I. RESULTS OF PRIOR SUPPORT

NSF Graduate Research Fellowship, $96,000, 08/01/2000-05/01/2005. ***Intellectual Merit*:** Research supported by this fellowship resulted in substantial theoretical, statistical, and empirical advances in the study of the spatio-temporal scaling of species diversity, geographic patterns of species richness, and the temporal dynamics of communities. This research included the development and analysis of multi-site, multi-taxon databases (Adler et al. 2005; White et al. 2006; including 8 ecosystem types and 9 major taxonomic groups). ***Broader Impacts*:** The primary broader impact of this fellowship was my training and development. This fellowship allowed me to develop quantitative and computational skills through formal course work in Multivariable Calculus, Differential Equations, Linear Algebra, Non-linear Dynamics, and GIS. I also participated in the Santa Fe Institute's Complex Systems Summer School and University of New Mexico's Biocomplexity seminar. ***Publications*** – 13 publications marked * in References, including papers in *American Naturalist*, *Ecology*, and *Ecology Letters*.

NSF Postdoctoral Fellowship in Biological Informatics (DBI-0532847), "Broad-scale patterns of the distribution of body sizes of individuals in ecological communities," $120,000, 11/01/2005-08/01/2007 and associated Research Starter Grant (DEB-0827826), "Understanding Multimodality in Animal Size Distributions," $50,000, 09/01/2008-08/31/2009. ***Intellectual Merit*** – This research focused on understanding macroecological patterns related to body size in ecological communities and on improving statistical methods for the evaluation of these patterns. This research (a) yielded substantial improvements in the statistical analysis of size distributions (White et al. 2008, Thibault et al. in review) and in the ability to compare the performance of models that make multiple predictions (Price et al. 2009); (b) helped establish linkages between major body size patterns and between those patterns and the species-abundance distribution (White et al. 2007, Morlon et al. 2009); and (c) involved the analysis of large databases of tree and bird communities showing that variability in size distributions among tree communities is driven largely by gap-phase dynamics, and that individual size distributions in terrestrial animal communities exhibit strong and general multi-modality (White et al. unpublished, Thibault et al. in review). ***Broader Impacts*** – The primary broader impact of this fellowship was my training and development in advanced computational and statistical methods for the analysis of bioinformatics data. I participated in formal short courses and workshops on: 1) model selection and multi-model inference; 2) modern statistical computation; 3) Bayesian statistical methods; and 4) structural equation modeling. In addition, through my sponsoring scientists, other researchers, and self-instruction I received informal training in: 1) database design and management in MySQL; 2) probability density estimation; 3) general probability theory; 4) likelihood based statistical methods; and 5) analytical and simulation based theoretical modeling. The Research Starter grant supported the mentoring of a postdoctoral researcher, a graduate student, and an undergraduate researcher (all female). We are currently preparing to publish a dataset of mammalian community composition for over 500 sites in *Ecological Archives*. ***Publications*** – 13 publications marked † in References, which include papers in *American Naturalist*, *Ecology*, *Ecology Letters*, and *Trends in Ecology & Evolution*.

II. INTRODUCTION

Climate change, range expansion of invasive species, and other anthropogenic perturbations often impact ecological systems at continental to global spatial scales. Factors operating at such broad spatial scales are challenging, if not impossible, to study using experimental manipulations. Macroecology is an approach which studies broad-scale processes by evaluating patterns in the distributions of important variables such as body size, abundance, and species richness, and the functional relationships between these measures (Brown 1995). Because of the strength of the

macroecological approach for studying broad-scale patterns, it has increasingly become an important tool in conservation ecology (e.g., Thomas et al. 2004, Cardillo et al. 2006, Kerr et al. 2007, Davidson et al. 2009, Keir et al. 2009).

There are a variety of important – and well studied – macroecological patterns (including the species-area relationship, species abundance distribution, size-spectrum, and diversity-productivity relationship), but most macroecological studies focus on only one of these descriptions of ecological structure (see e.g., Guilhaumon et al. 2008, Oberle et al. 2009). In addition, macroecological studies tend to utilize only a single major database (see e.g., Meehan et al. 2004, La Sorte et al. 2009). Consequently, progress in this field has been slow because a) research efforts are divided among many patterns, and b) most studies explore only a single dataset, requiring multiple publications to determine whether results are similar across different taxonomic groups and ecosystems. Macroecology would greatly benefit from approaches that facilitate sifting through the large number of patterns to focus on key relationships and from increases in the rate at which patterns are compared across ecosystems and taxonomic groups.

Recent work suggests that many macroecological patterns are actually closely related to one another (e.g., Reuman et al. 2008, Morlon et al. 2009). For example, the observed species-area relationship and species turnover can be determined by the form of the species-abundance distribution and population aggregation (He & Gaston 2002, Morlon et al. 2008), and combining the species-size distribution and the size-density relationship uniquely characterizes the species abundance distribution (Loehle 2006, Morlon et al. 2009) and the individual size distribution (Jonsson et al. 2005, White et al. 2007). One of the most promising attempts at pattern unification comes from an approach imported from statistical physics – Entropy Maximization. This method, also known as Maximum Entropy or simply MaxEnt, determines the most likely form of a pattern given the operation of a small number of constraints on the system in question (Box 1). For example, it can be used to determine the most likely form of the species-abundance distribution for a community given only the species richness and total number of individuals for that system (e.g., Banavar & Maritan 2007, Pueyo et al. 2007). This approach is superior to other methods of pattern unification in that it does not require assumptions about the specific functional forms of various relationships, or the detailed processes underlying the observed patterns; it simply requires the identification of the primary constraints (Harte et al. 2008, Frank 2009). Recent papers proposing species richness, total abundance, and total resource use as basic community constraints used this approach to predict many classic macroecological patterns including the species-area relationship, species abundance distribution, diversity-productivity relationship (including both the local unimodal and regional monotonic forms of the relationship), and the energetic equivalence rule (Table 1). These models have the potential to substantially reduce the current complexity of macroecology and focus researchers on studying a few relatively simple constraints instead of the entire breadth of macroecological pattern and process. This reduction in complexity would also have substantial benefits for conservation biology by reducing the complexities and assumptions necessary to extrapolate species richness estimates across scales (Harte et al. 2009) and by making predictions for the response of communities and broad-scale biotas to global change more tractable.

While MaxEnt is a promising framework for macroecology, thorough evaluation and development of these methods are needed. Currently, much of the work on these models is purely theoretical (Banavar & Maritan 2007, Dewar & Porte 2007, Pueyo 2007, Haegeman & Etienne in review). A few studies have compared MaxEnt models to ecological data, but only for a handful of taxonomically and geographically restricted datasets. Tests of these models have focused on whether they predict reasonable descriptions of natural patterns rather than whether MaxEnt models describe observed data better than other models (including null models; see e.g., Shipley et al. 2006, Harte et

al. 2009). Furthermore, the constraints used in these models were chosen because they are reasonable hypotheses for system level constraints, but the validity of the constraint selection has not been evaluated. Finally, almost all tests (but see Harte et al. 2009) have focused on whether the models can predict observed patterns at one or a few locations. This is the logical starting point for evaluating these models, but the true potential of this approach results from making predictions across many sites using the same specified constraints. Further research is clearly needed to fully evaluate the breadth of utility of MaxEnt approaches.

Box 1. A brief introduction to Maximum Entropy

Maximum Entropy is a method for making inferences based on partial or incomplete information. MaxEnt assumes that processes other than the major constraints on the variables of interest cancel out when combined to yield the overall pattern (due to aggregation and convolution). Therefore it selects the most likely configuration of the system that satisfies the primary constraints. This is achieved by maximizing a relative version of Shannon's entropy (*H*):

$$H = -\sum p(x) log \frac{p(x)}{m(x)}$$

where $m(x)$ is a prior distribution and $p(x)$ is the distribution of interest (e.g., the species-abundance distribution), subject to constraints on values such as the mean and variance of x.

Maximization is achieved with the method of Lagrangian multipliers, where the Lagrangian function takes the form

$$\Lambda = H - \psi(\sum p(x) - 1) - \sum_i \lambda_i (\sum f_i(x) p(x) - \langle f_i(x) \rangle)$$

with Lagrangian multipliers ψ and λ_i, and constraints $\sum f_i(x) p(x) = \langle f_i(x) \rangle$. Solving the Lagrangian function subject to these constraints yields the MaxEnt solution $p(x) = \frac{1}{Z} m(x) e^{\sum \lambda_i f_i(x)}$, where Z is the normalizing factor (Frank 2009).

An Ecological Example

Using the uniform distribution as a non-informative prior and two ecologically meaningful constraints (average species abundance and average energy use per species within a community), Harte et al. (2008) were able to derive the species abundance distribution with the above methodology. Their solution, which takes the form of Fisher's log-series distribution, accurately describes the species abundance distribution in their focal communities with no parameter fitting.

Here I propose an integrated research and education program with the goal of overcoming two of the major limiting factors in macroecological research - the treatment of a diverse array of patterns as independent research avenues and the limited usage of available datasets for determining the generality of patterns and processes. The education program focuses on removing barriers to the utilization of major ecological datasets, to facilitate more rapid progress in basic and applied macroecology. Effectively using and integrating these extensive datasets requires special knowledge and tools not currently available to most ecologists. I will produce a suite of online resources (including two online short courses, a wiki, and a forum system) and university courses (at the undergraduate and graduate level) to provide training in these areas. In addition, I will publish a series of computer programs designed to automate the more advanced aspects of database design and manipulation. The resulting knowledge gained from the development and subsequent use of these tools by the ecological community will be applied to over a dozen databases that will be used in concert to test and expand the use of entropy maximization as an approach to simplifying the focus of macroecological research. To accomplish this task I will address three major research objectives.

Research Objective 1: Evaluate the performance of current maximum entropy models. I will conduct extensive model evaluation of current MaxEnt models using 5 major community level databases including over 40,000 sites from 3 major taxonomic groups (birds, mammals, and trees), improving on current approaches by directly comparing the performance of different models.

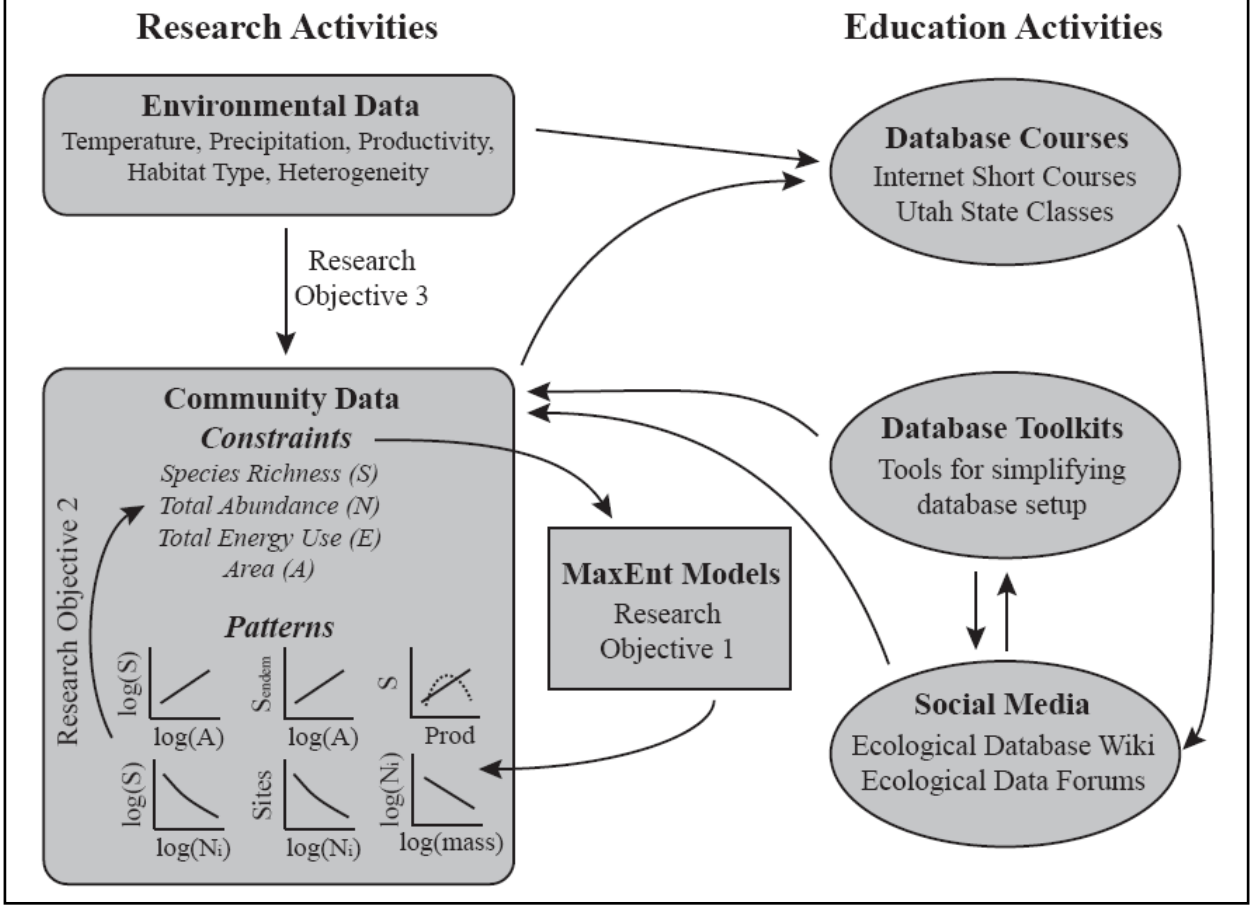
Research Objective 2: Identify underlying constraints generating observed patterns. I will use methods novel to ecological research (Inverse MaxEnt) to identify the most suitable constraints for future MaxEnt models and thus the important biological processes generating observed patterns.

Research Objective 3: Combine MaxEnt models with models of constraints to predict broad scale variation in community structure. I will develop a set of combined models that use environmental variables to model the ecological constraints. The modeled values of the constraints will then be used as input for the MaxEnt models to make predictions for continental to global scale variation in major macroecological patterns using only a small number of environmental variables.

The combination of maximum entropy methods to focus macroecological research on a more manageable set of questions, and informatics approaches to allow the generality of hypotheses regarding broad scale processes to be quickly evaluated, promises to improve the rapidity with which macroecology can address some of the key issues facing ecology.

III. INTEGRATED RESEARCH AND EDUCATION

Figure 1. Flowchart showing the research and educational plans and how they are integrated.



The research and education aspects of this proposal are inherently integrated. The research mission relies on the combination of five of the largest publically available community level databases, two organismal level databases, two taxonomy linking databases, and multiple environmental databases. One of the first phases of the proposed research is to configure and install these databases to allow their use in an integrated manner. This will be done using the proposed *Database Toolkits*, which will be made available as part of the education component of the proposal. While there are well established best practices for the use of a few of these databases, the rest have been used infrequently by macroecologists and therefore the idiosyncrasies related to their use are not well established. Working closely with my students, the postdoctoral researcher, and experts in plant, bird, and mammal macroecology, I will draft best practices on the proposed *Ecological Database Wiki*. These drafts will be modified by the broader ecological community through discussion on the forums and modification of the wiki. I will incorporate the resulting best practices into the research by modifying the treatment of the databases accordingly. The databases, toolkits, and social media will be integrated into the graduate, undergraduate, and internet based coursework. Students will be able to read papers published using the same data they are working with, and read and participate in discussions and development of best use practices for these databases via the social media tools.

IV. RESEARCH PROGRAM

In order to determine whether one or more of the proposed entropy maximization frameworks can successfully be used to simplify macroecology, substantial evaluation and expansion of current approaches is necessary.

RESEARCH OBJECTIVES

Objective 1: Evaluate the performance of current maximum entropy models and determine if there are situations in which MaxEnt performs poorly

All maximum entropy models published to date (Table 1) have succeeded at generating reasonable first cut predictions for ecological data. These tests have been confirmatory in nature, seeking simply to determine if the predicted patterns follow the same general form observed in empirical studies (e.g., Banavar & Maritan 2007, Dewar & Porte 2008), or fitting the model to data and determining if it reasonably approximates the empirical data based on visual inspection (e.g., Harte et al. 2009). While this approach is appropriate for initial evaluations, it is one of the weakest forms of evaluation possible (McGill 2003, McGill et al. 2006). In addition most of the comparisons of these models to data involve only a single dataset from a single taxonomic group. As a result it is difficult to determine whether the models perform well in general, or whether the quality of their fits is limited to the situation in which they were tested. Because MaxEnt models are supposed to be based on fundamental ecological constraints, if they perform well in one region or taxonomic group that they should perform well more generally. In order to evaluate this possibility I will test the published MaxEnt models using five continental to global scale databases (Table 2; see **Databases**) representing three major taxonomic groups (birds, mammals, and trees). These databases were selected because they contain all of the constraints included in the different models (either directly or via combination with other databases) and because they contain hundreds or thousands of sites from a broad array of habitat types. This should allow for strong general conclusions about the relevance of MaxEnt to macroecology.

I will compare individual patterns from the empirical data to the model predictions using chi-square tests on appropriately binned data and/or Kolmogorov-Smirnov tests (with appropriate corrections) when the predictions are for frequency distributions, and standard regression based methods when the predictions are for bivariate relationships (Sokal & Rolff 1995, Zar 2000). I will use False

Discovery Rate Control (Benjamini & Hochberg 2000, Verhoeven 2005) to properly control for multiple comparisons. I will also determine whether MaxEnt models perform better in some ecological contexts than others by comparing their performance in different habitats and taxonomic groups, and at different values of ecological constraints (e.g., high richness communities vs. low richness communities). I will compare the root mean square deviation for each of the models among different groupings using ANOVA (for categorical groupings), simple regression (for continuous factors; e.g., productivity), and if justified based on the data more complex approaches such as multivariate generalized linear models to deal with non-normal error structures and non-linear relationships.

Table 1. Maximum entropy models to be evaluated in this proposal, their constraints, and predicted patterns. For constraints, N is the total number of individuals, S is the total number of species, A is the total area, and E is the community-level resource use*. SAD, SAR, EAR, Spatial AD, Div-Prod and EER stand for species-abundance distribution, species-area relationship, endemics-area relationship, spatial abundance distribution, diversity-productivity relationship and energetic equivalence rule.

Models	Constraints	Predictions					
		SAD	SAR	EAR	Spatial AD	Div-Prod	EER
Banavar & Maritan 2007	N, S	✓					
Pueyo et al. 2007	N, S	✓					
Dewar & Porté 2008	N, S, E	✓				✓	✓
Harte et al. 2008	N, S, A, E	✓	✓	✓	✓		✓
Harte et al. 2009	N, S, A, E	✓	✓		✓		

*In some cases the constraints are combinations of these values (e.g., average species abundance).

In addition to these more rigorous evaluations of model fit I will compare the performance of the MaxEnt models to that of other models for the same individual patterns. I will compare MaxEnt models to three major classes of alternate models: 1) null models, such as the log-normal for the species-abundance distribution (e.g., McGill 2003) and the random placement model for the species-area relationship (e.g., Coleman 1981); 2) empirical models with fitted parameters (also known to ecologists as statistical models), such as the power-law for species-area (e.g., Arrhenius 1920) and diversity-productivity relationships (e.g., Wright 1983) and the log-series distribution for the SAD (Fisher et al. 1943); and 3) alternative theoretical models such as neutral theory for the SAD (e.g., Hubbell 2001, Etienne 2005) and competition models for the diversity-productivity relationship (Tilman et al. 1997). Comparison of MaxEnt and alternative models will be conducted using likelihood based model selection (e.g., Burnham and Anderson 2002). For alternative models that require fitted parameters the values of these parameters will be determined via maximum likelihood estimation, with appropriate bias corrections, or other appropriate methods (from Johnson et al. 1994, 1995, 2005, and more specific references for theoretical models with non-typical forms, e.g., Etienne 2005). If analytical solutions for the maximum likelihood values are not available they will be determined by numerical maximization. Models within 2 Δ_{AIC} units of the best model will be considered to have reasonable support and models differing from the best model by more than 10 Δ_{AIC} units will be considered to have no support (Burnham & Anderson 2002). Note, that in some cases MaxEnt models may predict identical distributions to either null or empirical models. In this case the models are equivalent and no comparison is necessary.

Objective 2: Identify underlying constraints generating observed patterns

Thus far the use of entropy maximization in ecology has focused on making predictions for community structure based on assumed constraints on ecological systems. In addition to allowing the development of theoretical models based on hypothesized constraints, maximum entropy can also be used in the opposite direction to attempt to understand what contraints are having the predominant influence on the system being observed. In other words, if the form of the observed pattern can be characterized in a relatively simple manner, then the form of this pattern can tell us about the relevant constraints (Kapur & Kesavan 1992, Frank 2009). I will use the Inverse Maximum Entropy approach (Kapur & Kesavan 1992) to identify the most likely constraints on each community and use the combined results across communities to make inferences regarding the most important overall constraints. This approach uses MaxEnt in the opposite direction of that used so far in ecological research This will provide insight into the processes governing ecological systems and potentially guide the development of future maximum entropy models by identifying the important constraints and how they may differ across ecosystems and taxonomic groups.

The inverse maximum entropy approach that I will take starts with a known probability distribution and a measure of entropy and then solves for the constraints that would produce the observed distribution. We will use Shannon's Entropy (Box 1) since it is the measure used in ecological studies (e.g., Shipley et al. 2006, Harte et al. 2008, 2009) and is considered by many to be the only appropriate entropy measure for these kinds of analyses (Frank 2009). The form of the observed probability distribution will be determined by fitting common statistical distributions to both the species abundance distribution and spatial abundance distribution for each community (other distributions may also be used, but these two distributions represent the easiest starting point). Of the fitted distributions, the probability distribution providing the best characterization of a given set of observed data will be determined using AIC based model selection (Burnham & Anderson 2002). All models within two Δ_{AIC} units will be considered reasonable characterizations of the distribution, so in some cases more than one set of constraints will be evaluated. The constraints yielding the best fitting probability distribution(s) will then be determined using the methods detailed in Kapur & Kesavan (1992; pages 382-399) or by using references of MaxEnt solutions for general constraints (Kapur 1989, Frank 2009). The prevalence of different underlying constraints within and among groups will be evaluated to determine if a single set of constraints to applies across or within taxonomic groups and regions, or whether constraints are idiosyncratic across communities.

Objective 3: Combine MaxEnt models with environmentally models of constraints to predict broad scale variation in community structure from a small number of environmental variables

Maximum Entropy models offer the potential to understand many aspects of community structure with information on a small number of ecological constraints. However, using these models to make predictions still requires that the constraints values be measured for every community of interest. While the datasets used in this proposal do provide information on some of the potential constraints at continental to global scales, it will typically be intractable to collect this type of information for all locations of interest. To overcome this limitation, I will generate environmental models to predict constraint values, which will then be used in MaxEnt models to predict macroecological patterns.

Based on the constraints used in maximum entropy models so far, it will be necessary to generate environmental models for three ecological values: species richness, total community abundance, and total community energy use (or resource use). The fourth constraint, area, is defined by the sampling design. Many of these values can already be modeled using statistical models (with mechanistic underpinnings) based on environmental variables. For example, for both breeding and wintering bird communities we can predict species richness fairly accurately using only measures of productivity

(NDVI) and elevational heterogeneity (Hurlbert & Haskell 2003). Likewise, total energy use in wintering bird communities has been shown to be directly proportional to productivity (Meehan et al. 2004) and the total sampled abundance (which is all that we need for evaluating these models) can be modeled effectively in breeding bird communities using a combination of a productivity measure and habitat type. Since the values of the constraints can be modeled with a small number of environmental variables, we can independently predict the values of the constraints and then use MaxEnt to predict the macroecological patterns.

The richness, abundance, and energy use of each community will be modeled using generalized linear models (GLMs) in order to accommodate discrete response variables, spatially autocorrelated error, and non-linear relationships among variables. A number of different environmental variables will be explored (see ***Databases***) and the best model will be determined by evaluating standard criteria for overall model fit (AIC comparisons, evaluation of residual structure, etc.). These comparisons will be done using cross-validation in order to avoid overfitting. Each dataset will be modeled independently as it is likely that at least the specific parameterizations, if not the overall structure, will vary among groups.

The predicted values of the constraints at each location from the environmental models will then be used as the input for the MaxEnt models to predict patterns of community structure at each site. The performance of these predictions will be evaluated in two ways. First, using simple univariate characterizations of the community structure such as evenness for the species-abundance distribution and the power-law exponent of the species-area relationship, predicted values will be regressed against observed values to determine if the slope is different from one and the intercept different from zero. The quality of the R^2 will also be evaluated. Second, the detailed analyses of MaxEnt model performance from **Objective 1** will be repeated using the predictions from the constraints for the environmental models. Cross-validation will be used with different sites used to fit the constraint models than to evaluate the MaxEnt predictions. If this approach is successful it will provide an avenue for making predictions about ecological responses to climate change and other perturbations through the modeling of the environmental variables that these perturbations are expected to influence.

DATABASES

Community data: Community level data for evaluating MaxEnt models will come from five major, publically available, continental scale databases (Table 2): Breeding Bird Survey (http://www.pwrc.usgs.gov/bbs/), Christmas Bird Count (http://www.audubon.org/Bird/cbc/), Mammalian Community Database (to be published in *Ecological Archives*), Forest Inventory Analysis (http://fiatools.fs.fed.us), and the Gentry Forest Transect Data (http://www.wlbcenter.org/gentry_data.htm). Each of these databases contains species-level abundance data for a major taxonomic assemblage (*sensu* Fauth et al. 1996; e.g., diurnal land birds or trees greater than 5 inches in diameter). I use the terms community and assemblage interchangeably as is common practice. This community level data provides information on the total abundance of individuals, and the total species richness for each site, two of the main constraints used in MaxEnt models. The area of the surveys, a third constraint in the models of Harte et al. (2008, 2009), is easily determined from the meta-data for each database. In the case of Breeding Bird Survey (stops within a route), Forest Inventory Analysis (circles within a plot), Gentry Tree Transects (individual transects within a site), and some sites in the Mammalian Community Database, multiple scales of sampling are available for individual sites. All databases will be used to evaluate species abundance distribution, regional scale diversity-productivity relationship, and energetic equivalence rule. The Christmas Bird Count will be excluded from analyses of the species-area relationship, endemics-area relationship, and spatial abundance distribution due to its lack of sub-site level sampling units. The

Gentry Tree Transects will be excluded from local scale diversity-productivity analyses because its sites are not close enough to one another to allow for aggregation into local scale groups.

Table 2. List of the community datasets to be used in the research components of this proposal, including the major taxonomic group for which the data is collected, the total number of sites available in the database, the spatial extent of the data and their availability for this proposal. Database installation scripts and templates will also be written for all of these databases as will pages on the Ecological Data Wiki.

Dataset	Taxon	Approximate # of sites	Scale	Availability
Breeding Bird Survey	Birds	5149	Continental	Public
Christmas Bird Count	Birds	5160	Continental	Via MoU*
Mammalian Community DB	Mammals	500	Continental	Owner/Public†
Forest Inventory Analysis	Trees	31364	Continental	Public
Gentry Forest Transect Data	Trees	242	Global	Public

*I am currently negotiating the Memorandum of Understanding with Audubon
†Developed by my lab; to be submitted to *Ecological Archives* as a Data Paper.

Metabolic rate data: As described above several of the MaxEnt models include a constraint on total resource use (also known as energy use) at a site, which will be estimated by summing over individual estimates of resource use based on published allometries for metabolic rate (see White et al. 2004, Morlon et al. 2009). ***Birds and mammals*:** For birds and mammals I will use empirically derived field metabolic rate allometries for the appropriate sub-group membership of each species (e.g., desert rodents) from Nagy et al. (1999). Because individual level size data is not available for the bird and some of the mammal communities I will assign the same metabolic rate to every individual of that species based on its average size (see **Size data**). Because we are only interested in the sum of metabolic rate, using the average size will have no influence on the results. ***Trees*:** The energetics of trees has been less well studied than that of vertebrates. It has been proposed on theoretical grounds that tree energy use (metabolic rate) should scale allometrically with the square of diameter (West et al. 1999). This has received support from surrogates of resource use such as biomass production, water consumption and respiration rates (Enquist et al. 1998, Allen et al. 2005, Meinzer et al. 2005) and foresters have traditionally assumed that total basal area is a good measure of total production or resource use. Therefore we will use this form of the mass scaling combined with the predicted temperature dependence of metabolic rate (Gillooly et al. 2001, Allen et al. 2005) to estimate the energy use of individual trees. Because these allometric relationships are less certain than those for vertebrates, sensitivity analyses will be conducted to confirm that the results of the analyses are not contingent on the specific choice of size and temperature dependence parameters (see Ernest et al. 2009).

Size data: Size data for the tree communities is available at the individual level in the form of diameter at breast height measurements as part of the community databases. In contrast, the bird and mammal community data do not include individual level size measurements. Therefore, to obtain sizes for estimating resource use constraints it is necessary to include organismal databases for birds and mammals that include measures of body mass. Mammalian body mass data will be obtained from the *Body Mass of Late Quaternary Mammals* database (Smith et al. 2003; http://esapubs.org/Archive/ecol/E084/094/). Data on bird masses will be obtained from the *CRC*

Handbook of Avian Body Masses (Dunning 2008). The script for manipulating the electronic version of this publication for use in database systems (already developed by my lab for use in Thibault et al. in review) will be made available on the Ecological Data Wiki (see below).

Taxonomy data: Linking community and body size databases is complicated by differences in the choice of taxonomy and species' coding used by the different databases. Bird taxonomic linkages will be managed using a taxonomy table developed by Dr. Allen Hurlbert's lab in collaboration with my lab (see letter of collaboration from Dr. Hurlbert). This table is based on eBird.com's taxonomy of the birds of North American (http://ebird.org/content/ebird/about/ebird-taxonomy), which is a hierarchical merging of three different authorities: American Ornithologist's Union (AOU), South American Classification Committee, and Clements (2007). The unique species codes used by each database are linked in this table to the specific or subspecific designations used by eBird, thereby standardizing taxonomy across databases to ensure comparability. Mammalian taxonomic linkages will be managed using a customized version of the Mammal Species of the World database (Wilson & Reeder 2005; http://www.bucknell.edu/MSW3/) that will be developed by my lab in collaboration with Dr. Morgan Ernest's lab as part of this proposal (see letter of collaboration from Dr. Ernest). These tables are designed to allow relatively seamless integration of community and organismal databases that use different taxonomic designations. These tables will be made publically available via the proposed Ecological Data Wiki (see below). A taxonomy table is not necessary for the tree data because individual size measurements are available within the community databases.

Environmental data: Environmental data will be used to develop models of the constraints used in the maximum entropy models. The data included in the environmental models will be chosen based on model selection, but I will explore the inclusion of

- temperature and precipitation (e.g., NOAA; http://lwf.ncdc.noaa.gov/)
- production (GLO-PEM; http://glcf.umiacs.umd.edu/data/glopem/)
- NDVI (AVHRR; http://eros.usgs.gov/products/landcover/ndvi.php)
- land cover (e.g., MLCC; http://www.epa.gov/mrlc/)
- biome type, (CEC/EPA; http://www.epa.gov/wed/pages/ecoregions.htm)
- elevation (based on a 30-arc second digital elevation model of North America (http://edc.usgs.gov/products/elevation/gtopo30/README.html)
- measures of elevational and land cover heterogeneity (calculated using spatial windows and the basic CEC/EPA and NALCCDB data).

Integration of environmental and community data will be done using ARCs Geographic Information Systems combined with Matlab's Mapping and Spatial Statistics Toolboxes.

V. EDUCATIONAL PLAN (BROADER IMPACTS)

Answering the kinds of questions described above, and in particularly addressing the generality of macroecological patterns and models, through the use of multiple sites and taxonomic groups, requires the ability to work with large quantities of data. However, most ecologists receive little training in how to use large databases. As undergraduates and graduate students most ecologists (and biologists in general) are required to take courses in mathematics and statistics, but the field lacks an equivalent emphasis on computer programming in general and database management in particular. As a result most ecologists rely on tools such as spreadsheets and high level computing languages (e.g., R, SAS) for managing and manipulating data. This represents a major challenge for ecological research, because many valuable datasets cannot be efficiently manipulated in these environments and in several cases they cannot even be opened in their entirety (e.g., the main tables of the *Forest Inventory Analysis*, ~12 million records, and the *Christmas Bird Count*, ~6 million records, and

Breeding Bird Survey, ~ 5 million records, cannot be opened in Excel). In addition, many of these databases contain extremely large numbers of tables and fields (e.g., CBC contains 126 fields distributed among 7 different tables) that cannot be effectively managed in anything other than a proper database environment. As such, my educational plan focuses on improving the training in, and reducing the barriers to, utilization of informatics infrastructure and large ecological databases.

PREVIOUS BROADER IMPACTS RELATED TO THIS PROPOSAL

I am a founding member of Weecology, a collaborative research group whose goal is to develop better communication and collaboration between empirical and quantitative scientists for the purpose of tackling major scientific challenges in ecology. This group aims to expose field scientists to advanced quantitative and computational methods and provide quantitative scientists with improved knowledge and understanding of ecological systems. In this context, and throughout my graduate school and postdoctoral periods, I have invested substantial effort in training non-technical users to use advanced computational tools. I currently have a postdoctoral researcher in my lab (Dr. Katherine Thibault) whose background is in field mammalogy. Since joining my lab she has received training in advanced database management using MySQL, BASH scripting, PHP scripting, and parallel computing. In addition, I have two undergraduates (1 female) involved in research projects building and utilizing databases, which required both of them to learn to use MS Access linked via ODBC with my labs MySQL database server. Weecology's four current graduate students (3 female, 1 Hispanic, 1 Asian) have all received training with this database system. This system recently enabled a student project using FIA data that could not be completed using R or MS Access because the main table is over 2 GB in size. I also have experience promoting engagement with social media such as wikis and blogs in scientific and academic settings. I set-up and trained on Weecology's wiki, which has had almost 50,000 page views in the one and a half years since its inception. It has served as a valuable collaborative resource for a number of projects within my lab and has facilitated the organization of multi-university collaborative teams related to NCEAS working groups and other collaborative endeavors (including DEB-0827826). In addition I use wikis and other social media tools including blogs in all of my courses (see **Biographical Sketch** for more details).

EDUCATIONAL OBJECTIVES

Educational Objective 1: Improve training for ecologists in both the basic use of relational databases and in more advanced methods including database servers and automation

Excellent short courses targeted towards ecologists have become increasingly available in mathematical modeling, advanced statistical analysis, and even end-user database interface methods that are still under development, but I am aware of nothing equivalent targeted at setting up and utilizing available databases with standard computational tools. In addition, courses on database use and management for non-computer scientists are not available at most universities. I will address this absence of available training in two ways.

Utah State University Courses: First, I will develop two courses at Utah State. The first will be an *Introduction to Computer Programming and Database Management for Biologists* course targeted towards undergraduates at all levels. The course will start with an introduction to simple programming using a high-level programming language (high-level languages are actually easier to learn despite the apparent implications of the name; e.g., R, Matlab, Python). The course will take a student-centered, active learning, approach to teaching this material. Class will be held in a computer lab and will typically consist of a short introduction to a biological problem and one or more affiliated programming techniques. The majority of class time will then be spent with students working in pairs to solve the problem using the relevant programming methods. Problems will be selected based on the specific interests of students in the course (determined using a survey at the

beginning of the semester). The second half of the course will focus on the use of relational databases using MS Access (the most intuitive relational database available). Students will work with existing biological data to build and utilize databases to solve problems related to biological topics. The graduate course will be targeted specifically towards ecologists – *Databases and Informatics for Ecologists*. The course will start with an accelerated version of the second half of the undergraduate course and then cover more advanced methods up through building database servers. The course will also include extensive introductions to major ecological databases and information on their proper utilization. Students will utilize and contribute to the *Ecological Data Wiki & Forums* (see **Educational Objective 3**) while developing independent research projects in small teams.

***Internet based short courses*:** One of the major weaknesses of the short courses, and distributed graduate seminars, currently being used to educate ecologists is that only a small number of individuals can participate in these opportunities. In order to reach a broader audience I will develop two internet-based short courses; an introductory course based on developing and using databases in MS Access and an advanced course on setting up and running a database server, working directly in SQL (with either MySQL or PostgreSQL), and writing scripts in BASH and PHP for quickly setting up databases and manipulating poorly structured ecological data. These self-paced courses will be available to anyone to work on at anytime, which should increase the adoption of proper database methods because individuals will be able to learn (or review) the necessary material while they are working on actual projects. This circumvents another problem with traditional short courses, where attendees often forget what they have learned prior to applying that knowledge to one of their own projects. These short courses will be composed of: 1) video-based introductions to key topics with examples of real world applications; 2) screencasts of specific approaches and methods; 3) problems using publically available ecological databases with answers available so that participants can determine if they have successfully completed the tasks; and 4) concise *How To* sheets to facilitate quick recollection and implementation of common database tasks. Access to additional help and feedback for course participants will be available in a *Database Software* section of *Ecological Data Wiki & Forums* (see **Educational Objective 3**).

Educational Objective 2: Increase the use of available ecological databases by simplifying initial database setup and configuration

Once trained in basic database utilization one of the major limitations to database usage is time. Designing database structure, manipulating poorly configured data, importing and checking the quality of the import involves substantial effort for many databases. In some cases this inconvenience will simply decrease the likelihood that a research project will address a question with five databases instead of one. In other cases, important resources may be ignored or overlooked because using an existing database may be prohibitively difficult or require extensive and laborious manipulation or even reentry. For example, the recently published revision of the CRC Handbook of Avian Body Masses includes an electronic version of the database on CD. Unfortunately the formatting of this version makes the simple import of the data impossible for most ecologists, even those with experience working with simple databases. What is needed to facilitate large cross-dataset analysis is the availability of simple, easy access to data without major investments of time and energy.

This goal is complicated by the fact that ecological databases are not available from a single source, or in a common format. Attempts have been made to centralize data access, but in part because data providers have understandably vested interests in their data, the centralization of major databases has not yet occurred. I propose to circumvent these complex issues by writing and providing *Database Toolkits* that when executed will download the relevant data from the data provider's site, build an optimized database structure on the user's computer, conduct any necessary manipulation of the raw data, and upload the result into the prepared database. Thus, each individual user actually downloads

the data themselves (thus providing an appropriate record of its access and hence importance to the data provider) and processes the data on their own machines. The database toolkits allow this set-up to occur in a rapid, standardized way that has been developed by experts in this type of data management. Thus data cleanup and import tasks that could take an individual days or weeks to perform by hand (e.g., reformatting the CRC handbook of avian body masses electronic files) will be performed in minutes by their computer. The toolkits will be made available on the appropriate pages of the *Ecological Data Wiki* so members of the ecological community can comment on and contributed to this effort. Toolkits will initially be developed in BASH and/or PHP on Linux, and then be translated for use on Windows based systems. Final versions of these toolkits will be built into packages to allow multiple ecological databases to be installed simultaneously. These tools will expose users to the utility of these relatively simple approaches for data management and hopefully inspire more individuals to acquire the requisite skills for developing their own scripts. This has the potential to result in a positive feedback where by the community rapidly develops tools for working with most publically available databases.

Educational Objective 3: Increase knowledge of how to best use ecological databases using social media – Ecological Data Wiki & Forums

The internet culture has now fully embraced the benefits of social media such as Wikipedia, Facebook, and Twitter. However, the sciences in general, and ecology in particular, have been slow to adapt these kinds of tools to facilitate scientific discourse. I will use social media to overcome the final limitation to the broad utilization of available ecological data – a lack of information available to users on how to properly treat the data for analysis, and a resulting lack of confidence on the part of new users that they understand the databases well enough to use them safely. While it is relatively straightforward to train someone to use the software associated with databases and to provide them with database toolkits to do the more difficult work of properly modifying and configuring publically available data, it is difficult for a single individual or even a group of individuals to educate someone in the best practices and idiosyncratic complexities of the numerous publically available ecological databases. To overcome this issue I will implement and host a website that combines a wiki and a set of discussion forums to allow the ecological community as a whole to develop these resources.

The traditional approach to educating users about the proper methods for working with a particular dataset is for the data provider to produce metadata that describes in detail every aspect of the database. One of the problems with this approach is that the metadata available for ecological data is highly variable. When metadata is not complete individual users typically fill in the gaps through correspondence with the data provider, but this results in redundant effort and haphazardly implemented approaches. A larger problem is that this model is a largely unidirectional discussion. The data provider publishes the metadata, which is read by a number of isolated end users. The end users rarely contact the data provider and any conversation among end-users is limited to informal discussions among colleagues. This proposed wiki and forum system opens up the conversation about appropriate use of publically available databases. It will facilitate dialog between data providers and users and among data users. These two groups approach databases from different perspectives, with data users learning things about the data that the providers would not have considered, and data providers possessing a deeper understanding of the methods and limitations of the data. Combining the knowledge of these groups in publically available repositories and providing avenues for questions about data use to be answered and discussions about best practices to occur, will reduce barriers to new users and improve the quality of science produced using publicly available data.

The wiki will (at least initially) be composed of two types of pages:

- ***Individual database pages*:** including descriptions of the data, links to more detailed metadata, lists of best practices, and code (including the *Database Toolkits* described above) written by users to facilitate easy use of the database
- ***Data type pages*:** that list and link to databases of a particular commonly used variety, e.g., time-series data, broad-scale data, mammalian data, etc.

The forums will also be composed of three major areas:

- ***Dataset questions*:** intended for questions related to the use of particular datasets, and the combination of different datasets, as well as for discussions of the proper use of different datasets to help the community reach consensus with respect to best practices
- ***Database design and management questions*:** intended for questions related to the use of relational databases, particular database software, and scripting methods
- ***Feedback and recommendations*:** to provide a the opportunity for users of the wiki, forums, and online short courses to provide feedback and make recommendations for new features or changes to existing features (see *Evaluation of Educational Impact* below)

To facilitate the development and growth of this wiki and forum system the PI, the postdoc, the two graduate students and three additional macroecologists and database providers (see letters of collaboration from Drs. Ernest, Hurlbert, and Kerkhoff) will develop starter articles on a number of major databases to provide enough structure to motivate participation by the broader community. Each of the three external collaborators contributes unique strengths to this project. Dr. Hurlbert is an expert in avian Macroecology, with an extensive background utilizing the BBS, CBC, and numerous avian atlas datasets (Hurlbert & Haskell 2003, Hurlbert 2004, Hurlbert & Jetz 2007). Dr. Ernest is both a prominent mammalian macroecologist and a major provider of ecological databases (Ernest 2003, Smith et al. 2003, Ernest et al. 2009). Dr. Kerkhoff has an extensive background in plant Macroecology (Kerkhoff et al. 2005, Kerkhoff & Enquist 2006, Kerkhoff et al. 2006) and also adds the perspective of a faculty member at an undergraduate teaching institution (Kenyon College). Once the basic site and starter articles are in place the site will be advertised by: 1) Ecolog-L the main ecology list server; 2) A post on White's blog (http://jabberwocky.weecology.org); and 3) direct email to the major database providers (including USGS, Audubon Society, U.S. Forest Service, Missouri Botanical Garden) and to the heads of labs that are expected to find the site beneficial.

Educational Objective 4: Training and interdisciplinary collaboration

Over the course of this proposal multiple postdocs and graduate students (including one graduate student involved in the development of this proposal – Xiao Xiao) will receive extensive training in the use of informatics methods in ecology, cross-disciplinary collaboration, the maximum entropy formalism, and the use of social media for improving dialog among scientists. Recruitment of postdocs and graduate students will occur across disciplinary boundaries. In addition to biologists, I will target recruitment towards individuals with backgrounds in physics, computer science, and mathematics who are interested in studying ecological questions. Undergraduate researchers with backgrounds in computer science will be recruited to work directly with the PI, postdoc, and graduate students on developing and translating database toolkits. The composition of this group will necessitate the development of interdisciplinary collaboration skills and I will work with all members of this group at learning to successfully communicate across disciplinary boundaries and how to teach and learn from collaborators from different backgrounds. I will provide training in advanced database methods and maximum entropy approaches. In addition the postdoc, graduate students, and

myself will have the opportunity to interact with, discuss, and work on projects related to maximum entropy with Dr. John Harte of UC Berkeley (see letter of collaboration from Dr. Harte). The graduate students and the postdoctoral research will be provided with opportunities and mentoring in education. Graduate students will teach one class each in the undergraduate course and the postdoctoral researcher will teach one or more classes in the graduate and undergraduate courses. The students and postdoc will develop the material for these classes with my guidance. This will expose them to the active learning approach to education and give them practice in developing course material. I will observe each of these classes and provide feedback on potential areas for improvement. In addition, the postdoc and the graduate students will be actively involved in generating content for the web based short courses and answering comments on forums. See the ***Postdoctoral researcher mentoring plan*** for more details on mentoring activities designed for the postdoc.

EVALUATION OF EDUCATIONAL IMPACT

Internet based educational activities: Evaluation of internet-based activities will be based on a combination of standard measures of website impact, surveys of users, and a discussion area in the forums to solicit user feedback and recommendations for modifications and additions to the sites. The success of different methods for attracting users will be based on monitoring changes in the number of new visits, page views, and bounce rates following announcement of these resources on different list-serves and blogs. Adoption rates will be based on the number of registered users and the satisfaction of these users will be assessed based on browsing habits (e.g., bounce rates, number of page views per visit, average time on site) and on short online surveys administered once after a user signs up and again after they have utilized the online resources 5-10 times. Usage metrics will be collected using Google Analytics (http://www.google.com/analytics). When appropriate I will setup alternative configurations of the sites designed to explore improvements in user experience. The performance of these different configurations will be experimentally determined using Google Website Optimizer (http://www.google.com/websiteoptimizer), which provides different configurations to different users and collects information on usage statistics for the different designs. The results of these evaluations will be used to iteratively modify approaches. To allow direct community involvement in the development of the social media resources I will implement a *Feedback and Recommendations* section in the forums where users can make recommendations for changes to the sites and requests for new features to be added.

Classroom activities: Evaluation of the Utah State University courses will be conducted with the help of two National Academies Summer Institute on Undergraduate Education teaching fellows, Dr. Gregory Podgorski and Dr. Lianna Etchberger (see the *Departmental Letter of Support*). This evaluation will be conducted in a standard project development/evaluation cycle where feedback from the evaluation results in modification to the course and the evaluation design. The evaluation will involve both student evaluation designed with the help of the NAS teaching fellows. I already include a mid-semester formative evaluation in all of my classes and I will continue to use this methodology in the proposed courses. This formative evaluation is a confidential online survey and request for feedback from my students, who are specifically informed that these recommendations play no role in the judging of my classes so that they should feel free to be completely honest about aspects of the course that should be changed to improve their classroom experience. A summative evaluation will be conducted at the end of the course. In addition, the NAS fellows will attend my class on multiple occasions during the semester and provide feedback and recommendations on the design and implementation of the coursework. Finally, I will evaluate the continued use of the online resources by members of the courses as an indication of continuing involvement in the subject matter.