**EAGER: Managing our expectations: quantifying and characterizing misleading trajectories in ecological processes**

**Introduction**

A fundamental problem in ecology is understanding how to scale discoveries: from patterns we observe in the lab or the plot to the field or the region, or bridging between short term observations to long term trends and trajectories (Levin 1992, Schneider 2001, Chave 2013). In this proposal, we describe a method to directly address the temporal aspects of scaling ecological observations by leveraging existing data produced at more than two dozen Long Term Ecological Research (LTER) sites, an NSF program in place since the early 1980s. LTER sites have produced time series data documenting various ecological phenomena, some going back nearly 40 years. Findings from these sites have been hugely influential in ecology because of their unprecedented longitudinal perspective (Hughes et al. 2017), yet shorter term studies that are more consistent in length with typical grant cycles and graduate program are still the norm.

We directly address bridging this gap between the short-term and the long-term with an automated approach: in short, we will repeatedly ‘sample’ moving windows of data from existing long-term time series, and analyze these sampled data as if they represented the entire dataset. We will compile typical statistics used to describe the relationship in the sampled data, through repeated samplings, and then use these derived data to gain insights to the questions, *how often are the trends observed in short term data misleading, and can we use characteristics of these trends to predict our likelihood of being misled?* This research presents several opportunities: first, findings will support our efforts to develop a deep understanding of temporal scaling in ecology, aiding in the interpretation of countless future short-term studies. Secondly, this work provides a clear, broad scale test-case of data reuse across an established research network, allowing us to examine if data reuse practices and policies in place are effective (Peters 2010). Finally, findings will have applicability across a variety of domains, including science funding policy, experimental design and interpretation, and data archiving.

**Background**

For many ecological systems (and systems under scientific investigation in general), measurements are taken over time, resulting in time series data. The shape a time series takes can provide meaningful information about the properties of the system- the rules that govern its variability and the trajectory the system is taking (Esling and Agon 2012). The question of trajectory over time is central in ecology, particularly as related to how ecological systems on which humans depend are responding to disturbance or will behave under future conditions (Sutherland et al. 2013). Trajectory is essential to our understanding of ecosystems, their management, and policy decisions, as we interact with our environment.

An ecologist’s interest in system trajectory transcends the ecological sub-disciplines, and the trajectory of a system is often examined in multi-year studies. For example, a three-year study of British ladybeetle communities concluded that native ladybeetle species were in decline, as was total ladybeetle abundance, following the introduction of an invasive species (Brown et al. 2011). Another three year study found that the richness and abundance of seeds in a soil seed bank were still in a recovery trajectory following a period of industrial pollution, which had occurred at the site ten years previously (Wagner et al. 2006). An adventive pest species was implicated in reducing carbon to nitrogen ratios, organic matter in soils of infested forests, all while increasing nitrogen mineralization rates in another three year study, thus substantially changing the ecosystem’s function over time (Orwig et al. 2008). These studies, representing very different ecological domains, have a common element of study duration, which is reflective of how ecological research is done. Patterns in publication (**Figure 1**) suggest that three-year studies dominate the ecology literature in recent years, although longer-term studies are becoming much more frequent. The three-year study duration, reflective of funding cycles or typical graduate program, may be fundamentally out of sync with the processes they aim to understand, from a temporal perspective (Birkhead 2014).



**Figure 1: Ecological studies are getting longer, but three-year studies are still (most) common.** To gain an estimate of typical ecological study length, we searched Google Scholar using the terms ‘“[X] year study” ecology’, bounded for 2003-2018 (solid bars) and 1988-2002 (shaded bars).

Ecological systems are inherently dynamic, and variations in the metrics humans collect about these systems can be driven by a variety of stochastic and deterministic processes (Suding and Gross 2006). Furthermore, short-term dynamics observed in an ecological system are not always indicative of long-term trajectory of that system (Carey and Cottingham 2016). In population processes, for example, density-dependent deterministic mechanisms couple with environmental perturbation to produce highly variable population numbers at any given time slice (Turchin 2003). Decoupling these processes can reveal the skeleton of a deterministic process interacting with external forces. However, to accomplish this disentangling from an empirical standpoint requires substantial data to be collected over time (Higgins et al. 1997).

We can illustrate this decoupling problem with a simple example: the case of fireflies at Kellogg Biological Station in southwestern Michigan (Hermann et al. 2016). In this project, Bahlai and students examined a 12-year time series with two questions in mind:   
  
*When does firefly activity peak?*   
*Are fireflies in decline?*

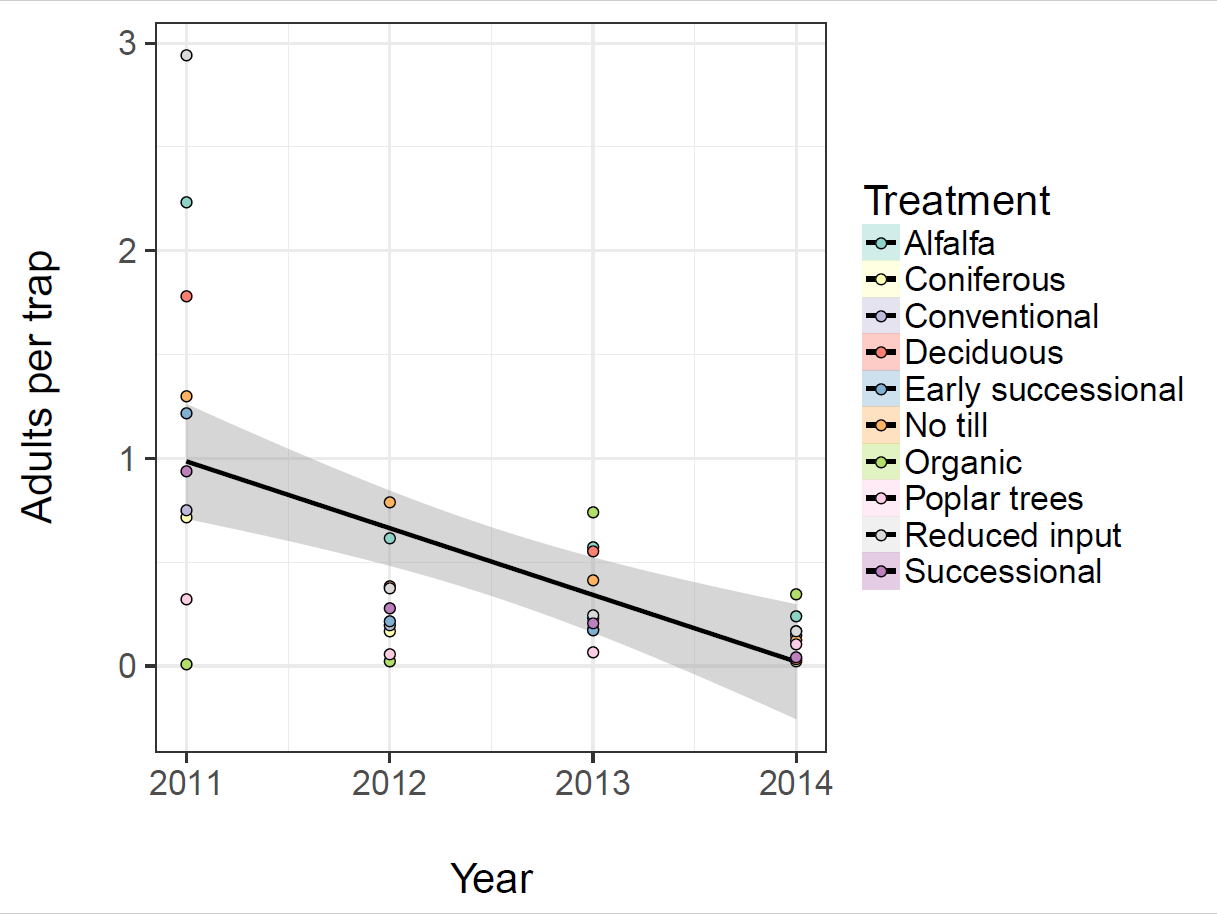
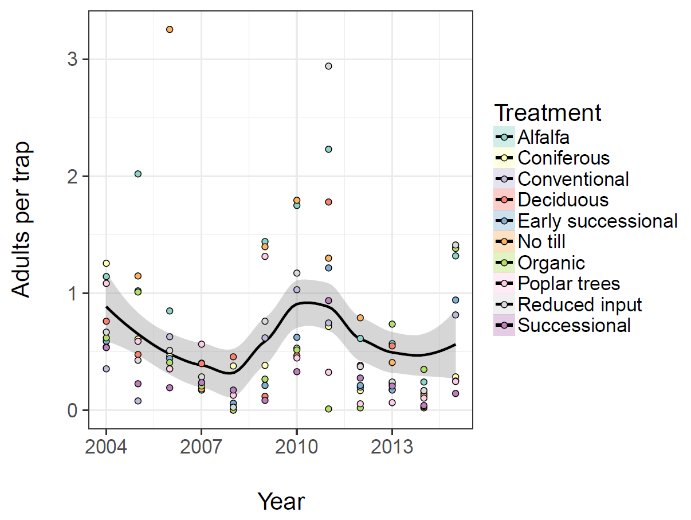
The first question was practical in nature: humans are generally interested in fireflies, and we wished to create a model that would tell us when we could expect the most firefly activity. The second was driven by some concerns raised in the literature that fireflies were indeed in decline (Gardiner 2009, Chow et al. 2014). Yet, we found no evidence of decline over the 12 years (**Figure 2**): there was no significant relationship between average captures and year (*p*=0.32) in the larger time series (2004-2015), and, indeed, there appeared to be evidence of a cyclical dynamic with a ~6y period (**Figure 2A**). However, we were compelled by the contrast we observed between short-term pattern and long-term trends in this system. For example, if we had conducted the study over the four-year period from 2011-2014, we would have had dramatically different conclusions (**Figure 2B**). In this four-year period, we observed significant decline of 0.32+/-0.07 adults per trap per year (*p*<0.0001), and would likely have concluded that fireflies were indeed experiencing a sharp, consistent decline at our study site. Simply, with less data (even with a slightly longer than typical study length), we would have made the wrong conclusions, and we would have been more confident in our wrong answer.

It is because of this phenomenon of “highly-confident wrong answers” that long-term studies are so valued in the ecological community. Ecology, until relatively recently, was a field defined by data scarcity: studies took place at local scales, over time periods manageable to small groups of researchers. The movement we observe towards longer-term studies we observe in **Figure 1** is, no doubt, partially derived from an understanding that many ecological processes observed in short-term studies cannot capture the full extent of these dynamics. Yet, this trend is likely to be influenced by the increasing availability of technology that enables automated collection and sharing of data products, and also, importantly, a function of the infrastructure availability and ‘maturity’ of the US (and international) Long Term Ecological Research networks (Brunt et al. 2002).

The benefits and advantages of long-term research cannot be overstated. Long-term studies are disproportionately represented in policy reports and in the ecological literature: studies involving long term observations are cited more often than studies of shorter duration (Hughes et al. 2017). Long-term ecological research provides insight into the inherent variability of natural systems (Lovett et al. 2007), and insights are thus often only apparent after many years of study (Knapp et al. 2012). Indeed, because biological systems are often defined by their variability, when studies are shown to be irreproducible, it is not necessarily due to poor research practice, but due to their inability to capture the full variability of the system within the limits of the study design (Jarvis and Williams 2016, Voelkl and Würbel 2016). Furthermore, long-term observational studies provide important baseline data: as the world itself changes, these data provide insight into how ecosystems function, instead of studying phenomena after they happen (Magurran et al. 2010).

A

B



**Figure 2: Same data, different observation periods, different conclusions.** Firefly populations monitored in ten plant community treatments at Kellogg Biological Station in southwestern Michigan cycle over an approximately 6 year period (panel A). Yet, if sampling had only occurred over a 4 year period, we would conclude the population underwent a steep (and statistically significant) decline in the four years from 2011-2014 (panel B). Data and figures adapted from Hermann et al (2016).

However, long-term studies themselves are not immune to uncovering misleading trends. For example, a 2002 study uncovered a significant multi-year cooling trend, from 1986-1999 in Antarctica’s McMurdo Dry Valleys (Doran et al. 2002). Yet, recent years have seen temperatures stabilize and increase, and correspondingly, increasing stream flow and decreases of thickness of ice in glacial lake systems (Gooseff et al. 2017). This highlights the importance not just of study duration, but of the selection of study starting and ending points: capturing an outlying data point or a high or low in a system’s natural variability near the beginning or end of the study period will be highly influential on the statistical outcome, and thus the conclusions reached. Understanding and characterizing these highly influential observations in the analysis process is essential to our interpretations of these ecological trajectories.

In general, scientists involved with long-term research make data resources available to the broader public because they understand the value of these practices. Although there are exceptions amongst the broader community of long-term researchers (Mills et al. 2015), data reuse is generally supported, encouraged and valued by LTER sites (Servilla et al. 2016), enabling a synthetic approach to ecological problems (Brazel et al. 2000, Kratz et al. 2003, Redman et al. 2004, Knapp et al. 2012, Kratz and Strasser 2015, Stoll et al. 2015, Swallow and Liu 2017). It is widely recognized that data sharing is necessary to support synthetic approaches (i.e. an approach that brings together multiple disparate resources), harnessing an integration of a wide variety of data sources is necessary to reconcile the issue of scale in ecology (Peters 2010). Furthermore, the challenge of scaling ecological findings has been hampered by a lack of facilitation of knowledge and data sharing to enable synthesis (Jackson and Füreder 2006).

Scaling between the short-term study and the long-term trajectory of a system is a fundamental problem in ecology, and is essential to maximize the utility of observations made in shorter-term studies. Patterns observed in local scale, short-term ecology tend to be dominated by stochastic forces, making generalizations, extrapolations and predictions difficult at larger scales, yet are essential to capture fine-scale understanding of system dynamics (Willis and Birks 2006, Chave 2013). To maximize the impact of research at any scale, it is essential to study the patterns that occur as we move between scales- so we can more meaningfully extrapolate between them. We have established the fundamental importance of long-term studies to the understanding of ecological systems. Yet, no one has yet empirically examined the converse: just how frequently are we misled by short term studies? And furthermore, can we use knowledge generated by studying the relationship between short- and long-term studies to bridge our interpretations of short-term data to long-term processes? We will use a synthetic, computational approach to address these and the following questions:

* What is the effect of observation period length on the likelihood of uncovering misleading trends?
* What is the effect of observation period starting point on the likelihood of uncovering misleading trends?
* How do system properties (landscape, site, seasonality, lifespan in the case of organisms, management regimes) affect the above?
* What are the characteristics of time periods that buck the greater trends observed in long term data, and conversely, what are the characteristics of time periods that are consistent with longer system trends?
* Can we separate trends in ecological systems from natural variability and underlying processes?

**Hypotheses**

Although the core of this work is pattern and question driven, we have several specific hypotheses that we will test:

*Shorter observation periods will increase the likelihood of observing misleading trends*

Because exogenous forces are of greater influence at smaller spatial and temporal scales, we predict that short time periods will be more variable due to these processes, and conversely do not capture the full extent of natural variability (Suding and Gross 2006, Lovett et al. 2007), so they are more likely to result in “highly-confident wrong answers.”

*Statistical metrics often used as a proxy for ‘confidence’ in short term trends (such as the p-value) will not be associated with an increased likelihood of capturing a time period consistent with long term trends.*

Following from the previous prediction, we predict that p-values will be inferior predictors of the ‘correctness’ of short-term trends in predicting longer term trajectory compared to other properties of the system. Better predictors may include statistical measures (slope, standard error), and system specific predictors (e.g. site, data type).

**Intellectual merit**

This study will result in the advancement of understanding of how data-driven approaches can be used to foster discovery in ecology, and thus has both data science and disciplinary applications. In ecology, it is widely recognized that understanding long term and broad scale processes are important, but scaling between the fine and the broad remains an open challenge (Wolkovich et al. 2014). Similarly, many scientists agree that data archiving has the potential to advance research, particularly in conservation and environmental sciences, where evidence-based decision-making is essential to human well-being (Bruna 2010, Hughes et al. 2017). This study will demonstrate a deeply synthetic approach to data reuse with applications in the practical understanding of ecological systems. We will specifically develop paths to data reuse, using best practices; and disciplinary, i.e.: how can we interpret trajectories in ecological systems, given the properties of the system? The disciplinary findings of this study have the potential to be transformative, and of broad influence, because they will provide an empirical framework for interpreting how we understand trends in ecology. The data synthesis aspect will provide a clear demonstration of downstream data use to data producers, and help demonstrate challenges in data reuse and identify best practices to support future synthesis.

**Approach**

We will approach this problem with an analytical approach consisting of four main steps:

1. Gather appropriate datasets

2. Clean and process data into a usable standard form

3. Develop a ‘moving window’ analysis algorithm

4. Examine trends in analysis algorithm outputs.

*1. Gather datasets*

We will use data collected at Long Term Ecological Research sites, which are, as a matter of network policy, publicly available (Karasti et al. 2006, Michener 2015). We will select data which has been taken longitudinally, and documents at least 12 years of continuous study. We will consider two domains of data in our initial synthesis: organismal abundance (numbers per observation, area occupied) and greenhouse gas emissions (CO2, NO3 concentration measurements). Organismal abundance is of interest for study because we expect the data has an underlying density-defendant structure (Turchin 2003). Greenhouse gas emission is of interest because its data tends to be highly variable (Steinmann et al. 2014), so there’s a high probability of uncovering misleading trends when captured in short time periods. Data from a single treatment, over time, will be considered an ‘observation’ for the purposes of this analysis. Repetitions (within a treatment) will be treated as subsamples. Site of collection will be recorded in the data compilation as a covariate.

*2. Clean and process data into a usable standard form*

Although many of these LTER-associated data sets are subjected to greater-than average quality control, significant ‘heterogeneities’ (as one study euphemistically described the difficulties with LTER data reuse) exist because of changing technology, personnel, and variations in management among the sites (Karasti and Baker 2008). We anticipate these issues will include both human and mechanical errors, lost samples or missing data, sample frequency variation, outliers, and other issues. Our goal for this step is to take data from this raw form to a format usable in our downstream analysis: a meaningful yearly metric for each response variable. The end product will be a derived dataset in a simplified (year, response) format, comparable across experiments.

Data manipulation will be achieved through a series of custom R scripts. All data will be downloaded directly from public sources and manipulated within R to document the full, repeatable data cleaning and preparation process. The scripted data manipulation approach will both enable future scientists to repeat or build on our analysis process, and provide an outline for the narrative describing our process for the purpose of data management outreach. We will write about our process for each data set used in Bahlai’s established data management blog, Practical Data Management for Bug Counters. Cleaned, derived data will be exported to a non-proprietary format and packaged with relevant metadata, and deposited in an appropriate data repository.

*3. Develop ‘moving window’ analysis algorithm*

The moving window analysis algorithm will be implemented as a series of functions written in R. Data will first be subjected to a standardization algorithm to normalize the data and make it possible to compare datasets with responses of very different magnitudes, and minimize the impact of measurement unit choice on the observed trends. Then, we will create a function that fits a linear model to the data and produces a vector with the particular summary statistics of interest, such as the slope of the relationship between the response variable and time, the standard error of this relationship, and then *R2*and *p*-values. Although *R2*and *p* are not measures of statistical confidence per se, they are often used by ecologists in this way (Yoccoz 1991, Nakagawa and Cuthill 2007), and thus can be used as a means to approximate ‘conclusions’ that a human researcher might make of the data. We will then create a moving window function that takes our data and iterates through it at defined intervals, feeding each interval to the fitting function described above, and compiling the fit statistics for each into a single object. We will define intervals of 3, 4, 5, 6, and 10 years, and subject the data to the moving window function for each of these intervals.

We will subject each of our derived data products to this analysis algorithm and package outputs with information that will be relevant for downstream analysis (such as site of collection, response variable being measured, landscape type, and similar). Fit data will then be will be exported to a non-proprietary format and packaged with relevant metadata, and deposited in an appropriate data repository.

*4. Examine trends in analysis algorithm outputs.*

Each output observation (that is, each start-point, interval, and dataset combination) will be associated with a long-term dataset. We will compare the output observations to trends in the long-term dataset in several ways. First, we will visualize the long term dataset to classify its patterns, if any. Then we will perform a regression analysis, as described above, on the entire long-term dataset, which will serve as our measure of overall trajectory and variability in the dataset’s response variable. For each short-term trend, we will compare the magnitude and direction of slope, the standard errors, and fit statistics to those same metrics for the long-term dataset. Essentially, we will create models that use the statistics produced by the analysis of long-term data as the response variables, and statistics produced by the analysis of the short-term data as predictor variables, with characteristics of the data, the short-term interval, and the system as covariates. We will subject these models to model selection to determine parameters and interactions of importance, and use the results from the model selection to identify the characteristics of short term datasets that cause them to be more (or less) predictive of long term trajectories.

**Results from Prior NSF support**

Bahlai, Stack Whitney and Perrone have not been the direct recipients of NSF funds, although Bahlai and Perrone have received funds indirectly through the DEB-LTER program as paid employees of a recipient institution (Michigan State University). Perrone served as a research technician partially appointed to LTER data collection and management, and Bahlai as a postdoctoral research associate based at Kellogg Biological Station, and as a delegate LTER scientist funded by the Network Communications Office to travel to the 2016 iLTER Open Science Meeting in Skukuza, South Africa. Bahlai’s LTER associated research has produced four publications examining ecosystem service delivery and temporal trends in human managed ecosystems (Bahlai et al. 2013, 2014, 2015, Hermann et al. 2016).

**Broader impacts**

**Enhanced infrastructure for research and education** The vast array of existing NSF data, such as those produced by the LTER program, is incredibly valuable for a wide variety of ecological investigations, but largely underutilized by scientists outside the organizations producing these resources. This underutilization likely stems less from a lack of awareness or lack of perception of value, and more from a lack of visible entry points to other scientists. Our goal, with the proposed work, is to provide a framework for downstream use of data produced by others, so that our investigation can meaningfully be used as a starting point for other scientists. Our commitment to ‘radical openness’ of data, intermediate data products, and analysis code not only will ensure reproducibility of the present work, but provide future scientists with both clear entry points, and a model for building their own open work with this communal data resource.

**Broaden dissemination to enhance scientific and technological understanding.** Bahlai and Stack Whitney are vocal and public proponents of open science approaches, and through this work, we will demonstrate the use of open analytical workflows (i.e. reproducible coding, open data, open note-booking) to mainstream academic biologists- particularly as the challenges arise from using data produced by others. In addition to technical documentation (as described in the Data Management Plan), we will also conduct specific outreach on the topics of data reuse. Specifically, Bahlai will use her blog, “Practical Data Management (For Bug Counters)” as an existing, established platform for dissemination of accessible, plain language documentation, commentary and engagement around the challenges arising data reuse. Bahlai’s blog has previously been used as the basis for lesson materials (i.e.: Data Carpentry’s Spreadsheets in Ecology lessons, much of Bahlai’s own Reproducible Quantitative Methods graduate class).

**Broaden participation of under-represented groups.** Funding to Bahlai, Stack Whitney and Perrone, all early-career women scientists, will enable us to rapidly establish high profile research programs in our respective areas of expertise. Additionally, computational and quantitative ecology are male-dominated fields, and the visibility of woman scientists is essential to the recruitment of people across genders and underrepresented minorities in these fields.

**Rationale for EAGER funding**

This project is an ideal fit for the EAGER Program, particularly as it intersects with *NSF 18-060: DCL-Advancing Long-term Reuse of Scientific Data- Data reuse track*. This project has a meta-scientific and computational theme not usually funded by other NSF programs: we propose to re-analyze data using simple computational methods. We use a novel and unusual approach that would appear too risky to conventionally funded programs because it collects no new data, involves synthesizing data from multiple disparate sources, and uses pattern-driven inferential methods better suited to questions of scale (Michener and Jones 2012). The proposed work does not guarantee a solution to the problem of predicting long-term trends from short-term observations. Indeed, it is possible that we may discover that there is simply no discernable patterns or consistent characteristics of short term trends. However, this work will allow us to directly quantify how frequently misleading trends occur in nature, helping to guide our interpretation of, and extrapolation from, shorter term studies, specifically addressing the issue of scale in ecology (Schneider 2001, Parsons et al. 2014, Spasojevic et al. 2016). Our proposed work specifically addresses the criterion to*“[i]nvolve, for data proposed for use, publicly-available data generated through NSF funding*” by harnessing data produced by LTER sites in our analysis. We have the infrastructure and expertise in place to address the call’s requirement to “[*a]gree to make public the details about their experiences reusing the data, including especially challenges associated with that reuse*” through Bahlai’s blog and outreach in the data management sphere.