Statement of authorship: SC and JH conceived the idea with encouragement from NMH. CB developed the moving window algorithm. SC and JH gathered datasets. SC ran algorithms, compiled data, and executed analyses. All the authors discussed the results and took part in writing the manuscript.

Data accessibility statement: Data are publicly available in the LTER portal (https://portal.edirepository.org/nis/home.jsp)

Title: How long do population level field experiments need to be? A meta-analysis across the 40-year old LTER network

Authors: Sarah Cusser[1]; Jackson Helms IV[1]; Christie Bahlai[1,2], Nick M. Haddad[1]

[1] W.K. Kellogg Biological Station, Department of Integrative Biology, Michigan State University, Hickory Corners, MI, USA

Sarah Cusser: sarah.cusser@gmail.com

Jackson Helms IV: jacksonhelmsiv@gmail.com

Nick M. Haddad: haddad@kbs.msu.edu

[2] Department of Biological Sciences, Kent State University, Kent, OH, USA

Christie Bahlai: cbahlai@kent.edu

Running Title: How long do experiments need to be?

Keywords: population dynamics, time series, data mining, trajectory, moving window, isothermality, long-term

Type of Article: Letters

Number of words in abstract: 149

Number of words in main text: 3,837

Number of references: 42

Number of figures: 3

Number of tables: 1

Corresponding Author: Sarah Cusser, Kellogg Biological Station, Michigan State University, 3700 Gull Lake Rd., Hickory Corners, MI, 49060; sarah.cusser@gmail.com

Conflict of Interest: Authors have no conflict of interest to declare.

ABSTRACT

Long-term experiments are important in evaluating ecosystem properties and processes that are slow to develop or require proper evaluation over an appropriately variable climate. We repurpose the wealth of data accessible through the forty-year-old Long-Term Ecological Research (LTER) network with a novel moving window algorithm and meta-analysis approach to ask if aspects of study taxa or environment alter the extent of research necessary to detect consistent results, or the proportion of spurious short-term trends. We found that experimental studies focused on plants, and those conducted in dynamic abiotic environments, were characterized by longer critical temporal thresholds and more spurious trends. Further, nearly half of the studies we investigated required 10 years or longer to reach a temporal threshold, and 4 studies (of 100) required longer than 20 years. We champion long-term data and argue that long-term experiments are more necessary than ever to understand, explain, and predict long-term trends.

1. INTRODUCTION

Long-term experiments are essential in the study of ecology: critical in isolating and understanding the ecological consequences of global land use and climate change (Del-Val & Crawley 2005; Haddad *et al.* 2015; Gonzalez *et al.* 2016; Hughes *et al.* 2017; Van Klink *et al.* 2020). Long-term data are especially important in evaluating ecosystem properties and processes that require proper evaluation over an appropriately variable climate or are slow to develop (Tilman *et al.* 1994; Rasmussen *et al.* 1998; Knapp *et al.* 2012). However, for a variety of reasons, short term experiments are the benchmark in ecology. Short term experiments, which are more consistent with typical grant cycles and graduate programs, are important for identifying ecosystem-related changes in a timely and cost-effective manner. Despite this, research conducted at constrained time scales has the potential to be misleading, either capturing spurious short-term trends or failing to detect trends at all (Bahlai *et al.* 2020; Cusser *et al.* 2020). If ecosystem properties, processes, or particular taxa are slow to change, develop, or become apparent to observers, lagged responses may lead to inappropriate assessments of experimental outcomes over short periods. As such, temporally restricted research may merely capture a snapshot of ecosystem properties as they gradually respond to manipulation (Hanski & Ovaskainen 2002; Helm *et al.* 2006; Knapp *et al.* 2012; Jarvis & Williams 2016; Voelkl & Würbel 2016). Rarely is data collected at time scales that can either be examined to instill confidence in proposed long-term trends or determine the extent of misleading short-term trends. One place where this is possible, and is the focus of our study, is in the forty-year old Long-term Ecological Research (LTER) network. The LTER network not only provides a ‘sandbox’ in which to examine long-term responses to experimental manipulation, it also allows us to

68 contextualize shorter term studies by parsing apart ephemeral, lagged or spurious responses
69 from those that are genuine changes in system behavior.
70 Long-term studies are essential in determining experimental outcomes in dynamic
71 environments that require evaluation over an appropriately variable climate (Ives & Carpenter
72 2007). Ecological systems are inherently dynamic, and variation can be driven by a variety of
73 stochastic and deterministic processes (Folke 2006; Suding & Gross 2006; Hastings 2010;
74 Beckage *et al.* 2011). The findings of short-term experimental studies may be the product of
75 these processes, and thus are not always indicative of the long-term trends of that system
76 (Turchin 2003; Carey & Cottingham 2016). For example, a four-year study of firefly populations
77 located in a dynamic Midwestern environment could have concluded that local populations
78 underwent drastic changes in abundance over a short period. Yet, with longer observations of
79 that same population, researchers found that variation was short lived and that populations
80 underwent no significant change over a longer period (Hermann *et al.* 2016; Bahlai *et al.* 2020).
81 In this case, a shorter study could have resulted in highly-confident, though thoroughly
82 misleading conclusions.
83 Further, population abundance may respond slowly to experimental manipulation, only
84 reaching a delayed response after some temporal threshold is met for that particular taxa
85 (Krauss *et al.* 2010). These temporal thresholds are likely to be closely linked to taxa specific
86 life-history traits, including generation time, dispersal and colonization ability, and dormancy
87 periods, among others. For example, if long-lived plants can survive initial experimental
88 disruption, changes in plant population abundance may take many generations to become
89 apparent, even if the immediate results are measurable in reduced individual fitness (Tilman *et*

al. 1994; Cousins *et al.* 2007; Ellis & Coppins 2007; Gustavsson *et al.* 2007; Jackson *et al.* 2009; Haddad *et al.* 2015). Also, a plant's seed bank may further prolong the lag in response, replacing individuals lost in the adult population following disturbance (Plie *et al.* 2017). Some animals, due to their relatively short generation time, high mobility, and potential to track resources in novel environments, may respond more rapidly to manipulation (Kuussaari *et al.* 2009; Krauss *et al.* 2010), and may consequently not require long experiments to confidently determine consistent results from manipulation.

We focus our quantitative synthesis on a single response in experimental studies: population level organismal abundance. While patterns of abundance are themselves a fundamental issue in ecology, they also underlie some of the most basic questions in the field and have been used to develop hypotheses concerning species response to climate change, identify probable locations of pest outbreaks, and choose the location of natural reserves (Elton & Nicholson 1942; Altieri *et al.* 1984; Pounds *et al.* 1999; Sagarin *et al.* 2006). Given that measures of organismal abundance are relatively quick to execute, consistently apparent between observers and years, and an intuitive measure of population condition in some systems, abundance is a regularly collected and relatively comparable metric between studies.

Here we make two hypotheses concerning organismal abundance: H1) If studies take place in highly variable environments, with increased system-specific abiotic variation, then studies of those systems will require longer periods of study to detect consistent results, and result in a higher proportion of spurious short-term trends, than those studies in environments with more consistent abiotic variables. H2) If taxa have long generation times or low dispersal and colonization abilities, then studies of those taxa will require longer periods of study to detect

consistent results, and result in a higher proportion of spurious short-term trends than taxa with shorter generation times, high mobility, and potential to track resources in novel environments. Specifically, we predict that studies taking place in highly variable abiotic environments, with large temperature and precipitation oscillations throughout the year, will require longer study periods to reach consistent results, and be more often characterized by misleading results than more abiotically stable systems. Second, we predict that experiments investigating plant taxa will require longer periods of study to form confident conclusions, and be more often characterized by high proportions of misleading short-term trends, than experimental studies focused on animal taxa. To test these predictions, we use a moving window algorithm and meta-analysis approach to repurpose the wealth of data across studies of organismal abundance accessible through the forty-year-old Long-Term Ecological Research (LTER) network. We champion the importance of long-term data and posit that long-term experiments are more necessary than ever to understand, explain, and predict long-term trends.

2. METHODS

2.1 Meta-Analysis and Moving Window Approach

We searched the forty-year-old Long-term Ecological Research database network portal (https://portal.lternet.edu/nis/home.jsp) to identify and repurpose relevant long-term experimental datasets reporting organismal abundance. We systematically explored each of the 6,957 unique datasets, from 30 locations that were available as of December 2018. Only datasets that met five requirements were included in our analysis: 1) research lasted ten years or longer; 2) included at least ten years of data, and data could be expressed as a summary

metric at a yearly resolution; 3) documented a press experiment (Bender *et al.* 1984) in which treatments could be divided into a 'control' and 'treatment' category; 4) treatment response was recorded as a measure of organismal abundance; and 5) the dataset included at least one statistically significant linear relationship over time, described below. Some study sites contained multiple datasets documenting organismal abundance time series, and some datasets quantified multiple taxa responses to the same experimental manipulation. Time series were divided into the finest taxonomic resolution available for analysis (i.e. order, genera, species, or morphospecies). Whenever possible, each organism within each dataset at each site was analyzed separately.

Because the fundamental response we sought to examine was the difference between treatments, we calculated effect size, yearly, for each organism time series by treatment pair. For each year of each dataset we calculated effect size as Hedges' g. As such, effect size was calculated as: $[x_t - x_c] / SD_p$, where x_t is the average treatment population size in that year, x_c is the average control population size in that year, and SD_p is the pooled standard deviation of that year. Hedges' g corrects for bias due to small sample size (Rosenthal *et al.* 1994).

To understand the period of time needed to identify long-term trends, we applied a moving window algorithm developed in R (Bahlai *et al.* 2020). First, we fit linear models to defined subsets of each dataset and produced summary statistics of interest (i.e. slope of the relationship between Hedges' g and time, standard error of this relationship, and p-value). The algorithm then iterated through each dataset at set intervals. We used moving windows of three-year periods or longer, fed each interval through the algorithm described above, and compiled resulting summary statistics for each study. Thus, we examined, and compiled, every

possible subset of at least three years duration or longer. As described above, only datasets that met our requirements were included in our analysis. To comply with our fifth requirement, that all datasets contain at least one linear relationship over time, we removed the 22 datasets that otherwise met our requirements but lacked any significant linear trend, as this situation would indicate there is no change in the difference between treatments over time for any possible study duration. No adjustments were made for multiple statistical comparisons in our analysis as each linear regression was considered in isolation, as a hypothetical observation period which an observer would use to reach conclusions regarding system behavior, from non-independent but still separate experimental durations. Conceptually, we were interested in the trajectory of the relationship between Hedges' g and time, and how linear regression model outputs vary with sample period duration.

2.2 Response Variables: Extracting temporal thresholds and percent spurious trends

With trends from each study plotted against corresponding window length, we extracted a critical temporal threshold from each moving window plot. First, we sorted studies into those with long-term trends (i.e. those with a significant trend for the entire dataset or an overall increase or decrease in abundance over the study period, Fig. 1a) and those without (i.e. those with no significant trend for the entire dataset, Fig. 1b). For each dataset with a long-term trend, we determined the minimum number of years until all trends agreed with the long-term trend (i.e. all trends of that duration are significant and in the same direction as the long-term trend). For each dataset lacking a long-term trend, we determined the minimum number of years to avoid all spurious results (i.e. all trends of that duration are not significant). As such, both datasets with and without long-term trends were scored for a critical temporal threshold.

We also calculated the percent of spurious results found in each dataset. For datasets with long term trends, spurious trends were those periods with significant trends in the direction opposite to the long-term. For datasets which lacked long term trends, any significant results were considered spurious. Percent spurious trends were determined for each study as the number of spurious trends / total number of trends (Fig. 1).

2.3 Categorical and Continuous Explanatory Variables

To address our first hypothesis, concerning the extent of abiotic variation of each study site, we extracted WorldClim v2 Bioclim variables for each LTER location (Busby 1991), including Isothermality, Precipitation Seasonality, and Annual Precipitation. We use the BIOCLIM variable of isothermality as a single measure of temperature fluctuation at each of our sites, as it quantifies how large the day- to-night temperatures oscillate relative to the summer- to-winter (annual) oscillations. As such, sites with low isothermality, were located in the most variable abiotic environments. To address our second hypothesis, concerning slow to change properties or processes, we divided datasets into two categories: those focused on plants and those focused on animals.

2.4 Analysis

We screened continuous explanatory variables (i.e. Isothermality, Precipitation Seasonality, and Annual Precipitation) for multi-collinearity using the 'vifstep' function in the R statistical package 'usdm' (Naimi 2015). Because climatic variables are inherently correlated, we chose a conservative theta value of VIF < 2 to eliminate collinearity from our models (O'Brien 2007). With any collinear variables removed, we use generalized linear mixed models (GLMMs) to determine the relationship between categorical (i.e. plant or animal) and continuous

200 explanatory variables and our two response variables (i.e. critical temporal thresholds and
201 percent spurious trends) using the 'glmer' function in the package 'lme4' (Bates *et al.* 2014). We
202 used explanatory variables as fixed effects and 'LTER dataset' nested within 'LTER Site' as a
203 random intercept. Continuous climatic variables were scaled around zero to account for
204 differences in magnitude and Poisson and Gaussian distributions were used for critical temporal
205 threshold and percent spurious results, respectively. We then used sample-size corrected
206 Akaike Information Criterion (AICc, Burnham & Anderson 2004) to select among all possible
207 combinations of the fixed effects using the 'dredge' function in the R package 'MuMIn' (Barton
208 & Barton 2015). To characterize the top models, we used the function 'model.avg' in the
209 package 'MuMIn' (Barton & Barton 2015) to average models within an AICc of less than 2. We
210 applied a Poisson GLMM, and tested for overdispersion, of which we found no evidence.

211 **3. RESULTS**

212 *3.1 Meta-Analysis and Moving Window*

213 We found 100 datasets from 28 distinct studies and 12 LTER sites that met our five
214 requirements (Fig. 2, ESM table 1). Another 22 datasets met our first four requirements, but
215 lacked any significant linear trend. Because these datasets essentially never reached our criteria
216 for stability regardless of the study duration, they were not likely to result in a consistent
217 difference between treatments over time. Before the removal of these 22 datasets, we
218 analyzed all 122 timeseries that met the first four requirements. Results followed identical
219 patterns with and without the 22 datasets, thus we felt confident in our decision to remove
220 them from analysis. A few of the datasets also had disjunct sampling schedules that included
221 outlying terminal data points, sampled several years after the most recent continuous sampling

year. To ensure continuity with other studies, in these cases, the last, sequentially disconnected, datapoint was removed from the time series before datasets were included in analysis. Experiments ranged from the exclusion of herbivores (Sevilleta and Short Grass Steppe) and predators (Plum Island), to manipulating moisture (Konza Prairie and Sevilleta), nutrients (Arctic, Cedar Creek, Hubbard Creek, and Plum Island), pH (North Temperate Lakes), and temperature (McMurdo Dry Valleys), as well as deliberately altering species diversity (Cedar Creek), or removing plants by fire, grazing (Konza Prairie), trimming (Luquillo), mechanical tillage (Kellogg), or some combination thereof. For a full list of LTER sites and experiments involved in our analyses see ESM table 1.

3.2 Response Variables: Extracting temporal thresholds and percent spurious trends

Of the 100 datasets, we found 24 studies with a significant trend for the full dataset (change in abundance over the full study period), and 76 studies without a significant trend for the full dataset. Whether significant or not, it took 9.66 years on average (SE: 0.52, range 3 to 32) to achieve consistent results. On average, 11.7% (SE: 1.1%, range 0.7 to 47%) of significant trends derived from subsets of sampling years were spurious, not agreeing with the long-term pattern of the data.

3.3 Categorical and Continuous Explanatory Variables

Across datasets, precipitation seasonality averaged 59.52 mm (SE: 2.39 mm) and ranged from 9.4 mm (Plum Island) to 102.9 mm (Santa Barbara). Annual precipitation averaged 616.87 mm (SE: 39.9 mm) and ranged from 68 mm (McMurdo Dry Valleys) to 2,115 mm (Luquillo) and Isothermality averaged 37.65% (SE: 1.22%) and ranged from 21.44% (McMurdo Dry Valleys) to 71.52% (Luquillo). Of the three continuous variables, ‘Precipitation Seasonality’ was found to be

collinear and was consequently removed from further analysis (Isothermality, VIF = 1.02; Annual Precipitation, VIF = 1.02). Of the 100 timeseries that contained at least one significant trend, 56 focused on animal abundance and 44 on plant abundance.

3.4 Analysis

Interpreting the results of our best performing models (Table 1), we found an interaction between our categorical variable (plant or animal) and one of our continuous variables (isothermality) in explaining the length of critical temporal thresholds (Fig. 3a). We found that plant studies had longer temporal thresholds, especially in highly dynamic environments, than animals. In terms of spurious results, our best model found that studies focused on plants were characterized by significantly more spurious results than those that focused on animals (Fig 3b). On average, 15% of significant plant trends were spurious, compared to only 9% of animal trends, nearly a two-fold increase. We found that abiotic factors did not significantly influence the percent of spurious results.

Addressing our first hypothesis, we found that both isothermality and the plant/animal distinction contributed to the length of critical temporal threshold (top model, AICc: 603.6). As the next best model had an AICc of 3.35 greater, the single lowest AICc model is our best for explaining critical temporal thresholds (Table 1). Addressing our second hypothesis, we found that the plant/animal distinction was the best predictor of percent spurious trends. The next best model had an AICc value of 8.97 greater than the top model and, as above, the single lowest AICc model is our best model for explaining percent spurious trends (Table 1).

4. DISCUSSION

We found support for most of our predictions: experimental studies focused on plants, and those in dynamic abiotic environments, were generally characterized by longer critical temporal thresholds and a greater proportion of spurious trends. We also championed the importance of long-term data. First, for every 1% increase in abiotic variation (1% decrease in isothermality), we saw a 0.1-year (1.2 months) extension of the critical temporal threshold across taxa. Interestingly, we found that increased isothermality did not increase the proportion of spurious results, as we had expected. Second, we show that plant studies require longer critical temporal thresholds than animals, especially in highly dynamic (low isothermality) systems and that plant studies were characterized by a nearly two-fold increase in the proportion of spurious results, with 6% more misleading trends on average. Most importantly, we underscore the importance of long-term data. We see that nearly half (46/100) of the studies we investigated require 10 years or longer for relationships between treatments to reach a temporal threshold where stable relationships occur, and 4 studies required longer than 20 years.

We found that studies taking place in highly variable abiotic environments required the longest periods of study to reach consistent results. As such, those sites located in the most dynamic abiotic environments (those with low isothermality) required the longest periods of evaluation. For example, studies undertaken at the Cedar Creek and Arctic LTERs, which are characterized by the strongest seasonal extremes in our study, also had the longest critical temporal thresholds (32 and 16 years, respectively). Given their abiotic variation, these systems may have required longer sampling efforts to capture the entire range of climate variation. In fact, some of the datasets that lacked long term trends may merely have been the product of a truncated sampling effort, and that as the LTER network continues to age, these

trends may emerge with the continued collection of appropriate data. For example, while only 12 of 52 studies sampled less than 15 years were found to have consistent long-term trends, we confirmed consistent trends in more than half of the studies that lasted longer than 25 years. We also show that every study investigated contained at least one spurious trend, and most studies (63%) had more than the expected number of false positives, or type I error, expected at the traditional 0.05 alpha threshold, that is, the expected error rate on a linear regression applied to independent observations. Although we acknowledge that use of time series tools would mitigate the likelihood of these assertions, ecologists frequently do apply linear statistical models to temporal processes, increasing the likelihood of spurious interpretations of these statistical patterns (Yoccoz 1991, Nakagawa and Cuthill 2007, Bahlai et al 2020)

We found that experiments investigating plant taxa require longer periods of study to form confident conclusions, and were more often characterized by high proportions of misleading short-term trends than those studies focused on animals. We hypothesize that our findings reflect specific life history traits of both plants and animals. Some animals, due to their relatively short generation time, high mobility, and potential to track resources in novel environments, may respond rapidly to experimental manipulation (Kuussaari *et al.* 2009, Krauss *et al.* 2010), and consequently not require long experimental periods to confidently determine results from manipulation. Plants on the other hand, with potentially longer generation times, lower dispersal and colonization abilities, and long dormancy periods, may respond more slowly to experimental manipulation and be more characterized by spurious results, only reaching a consistent, delayed response after some temporal threshold is met (Krauss *et al.* 2010). While we do not directly measure the life history traits that may prove most important in altering the

rate of response to manipulation (i.e. dispersal ability, generation time, dormancy period, etc.), as a *post hoc* analysis, we determined the average size of each organism under study (height of each plant and length of each animal). We investigated whether organismal size could serve as a proxy for the life history traits that may contribute to the rate of experimental response. While we found that plants were three times larger than animals on average (ANOVA, F value = 20.65, P <0.001), we did not find that size was a predictor of either temporal threshold or percent spurious trends.

Delayed reactions are critical to consider from a conservation or management perspective, as slow to detect results following experimental manipulation may lead to inappropriate assessments of the status of a population's abundance. For example, a macro-alga (*Stephanocystis osmundacea*) at the Santa Barbara Coastal LTER, only responded to experimental manipulation after six years of continuous plant removal, and only became consistent in the direction of its response after eight years of manipulation. In the presence of these delayed reactions, researchers may either over (or under) estimate the effects of experimental manipulation on organismal abundance in habitats that may not support them in the long-term (Hanski & Ovaskainen 2002; Helm *et al*. 2006). In the case of macro-algae, researchers may have concluded that plant removal had no effect on population abundance if research had not continued until the eighth year.

Ecologists often work at five broad levels: organismal, population, community, ecosystem, and biosphere. While the focus of this meta-analysis is on the population level metric of organismal abundance, our technique is applicable to higher level community or ecosystem processes. For instance, future meta-analyses should focus on taxonomic or functional richness,

diversity, or evenness at the community level, or biogeochemical processes at the ecosystem level, all of which are available in the forty-year-old LTER network portal.

Given the extent of ongoing global land use and climate change, long-term experiments are more necessary than ever to understand, explain, and predict long-term trends. With global climate change increasing abiotic variability worldwide, results from short term studies may become increasingly unreliable in the face of global climate change. New efforts should work in parallel, coordinating network wide experiments and syntheses across ecosystems and climates. Understanding the relationship between transient and long-term dynamics is a significant challenge that ecologists must tackle, and long-term experiments will be essential for relating observation to theory now, as well as in the future.

5. ACKNOWLEDGEMENTS

Support for this study was provided by the National Science Foundation Long–term Ecological Research Program (DEB 1832042) at the Kellogg Biological Station, the National Science Foundation Directorate for Computer and Information Science and Engineering (OAC 1838807), USDA National Institute on Food and Agriculture, and by Michigan State University AgBioResearch.

REFERENCES

Altieri, M.A., Letourneau, D.K. & Risch, S.J. (1984). Vegetation diversity and insect pest outbreaks. *CRC Crit Rev Plant Sci*., 2, 131-169.

Bahlai, C., White, E., Perrone, J., Cusser, S. & Whitney, K.S. (2020). An algorithm for quantifying and characterizing misleading trajectories in ecological processes. *BioRxiv*

Barton, K. & Barton, M.K. (2015). Package 'MuMIn'. Version 1.18.

353 Bates, D., Mächler, M., Bolker, B. & Walker, S. (2014). Fitting linear mixed-effects models using
354 lme4. *arXiv preprint arXiv:*1406.5823.
355 Bender, E.A., Case, T.J. & Gilpin, M.E. (1984). Perturbation experiments in community ecology:
356 theory and practice. *Ecology*, 65, 1-13.
357 Burnham, K.P. & Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in
358 model selection. *Sociol. Methods Res.*, 33, 261-304.
359 Busby, J. (1991). BIOCLIM-a bioclimate analysis and prediction system. *Plant Prot. Q.*, 61, 8-9.
360 Carey, C.C. & Cottingham, K.L. (2016). Cross−scale Perspectives: Integrating Long−term and
361 High−frequency Data into Our Understanding of Communities and Ecosystems. *Bull.*
362 *Ecol. Soc. Am.*, 97, 129–132.
363 Cousins, S.A., Ohlson, H. & Eriksson, O. (2007) Effects of historical and present fragmentation on
364 plant species diversity in semi−natural grasslands in Swedish rural landscapes. *Landsc.*
365 *Ecol.*, 22, 723– 730.
366 Cusser, S., Bahlai, C., Swinton, S.M., Robertson, G.P. & Haddad, N.M. (2020). Long−term
367 research avoids spurious and misleading trends in sustainability attributes of no−
368 till. *Global Change Biol*.
369 Del-Val, E.K. & Crawley, M.J. (2005). Are grazing increaser species better tolerators than
370 decreasers? An experimental assessment of defoliation tolerance in eight British
371 grassland species. *J. Ecol.*, 1005-1016.
372 Ellis, C.J. & Coppins, B.J. (2007) 19th Century woodland structure controls stand scale epiphyte
373 diversity in present day Scotland. *Diversity and Distributions*, 13, 84– 91.

Elton, C. & Nicholson, M. (1942). The ten-year cycle in numbers of the lynx in Canada. *J. Anim. Ecol.*, 215-244.

Folke, C. (2006). Resilience: The emergence of a perspective for social–ecological systems analyses. *Glob. Environ. Change*, 16, 253-267.

Gonzalez, A., Cardinale, B.J., Allington, G.R., Byrnes, J., Arthur Endsley, K., Brown, D.G. *et al*. (2016). Estimating local biodiversity change: a critique of papers claiming no net loss of local diversity. *Ecology*, 97, 1949-1960.

Gustavsson, E., Lennartsson, T. & Emanuelsson, M. (2007) Land use more than 200 years ago explains current grassland plant diversity in a Swedish agricultural landscape. *Biol. Conserv.*, 138, 47– 59.

Haddad, N.M., Brudvig, L.A., Clobert, J., Davies, K.F., Gonzalez, A., Holt, R.D. *et al*. (2015). Habitat fragmentation and its lasting impact on Earth's ecosystems. Sci. Adv.

Hanski, I. & Ovaskainen, O. (2002). Extinction debt at extinction threshold. *Conserv. Biol.*, 16, 666-673

Hastings, A. (2010). Timescales, dynamics, and ecological understanding. *Ecology*, 91, 3471-3480.

Helm, A., Hanski, I. & Pärtel, M. (2006). Slow response of plant species richness to habitat loss and fragmentation. *Ecol. Lett.*, 9, 72-77.

Hermann, S.L., Xue, S., Rowe, L., Davidson-Lowe, E., Myers, A., Eshchanov B. & Bahlai, C.A. (2016). Thermally moderated firefly activity is delayed by precipitation extremes. *R. Soc. Open Sci.*, 3, 160712.

Hughes, B. B., Beas-Luna, R., Barner, A.K., Brewitt, K., Brumbaugh, D.R., Cerny-Chipman E.B. *et al*. (2017) Long-term studies contribute disproportionately to ecology and policy. *BioScience*, 67, 271-281.

Ives, A.R. & Carpenter, S.R. (2007). Stability and diversity of ecosystems. *Science*, 317, 58-62.

Jackson, S.F., Walker, K. & Gaston, K.J. (2009). Relationship between distributions of threatened plants and protected areas in Britain. *Biol. Conserv.*, 142, 1515-1522.

Jarvis, M.F. & Williams, M. (2016). Irreproducibility in Preclinical Biomedical Research: Perceptions, Uncertainties, and Knowledge Gaps. *Trends Pharmacol. Sci.*, 37, 290–302.

Knapp, A.K., Smith, M.D., Hobbie, S.E., Collins, S.L., Fahey, T.J., Hansen G.J.A. *et al*. (2012). Past, present, and future roles of long-term experiments in the LTER Network. BioScience. 62, 377–389.

Krauss, J., Bommarco, R., Guardiola, M., Heikkinen, R.K., Helm, A., Kuussaari, M. *et al*. (2010) Habitat fragmentation causes immediate and time delayed biodiversity loss at different trophic levels. *Ecol. Lett.*, 13, 597– 605.

Kuussaari, M., Bommarco, R., Heikkinen, R., Helm, A., Krauss, J., Lindborg, R. & Öckinger, E. (2009) Extinction debt: a challenge for biodiversity conservation. *Trends Ecol. Evol.*, 24, 564– 571.

Naimi, B. (2015). usdm: Uncertainty analysis for species distribution models. *R package version*, 1, 1-12.

Nakagawa, S. & Cuthill, I. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* 82, 591–605.

O'Brien, R.M. (2007). A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.*, 41, 673-690.

Plie, L.S., Wang, G.G., Stovall, J.P., Siemann, E., Wheeler, G.S. & Gabler, C. A. (2017). Mechanisms of Chinese tallow (Triadica sebifera) invasion and their management implications–a review. *For. Ecol. Manag.*, 404, 1-13.

Pounds, J.A., Fogden, M.P. & Campbell, J.H. (1999). Biological response to climate change on a tropical mountain. *Nature*, 398, 611-615.

Rasmussen, P.E., Goulding, K.W.T., Brown, J.R., Grace, P.R., Janzen, H.H. & Korschens, M. (1998). Long-term agroecosystem experiments: Assessing agricultural sustainability and global change. *Science*, *282*, 893–896.

Rosenthal, R., Cooper, H. & Hedges, L. (1994). *Parametric measures of effect size. The handbook of research synthesis.* Russel Sage Foundation, New York City, 621, 231-244.

Sagarin, R.D., Gaines, S.D. & Gaylord, B. (2006). Moving beyond assumptions to understand abundance distributions across the ranges of species. *Trends in Ecol. Evol.*, 21, 524-530.

Suding, K.N. & Gross, K.L. (2006). The dynamic nature of ecological systems: multiple states and restoration trajectories. *Foundations of Restoration Ecology*. Island Press, Washington D.C. 190–209.

Tilman, D., May, R.M., Lehman, C.L. & Nowak, M.A. (1994). Habitat destruction and the extinction debt. *Nature*, *371*(6492), 65-66

Turchin, P. (2003). *Complex population dynamics: a theoretical/empirical synthesis*. Princeton University Press.

van Klink, R., Bowler, D.E., Gongalsky, K.B., Swengel, A.B., Gentile, A. & Chase, J.M. (2020). Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. *Science*, 368, 417-420.

Voelkl, B. & Würbel, H. (2016). Reproducibility Crisis: Are We Ignoring Reaction Norms *Trends Pharmacol. Sci.*, 37, 509–510.

Yoccoz, N.G. (1991). Use, Overuse, and Misuse of Significance Tests in Evolutionary Biology and Ecology. *Bull. Ecol. Soc. Am*. 72, 106–111.

6. **TABLES**

Table 1: Parameter estimates for the best-performing models explaining critical temporal threshold and percent spurious trends

Critical Temporal Threshold

Fixed effects	Estimate	Std. Error	Z Value	P Value
Intercept	2.0092	0.1131	17.76	<0.001
Isothermality	-0.208	0.0791	-2.629	0.009
Plant/Animal	0.4876	0.1414	3.449	<0.001

Percent Spurious Results

Fixed effects	Estimate	Std. Error	df	T Value	P Value
Intercept	0.07891	0.02086	4.46615	3.783	0.016
Plant/Animal	0.08541	0.0288	10.93134	2.966	0.013

7. FIGURES

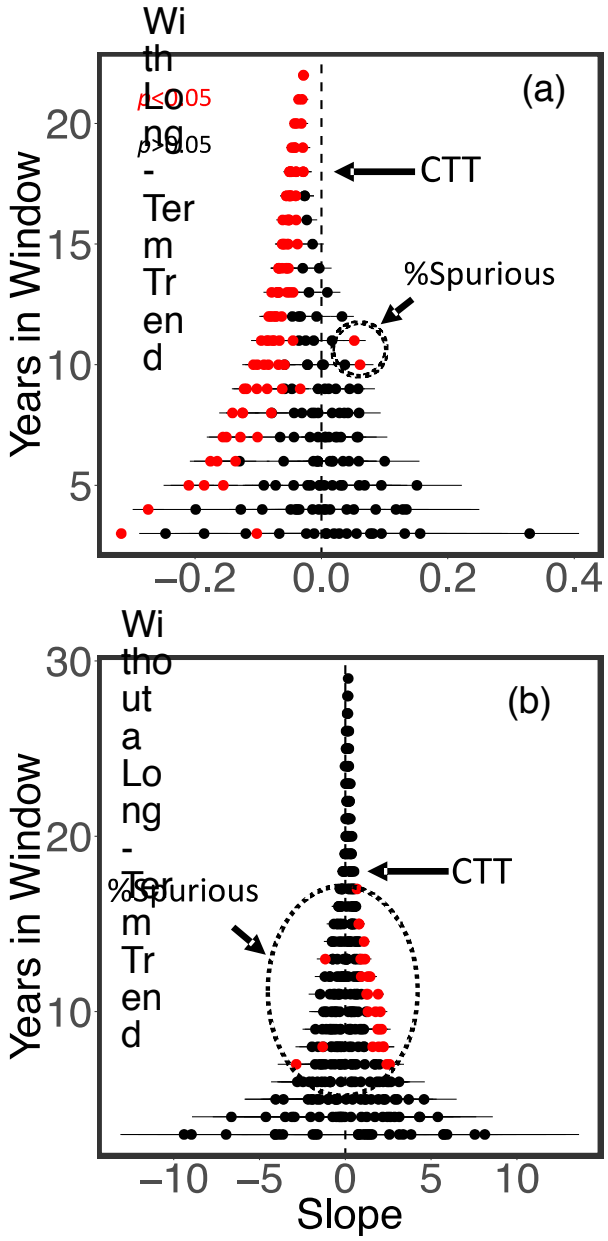

Figure 1: Example moving window plots showing critical temporal thresholds and percent spurious results from a dataset with a long-term trend (a) and dataset without a long-term trend (b). Each plot represents a single experimental study tracking organismal abundance. Red dots show significant trends at the alpha = 0.05 level. Black dots represent non-significant trends. Positive regression slopes indicate that organismal abundance increased in the control

458 relative to treatment while negative slopes indicate the opposite. Panel (a) shows data from the
459 Konza Prairie LTER (knz.72.8) *Andropogon gerardii* response to moisture manipulation. Panel (b)
460 also shows data from the Konza Prairie LTER (knz.26.10) *Dickcissel* response to plant removal by
461 fire.
462

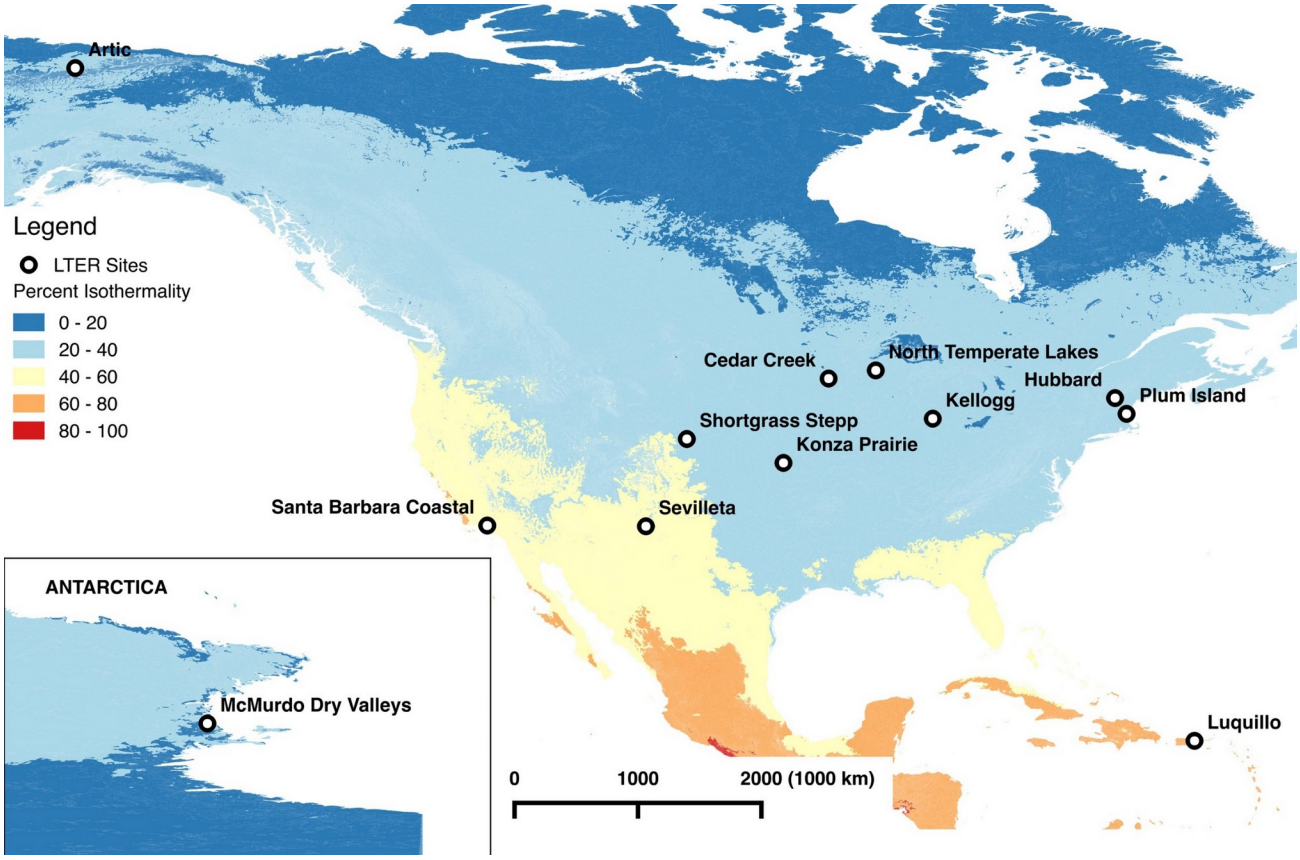

Figure 2: Map showing 12 LTER sites with data in our study across North America and in Antarctica (inset). Colors represent climate variability as determined by isothermality. (BIOCLIM Variable 3). Lower isothermality (cooler colors) indicate higher annual climate variability. Higher isothermality (warmer colors) indicate lower annual climate variability.

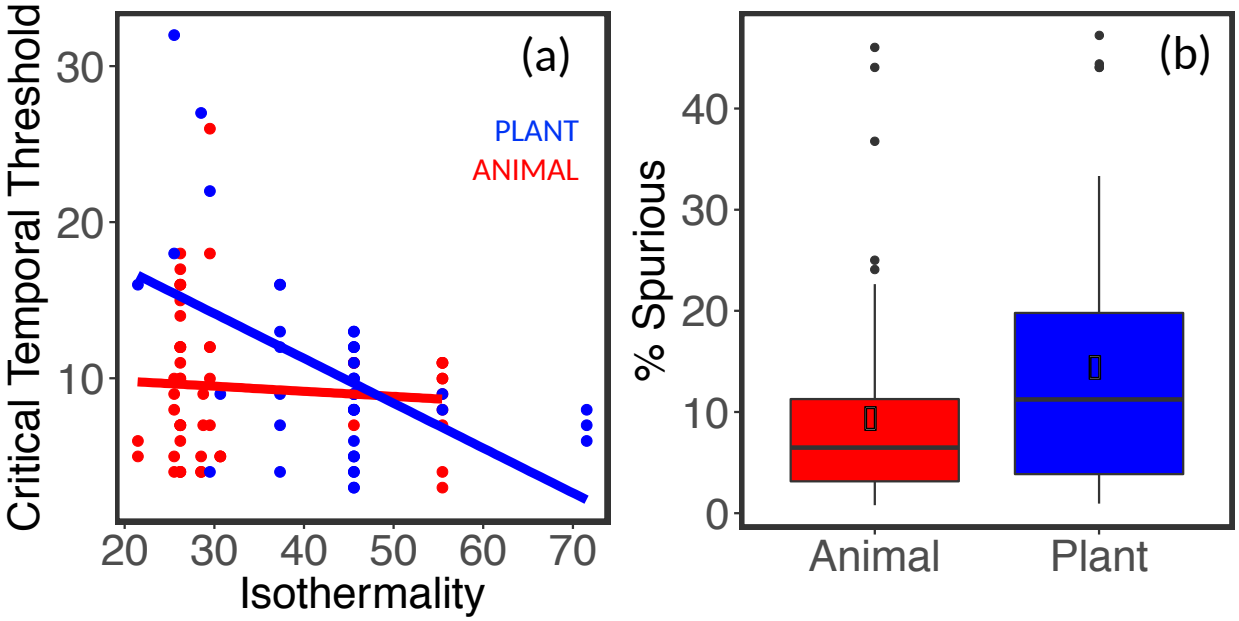

Figure 3: Graphical depictions of our best performing models: (a) Model showing negative relationship between isothermality (%) and critical temporal threshold (years), which is more apparent in plants (blue) than in animals (red). (b) Boxplot comparing percent spurious results between studies of animals and plants. The central bar gives group median, boxes give the 1st and 3rd quartiles, closed circles show outliers, and open circles show group mean.