

Response to Previous Reviews

The previous full proposal panel was enthusiastic that the research would advance the field of community ecology, but concerned about methodological details and whether the scope justified the budget. We addressed these issues by being more explicit about how we will evaluate environment-richness relationships, elevating the examination of spatial scale to a full Aim, eliminating analyses of phylogenetic community assembly, and by reducing the budget to that of a Small Grant. The response of the pre-proposal panel to these changes was extremely positive (all Excellents). The only challenge highlighted by the panel was compiling sufficient data to generalize our analyses across taxonomic groups. We have provided a detailed list of the thousands of communities we will use to accomplish this generalization in Table 1 and the Data Sources section of Aim 2.

I. Conceptual Framework and Specific Aims

Determining the processes governing community assembly and diversity is crucial to understanding and managing ecological systems. However, most studies investigating patterns of species richness fail to recognize a critical insight made 30 years ago by Shmida and Wilson (1985): species at a site typically fall into two distinct groups. *Core species* maintain self-sustaining populations at a site, while *transient species* are poorly suited to the site but are present due to immigration from neighboring source areas. Because the dynamics and diversity of these two groups are expected to be driven by different processes, accurately inferring the importance of those processes requires the explicit consideration of each group separately (Coyle *et al.* 2013). Furthermore, since the ratio of transient to core species within a community may vary geographically and across taxonomic groups, this distinction may provide the key to reconciling conflicting evidence from previous studies regarding the relative importance of local versus regional, and niche versus neutral processes. Comparative, data-intensive studies of the impact of the core-transient distinction are critical for advancing our understanding of diversity and community assembly.

Here, we develop a unique compendium of community time series datasets with which we will distinguish core and transient species and analyze patterns of diversity. Our primary aims are:

Aim 1. Evaluate the generality of the distinction between core and transient species across taxa and ecosystems. We will use data from a wide range of taxonomic groups and ecosystems to determine how core-transient patterns vary across taxa, dispersal mode, and landscape context.

Aim 2. Advance models of species richness by treating core and transient species separately. We will develop models of species richness that allow local and regional variables to exert differential effects on core and transient species. We will compare the resulting parameter values and predictive abilities of these models to traditional approaches that ignore differences between the two groups.

Aim 3. Determine the spatial-scaling of core and transient designations. Understanding differences in the proportions of core and transient species between taxonomic groups and ecosystems will hinge on understanding how the prevalence of these groups varies with spatial scale.

By addressing the important differences in the processes driving core and transient species patterns, this research will produce a better understanding of the linkages between local and regional scale processes in driving patterns of species richness, and how the relative importance of those processes varies across taxonomic groups, ecosystems, and spatial scales.

II. Background and Significance

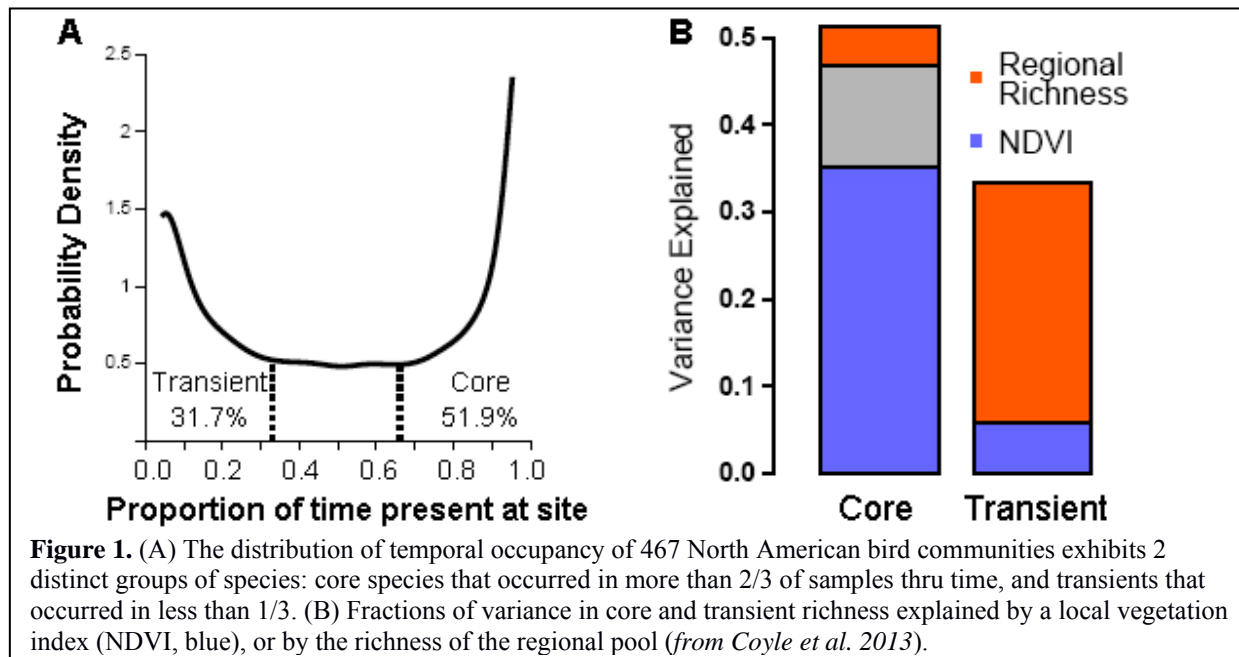
For decades, the most important determinants of species richness were thought to be purely local processes such as competition, predation, and disturbance (Paine 1966; MacArthur 1972; Connell

1978). Even in recent years, many studies have been conducted seeking to characterize the extent to which local biotic or abiotic conditions dictate an equilibrium level or limit to the number of species that might coexist (Srivastava & Lawton 1998; Brown *et al.* 2001; Hurlbert & Haskell 2003; Sanders *et al.* 2007). However, an alternative perspective emerging over the past twenty-five years highlights the importance of regional processes where the richness of local communities primarily reflects the availability of colonists from the regional species pool and the regional-scale variables that affect speciation, colonization, and extinction dynamics (Ricklefs 1987, 2007; Karlson *et al.* 2004). While it is increasingly well accepted that processes operating at both local and regional scales contribute to local richness patterns (Freestone & Harrison 2006; Harrison & Cornell 2008; Hortal *et al.* 2008; White & Hurlbert 2010), exactly how these two scales of processes combine to structure communities and determine species number remains poorly understood. We highlight an important distinction between species that yields novel insights into how local and regional processes combine to structure ecological communities.

Core versus transient species

Species within a community can be divided into core and transient species based on their temporal persistence and ability to maintain viable populations. These two groups are expected to differ substantially in the processes influencing their presence and abundance in ecological systems (Shmida & Wilson 1985; Grime 1998; Novotný & Basset 2000; Magurran & Henderson 2003). In order to maintain a viable population, core species must be able to successfully compete in the local abiotic and biotic environment, and are therefore expected to be strongly influenced by deterministic processes such as environmental filtering and competition. On the other hand, transient species, which do not maintain viable populations, are not expected to exhibit a strong ecological match to their environment. As such, their presence should be primarily influenced by regional processes governing the prevalence of species available to disperse to the site from a broader species pool.

Since the processes influencing core and transient species are expected to differ so strongly, we hypothesize that when these groups can be distinguished, modeling them separately will improve both our understanding of ecological processes and the predictive power of related ecological models (Magurran 2007). This hypothesis is supported by empirical research showing that core and transient species exhibit distinct forms of the species abundance distribution, and modeling the two groups



separately leads to better fits than a single fit to the community as a whole (Magurran & Henderson 2003; Ulrich & Ollik 2004). In addition, regional diversity has a stronger effect on transient species richness than on core species richness (Belmaker 2009). Recently, we conducted the first study of the core-transient influence on richness-environment relationships (Coyle *et al.* 2013). We found that across 467 North American bird communities, most species could be reliably assigned to core and transient categories based on how frequently they occur in the community (Fig. 1A). More importantly, we found that core species richness and transient species richness were best predicted by completely different variables: the number of core species was most strongly related to a local measure of primary productivity in the breeding season, while the number of transient species was most strongly related to regional measures of habitat heterogeneity and the richness of the regional species pool (Fig. 1B). These contrasting patterns are consistent with the predicted differences in the way local and regional processes should affect species richness. If these results apply broadly across ecological systems, then distinguishing between core and transient species has the potential to: 1) change the way we model diversity patterns; and 2) reconcile contradictory views regarding the relative importance of local versus regional influences (Ricklefs 1987; Harrison & Cornell 2008) and niche versus neutral processes (Holyoak & Loreau 2006; Vergnon *et al.* 2009).

Identifying core and transient species

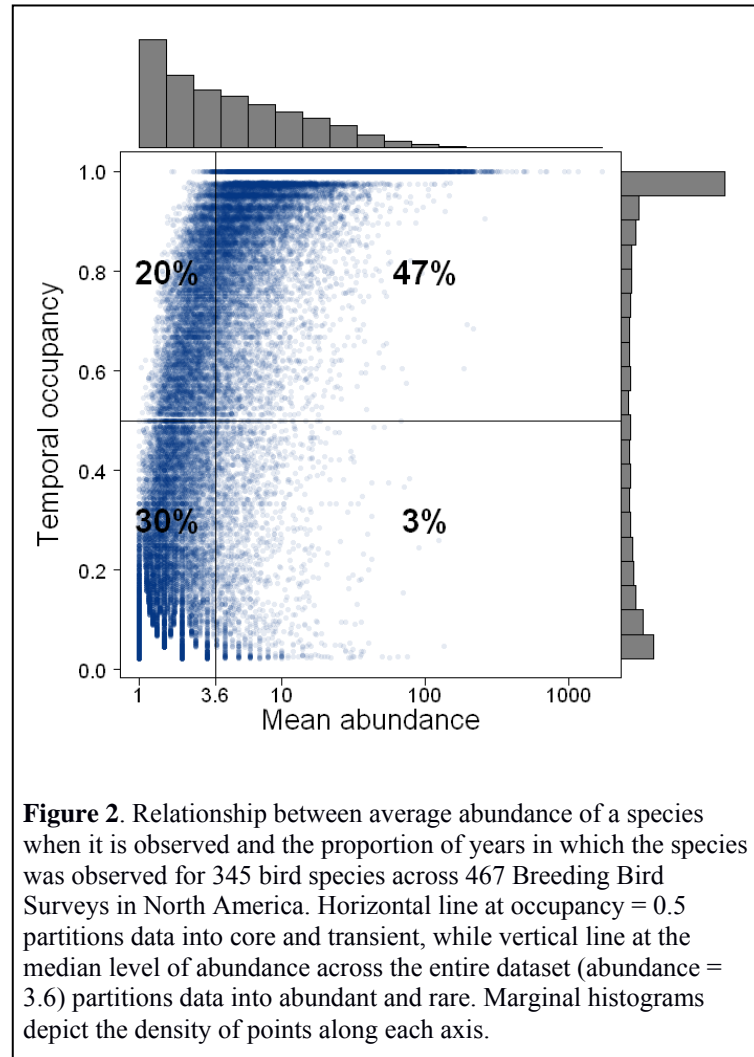
The most common approach for identifying core and transient species uses information on temporal occupancy. The frequency with which a species is observed at a sampling location through time is quantified and then species are divided into groups with high values of occupancy (core species) and low values of occupancy (transient species) (Costello & Myers 1996; Magurran & Henderson 2003; Ulrich & Ollik 2004; Vergnon *et al.* 2009). The distribution of occupancies for North American birds is clearly bimodal (Fig. 1) leading to a natural division of species into two groups. The simplicity of this empirical approach makes the distinction between core and transient species possible for any dataset with a sufficiently long community time-series. This general approach has been validated using independently derived habitat association data in other systems (Magurran & Henderson 2003) and analyses of richness using non-occupancy based designations yield similar results to those using occupancy (Belmaker 2009). We have further validated this approach by comparing the statuses of bird species based on occupancy distributions to independently collected breeding status data from a statewide breeding bird atlas project. Comparing 30 Breeding Bird Survey routes (BBS; see Aim 2) from New York state to high quality blocks in the New York Breeding Bird Atlas (Andrle & Carroll 1988) shows that on average 85% of both core and transient species identified using the BBS occupancy distributions were similarly classified by the atlas data (i.e., as probable or confirmed breeders as opposed to transients).

Related ideas

Core-satellite. Temporal site occupancy shares conceptual overlap with spatial occupancy across a region, an idea which has been investigated extensively in the literature (Raunkiaer 1934; Hanski 1982; Collins & Glenn 1990; McGeoch & Gaston 2002). Frequently, the spatial occupancy distribution is observed to be bimodal as well, prompting the development of the well-known core-satellite hypothesis (Hanski 1982) and other explanations (see review in McGeoch & Gaston 2002). However, spatial and temporal occupancy differ in important respects making these very different areas of research. Under the spatial framework a species is designated as core or satellite over an entire region or continent and this designation is a characteristic of the species. In contrast, the designation of a species as core or transient occurs at the local scale, may vary from site to site, and is a characteristic of a species at a particular location. Narrowly distributed satellite species can still be ‘core’ community members at the sites where they occur, and widely distributed species may be transient visitors to some of the sites at which they are observed. As such, the core-transient distinction is

more directly tied to the dynamics of a local population and its ability to persist in a particular environment.

Abundant-rare. A second distinction related to the core-transient dichotomy is that between abundance and rarity. Temporal occupancy is correlated positively with mean abundance in the North American Breeding Bird Survey (BBS), but imperfectly so (Fig. 2). Preliminary analysis shows that 40% of the variance in occupancy in our data is not associated with abundance (Fig. 2). While many core species are abundant and many transient species are rare, a substantial proportion of species are numerically rare but persistently occurring species, and a noticeable fraction of points also indicate transient species that occur infrequently but are abundant when they do occur (Fig. 2). The most obvious difference between the abundant-rare and core-transient distinctions is that abundance distributions tend to be unimodal (Fig. 2, top histogram), and thus any categorical distinction between abundant and rare relies on choosing a somewhat arbitrary abundance threshold. In contrast, occupancy distributions are often strongly bimodal (Fig. 1; Fig 2, right histogram) lending greater confidence in the existence of two biologically distinct groups.



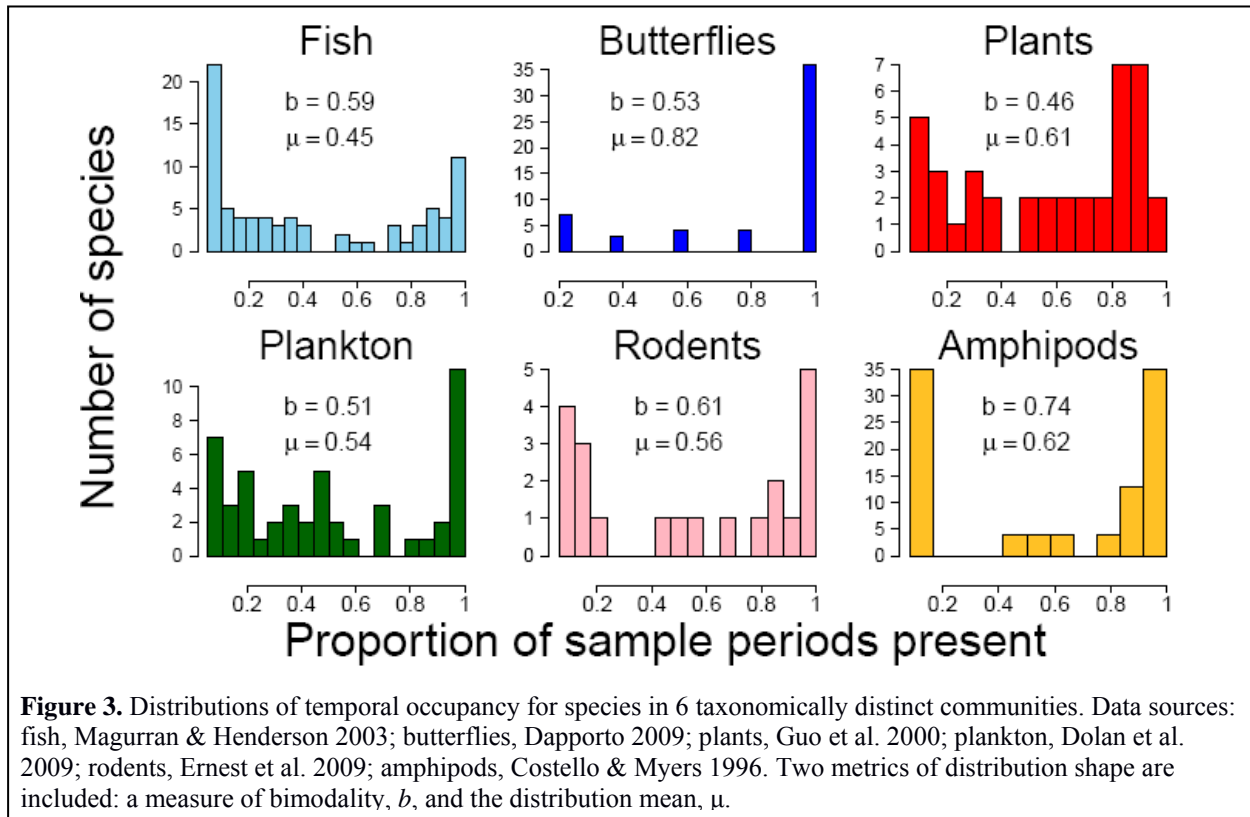
III. Research Aims and Approach

Aim 1. Evaluating the generality of the distinction between core and transient species across taxa and ecosystems

Preliminary data suggests that the bimodal form of the temporal occupancy distribution is consistent across a variety of systems (Fig. 3), and that core and transient species are generally identifiable as distinct groups. However, the exact shape of this distribution and the relative representation of core and transient species varies. We will compile an extensive database of temporal occupancy distributions across a wide range of taxonomic groups and ecosystems to assess general patterns in the shape of the occupancy distribution and refine its use for designating core and transient species.

1a. Does the shape of the occupancy distribution vary predictably with taxonomic group, dispersal ability, or landscape context?

We have already identified approximately 60 datasets with community time series spanning over 5,000 communities (Table 1). These datasets are either publicly available or the PIs have established



relationships with the data holders and a history of access to the data via memoranda of understanding. We will add to this set through targeted literature searches. We will use all communities which have been sampled in at least 10 time intervals (usually years, but potentially months or weeks depending on the taxon). These data, or pointers to them, will be made available through the Knowledge Network for Biocomplexity and the Ecological Data Wiki (see Data Management Plan).

We will measure two critical aspects of the shape of temporal occupancy distributions with the goal of gaining insight into the generality of the core-transient distinction. The first is a measure of bimodality, b , using the variance in occupancy scaled by the maximum possible variance. When all species are evenly split between the highest and lowest occupancy bins, the scaled variance will be 1, while if all species have identical occupancy, the value will be 0. Most of the distributions in Figure 3 are strongly bimodal ($b > 0.5$). The significance of this bimodality will be assessed using methods by Silverman (1981) and Tokeshi (1992). A second measure is whether the weight of the distribution falls towards the core or the transient end, which is reflected in the overall mean occupancy. The fish and rodent datasets in Figure 3 provide examples of two distributions with a similar degree of bimodality but with a notable difference in the side to which occupancy values are skewed.

Our aim here is necessarily descriptive, as we will be the first to explore generalities in the shape of temporal occupancy distributions. We will employ a mixed modeling approach (as in Soininen *et al.* 2007) that examines how the above shape parameters vary as a function of taxonomic group, geographic location, dispersal mode, and environmental variables. We expect the proportion of core species in a community to be related to species traits and environments that reduce the importance of dispersal. Specifically the proportion of core species should be higher for: 1) taxonomic groups that are active rather than passive dispersers; 2) communities in homogeneous rather than heterogeneous environments; 3) for taxonomic groups with lower reproductive output; and 4) for groups in which local community richness is a greater proportion of the regional species pool.

Table 1. Available datasets for investigating temporal occupancy and the core-transient dichotomy.

Taxonomic Group	Number of communities	Maximum time span	Aims	Data sources
Terrestrial birds	~1,700	47 years	1, 2, 3	Beven 1976, Diamond & May 1977, Kendeigh 1982, Williamson 1983, Hall 1984, Svensson et al. 1984, Vickery & Nudds 1984, Knapp et al. 1998, Holmes et al. 2009, Sauer et al. 2011, National Audubon Society 2012
Plants	~2,200	41 years	1, 2, 3	See Table 2
Rocky Intertidal	~1,000	25 years	1, 2, 3	Petraitis & Vidargas 2006, Petraitis et al. 2008, Raimondi et al. 2009
Fish	~120	26 years	1, 2, 3	Grossman 1982, Grossman et al. 1982, Magnuson & Bowser 1990, Henderson & Holmes 1994, Pigg et al. 1998, Sosebee & Cadrin 2006, Sweatman et al. 2008
Terrestrial invertebrates	~130	25 years	1, 2, 3	Novak 1983, Taylor et al. 1990, Pollard et al. 1986, Pollard 1991, Bloch et al. 2007, Ernest et al. 2009, NABA 2009
Small mammals	~30	32 years	1, 3	Ernest et al. 2000, Kaufman et al. 2000, Ernest et al. 2009, Merritt 2009, Stapp 2009, Thibault et al. 2011, Kelt et al. 2013
Plankton	~70	60 years	1, 2, 3	Magnuson & Bowser 1990, Hampton et al. 2008, Yan et al. 2008

1b. Refining core-transient species designations.

Currently we define transient species operationally as those occurring in the bottom third of the occupancy spectrum (≤ 0.33), and core species as those occurring in the top third (≥ 0.67). The small fraction of species in the middle of the occupancy distribution for which there is equivocal evidence and for which classification errors may be greatest are excluded from analyses. This general approach to distinguishing between core and transient species performs well (see Background & Significance) and yields richness model results that are robust to specific occupancy thresholds or exclusion of intermediate species (Coyle et al. 2013). However, no effort has been made to further improve this approach since it was first described over a decade ago. We will further refine the methods for determining the richness of core and occasional species using simulation modeling.

The current approach ignores the fact that some species which maintain viable local populations may be designated as transient because they are rare or otherwise difficult to detect (Coyle et al. 2013). For this reason, estimates of transient richness are expected to be biased high while estimates of core richness are expected to be biased low. We will quantify the degree of misclassification using simulation models and use these results to adjust the estimates of core and transient species at a site. A simulation model for this problem requires 1) distinct suites of species with particular habitat affiliations, 2) long-term persistence at a site dependent on the match between a species and its habitat, and 3) the dispersal of propagules across the landscape. We will use the individual-based model of Smith & Lundholm (2010; available as part of the “neutral.vp” R package) in which individuals reproduce, disperse, and die within a heterogeneous landscape. Individual grid cells will be coded as being one of two habitat types, and species will be assigned “trait values” *a priori* such that they are more suited to (i.e., core species in) one habitat type over the other. Birth and death rates

are functions of the match of an individual to its environment, and dispersal operates as either a random walk (Smith & Lundholm 2010) or a random walk biased toward suitable habitat type. Each grid cell supports a fixed carrying capacity, and if after reproduction and dispersal the total number of individuals exceeds that level, excess individuals are removed at random (Smith & Lundholm 2010). The simulated landscape will be 100 x 100 cells and implicitly lies within a broader metacommunity from which individuals disperse at a fixed migration rate. Each cell will be initialized with an equal number of individuals of each species. After 10,000 time steps, we will examine the temporal occupancy of species within local community samples over a 25-time step period. We will use standard parameter values from Smith and Lundholm (2010), which produce suites of realistic ecological patterns.

We will simulate 10,000 sets of communities and compare core-transient assignments based on the occupancy distribution to their known state. This comparison will yield quantitative estimates of the magnitudes by which transient species richness is overestimated and core richness is underestimated using traditional approaches, and will indicate the range of potential bias over which to conduct sensitivity analyses in Aim 2.

Aim 2. Advance models of species richness by treating core and transient species separately

The status of a species as core or transient is a reflection of the extent to which that species is adapted to the local environment, and also an indication of whether that species is a predictable member of the community and subject to local stressors and resource constraints (Shmida & Wilson 1985; Magurran & Henderson 2003; Ulrich & Ollik 2004). This biological distinction between core and transient species yields five specific predictions about drivers of species richness of the two groups. Our initial work on birds was supportive of the first four predictions (Fig. 1b, Coyle et al. 2013).

H1a) The number of core species at a site will be best predicted by local scale environmental variables such as temperature, productivity, or soil nutrients.

H1b) The number of core species will *not* be directly influenced by regional variables reflecting broad-scale habitat heterogeneity and the diversity of the species pool.

H2a) The number of transient species at a site will be best predicted by regional variables reflecting habitat heterogeneity and the diversity of the species pool.

H2b) The number of transient species will *not* be directly influenced by local environmental variables.

H3) Modeling species richness as the sum of two sets of processes operating distinctly on core versus transient species will yield greater predictive power than approaches that ignore this distinction.

Data sources

In order to assess patterns of species richness, we focus on the subset of studies listed in Table 1 that include data for at least 30 locations sampled in an equivalent manner across an extent of at least 1,000 km. For each dataset, we will use information from previous studies to select variables that best characterize the local environment and regional heterogeneity for that taxon from a suite of remotely sensed environmental data (Table 2), and where appropriate, site-specific meteorological stations. We will also use taxon specific sources for data on the richness of the regional species pool. A brief summary of four terrestrial and four marine systems that will be investigated is presented below.

Table 2. Remotely sensed environmental data for Aim 2. Data will be averaged over the relevant temporal (e.g., seasonal or annual) and spatial scales to match the scale of community data.

Variable	Resolution	Source
<i>Terrestrial</i>		
Temperature	1 km	WorldClim (Hijmans et al. 2005)
Annual precipitation	1 km	WorldClim (Hijmans et al. 2005)
Normalized difference vegetation index	250 m	MODIS/Terra, MOD13
Actual evapotranspiration	1 km	MODIS/Terra, MOD16
Elevation	1 km	USGS GTOPO30
Land cover	30 m	National Land Cover Database (Fry et al. 2011)
Soil type	10 m	Gridded Soil Survey Geographic Database
<i>Marine</i>		
Sea surface temperature	1 km	MODIS/Aqua, MOD28
Chlorophyll- <i>a</i>	1 km	MODIS/Aqua
Organic pollutants	1 km	Halpern et al. (2008)
Inorganic pollutants	1 km	Halpern et al. (2008)
Reef area	1 km	Global shallow bathymetry from SeaWiFS
Ocean currents	1/3 degree	NOAA OSCAR, Bonjean & Lagerloef (2002)

Birds. The North American Breeding Bird Survey (BBS; Bystrak 1981) and Audubon Christmas Bird Counts (CBC; Bock & Root 1981) are long-term, large-scale monitoring programs that provide an unparalleled resource for examining spatial and temporal variation in avian populations and communities during the breeding and wintering seasons, respectively. 467 breeding surveys (25 km² roadside routes) and ~1,200 Christmas counts (452 km² circles) meeting *a priori* quality control criteria have been surveyed continuously between 1996 and 2010 (Fig. 4a). *Local environmental variables* include long-term means for seasonal and annual temperature, precipitation, and vegetation indices as described in Coyle et al. (2013). *Regional heterogeneity* is evaluated by calculating the spatial variance of these local variables over a 100 km radius circle centered on the community of interest. This scale is based on the typical dispersal distances of 100 bird species from two continents (Paradis *et al.* 1998; Tittler *et al.* 2009). The *regional species pool* of a site is determined independently of the survey data using the number of species with range maps that overlap this region (data from NatureServe, Ridgely *et al.* 2007).

Butterflies. The North American Butterfly Association's Fourth of July Butterfly Count (North American Butterfly Association 2009) is a continent-wide monitoring program with ~100 count circles (452 km² each) with at least 10 years of continuous data meeting minimum survey effort criteria (Fig. 4a). *Local environmental variables* expected to be most important for explaining butterfly richness include growing season temperature (White & Kerr 2006) and actual evapotranspiration (Hawkins & Porter 2003). *Regional heterogeneity* measures include measures of land cover diversity (Kerr *et al.* 2001) and spatial variance of local variables over the 12-km radius count area which spans typical dispersal distances (Stevens *et al.* 2010). The *regional species pool* for each site is derived from county level range maps (Opler *et al.* 2013).

Plants. We will compile an extensive set of grassland and shrubland plant community quadrat time series (Fig. 4a, Table 3). When community data are collected as part of an experiment, we will use unmanipulated control plots only. *Local environmental variables* include seasonal and annual measures of precipitation and temperature (available from site-specific weather stations or from sources in Table 2), and for many of the datasets, soil variables such as pH, conductivity, % organic

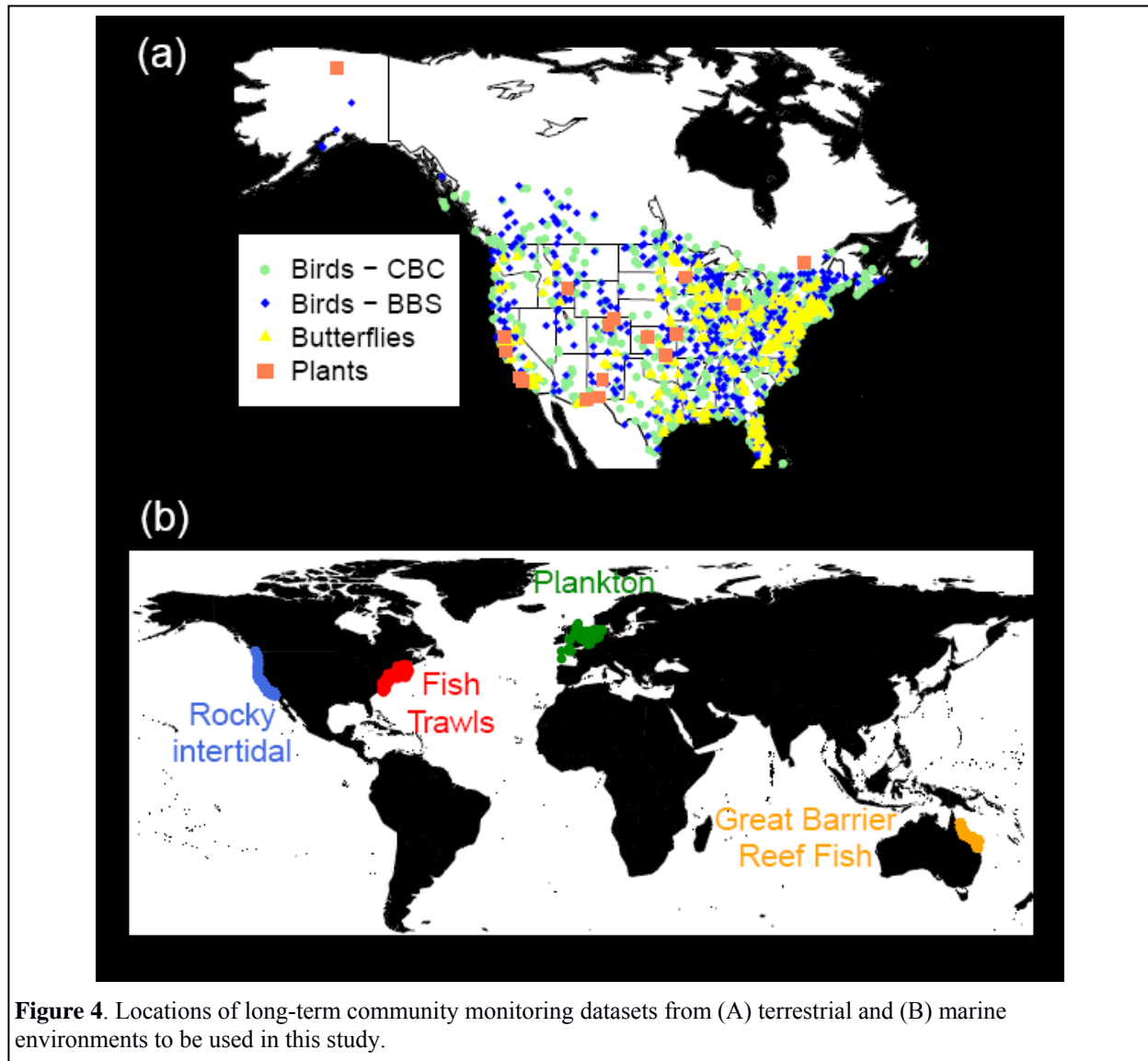


Figure 4. Locations of long-term community monitoring datasets from (A) terrestrial and (B) marine environments to be used in this study.

matter, and mineral content. *Regional heterogeneity* measures include land cover diversity, topographic diversity, and soil diversity from the Natural Resources Conservation Service soils database (Soil Survey Staff 2013), and spatial variance in local environmental variables within a 5 km radius of each focal site. Five km is greater than the dispersal distance of >95% of terrestrial plants surveyed in a recent meta-analysis (Kinlan & Gaines 2003). Finally, the *regional species pool* of a site will be compiled from county-level lists from the USDA Plants Database (USDA, NRCS 2012).

Fish. Fish communities will be examined using two datasets at different spatial scales. The Australian Institute of Marine Science has conducted 20 years of visual fish censuses on 46 reefs spanning ~1,200 km of the Great Barrier Reef (Sweatman *et al.* 2008; Fig. 4b). Each reef is characterized by fifteen 50 m transects at 6-9 m depth. In addition, the Northeast Fisheries Science Center (NEFSC) Bottom Trawl Survey has monitored the species composition and abundance of fish in the Northwest Atlantic from Cape Hatteras to the Gulf of Maine at hundreds of sampling stations since 1948 (NEFSC 1988; Fig. 4b). Data are aggregated into regional strata (~10³ km²) as defined by the NEFSC (Sosebee & Cadrin 2006), with a focus on annual autumnal surveys in offshore areas. *Local*

Table 3. Preliminary list of plant community time series datasets.

Site	Citation	No. plots	Plot size	Years of data	Soil data?
Arctic LTER, AK	Shaver and Chapin 1991	20	0.04 m ²	20	Y
Carpinteria salt marsh, CA	Cleland et al. 2008	25	0.25 m ²	7	N
Cedar Creek LTER, MN	Tilman 1987	36	16 m ²	30	Y
Central Plains Exptl Range, CO	Chu et al. 2013	24	1 m ²	14	N
Channel Islands, CA	Johnson & Rodriguez 2001	163	30 m transect	17	N
Fort Hays, KS	Adler et al. 2007	51	1 m ²	41	N
Jasper Ridge, CA	Zavaleta et al. 2003	96	0.8 m ²	8	Y
Jornada LTER, NM	Nelson 1934, Huenneke et al. 2001	751	1 m ²	32	N
Kellogg Biological Station, MI	Huberty et al. 1998	12	1 m ²	23	Y
Konza Prairie LTER, KS	Gibson & Hulbert 1987	120	10 m ²	25	Y
Lac Croche, Quebec	Paquette et al. 2007	43	400 m ²	9	Y
McLaughlin Reserve, CA	Elmendorf & Harrison 2011	355	1 m ²	10	Y
Niwot Ridge LTER, CO	Theodose & Bowman 1997	40	4 m ²	11	Y
Portal, AZ	Ernest et al. 2009	384	0.25 m ²	30	N
Sagebrush Steppe, ID	Zachmann et al. 2010	26	1 m ²	35	N
Sevilleta LTER, NM	Baez et al. 2006	36	1 m ²	10	Y
Sonoran Desert Lab, AZ	Rodriguez-Buritica et al. 2013	30	100 m ²	20	N
Tall Grass Prairie Preserve, OK	McGlinn et al. 2010	80	1 m ²	12	Y

environmental variables for both datasets include annual mean estimates of sea surface temperature, salinity, nitrate and ocean productivity (estimated from chlorophyll-*a*) which have all been found to be important drivers of fish richness (Mellin *et al.* 2010). In addition, for the coral reef dataset, percent coral cover recorded along the fish transects will be used as a local variable. *Regional heterogeneity* measures will be assessed over a circle of radius 30 km in the Great Barrier Reef, and over a circle of radius 150 km in the North Atlantic. These differing regional scales account for extended larval duration at colder temperatures, and hence greater distances that larvae may be passively dispersed (O'Connor *et al.* 2007), and are also in accordance with published estimates of larval dispersal in the two regions (Kinlan & Gaines 2003; Planes *et al.* 2009). Heterogeneity measures include the spatial variance of local variables over these regions, as well as total regional reef area for the Australian sites. The *regional species pool* will be estimated for each location by overlaying distribution maps of the Australian ichthyofauna provided by FishMap (<http://fish.ala.org.au/>), and by expected species richness based on latitudinal extents for fish in the Northwest Atlantic from FishBase (Froese & Pauly 2012).

Plankton. The Sir Alistair Hardy Foundation for Ocean Science manages the Continuous Plankton

Recorder which has collected data on both phyto- and zooplankton from sample tows at ~10 m depth from throughout the North Atlantic for over 60 years (Richardson *et al.* 2006; Fig. 4b). We will characterize plankton assemblages in the sixty-six 0.5 degree latitude-longitude cells that have at least 10 consecutive years of data with at least 10 sample tows in each year. Each sample tow spans 10 nautical miles. *Local environmental variables* include annual mean estimates of sea surface temperature, chlorophyll-*a*, salinity, and measure of organic and inorganic pollutants from the atlas of human impacts on marine systems (Halpern *et al.* 2008). *Regional heterogeneity* measures include spatial variance in local variables and variance in the source of incoming current directions within a 2-degree block centered on the focal cell. The *regional species pool* for each cell will be defined as the set of species encountered within the regional 2-degree block over the previous 25 years.

Rocky Intertidal. We will use data on the abundance of marine invertebrates from a set of over 1,000 long-term 50 x 75 cm photoplots from 70 sites along the Pacific coast of North America, monitored by the Marine Rocky Intertidal Network and its partners for up to 18 years (Raimondi *et al.* 2009; Fig. 4b). We will conduct analyses separately for invertebrates and plants, as well as an analysis that considers them together. We will use *local environmental variables* following Cruz-Motta *et al.* (2010), including sea surface temperature, chlorophyll-*a*, and organic and inorganic pollution. *Regional heterogeneity* measures include spatial variance in these local variables within 100 km of each site based on published estimates of dispersal distances (Kinlan & Gaines 2003). The *regional species pool* for each site will be based on the total number of taxa expected based on known latitudinal extents of each species as well as habitat preferences (Ricketts 1985).

Statistical methods

We will fit a series of regression models to predict core, transient, and total richness in each dataset based on local environmental variables, regional heterogeneity variables, and all variables combined. We will use random forest modeling (Olden *et al.* 2008) to identify non-linearities in the response of richness to environmental variables as well as important variable interactions. These results will inform the structure (i.e., inclusion of interaction and quadratic terms) of conditional autoregressive models (Wall 2004) that explain spatial patterns of species richness while accounting for the spatial autocorrelation inherent to these data. For datasets that are hierarchically structured in space (e.g., the plant data that has large numbers of quadrats at a single site) we will use hierarchical models to properly account for within versus among site variation (Gelman & Hill 2007). All independent variables will be normalized to z-scores (subtracting the mean and dividing by the standard deviation) to facilitate the comparison of effect sizes for variables with disparate units and values. We will use cross-validation to avoid overfitting in the conditional autoregressive models by evaluating model performance on data that have not been part of the model fitting process (Geisser 1993).

Assessing the relative importance of local and regional variables (Hypotheses 1-2): We will evaluate the importance of different processes for determining core and transient richness using regression models based on only local variables, only regional variables, and all variables combined. All three models will be fit to core and transient richness separately and compared in two ways. We will use the coefficients of determination for the local, regional, and combined models as inputs for variance partitioning analyses (Legendre & Legendre 1998) to determine the relative explanatory power of local and regional variables (White & Hurlbert 2010, Coyle *et al.* 2013). This provides information on the relative importance of the different sets of processes. We will also conduct a stronger test using information theoretic-based model selection, by selecting the best fitting model for each dataset using AIC (Burnham & Anderson 2002). This will determine whether variables that are predicted not to influence a particular group's species richness have no meaningful effect or are simply less important than other variables. In cases where more than one model provides a good fit to the data we will also use model averaging to assess the relative importance of local and regional variables

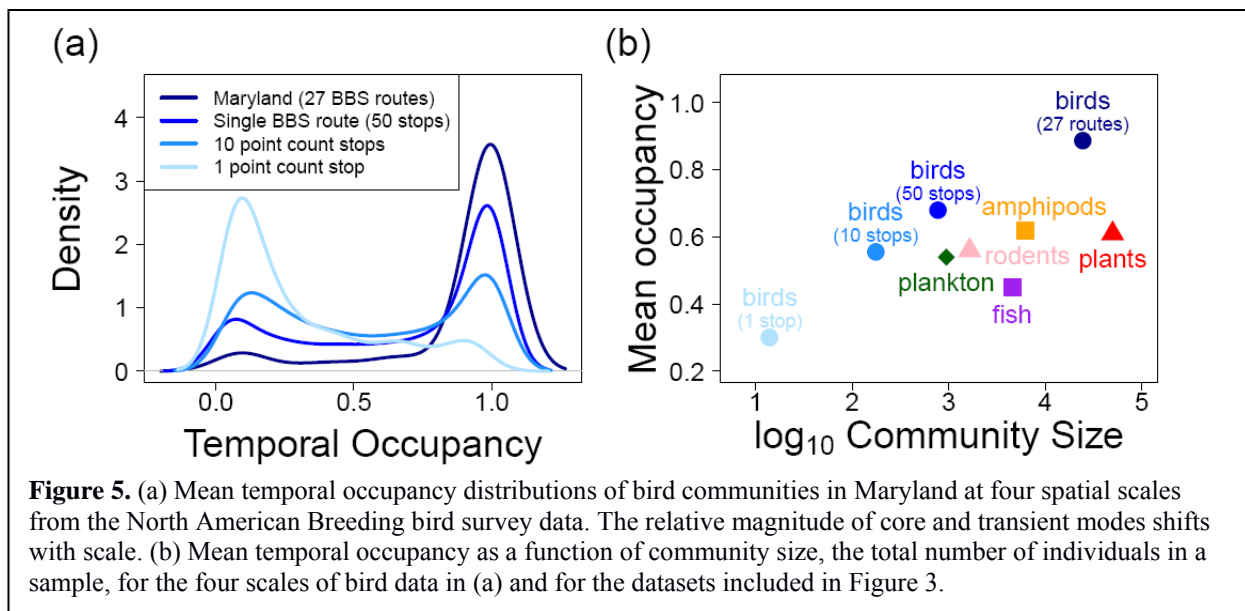
from an AIC perspective (Burnham & Anderson 2002). We will use the results of *Aim 1b* to re-evaluate model performance and variance partitioning using core and transient richness estimates that have been adjusted for sampling bias.

Does the core-transient distinction improve predictive power? (Hypothesis 3): We will test the hypothesis that modeling the two groups separately yields better predictions for overall species richness by modeling total species richness using a combination of local and regional variables and comparing the predictions of this model to those obtained by modeling core and transient species richness separately and summing the predicted richness values. Comparisons will be done using leave-one-out cross validation (LOOCV; Arlot & Celisse 2010). Cross validation is used to assess how accurately a predictive model will perform in practice, and reduces the likelihood of model overfitting leading to superior performance. In LOOCV, the value of species richness for each data point is validated against the prediction from a model based on all data excluding that data point. Models of overall species richness and core + transient species richness will be compared using the mean squared error of cross validation analyses, and analyses will be conducted with and without adjustments for sampling bias.

Aim 3. Determine the spatial-scaling of core-transient designations.

Patterns of diversity change with the spatial scale of analysis due to shifts in the processes that dominate at different scales (Chase & Leibold 2002; Hurlbert & Haskell 2003). The identification of core and transient species will also depend on spatial scale. At continental scales, nearly every species can be considered to have a sustainable population somewhere, and hence all will be core species. Once spatial scales fall below the average home range size of an individual, then species will be inadequately censused in any given year and will all appear to be transients. This calls for a deeper examination of how the shape of the temporal occupancy distribution varies with spatial scale, which we will pursue using both empirical and simulation-based analyses.

A number of the datasets in Table 1 were collected in a spatially hierarchical manner allowing for the natural aggregation of community data at different spatial scales. For example, plant quadrats are frequently arrayed in grids, plankton samples can be aggregated at coarser and coarser resolutions, and coral reef fish can be examined at the scale of a single transect, a group of transects at a site, and the set of sites sampled within a reef. Figure 5 illustrates how occupancy distributions of bird



communities vary with spatial scale using the BBS dataset. A single BBS survey route describes the species seen over 50 point count stops, so we calculated occupancy at the individual stop scale, over 10 adjacent stops, the entire survey route, and over the 27 survey routes within the state of Maryland. At each scale, the distribution is bimodal, but the frequency of core species increases with spatial scale as expected. Within datasets we will characterize this shift quantitatively by using AIC based model selection to identify the best fitting spatial scaling function for both the mean occupancy and bimodality metrics described in Aim 1.

Understanding how occupancy is influenced by scale is particularly important because different taxonomic groups are sampled at different absolute scales, as well as at different scales relative to average body sizes and dispersal distances. We will explore potential differences among groups by comparing the form of the spatial scaling relationships among different taxonomic groups and datasets. We will use these results to help control for scale in comparisons of core and transient designations across taxa. Preliminary analyses suggest that average community size (i.e., number of individuals) serves as an indicator of the taxon relevant spatial scale. We will plot occupancy as a function of the total number of individuals encompassed by the sample (Fig. 5b). In this way, a community of 1,000 plankton individuals in a liter of seawater can potentially be compared to a community of 1,000 birds over 25 km². We will validate this approach (and fine tune it if necessary) using simulations. We will use the simulation model described in Aim 1b to establish a null expectation for the relationship between scale and the proportions of core and transient species in a community. After simulating spatially explicit communities we will examine the temporal occupancy of species within focal regions spanning a 100-fold range of area over a 25 time-step period. We will also conduct simulations in which the spatial scale of analysis is held constant but organism density is varied. In so doing, we will be able to assess the independent contributions of spatial scale and community size in driving occupancy scaling relationships.

Understanding the impacts of spatial scale on occupancy distributions will improve our ability to evaluate differences in these distributions among ecosystems, taxonomic groups, and datasets as described in Aim 1. Characterizing the amount of between dataset variation in occupancy that results from absolute and relative spatial scale differences allows us to focus more directly on environmental and biological drivers of variation in occupancy distributions. To assess the benefits of explicitly considering spatial scale in occupancy distribution based modeling we will rerun analyses conducted in Aim 1a after accounting for spatial scale and evaluate how this influences conclusions about differences among groups and underlying processes.

IV. Project Management and Milestones

Hurlbert and White have a long history of successful collaboration on macroecological research involving large datasets (e.g., Hurlbert & White 2005, 2007; White & Hurlbert 2010; Coyle *et al.* 2013). As such, the PIs will be able to accomplish this substantial ecoinformatics undertaking quickly and productively within the proposed two-year award period. As the lead PI, Hurlbert will oversee the training and supervision of the graduate student, data collection and analysis, and writing of manuscripts and annual reports. Co-PI White will oversee database development, ecoinformatics, and statistical analysis. White will also run a scientific programming workshop each year. The PIs and project graduate student will hold monthly skype conferences to discuss project status and to troubleshoot any obstacles. In addition, White will travel to UNC once per year for more in depth collaboration on analysis and manuscript writing. The graduate RA will be in charge of compiling the data in Tables 1 and 2, searching for additional community datasets, and ensuring all datasets are registered and accessible, where permitted, in standard repositories and data registries. In the final year of the project, the RA will develop and plan the Pre-College Saturday Academy (see Broader

Impacts), and ensure that the developed curriculum is disseminated through the Environmental Educators of North Carolina (www.eenc.org) and the National Environmental Education Foundation (www.neefusa.org). A summary of the expected project timeline is below.

Table 4. Project timeline.

Project task	F 2014	S 2015	F 2015	S 2016
Cross-taxon database compilation	x	x		
Aim 1 - generality of core/transient	x	x		
Aim 2 - species richness modeling		x	x	
Aim 3 - effect of spatial scale			x	x
Software Carpentry workshops		x		x
Pre-College Saturday Academy				x

V. Broader Impacts

Previous activities

PI Hurlbert has successfully mentored four undergraduates (3 of them females from underrepresented groups) in independent research, one of whom recently published results from her project in *PLoS One*. In addition to his graduate mentoring (3 students, 2 female), he co-leads a graduate seminar on using R for data manipulation and analysis in ecology each year. Hurlbert has also demonstrated a commitment to public outreach through Science Day events at local middle schools, and public talks at venues such as the North Carolina Museum of Natural Sciences, local Audubon Society chapters, and the Chapel Hill Bird Club. Co-PI White has been actively involved in providing computational training to biologists. He teaches courses in computational skills to both undergraduate and graduate students at Utah State University and is both an instructor and member of the steering committee for the Software Carpentry project (<http://software-carpentry.org>). As such the researchers involved in this proposal are well suited to training students how to work with large datasets effectively.

Planned activities

Recruitment and training. The work will involve the training of one graduate student in biodiversity science and the use of big data in ecological research. We are committed to recruiting and training young scientists from underrepresented groups for these positions. To insure that a diverse pool is reached, advertisements for the graduate position will be placed with groups focusing on diversity in science, including: ESA SEEDS, the Society for the Advancement of Chicanos and Native Americans in Science (SACNAS), and American Women in Science. The graduate student will receive both formal and informal professional development training at UNC, including a semester-long seminar on Professional Development Skills for Ecologists offered by the department in addition to weekly meetings with PI Hurlbert.

K-12 outreach. We will also engage in outreach activities targeted toward younger students. To facilitate this outreach, in collaboration with the UNC Center for Mathematics and Science Education, we will run a special session of the Pre-College Program's Saturday Academy. The Pre-College Program is an inquiry-based program of enrichment and encouragement in science and mathematics for women and minorities underrepresented in scientific and mathematical careers. Our Saturday Academy session will bring in 200 middle school students from the Chapel Hill-Carrboro City Schools, Durham Public Schools, and Orange County Schools to experience hands-on learning in math and science. In this session we will engage middle school students with their natural surroundings, introduce them to the idea that ecological systems are inherently dynamic, and identify

the forces, such as global climate change and human disturbance, which might contribute to those dynamics. In conjunction with Hurlbert, the project graduate student will develop activities that engage middle school students with their natural surroundings at the North Carolina Botanical Gardens (NCBG), introducing them to the concepts of biodiversity, biotic versus abiotic interactions, and global change. Funds are included in the budget to support these activities.

Computational skills for working with big data. Working with the scale of data in this proposal requires computational and data management skills that most ecologists lack. White is a member of the Software Carpentry team (<http://software-carpentry.org>) that trains scientists in core computational skills and tools, and also teaches computational classes targeted at biologists. This proposal will provide funding (including travel for instructors and some participant support costs) for White to deliver one workshop each year, one at UNC and one at USU.

The project also involves the compilation of a large number of community and trait datasets. To facilitate the use and discovery of these data by other scientists, we will catalog them on both the Knowledge Network for Biocomplexity (<http://knb.ecoinformatics.org/>) and the Ecological Data Wiki (<http://ecologicaldata.org>), a wiki site developed by White for the identification of datasets and the sharing of best practices in their analysis (see Data Management Plan). Public datasets that are not already part of the EcoData Retriever (software development by White's lab for simplifying the use of large datasets; <http://ecodataretriever.org/>) will be added to allow other researchers to quickly download and analyze the data.

VI. Results of prior support

Ethan P. White: NSF CAREER: Advancing Macroecology Using Informatics and Entropy Maximization (0953694; \$657,499, 2010-2015). **Intellectual Merit** – This research develops and evaluates general theories of macroecology to establish linkages between patterns and simplify efforts to make predictions about ecological systems at large scales. Using one of the largest and most diverse datasets ever assembled in community ecology this research has: 1) shown that entropy maximization theories can predict commonness and rarity across ecosystems and taxonomic groups; 2) identified weaknesses with the current theories using strong tests evaluating numerous theoretical predictions simultaneously; and 3) developed new methods for contextualizing the agreement between macroecological patterns and theoretical predictions. **Broader Impacts** – This grant provides training and computational tools to allow ecologists to take advantage of the large amounts of ecological data that are available for analysis and include: 1) software to automatically download, cleanup, and install most of the major macroecological databases (this software is Highly Recommended, Highly Discussed, and Highly Cited based on Impact Story metrics); 2) an advanced wiki-based system to allow crowd-sourced identification and discussion of ecological datasets; and 3) courses taught at Utah State University and nationally through Software Carpentry to train ecologists in cutting edge computational skills. The grant has also supported the training of 2 postdoctoral researchers, 5 graduate students, and an undergraduate researcher. **Publications** – 8 publications (already cited over 30 times) including papers in *PLOS Biology*, *Ecology Letters*, *Ecology*, *American Naturalist*, and *Philosophical Transactions of the Royal Society B*.