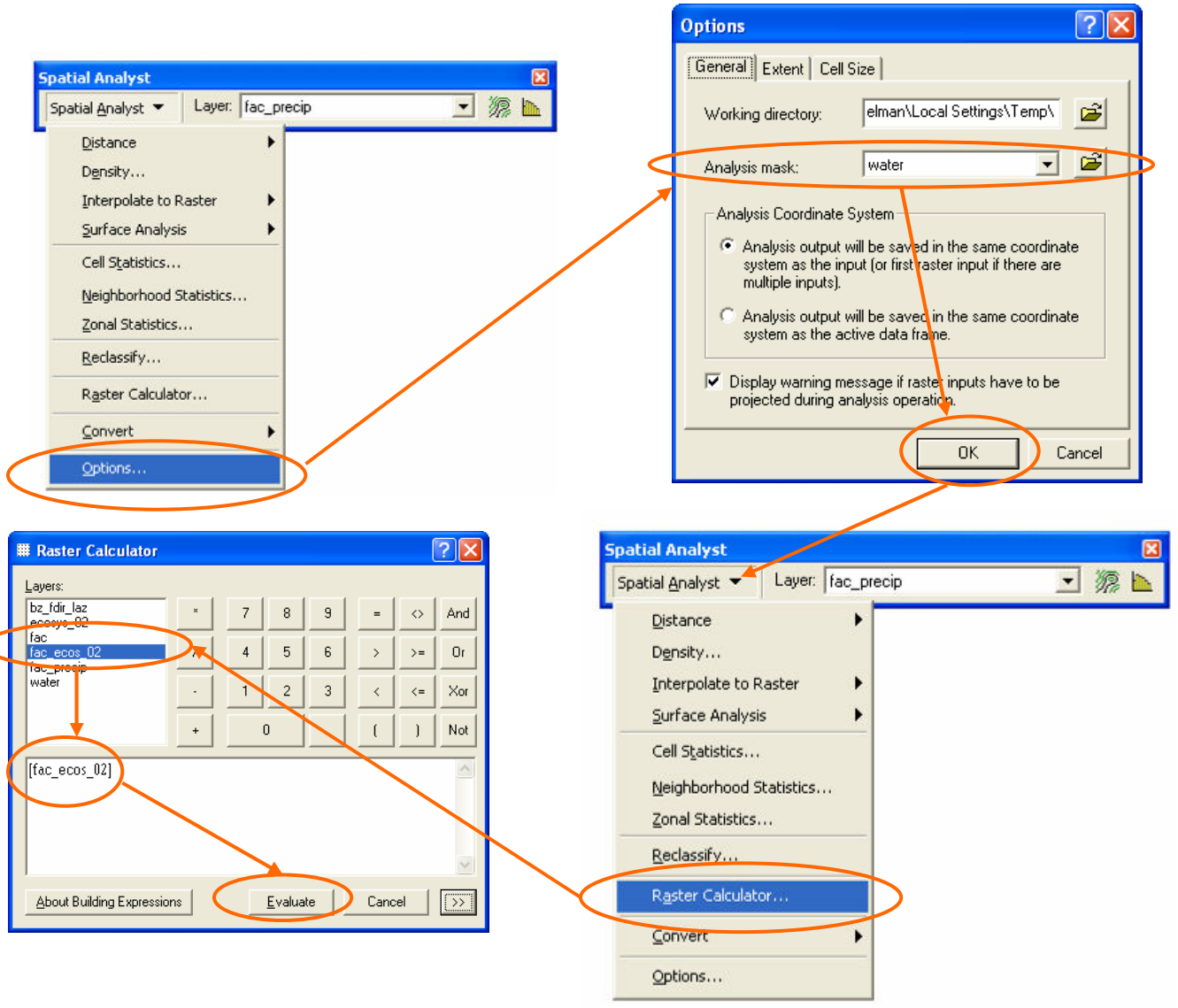
layer and your flow direction grid line up precisely. The best way around this problem is to use the Synthetic Streams posted on the IABIN-DGF webpage as your stream lines, which were derived directly from the IABIN-DGF flow direction grids and thus match perfectly.

GIS: Apply mask to rasters

Now you should have a water mask that contains values in the pixels where streams and rivers are located, and no values in other pixels. There are several ways to use this new information to constrain your environmental data sets to only those pixels. The first way is to adjust the Environments settings when creating any new raster datasets in the initial processing of your data. Clicking the Environments button in any raster creation tool window yields the exact same Environments window each time. To set a mask for the output raster and limit it to only water cells you need to load your new river mask layer into the "Mask" box under Raster Analysis Settings (see graphic). Be sure to also set the other Environments to match the extent and cell size of your Reference Grid. The resulting layer will only contain values in the pixels you defined as water.



If you want to apply your river mask to rasters that you already created that are already formatted for your species modeling effort, you can use the options within the Spatial Analyst Toolbar. This is the toolbar that opens when you initially activate the Spatial Analyst Extension (you can open the toolbar under View→Toolbars→Spatial Analyst in the Main Menu). By selecting Options… from the Spatial Analyst dropdown menu, you can set the "Analysis mask" equal to your water mask (see graphic below). After hitting OK in the options window, next open the Raster Calculator from the Spatial Analyst dropdown menu. This will open the Raster Calculator window. To apply the mask to each layer using the raster calculator, just double click the appropriate layer in the Layers box and then press the Evaluate button. A temporary raster will be generated with the name Calculation. To make this a permanent layer, you should right click on the Calculation output and select the option "Make Permanent" from the menu. This will allow you to save the constrained raster to an appropriate location.



GIS: Convert grids to ASCII (.asc) format

After you have created all of your constrained grids, the final step to prepare your environmental data for use in constrained local prediction models is to convert all of your grids to the format required for import to the MaxEnt software. This process is exactly the same as the one presented on page 20.

D. Constrained watershed prediction. A constrained watershed prediction model estimates suitability of all water pixels for a species using environmental data representing factors in the watershed upstream of each pixel. Local variables can also be mixed into these models. Constrained watershed models require the calculation of cumulative upstream influence on each pixel. To prepare data for these models you must:

1. Convert all data to raster grids with the exact same extent and cell size
2. Accumulate values for watershed based variables

3. Normalize watershed variables (to give proportion and average data)
4. Apply a water mask to the grids to exclude terrestrial cells
5. Save each grid to the format required to be imported by MaxEnt (ASCII text file)

All of the five steps for preparing for a constrained watershed model are detailed in the text above. The only difference between this type of model and the prior one is that watershed values are used, rather than only local scale variables. As with the prior type of model, you must be certain that the stream line feature precisely matches the areas of maximum flow accumulation in your flow accumulation grids. When using the IABIN-DGF flow direction and flow accumulation grids, your best choice is to use the synthetic stream coverages provided on the IABIN-DGF download page.

GIS: Convert all data to raster grids with the same extent and cell size
This topic is covered above (page 19). The only difference in making the conversions is that you should set a mask to all datasets for which you must create a new grid (see "Apply mask to rasters" above, page 26).

GIS: Accumulate values from watershed based variables
See page 23 for specific instructions on how to accomplish this step.

GIS: Normalize watershed variables
See page 24, but use a normal flow accumulation grid (not flow accumulation plus 1).

GIS: Create a river grid mask
See page 25.

GIS: Apply mask to rasters
See page 26.

GIS: Convert grids to ASCII (.asc) format
The final step to prepare your environmental data for use in constrained watershed prediction models is to convert all of your grids to the format required for import to the MaxEnt software. This process is exactly the same as the one presented in the directions in a previous model type on page 20.

Locate, evaluate, and format species occurrence data
Assembly of a good point occurrence dataset for the target species in your investigation is one of the most important steps in the modeling process. There are two fundamental challenges that you must overcome before you can prepare species sample data for use in your MaxEnt models: (1) locating sufficient data and (2) cleaning your data to ensure that you have a species occurrence dataset that aligns well with any flow accumulation grids you will use.

Locating information is the first challenge. The species data likely to be most commonly available are going to be records of specimens captured during species inventory efforts at various times in the past. Locating sufficient data can be difficult, particularly in developing countries where online specimen databases may not currently be available, though in Central America, through IABIN and other efforts like it, specimen, species, and ecosystem data are increasingly available online. While there is no one rule of thumb saying that you need *x* number of data points to develop successful models, more data will make your model results

more robust and reliable. You need to find information on where specimens of your desired species were observed or captured, complete with the species name and the latitude and longitude of each record. Such data may be housed in local universities or museums, but internet-accessible databases are increasingly powerful. Some highly useful online data sources for locating species information can be found at the bottom of Table 1. These online data servers (e.g., GBIF; Figure 10) allow you to query the scientific and/or common names of species and often allow you to specify your geographic area of interest. GBIF's website also returns records for taxonomic synonyms (previous names used for a species whose taxonomy has been revised), gives detailed source information for where the data are being drawn from, and sorts the data between those records with geographic coordinates, and those without. The only data that will be of use to your modeling effort are data with geographic coordinates (latitude and longitude). If data are extremely sparse, then you may be able to use detailed descriptions of the locations where specimens were collected to assign provisional locations to each record. The assignment of provisional locations is not ideal because it compromises the accuracy of your data, and hence the reliability of your resulting models. However, when these are the best data available, you can consider using them on an exploratory basis.

Once you have located data you should carefully assess whether or not the data meet your desired purpose. Answer the following questions about the data that you do locate to help you make a judgment call about whether or not the data are appropriate:

- *Do the data have latitude and longitude coordinates? If not are the locality data detailed enough to map them manually?* Without latitude and longitude you cannot use the data, unless you can confidently recreate the geographic coordinates from the descriptive locality information.
- *When were the data collected?* If you want to use current land use information, you will not be able to use specimens collected many years ago. Or if you want to map the "historical range" of an organism, you will benefit from having some records from before substantial human impact occurred.
- *Who collected and identified the specimens?* Are they reputable professionals in their field or amateurs? Usually, data drawn from large university specimen collections are excellent data sources, because of the high taxonomic and field standards of university workers. Data from amateur naturalists should be considered suspect because the likelihood of taxonomic mis-identifications is greater.
- *Were the specimens collected from within their native range?* Specimens collected from populations outside of their native range deserve more scrutiny, because they may have different environmental limits in a novel ecological context, or they may be the result of an active stocking program without which the species would not occur. Consider whether non-native species records are appropriate for your study.

It is up to you to make judgment calls about which data to use, and to justify these when you present your results. Once you have located data with latitude and longitude information, you should map the points and examine the position information. This involves utilizing a spreadsheet program and ArcGIS (or another GIS program).

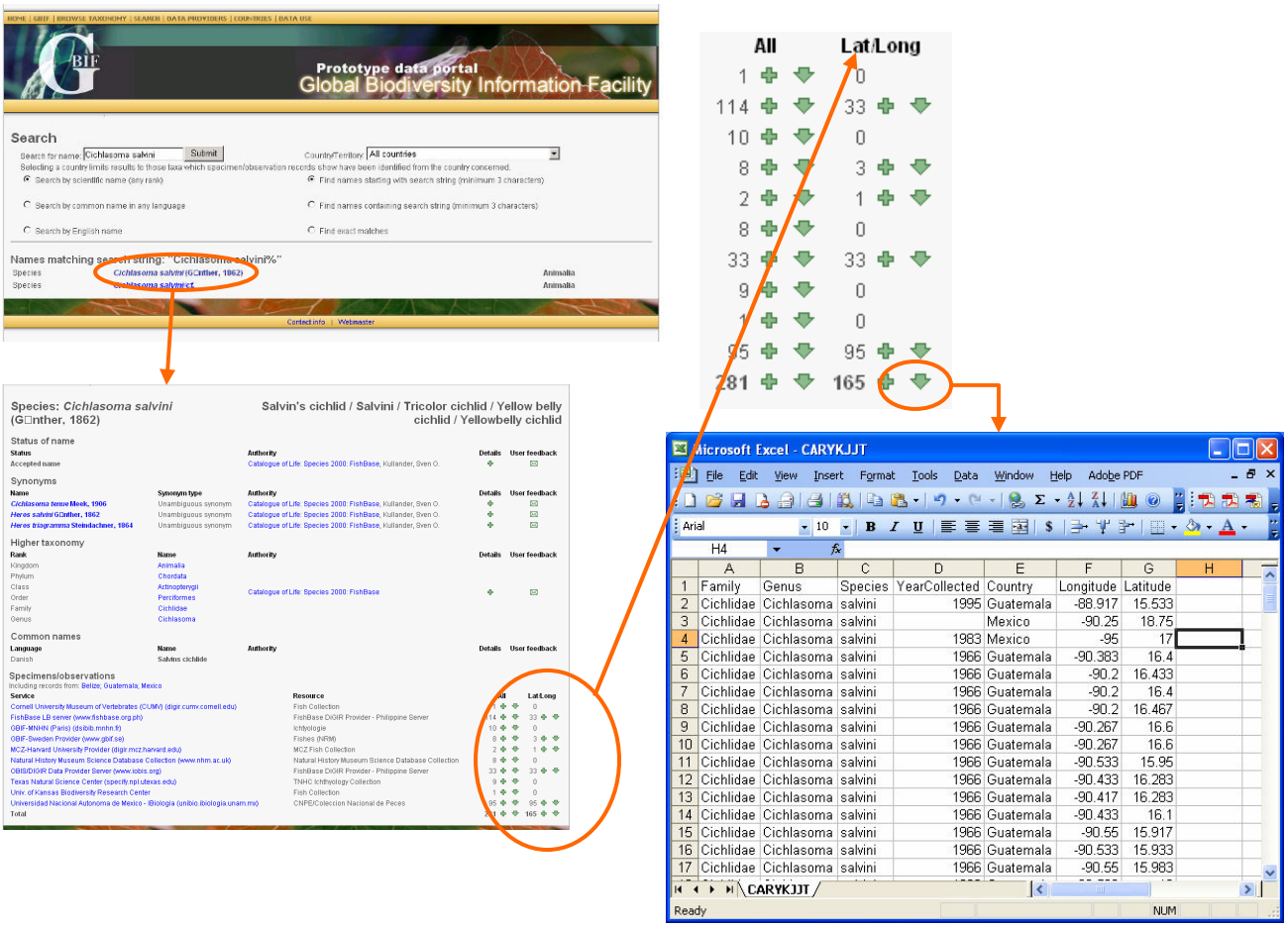
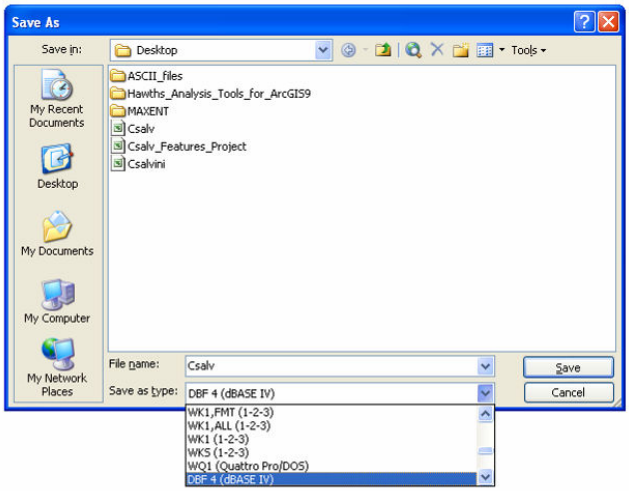


Figure 10. GBIF is an excellent online data portal to search for information on any species. The query tool searches many databases and returns information about the locations that have data on the species you are interested in. The query engine sorts data that have latitude and longitude coordinates from those without (top right) allowing you to simply download data as tab-delimited text, in this case 165 localities for *Cichlasoma salvini* with latitude and longitude, which is easily imported into Microsoft Excel (bottom right).

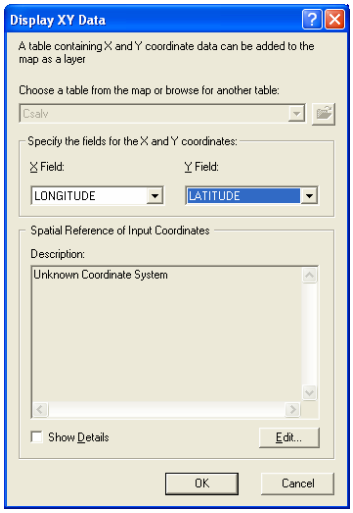
<u>GIS: Tabulate data and project to map</u>

You should compile a spreadsheet with the species name, and latitude, and longitude of the specimen records you want to use. If you are mixing records from different sources, be sure that you know the coordinate system of the data. If you have data with different coordinate systems, you should prepare one spreadsheet for the points in each coordinate system. In Excel, be sure that you format the longitude and latitude columns as Numeric with as many decimals as your data are in (for GBIF data this



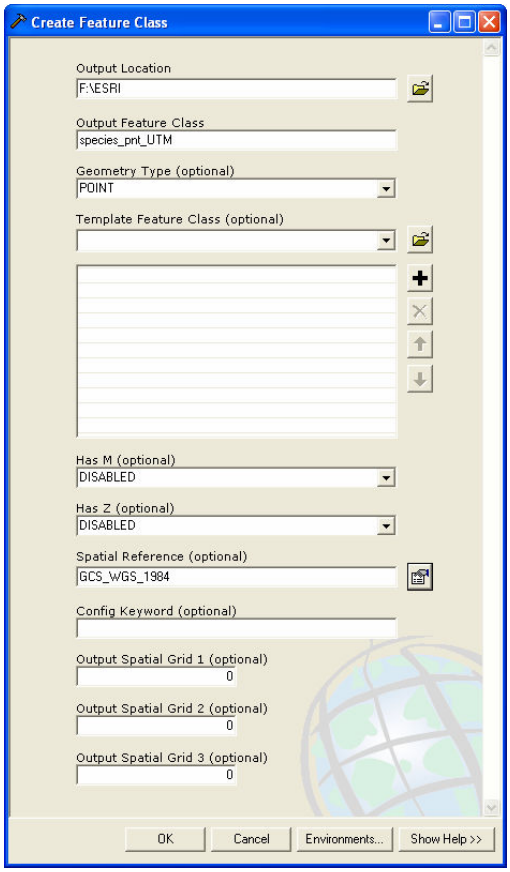
is no less than 3 decimal places). This is accomplished by selecting the latitude and longitude columns in Excel, right clicking and selecting Format cells, then under the Number tab, selecting Number as the category, and assign the number of Decimal places. Once you have done this, close the Format cells window by clicking OK. Then go to File→Save as... and select the DBF 4 (dBASE IV) format from the "Save as type" dropdown menu.

Next in ArcMap, click the Add Data button ( ), and add the DBF file that you just saved. It will appear in the ArcMap Table of contents as a Table (Csalv). Right click on this and select Open to view the table. Be sure that the longitude and latitude numbers and columns all imported correctly (sometimes DBF files make strange errors; if the numbers did not save correctly, try clicking any numeric cell in Excel before saving next time, this sometimes helps). If the numbers look correct, right click the DBF file again and this time select Display XY Data. This will open the Display XY Data window. Be sure to set the X Field option to Longitude and the Y Field option to Latitude. You can then click the Edit button and define the coordinate system, then hit OK. Your data should be converted to a point coverage displaying the locations of your specimen records.



GIS: Digitize data manually

If you are in the position where you only have good descriptions of localities and want to use these to model species, you will need to use the Edit functions in ArcMap to create a new point layer. First however, you should create a new Feature class using the Create Feature Class tool (Create Feature Class) under ArcToolbox→Data Management Tools→Feature class. This will open the Create Feature Class window. Define the Output location, name the Output Feature Class, select Geometry Type = "POINT", and define the Spatial Reference (leave all the other options in their default condition). Click OK to create the new Feature class; it will appear in the Table of Contents. You should also add your best river coverage and any other files that may be of use to interpret a map of species localities (e.g., roads, towns, etc.).



Next we will add points using the Editor Toolbar, which can be accessed from View→Toolbars→Editor. Click on the Editor button (Editor) on the Editor Toolbar and select Start Editing (Start Editing). Click the Target Layer button on the Editor Toolbar ( )

and be sure your new Feature Class is selected. Next select the Sketch Tool ( ) and place the points in place to estimate locations of sampling sites. When you are finished editing Click Stop Editing (Stop Editing) under the Editor button and save your edits when prompted. You now have a point feature of estimated sampling locations.

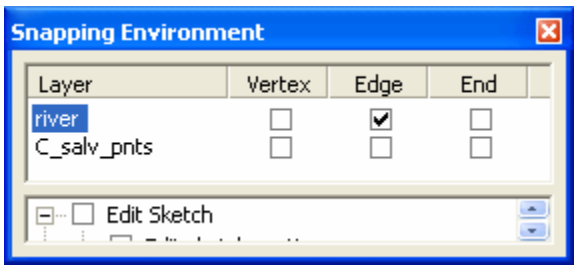
GIS: Delete unwanted points

You can use the ArcEditor Toolbar to delete any unwanted points from your species records that you deem unreliable or inaccurate. Go to View→Toolbars→Editor. Click on the Editor button (Editor) on the Editor Toolbar and select Start Editing (Start Editing). Click the Target Layer button on the Editor Toolbar ( ) and be sure that the point layer you want to edit is selected. You can use the Edit Tool ( ) to select points and delete them with the Delete key. Be sure to Stop Editing and Save Edits when you are done.

GIS: Convert data to common projection

Once you have mapped and "cleaned" up any errant data points, you need to create a unified point coverage by merging any point feature classes that are in different projections. You need to put your point data in the same coordinate system as the environmental layers. To do this, you may need to reproject your point data (see page 16 for instructions). Once you have them in a common coordinate system, you can again use the Editor Toolbar to merge the data. To do this, use ArcCatalog to create a copy of the point file with the most points in it and rename it (ArcCatalog can be used to do this). Start Editing the copied feature class in ArcMap making sure it is selected as the Target Layer ( ). Turn on the other point layers one by one so they appear by themselves on the screen, then use the Edit Tool ( ) to click and drag and select all of the points. Once points are selected, right click the selected points and click copy. By then right clicking anywhere in the blank portion of the screen and selecting Paste, the points will be pasted into the Target Layer. Once points have been added from each different point dataset, stop editing and save your edits.

GIS: Snap points to stream lines



If you are doing fine resolution modeling, it is a good idea make sure that your species points fall on a stream line in the river layer that matches your flow direction grid. You need to snap points to lines manually with the Edit Tool ( ), but it is helpful to first set the Snapping Environment, which will make the points jump to the nearest stream line like a magnet as you drag it within a specified number of pixels. To set the snapping environment click on the Editor button (Editor) on the Editor Toolbar, and select "Snapping…". This will open the Snapping Environment window with each features listed with check boxes labeled Vertex, Edge, and End next to it. To enable the snapping function, Check the Edge box for your river layer. Now when you use the Edit Tool to select and drag an individual point close to a river line, the point will snap to the line when it gets close enough. You can set the "Snapping Tolerance" (e.g., the distance from the line after which the point will snap) by selection Options in the Editor menu, clicking the General tab and typing in the desired number of units in the Snapping tolerance text box. A larger number will

make the point snap to the line at a greater distance from it. When you are done snapping all the points to the locations of the river lines, stop editing and save your edits.

Now you should have a "cleaned" point coverage that exactly matches the location of the river lines in your environmental data, and which is almost ready to use in MaxEnt.

GIS: Add x/y fields and edit attribute table

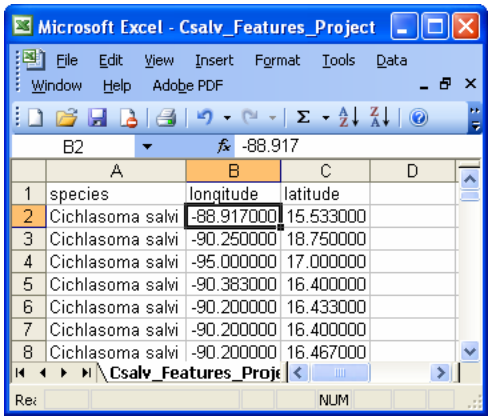
The next step is to make sure that all of your points have correct Latitude and Longitude information associated with them. This is an easy operation in ArcToolbox→Data Management Tools→Features→Add XY Coordinates. Clicking this tool will open the Add XY Coordinates window which only has one entry window for Input Features. Enter your master point coverage into this text box, click OK, and two columns will be added to the attribute table of the point shapefile, POINT_X and POINT_Y.

GIS: Save data to CSV file

MaxEnt requires that your data be formatted in Comma Separated Value (.csv) format with only three columns (species, longitude, latitude), so that your final data look like this:

species, longitude, latitude
C_salvini, -65.4, -10.3833
C_salvini, -65.3833, -10.3833
C_salvini, -65.133, -16.9

Use Microsoft Excel or another spreadsheet program to format your data this way. In Excel open the dBASE (.dbf) file associated with your species point occurrence dataset. Modify this table to contain only the three columns necessary, and then Save as a CSV file. First, to avoid damaging the species point file that you worked so hard on, Save a copy from the DBF file to a spreadsheet with a different name, using the Microsoft Office Excel (.xls) format. When this is done (and the dbf file associated with your point data is safe), delete any unwanted columns leaving only the columns: species name, POINT_X and POINT_Y. Change POINT_X to read Longitude, and change POINT_Y to read Latitude. Be sure that the column with your species name in it is labeled Species in the top cell. Once this is done, you should Save the file normally, and then save it as a CSV file by selecting "CSV (MS-DOS)" from the Save as type dropdown menu.



Step 4. Develop Models

In Step 4 all of the hard work that you have put into forming a conceptual model and preparing datasets will come to fruition. MaxEnt will receive your species occurrence point dataset and your environmental layers to generate predictions of the suitability of each pixel for the species to occur. It is important for you to make certain decisions prior to developing a model if possible. In particular, you should decide which variables will be used in the model.

It is a normal process to develop a large pool of variables, and then to reduce this variable set to a few predictors that carry unique information and are most easily interpreted. There is no set rule of thumb for how many variables you should include in your model. However, less variables will allow you to more easily interpret the results. In the MaxEnt literature, the number of predictor variables entered into the models ranged from 7 to 20. There are many ways that you can accomplish this reduction, but a common method is to statistically analyze the correlation strengths between variables to identify redundant predictors that carry essentially the same information[16]. Alternatively, an exploratory approach using multiple logical subsets of variables from your predictor pool to create models can be employed. The models resulting from this approach can be compared using the model evaluation criteria described below.

You will be guided through the entire modeling process step-by-step below.

Install software

A user friendly software for using MaxEnt in species modeling has been developed and is available free of cost at:

http://www.cs.princeton.edu/~schapire/MaxEnt/

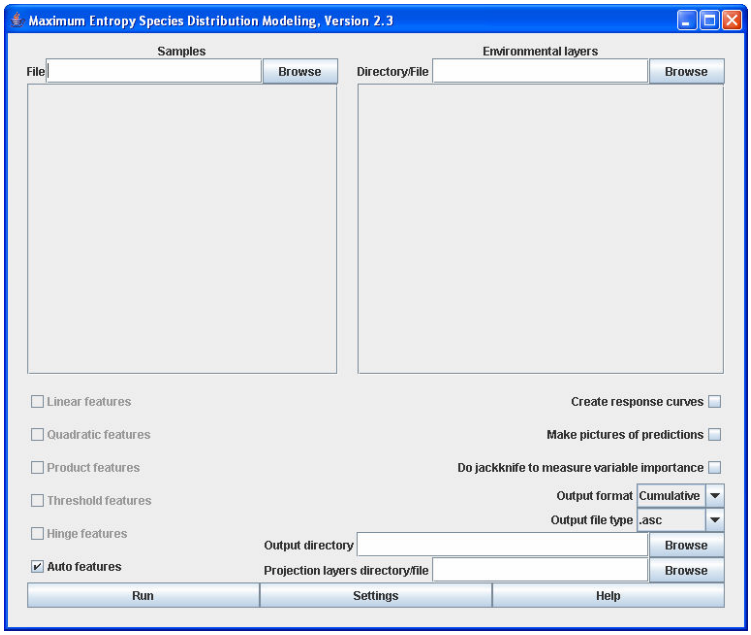
The download site has a readme.txt file that guides your through the installation procedure. The software requires that you have Java version 1.4.2 or later installed on your computer. The most recent Java software can be downloaded at:

http://www.java.com/en/download/index.jsp (English language version)
-or-
http://www.java.com/es/ (Spanish language version)

Open program

Open the MaxEnt program by double clicking on the file named <MaxEnt.bat>. This will cause a MS-DOS window to open followed by the MaxEnt Java program. You should leave the MaxEnt.bat file open to utilize the modeling software (minimize it). We will begin working through the MaxEnt window clockwise from the upper left, starting with the Samples box.

[16] Predictor variables with very high correlation strengths carry almost the same information, and thus are redundant. When you identify redundant variables, you should retain only those variables that are most easily interpreted, and eliminate the others from the final pool of variables that you will use in your model. Correlation analysis can be done using the Analysis ToolPak in Microsoft Excel in conjunction with the environmental data in the samples SWD type file described below. See Excel's help documentation for more information on how to run the statistical analysis.

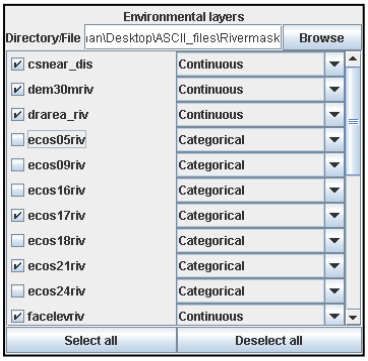


<u>Load species data (.csv)</u>

The Samples box is where we will load our species data. Click the browse button next to the File text box under Samples, and navigate to the location of your species .csv file. Open this file and you will see your species name appear in the Samples window with a check box next to it. If you have a CSV file with more than one species listed in it, you will have the option to select any of the species listed in the Species column of your CSV file.

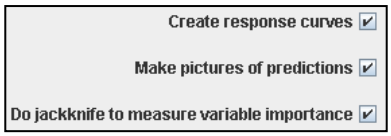
<u>Load folder with environmental ASCII files</u>

Moving to the right across the top of the screen, we will now load our Environmental layers. You should have saved all of the ASCII files (.asc) representing your environmental variables in the same folder. Use the Browse button to navigate to this folder now. When you find it, click the folder and hit the Open button. The names of all the ASCII files that you saved to that folder should appear with check boxes next to them in the Environmental layers boxes. The check boxes allow you to select which variables to include in your model. Notice the "Select all" and "Deselect all" buttons under the list of environmental layers. Also notice that each of the environmental variables has a pull down menu to its right with the word "Continuous" as the default value. This is the variable type pull down menu. The two options are Continuous and Categorical. Continuous variables are those that have an unlimited number of potential values. Categorical variables are numeric representations of categories that have a limited number of classes represented by numeric codes. For instance, in the graphic the variable "ecos05riv" represents a layer with only two values 1 and 0 (binary data). Thus it is classified as Categorical. You should classify any of your categorical variables now.



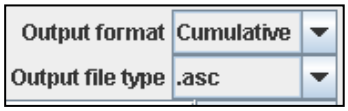
Select check box options

Moving down the right side of the MaxEnt window, there are three check box options available to you. "Create response curves" tells the program to show you how each variable influences the model outcome. "Make pictures of predictions" tells the program to create a graphic of the model output. "Do jackknife to measure variable importance" tells the program to evaluate how important each variable is by rerunning the model without each variable in turn, and running the model with each variable alone. These tests help you interpret the results. You should check each of these boxes to receive the benefits of these features in the program.



Set format and output type

The next choices in the Model software are the type of output that will be created, and the file format of the spatial predictions that are made by the program. Output format options include Cumulative (the default) and Raw. Cumulative output means that the pixel values in the output map represent the percent of pixels with values less than each particular pixel. Raw format gives a number representing the tiny fraction of the probability distribution (adding to 1) that is contributed by each pixel. Cumulative is the best choice because it allows for easier interpretation. Output file type options include three formats: ASCII (.asc), Macro ExpressMindex Effect (.mxe), and DIVA-GRD (.grd). The preferred format is the default, ASCII (.asc).



Set output directory

The output directory is the place where all of the files created at the end of a MaxEnt run are placed. This will include an HTML file, various image files, and text-based numeric results files. It is a good idea to create a new directory for outputs and create a new folder for each iteration that you run so you can go back and look at the results from different models. Use the browse button to open the folder you want the results saved to.

Define projection layers directory

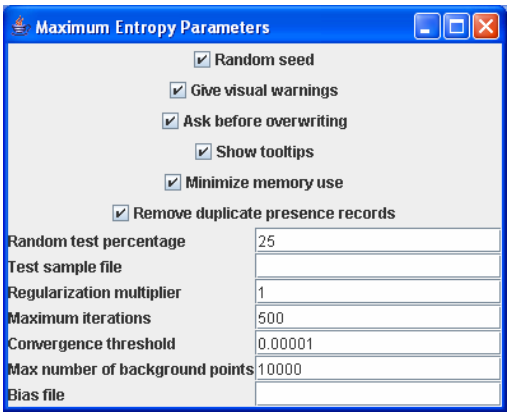
If you project your results, this means that you take the model created on one set of data and apply it to a new set of data. The Projection layers directory is the location of the folder with the layer files (.asc files) in it for the geographic area the results will be projected to. This is a useful option in different situations such as when you want to model the range of an invasive species in its native habitat and project the results to the rivers that are being invaded. You will also use this option when you run SWD models (see below). For now, leave it blank.

Help

The help button opens a help file with general advice on running a model and solving problems that might arise.

Settings

Clicking on the settings button at the bottom center of the modeling window opens the Maximum Entropy Parameters window which has some important options that control the way your model runs. There are six checkboxes at the top of the window that should be set as follows until you are familiar with the software:



- Random seed. This tells the program to select a different set of background points from the environmental grids and different test cases from your species data for each run. This should be selected when running the model.
- Give visual warnings. Clicking this will tell the program to warn you about any potential problems with the input data in a popup window. Make sure this box is checked.
- Ask before overwriting. If output files already exist for a species being modeled, checking this will tell the program to ask you before overwriting any of these files.
- Show tooltips. Checking this box will make tips appear as you roll your mouse over different parts of the MaxEnt modeling window.
- Minimize memory use. Checking this option makes the program run a little bit more slowly, but it helps the computer handle large data sets that may cause the program to run out of memory and crash.
- Remove duplicate presence records. Checking this removes any presence records with exactly the same coordinates.

The next set of options involves 7 text boxes:

- Random test percentage. This box allows you to define what percentage of all of your species points will be used to test the performance of your model. If test cases are selected the software will use these to evaluate model performance. A recommended test percentage is 25% of your samples. If you have an extremely small variable set (<15 samples), then you will probably want to use all of your points in training. If you have a very large data set (>500 samples) you can probably use up to 50% of your samples as test cases without affecting the performance of the final model. The more test cases you use, the more reliable your evaluation measures will be.
- Test sample file. This option allows you to manually define cases that the program will use to test the model. If you wanted to define test cases in this way you should separate them from your Sample file and put them in a new CSV file, and define the path name to that file in this box. If you define a test file, the Random test percentage will be ignored. It is recommended that you leave this blank unless you see a specific need for it.
- Regularization multiplier. This setting affects how focused the output distribution is—a smaller number will result in a more localized distribution that will fit the data better, but may lead to “overfitting”. Overfitting is undesirable and occurs when the model approximates the training data so closely that it loses its ability to generalize to other cases. It is recommended that you leave the multiplier at the default value of 1.
- Maximum iterations. This number tells the program how many iterations it should make while optimizing the model before stopping. Set this to the default value of 500.
- Convergence threshold. This is another stopping rule that tells the program to stop when the drop in log loss per iteration is less than this number. This stops the program when the improvement in the model reaches an acceptably small amount. Leave this value at the default value of 0.00001.
- Max number of background points. Background points are the pixel locations where the model draws values to determine parameters in the maximum entropy statistical process

- (e.g., such as the empirical average of each environmental layer, a crucial parameter to calculating the final probability distribution). Leave this value at the default of 10,000.
- Bias file. This choice is for advanced users, and should be left blank.

The environmental layers are used to produce the features that constrain the probability distribution being computed. The available feature types are linear, quadratic, product, threshold and discrete. Using "auto features" allows the set of features used to depend on the number of presence records for the species being modeled using general empirically-derived rules (see the MaxEnt Help file for more details). It is recommended that you use Auto features.

Run model

Clicking the Run button will initiate the training, testing, and jackknifing processes, and progress bars will be shown on-screen as the program works through the analysis. If the model runs successfully, it will return to the modeling window without giving you any error messages. Large data sets may take hours to run on a slow computer.

The outputs of a successful model are written in the output directory that you defined before running the model. Outputs include the following:

- [speciesname].html – the main output file in web page format, containing statistical analyses, plots, pictures of the model, and links to other files. It also documents parameter and control settings that were used to do the run. This is the first file you should view after doing a run to get a summary of the results.
- MaxEntResults.csv – a summary of the MaxEnt run listing the number of training samples used for learning, values of training gain and test gain and area under the curve (AUC). Test gain and AUC are given only when a test sample file is provided or when a specified percentage of the samples are set aside for testing. If a jackknife is performed, the regularized training gain and (optionally) test gain and AUC for each part of the jackknife are included here.
- MaxEnt.log – records the parameters and options chosen for the run, and some details of the run that are useful for troubleshooting. This file can be opened using a text editing program.
- [speciesname].asc (or.grd or .mxe) – contains the probabilities or cumulative predictions in ESRI ASCII grid format.
- [speciesname].lambdas— contains the computed values of the constants c1, c2, ... (described above on page 8).
- [speciesname]_omission.csv – describes the predicted study area, and training and test omission for various cumulative thresholds. These are the data used to help determined thresholds for occurrence.
- [speciesname]_samplePredictions.csv – predicted raw and cumulative values for training and test points.
- Plots subdirectory – contains the various plots generated for jackknife and response curves, as well as [speciesname].png, a picture of the prediction map that can be imported to written reports of results.

"Samples With Data" (SWD) analysis

Memory limitations can make it impossible to use the "normal" modeling procedure described above with very large data sets typical of high resolution models. In a model using a 30m grid size over an area slightly larger than the country of the Belize, each environmental grid contained over 80 million cells and the ASCII files associated with these grids each required

between 300 - 500 Mb of memory. The MaxEnt software program has an alternative implementation that allows the user to make predictions to higher resolution data sets.

The alternative implementation, called "Samples With Data" allows you to feed only specially prepared CSV data into the model in the Samples and Environmental layers text boxes. The basic idea of the SWD approach is to relieve the software of the memory burden of loading and taking a sample of background points randomly from across very large grids. Instead you give the program the environmental values for the species samples and for 10,000 background points. A model is trained from these data and the results are then projected (using the Projection layers directory option) to get a map of the predictions.

To use SWD input, you must format your Samples to contain information about the associated environmental data under each point, and you must create a random selection of 10,000 background points which represent the range of background environmental conditions where samples are not available. Thus your SWD species file must have column names across the top as before, but with environmental data, like this (in CSV format):

```
species,longitude,latitude,cld6190_ann,dtr6190_ann,ecoreg,frs6190_ann,h_dem,pre6190_ann,pre6190_l
10,pre6190_l1
C_salvini,-65.4,-10.3833,76.0,104.0,10.0,2.0,121.0,46.0,41.0,84.0,
C_salvini,-65.3833,-10.3833,76.0,104.0,10.0,2.0,121.0,46.0,40.0,84.0
C_salvini,-65.1333,-16.8,57.0,114.0,10.0,1.0,211.0,65.0,56.0,129.0
C_salvini,-63.6667,-17.45,57.0,112.0,10.0,3.0,363.0,36.0,33.0,71.0
C_salvini,-63.85,-17.4,57.0,113.0,10.0,3.0,303.0,39.0,35.0,77.0
```

Your environmental data must represent randomly selected points from across the study area that appear like this:

```
species,longitude,latitude,cld6190_ann,dtr6190_ann,ecoreg,frs6190_ann,h_dem,pre6190_ann,pre6190_l
10,pre6190_l1
background,-61.775,6.175,60.0,100.0,10.0,0.0,747.0,55.0,24.0,57.0
background,-66.075,5.325,67.0,116.0,10.0,3.0,1038.0,75.0,16.0,68.0
background,-59.875,-26.325,47.0,129.0,9.0,1.0,73.0,31.0,43.0,32.0
background,-68.375,-15.375,58.0,112.0,10.0,44.0,2039.0,33.0,67.0,31.0
```

Notice that the species name in this file has been replaced with the word "background".

As might be expected, SWD sampling requires several additional GIS steps to those listed for the normal runs above. These steps are described below, followed by instructions for performing an SWD run with the MaxEnt software.

GIS: Attributing species sample points with environmental data

For each environmental layer that you will use in your model, your SWD species data must have a column header and a value. This can be done in an easy process in ArcGIS 9 with a free downloadable set of extensions called Hawth's Tools.

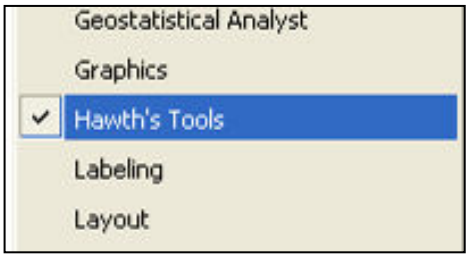
First download and install Hawth's Tools from:

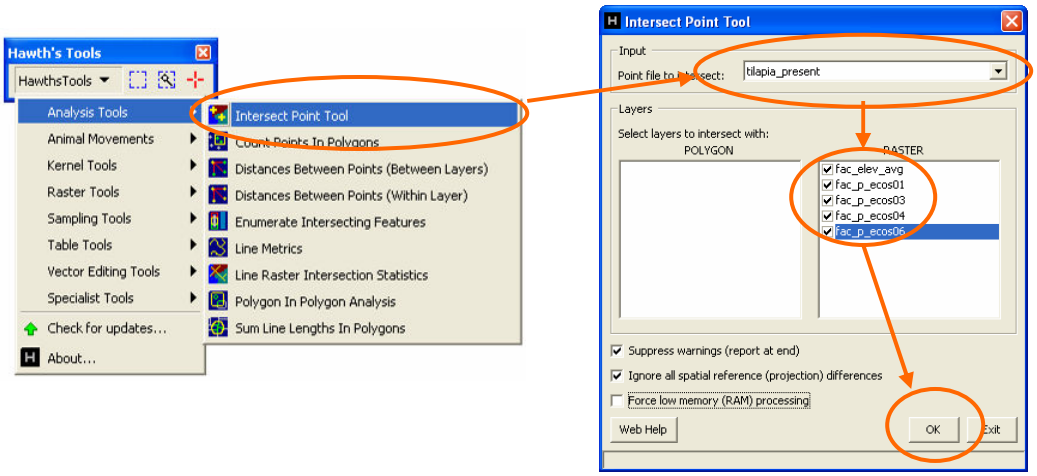
http://www.spatialecology.com/htools/download.php

Once Hawth's tools are installed, open ArcMap and load your species samples point data and all of the environmental data raster datasets into the same map. You will

modify your point dataset using Hawth's tools. Before doing this, you may want to create a copy of your point dataset, by right clicking on it, selecting Data…→Export data, defining the "Output shapefile or feature class" in the Export Data window and clicking OK. Add the new layer to the map view.

Next, activate the Hawth's tools extension by right clicking the blank part of the Main Menu (next to Help) and selecting Hawth's tools from the pop up window. This will activate the Hawth's tools menu bar. Click on the menu bar where it says Hawth's Tools and select Analysis Tools→Intersect Point Tool (see graphic below). This will open the Intersect Point Tool window. In the "Point file to intersect" box, select your species samples file from the dropdown menu. In the RASTER box, any raster layers you have loaded will appear. You should verify that only the environmental layers that you are interested are checked. At the bottom, check Suppress warnings and Ignore all spatial reference differences, and click OK. For each environmental layer, a new field will be added to the attribute table of the point dataset with the information from the cells lying underneath each point. The new fields that are added to your attribute table will be named using the first 6 letters of each raster name.





Once the intersect point process is complete, open the attribute table of your point feature to make sure that the columns were added correctly. Do not be concerned if the new field names are incorrect. We will change those next.

Remove your modified point data set from the map, and open Microsoft Excel. Use Excel to open the dBase file for the modified point dataset (File→Open→Files of type: dbase files). Once this is open you should change the new column names to exactly match the names of the ASCII environmental layers that are stored in your ASCII folder and will be loaded into the MaxEnt software as projection files—the names must match perfectly. Save the dBase file.

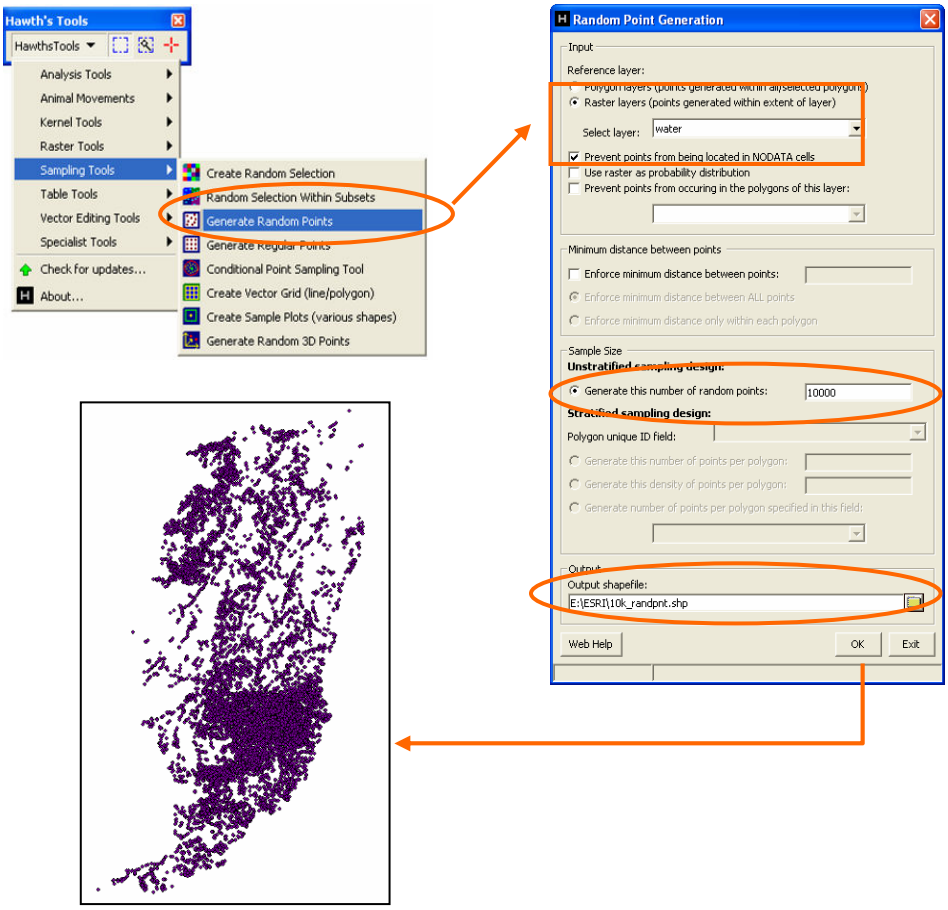
Now you will format the data to be a CSV file. To do this, first save the file in Microsoft Office Excel Workbook format so that you do not disturb the .dbf associated with your point shapefile. Then modify the columns so that only species, longitude, latitude, and

all of your environmental layers exist. Save this in Excel Workbook format again, and then save a copy in CSV format. Now your SWD Samples dataset is complete.

GIS: Create sample of 10,000 background points

Next you will create a background sample of 10,000 randomly selected points. To do this you will use Hawth's tools to select 10,000 points from an appropriate raster grid, and then attribute each of these 10,000 points with data for each of your environmental layers (as in the previous step), and finally format it in Excel and save it as a CSV file.

First, you must select an appropriate background raster dataset. If you are doing continuous modeling this should consist of a continuous raster from your study area. If you are doing constrained models, this should be your water mask. Load this layer into a new ArcMap document. Open Hawth's tools and select HawthsTools→Sampling Tools→Generate Random Points. In the Random Point Generation window, select your raster grid where it says Select layer and make sure that the "Raster layers (points generated within extent of layer)" radio box is checked (see graphic below). Also check the box that says "Prevent points from being located in NODATA cells". Under Unstratified sampling design→Generate this number of random points, select 10,000, then define the Output shapefile and click OK. At the end of the process, you will have a new point shapefile with 10,000 random background points.



Now that you have background points you need to assign XY coordinates to each of these (ArcToolbox→Data Management Tools→Features→Add XY Coordinates), and then use Hawth's tools to intersect these points with the exact same layers as you did with your species sample points, and using the same process (see page 39). Once the intersect point process is complete, open the attribute table of your 10,000 point feature to make sure that the columns were added correctly. Do not be concerned if the new field names are incorrect. We will change those now.

Remove your modified point data set from ArcMap, and open Microsoft Excel. Use Excel to open the dBase file for the 10,000 point dataset (File→Open→Files of type: dbase files). Once this is open you should change the new column names to exactly match the names of the ASCII environmental layers that you plan to project your model results to. Save the dBase file.

Now you will format the data to be a CSV file. To do this, first save the file in Microsoft Office Excel Workbook format so that you do not disturb your point shapefile. Then modify the columns so that only species, longitude, latitude, and all of your environmental layers exist. In the species column under the header make sure all of the entries have "background" typed in them. Save this in Excel Workbook format again, and then save a copy in CSV format. Now your SWD samples dataset should ready to enter into the MaxEnt software for analysis.

Run a SWD model

The process for running a SWD model is very similar to running a normal model. You should keep all the settings the same as a normal model run except:

1. Load your SWD species dataset in the Samples box instead of your species dataset without environmental data
2. Load your SWD background point dataset instead of the folder with the ASCII files in the Environmental layers box.
3. Load your folder with ASCII files in the "Projection layers directory/file" text box.
4. In Settings, do not check Random seed.

When you hit run, the model will be trained, and then the results will be projected to the ASCII files in the directory that you defined as your Projection layers. Output files from the SWD model are the same as in a normal run of the model.

Step 5. Interpret and Display Results

You will have completed much work to successfully reach Step 5, and now is the payoff. It is time to examine your results and interpret the performance of the MaxEnt model you have generated. Fortunately, the MaxEnt software outputs are very useful for interpreting model performance, and for examining how each predictor variable influences your model. This allows you to interpret the general quality of your predictions, and to begin thinking about what factors are important to the species you are modeling. Very little extra analysis is needed once you have the MaxEnt output in hand.

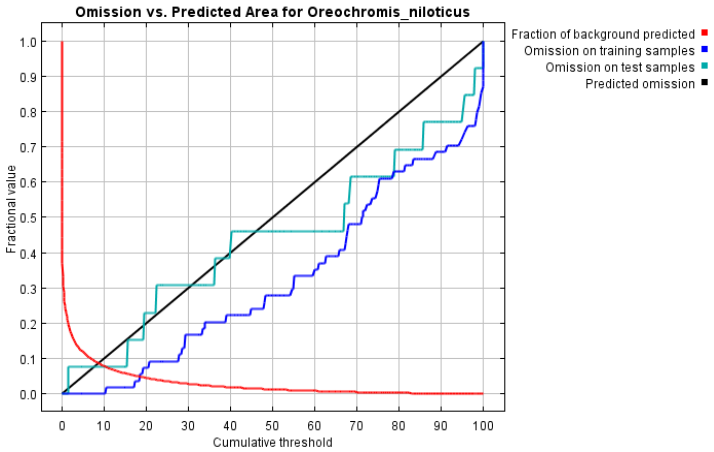
Understanding the MaxEnt output files

The output of MaxEnt is 7 different files and one folder with graphics. Two of these files are particularly important for interpretation and display: the [speciesname].html file and the [speciesname].asc file. The HTML file summarizes all of the performance tests of your model and can be viewed from any web browser, and the ASCII file is the basis for GIS maps of pixel suitabilities for the species.

It is usually best to review the HTML file first. In order of appearance, the following elements will be found:

Analysis of omission/commission

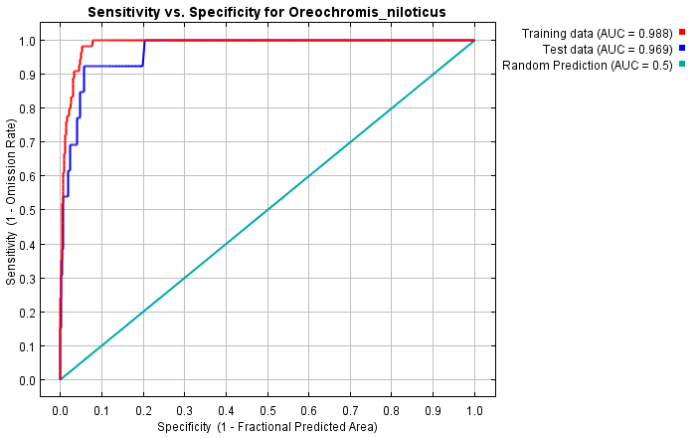
The first graphic in the output file, though not very useful, is a great starting point to introduce several important concepts that you need to understand to interpret your results. The Y axis represents the fraction of all pixels in the study area represented on a scale of 0 to 1. The X axis represents the cumulative prediction output on the scale from 0 to 100. Remember that the cumulative prediction represents the percent of all pixels that have a value of the number in that pixel or less. As cumulative predictions get closer to 100 the suitability of each habitat unit for the species increases.



There are two kinds of errors that can result from incorrect prediction of species occurrences: omission error and commission error. Omission errors are the failure to predict a location where a species in known to be present ("false negative prediction"). Commission errors occur when a species is predicted to be present when none actually exist ("false positive"). The X axis says cumulative "threshold", because omission and commission can only be estimated when we make a judgment about what cumulative value represents the break point below which a species is judged absent and above which a species is judged present. By thinking about cumulative value as a sliding scale, where any point can represent a species presence threshold, we can then chart the fraction of omissions at any given cumulative value for the Sample data we used. When the threshold is at zero, every single pixel in the area is suitable for the species to be present, thus the omission rate is zero (no sites of known presence are predicted as absent). However, as we slide the threshold toward 100 we get an increasing fraction of known present samples classified as absent (below the threshold) and so the omission rate increases and the line slopes upward. At a threshold of 100%, all pixels are

classified as not suitable for the species and all training and test cases of known presence are classified as absent so the fraction omitted is 1. The theoretical relationship between the cumulative threshold and the fractional omission error is a straight line where, a 50% threshold classifies one half (fraction = 0.5) of the known presence values incorrectly, hence the black line drawn through the middle of the graph. The red line indicates the fraction of all the pixels within the modeling area ("background" pixels) predicted at different cumulative values. This shows that the major fraction of all background pixels have very low cumulative values.

The next graphic in the HTML file shows the "receiver operating characteristic" (ROC) plot. To understand the ROC plot you must first understand the definitions of "sensitivity" and "specificity". Sensitivity is the fraction of samples with species known to be present where they are predicted present (true positive fraction). Specificity is the fraction of samples with species known to be absent where they are predicted absent (true negative fraction). A ROC plot is created by plotting the true positive fraction (sensitivity) against 1-specificity (which is the same as commission error; e.g., the false-positive fraction) across all available thresholds. A curve that maximizes sensitivity at low values of the false-positive fraction is considered a good model and can be quantified by calculating the area under this curve (AUC)[17]. The AUC is considered a measure of the model's overall performance and usually has values ranging from 0.5 to 1.0, where a score of 0.5 implies that the model predictions do not discriminate any better than a randomly generated probability distribution, and 1 indicates that the model can discriminate between true and false positive occurrences perfectly. While generally used when presence *and* absence data are available, ROC plots can also be generated with presence and random background points instead of absences (also called "pseudo-absence" data)[18]. An AUC derived from pseudo-absence data should be interpreted as a measure of the ability of the model to discriminate between a suitable environmental condition and a random background pixel, rather than between suitable and unsuitable habitat[19]



Moving down the HTML file, you will next find a table with threshold values and associated performance measures for each one. In order to utilize the performance values to the right of the table (described below) you must first select a threshold. A survey of the scientific literature reveals different opinions on what threshold value is best. Some papers use the minimum training presence threshold[20], while others use the maximized sum of training sensitivity and specificity[21], and an approach based on maximizing Cohen's kappa statistic (a method that tests the improvement of predictions over chance)[22]. Of these statistics the only one automatically calculated by MaxEnt is the minimum training presence, which is a logical choice and does not

[17] Fielding and Bell 1997
[18] See Phillips et al. 2006 for details.
[19] Hernandez et al. 2006
[20] Phillips et al. 2006, Pearson et al. 2007
[21] Hernandez et al. 2006
[22] Graham and Hijmans 2006

require use of an additional statistics program[23]. Minimum training presence uses the lowest predicted cumulative value of all the pixels occurring underneath the training samples, and can be interpreted as the value that "is at least as suitable as those where the species has been recorded present".

Once a threshold has been selected, you gain several additional performance metrics as follows:

- Test prediction success—By subtracting number in the "Test omission rate" column from 1, you end up with the proportion of sites that were successfully predicted by the model at the given threshold. A value of 1 means that 100% of the sites were predicted successfully.
- One-tailed binomial test ("P-value" column)—The last column in the threshold table has a probability value that represents the significance of a binomial test that determines whether a model predicts the test localities significantly better than random. The lower the number the stronger the test. If the P-value is greater that 0.05 then your model failed to significantly predict better than random at the 95% confidence level.

In summary, there are three tests available in the MaxEnt output to evaluate the performance of your models. The first is the threshold-independent Area Under the Curve (AUC) of the ROC plot; the second in the test prediction success; and the third is the one-tailed binomial test. These provide solid footing for examining your model's performance on the data you provide it. You should report all of these in your results.

Pictures of the model

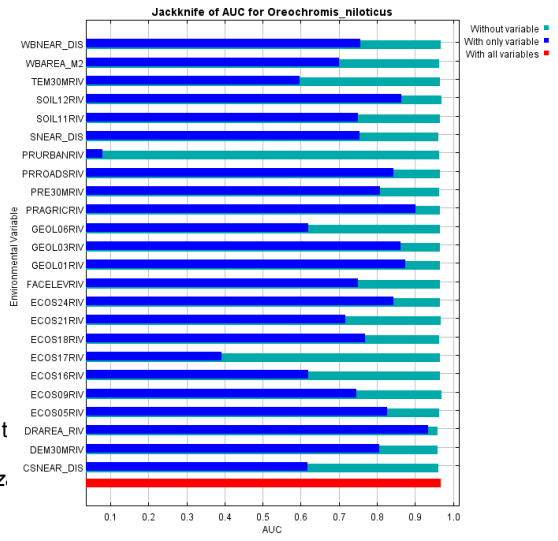
A picture of the model is the next feature provided in the HTML file. This gives you a first look at the model and can be expanded in the web browser by clicking on it. Below you will learn how to import and display the maps in GIS.

Response curves

The raw MaxEnt model has the form exp(…)/constant. The response curves show how the exponent (…) changes as each environmental variable is changed keeping all other environmental variables at their average sample value. These curves allow you to examine the effect that an individual variable has on the final model across its entire range. This can help you identify the whether each variable has a positive or negative influence on the model outcome and whether there are strong thresholds or non-linearities in the response[24].

Analysis of variable importance

The final outputs help you interpret which of the variables that you entered into your model are most influential in the model. When you use these results in conjunction with the individual charts of response curves you can get a good idea of which environmental variables influence your model the most and how. The jackknife operation runs the model over again *without* each environmental



[23] See the afore mentioned papers and Liu et al. 2005 for a discussion of t approach that makes use of available presence/absence datasets.

[24] Read the MaxEnt help file section entitled *Feature Types* and *Regulariz* about the shapes of these curves.

variable, and then with each variable *alone*. Running the model without the variable (green bar in chart) reveals which environmental layer contains the most information that is not present in other variables (the further the green bar drops the more that variable contributed to the model). Running the model with each variable alone (blue bar) shows which variables appear to have the most useful information by themselves. This process is performed for training gain[25] (the larger the gain the greater the influence on the model), testing gain, and for the Area Under the Curve. It is recommended that you compare both the training gain plot and the AUC plot, but report the AUC plot.

Mapping your data

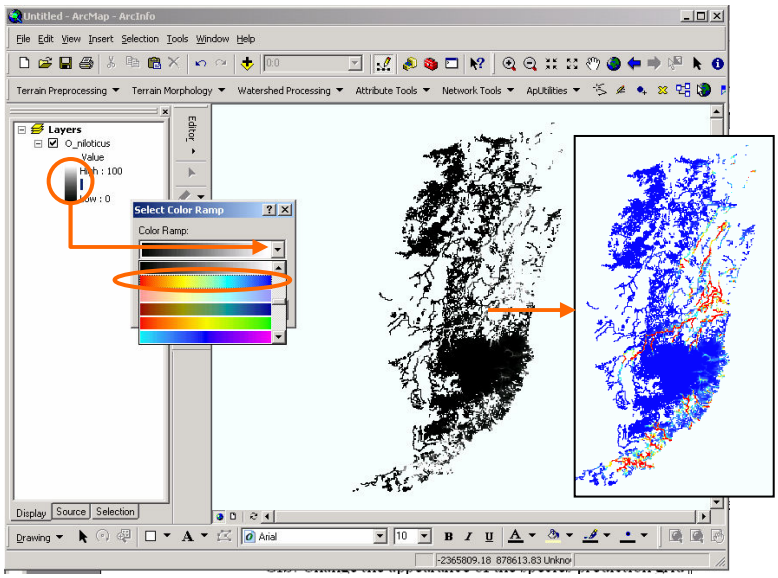
One of the most important outputs of MaxEnt is a spatial dataset that can be directly imported to ArcMap. There are at least two important ways to use this: (1) to display and communicate your results, and (2) to classify the output values by a threshold value to display a prediction of species presence or unknown status. In either case you must first convert the output file from ASCII grid format to a raster grid usable by ArcMap, and then change the way it looks.

GIS: Converting predictions from ASCII to Raster

From ArcMap or ArcCatalog, navigate to ArcToolbox→Conversion Tools→To Raster and select the ASCII to Raster tool (ASCII to Raster). Double click the tool to open the ASCII to Raster window. In the "Input ASCII raster" text box, select the ASCII file located in your MaxEnt output directory (named [speciesname].asc). In the "Output raster" text box, give the file an appropriate name. Select FLOAT from the Output type pull-down menu, and then hit OK. The ASCII file will be converted to a raster grid that can be easily viewed and manipulated in ArcGIS.

GIS: Change the appearance of the species prediction grid

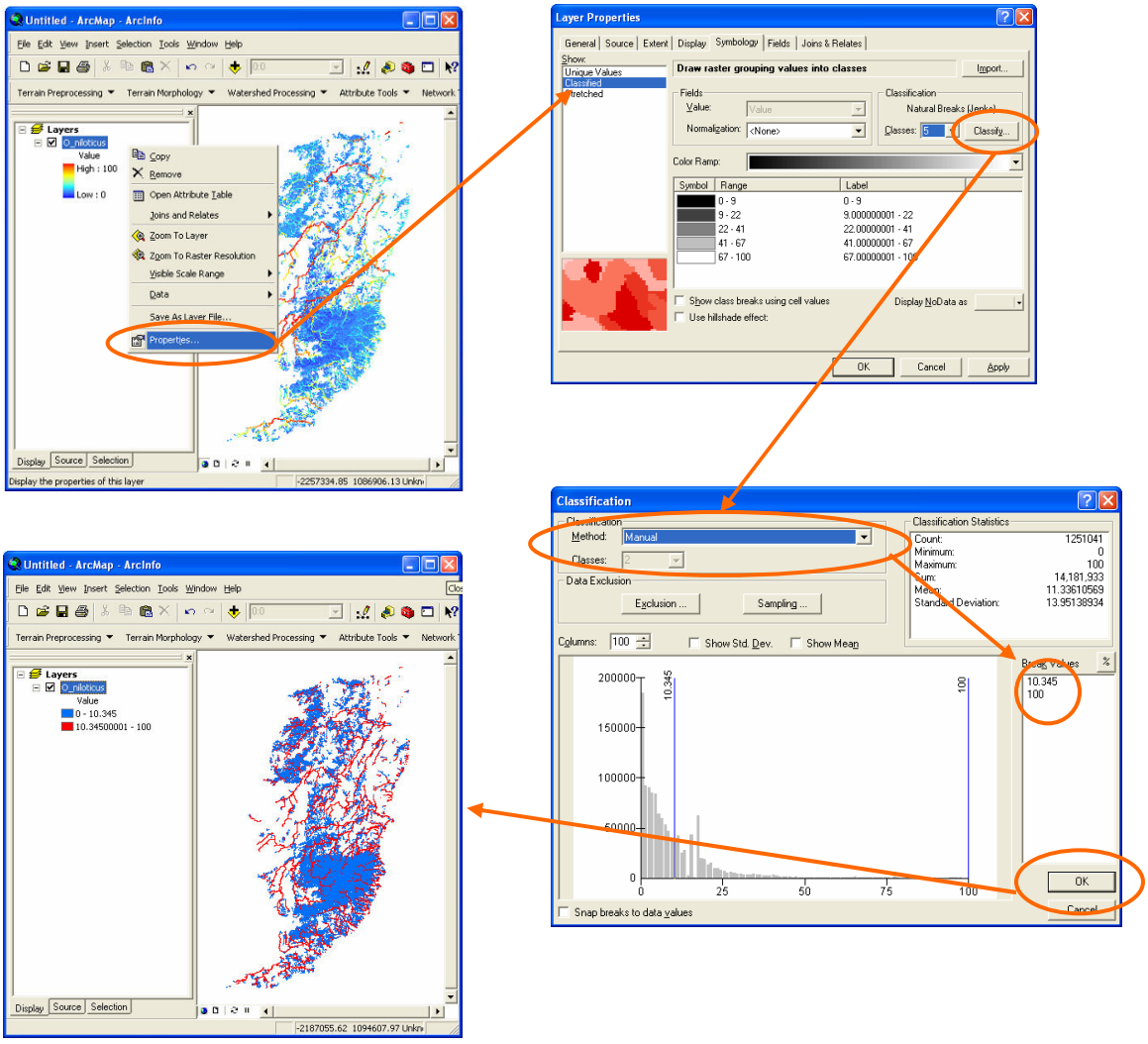
When you convert your prediction grid from ASCII to Raster and bring the data into ArcMap, it is important to be able to display the data well. The color of the value scale can be adjusted very easily from the table of contents by left clicking the color ramp, which will cause the Select Color Ramp window to open. Open the pulldown menu to see your options; Red to Yellow to Blue is a good choice—select this and click the Invert check box to cause the reds to appear on the higher habitat suitabilities and the blue to fall on less suitable habitats (see graphic).



[25] Gain is the log of the number of grid cells minus the log loss (average of the negative log probabilities of the sample locations)

GIS: Classifying your raster with a species presence threshold

Open ArcMap and load the raster layer of your MaxEnt predictions. Right click on the raster file name in the Table of Contents and select Properties (see graphic below). This will open the Layer Properties window. Select the Symbology tab at the top of this window and select Classified in the "Show:" menu to the left. This will show your data range as categories instead of as a continuous scale. Next in the "Classification" area of the window click on the Classify… button. This will open the Classification window. You need to limit your data to two classes: "predicted present" and "uncertain". Do this by selecting 2 from the Classes dropdown menu near the top. Then under Method, select Manual. Now you can type your desired threshold value (taken from the Cumulative value column of the MaxEnt threshold table in the HTML file) in as the lower number in the Break Values box on the lower right of the Classification window. Make sure that the upper value is at 100 and then press OK. This will take you back to the Layer Properties window where you should now see only two categories with the break point that you determined. You can adjust the color of these by double left clicking the Symbol colors and selecting a new color. When you hit OK now, the map will be classified as predicted present and predicted absent (see graphic below).



GIS: Transfer raster predictions to vector streamlines for display. For display purposes, one final GIS process is useful. The raster representation of streamlines is not particularly flexible for display. This can be corrected by transferring the values from the raster predictions to vector streamlines. To do this, open the Spatial Analyst toolbar and select Zonal statistics from the pull down menu. In the Zonal Statistics window, choose your stream layer as the "Zone dataset", allow the "Zone field" to populate itself, and select the output from your MaxEnt model as the "Value raster". The rest of the parameters should be left the same. When you hit OK and run the process, a temporary data table will be generated and displayed. Right click the stream layer you would like to use to display the data and select Joins and relates→Joins… In the Join Data table you want to Join data from a table, select the FID field to base the join on (1), select the temporary table as the table to "join this layer to", and select the OID field as the "field to base the join on". Clicking OK will join the two tables. To finalize the addition of the new data to the stream layer, you will need to then export the stream shapefile (Right click the file, select Data→Export data…). When you examine the new shapefile, you will see new fields that include summary statistics for all of the pixels that underlay each stream segment. You will want to use the "MEAN" column to represent the habitat suitability values for each segment. If you classify this by the same threshold value mentioned in the previous instruction you will find it easier to display your results.

Potential Model Applications

Having the ability to accurately predict species distributions has many potential applications to conservation and natural resource management. Several of these are mentioned below:

Invasive species modeling

Non-indigenous species are one of the primary threats to freshwater ecosystems around the world[26]. Having the ability to predict the distributions of non-indigenous species can offer distinct advantages to natural resource agencies and conservation groups interested in developing strategies to combat current and future invasions.

There are at least two basic ways to approach modeling of non-indigenous species. Which one you use will depend on what types of data are available. The first way is appropriate when data of an invasion are available from *within* the geographic area where predictions are being made. In this case, data on an invasive fish species collected in several watersheds, for instance, can be used to predict the possible geographic boundaries of the invasion while it is happening[27]. The map of predicted habitat suitability that results becomes a guide for thinking about what the likely ranges of the species are if it continues to expand its range to occupy all suitable habitats. This map of invasive species habitat suitability can be used to identify possible conservation conflicts, such as overlap of the predicted range with areas of extremely high native biodiversity, or with populations of rare or endangered native species.

The second scenario is when no data exist from within the study area for the non-indigenous species of concern. In this case, models must be developed from data in other locations—perhaps from specimens collected throughout the species' native range elsewhere—and projected to the new location (Figure 11)[28]. MaxEnt can easily accommodate such prediction problems assuming consistent datasets are available across the native range and non-

[26] Sala et al. 2000
[27] See Drake and Bossenbroek 2004 for a similar application using the GARP approach with presence only data.
[28] See Iguchi et al. 2004 for an example of this using the GARP algorithm.

indigenous prediction area. You must develop the models based on samples and background points from one location, and then project the model results to data prepared for the second location by creating a Projection layers directory with the non-indigenous range of data included in it.

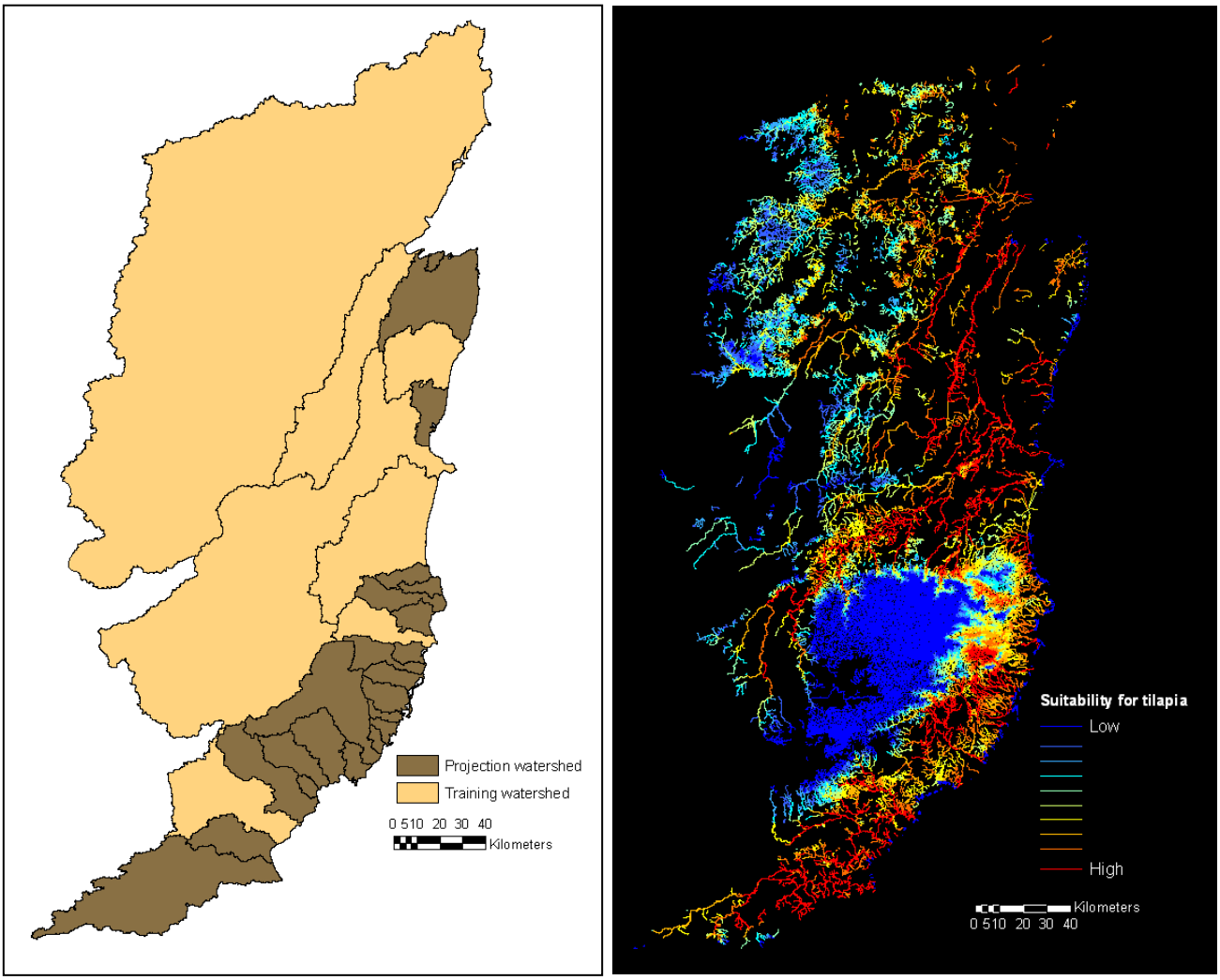


Figure 11. Models of invasive African tilapias were trained with data from watersheds in Belize where tilapias were known to have been established for at least 5 years (a). The resulting model was then used to project habitat suitability in the watersheds with no known tilapia presence or where tilapias had been introduced only recently (b). The predictions to the un-invaded watersheds represent an estimate of 'vulnerability' to future invasion by tilapias.

In either case, the model outputs represent valid hypotheses about where the species might occur as it invades the novel ecosystems where it has been—or could be—introduced. Having the ability to visualize the future in this way, though it is held in one simple map, is a powerful tool if utilized by an effective conservation team.

Rare/endangered species prediction
Predicting the possible locations of rare or poorly known taxa is a very challenging problem. However, MaxEnt has shown promise for revealing where similar environmental conditions might be found for a species, even if only very small numbers of point occurrences are available (e.g., 5 to 10 sample points). The real applications for this are to guide species inventory teams to locations where they might find more of the species, or in the case of species reintroduction programs, to identify locations with suitable habitat where the species might be able to live if it were reintroduced there in sufficient numbers by an ecosystem restoration team.

The main challenges involved in this type of application are to evaluate the performance of models, and to interpret the results conservatively. When you only have 10 sample points for a species all of your data should be used to train the model, leaving no data to test the model. Some workers have suggested ways around this problem that have validity[29]. With so few training data it is unlikely that the models will be as robust as a model with 50 or more training points. Thus the predictions should be considered tenuous predictions of where to find similar environmental conditions to those where the species is known to occur. However, tenuous predictions may be good starting point to guide sampling teams to new locations so they can at least target their search for find rare or endangered species.

Predicting species richness
Most conservation organizations want to conserve the most species using the minimum amount of resources, but without good estimates of species richness it is difficult to be strategic in trying to do this. Applying species occurrence thresholds with MaxEnt results allows you to predict species presence[30] for any species for which point occurrence records are available. Using the exact same environmental data, it is a very feasible exercise to predict ALL species known from an area, to apply a presence/absence threshold, and then to layer these results into a cumulative estimate of species richness. This is a method that has been applied to aquatic ecosystem challenges already[31], and which has much potential for guiding conservation thinking.

As always it is important to consider the outputs of such an exercise as a hypothesis of reality rather than reality itself. Prior to the implementation of expensive conservation strategies, it is always best to validate species richness predictions with independent data. However, modeling of species richness is a great way to create necessary base layers for thinking about the adequacy of protected area networks, and is an approach that has been often done in terrestrial ecosystems, and rarely done in the aquatic realm.

Targeting of Rapid Ecological Assessments
Rapid ecological assessments are a commonly used tool for rapidly documenting the biodiversity that exists in a place and to shed light on the ecological status of a location. A predicted species richness map is the ideal tool for guiding targeted assessments of areas predicted to have very high species richness, and maps of single species for guiding the search for species populations of special value for conservation.

[29] See Pearson et al. 2006 for an example of an "n-1" validation procedure.
[30] Though not true absence because no absence data were used to make the model.
[31] Dominguez-Dominguez et al. 2006

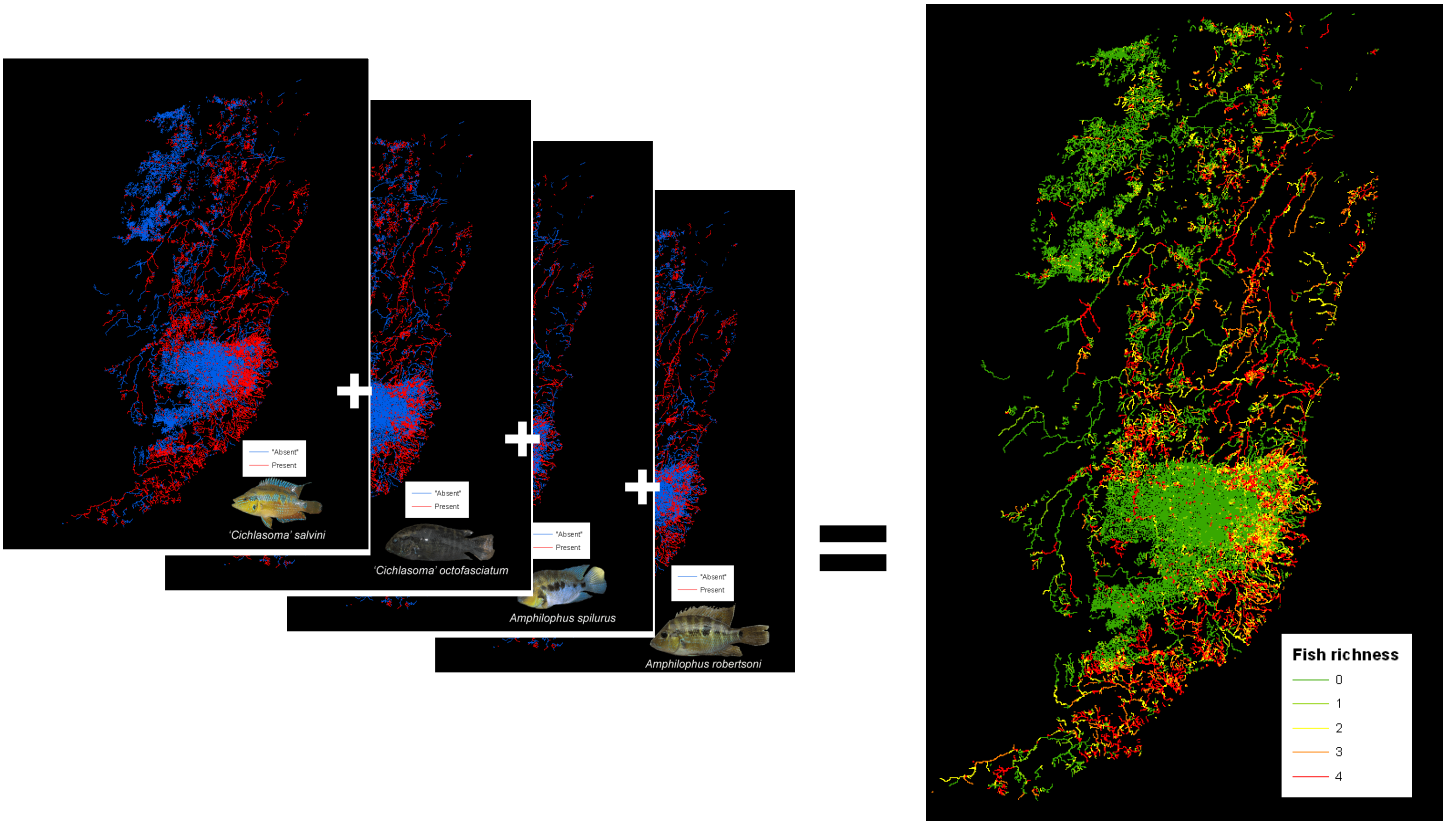


Figure 12. An example of how species presence predictions can be added together to created predicted species richness surfaces.

Prediction of species responses to climate change

Ecological niche models like those presented in this tutorial are well suited for predicting possible future species ranges under climate change scenarios. Though this type of exercise has not commonly been done for aquatic ecosystems[32], future climate prediction datasets are easily accessible (from the DIVA-GIS website for instance) for the globe. The models would have to be trained from current climate data, and the results then projected onto the future climate data. Such an application is very tenable with existing datasets, and responds to a clear and urgent threat to globally biodiversity.

[32] See Peterson et al. 2001 for an example in a terrestrial ecosystem.

Literature Cited

Dominguez-Dominguez, O., E. Martinez-Meyer, L. Zambrano, G. Perez-Ponce de Leon. 2006. Using ecological-niche modeling as a conservation tool for freshwater species: live-bearing fishes in Central Mexico. Conservation Biology 20: 1730 - 1739.

Drake, J.M., and J.M. Bossenbroek. 2004. The potential distribution of zebra mussels in the United States. BioScience 54: 931 - 941.

Elith, J., C.H. Graham, R.P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. Overton, A.T. Peterson, S.J. Phillips, K. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soveron, S. Williams, M.S. Wisz, and N.E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151.

Fielding, A.H., and J.F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24: 38 – 49.

Graham, C.H., and R.J. Hijmans. 2006. A comparison of methods for mapping species ranges and species richness. Global Ecology and Biogeography 15: 578 – 587.

Guisan, A., and N.E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modeling 135: 147 – 186.

Hernandez, P.A., C.H. Graham, L.L. Master, and D.L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29: 773 – 785.

Iguchi, K., K. Matsuura, K.M. McNyset, A.T. Peterson, R. Scachetti-Pereira, K.A. Powers, D.A. Vieglais, E.O Wiley, and T. Yodo. 2004. Predicting invasions of North American basses in Japan using native range data and a genetic algorithm. Transactions of the American Fisheries Society 133: 845 – 854.

Joy, M.K., and R.G. Death. 2004. Predictive modeling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology 49: 1036 – 1052.

Karr, J.R., K.D. Fausch, P.L. Angermeier, P.R. Yant, and I.J. Schlosser. 1986. Assessing Biological Integrity in Running Waters: A Method and Its Rational. Special Publication 5, Illinois Natural History Survey, Champaign.

Liu, C., P.M. Berry, T.P. Dawson, and R.G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28: 385 – 393.

Parrish, J.D., D.P. Braun, and R.S. Unnasch. 2003. Are we conserving what we say we are? Measuring ecological integrity within protected areas. BioScience 53: 851 – 860.

Pearson, R.G., C.J. Raxworthy, M. Nakamura, and A.T. Peterson. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34: 102 – 117.

Peterson, A.T., V. Sanchez-Cordero, J. Soberon, J. Bartley, R.W. Buddemeier, and A.G. Navarro-Siguenza. 2001. Effects of global climate change on geographic distributions of Mexican Cracidae. Ecological Modelling 144: 21 – 30.

Phillips, S.J., M. Dudik, and R.E. Schapire. 2004. A maximum entropy approach to species distrution modeling. Pages 655-662 in Proceedings of the Twenty-First International Conference on Machine Learning.

Phillips, S.J., R.P. Anderson, and R.E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190: 231 – 259.

Richter, B.D., J.V. Baumgartner, J. Powell, and D.P. Braun. 1996. A method for assessing hydrologic alteration within ecosystems. Conservation Biology 10: 1163 – 1174.