1 **The Broken Window: An algorithm for quantifying and characterizing**
2 **misleading trajectories in ecological processes**

3

4 **Christie A. Bahlai[1,2,*], Easton R. White[3,4], Julia D. Perrone[1,5], Sarah Cusser[2] and Kaitlin Stack**
5 **Whitney[6]**

6

7 1 Department of Biological Sciences and Environmental Science and Design Research Institute, Kent
8 State University

9 2 Kellogg Biological Station Long Term Ecological Research Site, Michigan State University

10 3 Department of Biological Sciences, University of New Hampshire

11 4 Gund Institute for Environment, University of Vermont

12 5 School of Information, Kent State University

13 6 Science and Technology Studies Department, Rochester Institute of Technology

14 * Corresponding author cbahlai@kent.edu

15

16

17 **Short title: An algorithm for quantifying and characterizing misleading trajectories**

Abstract

A core issue in temporal ecology is the concept of trajectory—that is, when can ecologists have reasonable assurance that they know where a system is going? In this paper, we describe a *non-random resampling* method to directly address the temporal aspects of scaling ecological observations by leveraging existing data. Findings from long-term research sites have been hugely influential in ecology because of their unprecedented longitudinal perspective, yet short-term studies more consistent with typical grant cycles and graduate programs are still the norm. We use long-term insights to create 'broken windows,' that is, reanalyze long-term studies from short-term observational perspectives to examine discontinuities in trends at differing temporal scales.

The broken window algorithm connects our observations between the short-term and the long-term with an automated, systematic resampling approach: in short, we 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 then 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: 1) *how often are the trends observed in short-term data misleading, and* 2) *can characteristics of these trends be used to predict our likelihood of being misled?* We develop a systematic resampling approach, the 'broken_window algorithm, and illustrate its utility with a case study of firefly observations produced at the Kellogg Biological Station Long-Term Ecological Research Site (KBS LTER). Through a variety of visualizations, summary statistics, and downstream analyses, we provide a standardized approach to evaluating the trajectory of a system, the amount of observation required to find a meaningful trajectory in similar systems, and a means of evaluating our confidence in our conclusions.

Highlights

Trends identified in short-term ecology studies can be misleading.

Non-random resampling can show how prone different systems are to misleading trends

The Broken Window algorithm is a new tool to help synthesize temporal data

This tool helps to understand how much data is needed for forecasting to be reliable

It can also be used to quantify how likely it is that an observed trend is spurious.

KEYWORDS: Population, time series, data mining, scaling, trajectory, firefly, lampyridae

1 Introduction

A fundamental problem in ecology is understanding how to scale discoveries: from patterns observed in the lab or the plot to the field or the region, or bridging between short-term observations to long term trends and trajectories (Chave, 2013; Levin, 1992; Schneider, 2001). While shorter-term studies (i.e. those where data collection occurs for less than ~5 years) that coincide with length of typical grant cycles and graduate programs are still the norm, these human constraints do not necessarily capture the ecological phenomena they seek to measure, particularly their temporal dependencies (Hastings, 2004; Wood et al., 2020). This unfortunate mismatch of scales has the potential to limit our understanding of ecological trajectories- that is, the direction a system is going through time, and can undermine our efforts towards a predictive ecology (Evans et al., 2012).   Understanding where and how short term patterns fit into broader trajectories, and how to interpret short-term patterns in the context of a system's trajectory remains an open question (Wauchope et al., 2019; White, 2019). This is illustrated by the recent insect decline controversy, where several high profile papers have observed precipitous declines in insect populations have been subsequently shown to use inappropriate methods for synthesizing the data (Daskalova et al., 2021; Grames et al., 2019; D. L. Wagner, 2020). For example, it is inappropriate to simply combine multiple short term studies and extrapolate (e.g.: Sánchez-Bayo & Wyckhuys, 2019), particularly without explicitly considering the underlying temporal dependencies in the data (Didham et al., 2020).

However, simply recommending that scientists collect more data, for longer, is not necessarily practicable. While long term studies are hugely influential in ecology, they require long-term access to research resources and infrastructure and thus their unprecedented longitudinal perspective is not typical (Hughes et al., 2017). Furthermore, short-term studies, given their prevalence and more limited temporal commitments, can provide a more spatially distributed and potentially richer and more nuanced view into a specific phenomenon at a point in (or shorter period of) time. The key to meaningful synthesis of this vast resource of short-term studies is linking the two extremes of scale. Thus, long-term data, particularly those produced in networked, uniform approaches like those offered by the United States Long Term Ecological Research Network (LTER), present a fundamental opportunity to bridge short and long-term trends through data mining. With long term data, ecologists can systematically investigate the presence and prevalence of short-term trends and compare them to the long-term system trajectories these data document.

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, as well as by sampling error or other research-induced effects (Suding & Gross, 2006). Short-term dynamics observed in an ecological system are not always indicative of the long-term trajectory of that system (Carey & Cottingham, 2016), and furthermore, shorter observation periods can lead to spurious observations because of sampling error variance (Daskalova et al., 2021). In population processes, for example,

84   density-dependent deterministic mechanisms, combined with environmental perturbations, can produce

85   highly variable population numbers over various time scales (Turchin, 2003). Decoupling these processes

86   can reveal the skeleton of a deterministic process interacting with external forces (Bahlai & Zipkin, 2020).

87   However, to disentangle these drivers from an empirical standpoint generally requires a substantial

88   amount of data to be collected over time (Cusser et al., 2020; Higgins et al., 1997). Indeed, in a recent

89   study, White (2019) found that 72% of vertebrate population monitoring programs required at least a

90   decade of observation before the overall trajectory of the population could be detected statistically. A

91   recent study of trends in water bird populations found that short term trends were generally reflective of

92   longer-term patterns (Wauchope et al., 2019), but varied by the generation length of the organism under

93   study. However, they found that, similar to the White (2019) study, greater than two decades of

94   observations would be required to reliably detect a change of 1% per year. Conversely, a study of

95   population viability modelling in snails determined that although longer time series were generally better

96   for establishing the population's trajectory, diminishing returns in precision were observed after about 10-

97   15 years of data were collected (Rueda-Cediel et al., 2015). It is unclear how these findings can be

98   generalized across organisms with differing lifespans, reproductive strategies and life histories, or other

99   environmental processes.

100   The question of trajectory over time is central in ecology, particularly as related to how ecological

101   systems on which humans depend are responding to disturbance or will behave under future climate or

102   environmental conditions (Sutherland et al., 2013). Trajectory is essential to our understanding of

103   ecosystems, their management, and policy decisions, as we interact with our environment. Analytic

104   approaches to time series data have long been a focal area of research in ecology, allowing practitioners

105   to examine temporal dependencies in a variety of processes. The shape a time series takes can provide

106   meaningful information about the properties of the system, the rules that govern its variability, and the

107   trajectory that the system is taking (Esling & Agon, 2012). But when insufficient data exists to apply (or

108   even select) an appropriate time-series approach, a scientist may resort to simpler statistical tools, such

109   as linear models, to describe the patterns observed in the data through the study's window of

110   observation. It is not uncommon for a shorter-duration multi-year ecological study to extrapolate from its

111   data, using the trends observed within their sampling window to draw conclusions about a system's

112   apparent trajectory. For example, a study of British ladybeetle communities concluded that native

113   ladybeetle species were in decline, as was total ladybeetle abundance, following the introduction of an

114   invasive species (Brown et al., 2011). Another found that the richness and abundance of seeds in a soil

115   seed bank were in a recovery trajectory following a period of industrial pollution (M. Wagner et al., 2006).

116   An adventive pest species was implicated in reducing carbon to nitrogen ratios, organic matter in soils of

117   infested forests, thus substantially changing the ecosystem's function over time (Orwig et al., 2008).

118   These examples, representing very different ecological domains, have a common element of a three-year

119   study duration. Yet these inferences may be out of temporal sync with the processes they aim to

120   understand (Birkhead, 2014).

121   A vexing problem arises when shorter term studies apply statistical tools at time scales that are

122   not matched with the underlying processes to make inferences about trajectory: not only may spurious

123   trends be observed, but because only a portion of the underlying process variability is captured, a higher

124   degree of statistical confidence in the result will be found. For example, Bahlai and students examined a

125   12-year time series of firefly captures from Michigan (Hermann et al., 2016). Concerns had been raised

126   about the status of fireflies in eastern North America (Chow et al., 2014), however, for that population, the

127   authors found no evidence of decline over the 12 years (**Fig. 1**): there was no linear relationship between

128   average captures and year (p=0.71, R^2=0.002), and, indeed, there appeared to be evidence of a cyclical

129   dynamic common to many populations near their carrying capacity (**Fig. 1A**). However, students

130   remarked that if the study had been limited to, for example, the 4 years from 2005-2008 (**Fig. 1B**),

131   dramatically different conclusions would have been made. A linear regression of these data would very

132   likely have been interpreted as 'strong evidence' that a decline was occurring in this population (slope -

133   0.31±0.05, p=0.000003, R^2=0.633). Simply, with less data, we would have made the wrong conclusions,

134   and we would have been very confident in our wrong answer. This connection between shorter

135   observation periods and more pronounced patterns is supported by observations made in synthesis

136   efforts: in a compilation of insect biodiversity studies, the shortest time series were more likely to show the

137   most extreme trends (Daskalova et al., 2021).

138   It is because of this phenomenon of "highly-confident wrong answers" that long-term studies are

139   so valued in the ecological community. Indeed, because biological systems are often defined by their

140   variability, when studies are shown to be irreproducible, it is not necessarily due to poor research

141   practice, but due to their inability to capture the full variability of the system within the limits of the study

142   design (Jarvis & Williams, 2016; Voelkl & Würbel, 2016). Long-term ecological research provides insight

143   into the inherent variability of natural systems (Lovett et al., 2007), and insights are thus often only

144   apparent after many years of study (Knapp et al., 2012). Beyond this, there are many other inherent

145   benefits to long-term studies. Long-term studies are disproportionately represented in policy reports and

146   in the ecological literature: studies involving long term observations are cited more often than studies of

147   shorter duration (Hughes et al., 2017). Furthermore, long-term observational studies provide important

148   baseline data: as the world itself changes, these data provide insight into how ecosystems function,

149   instead of studying phenomena after they happen (Franklin et al., 1990; Hastings, 2004; Magurran et al.,

150   2010).

151   Although the importance of long-term studies is clear, empirical examinations of the converse are

152   rare: just how frequently are scientists misled by short-term studies? Can knowledge generated by

153   studying the relationship between short- and long-term studies to bridge the interpretations of short-term

154   data to long-term processes? In this study, we describe a synthetic, computational approach to create a

155   framework to address two hypotheses:

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

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

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

163   Following from the previous prediction, we predict that p-values will be inferior predictors of the
164   'correctness' of short-term trends in predicting longer term trajectory compared to other properties
165   of the system. Better predictors may include statistical measures (slope, standard error), but
166   trends are likely moderated by system specific predictors (e.g. site, data type).

167   The Broken Window Algorithm is a suite of tools which will allow ecologists to leverage existing
168   data to make inferences about system behavior and data needs to characterize system trajectories using
169   an automated, non-random resampling approach (White & Bahlai, 2020): in short, our algorithm
170   repeatedly 'samples' sequential moving windows of data from existing long-term time series, and
171   analyzes these sampled data as if they represented the entire dataset that is, using knowingly limited
172   'windows' of observation to determine how temporal dependencies in a time series affect the likelihood of
173   a short time making a spurious conclusion about how a process varies in time. The tool then compiles
174   typical statistics used to describe the relationships in the sampled data, through repeated samplings, and
175   then use these derived data to gain insights to the questions, *how often are the trends observed in short-*
176   *term data misleading, and can the characteristics of these trends be used to predict our likelihood of*
177   *being misled?* Findings from this work will support the development of a deep understanding of temporal
178   scaling in ecology, aiding in the interpretation of countless future short-term studies. Secondly, and more
179   broadly, our findings have applicability across a variety of domains. Results from this approach will have
180   the opportunity to guide science funding policy, experimental design and interpretation, and data
181   archiving.

182   **2 Materials and Methods**

183   *2.1 Developing the 'broken_window' analysis algorithm*

184   The broken_window algorithm breaks a time series dataset into all possible sequential subsets
185   and then fits a linear model to each of these subsets and compiles the resulting summary statistics,
186   allowing a user to identify and quantify spurious trends within their data. The algorithm is implemented as
187   a series of functions written in R. The algorithm requires a user-inputted two variable data frame with a
188   regular measurement interval as the first variable, and a response variable as the second variable. For
189   the purpose of this study, we assume a yearly measurement interval and some integrative response
190   metric (captures of organisms per trap, average reading, total yield). Data are first subjected to a
191   standardization function which converts the response metric to a unitless Z-score to normalize the data

192   and make it possible to compare datasets with responses of very different magnitudes, and to minimize

193   the impact of measurement unit choice on the observed trends.

194

195   A function that fits a linear model to the data and computes an output vector with the number of

196   observations, the number of years in the study, and particular summary statistics of interest, namely, the

197   slope of the relationship between the response variable and time, the standard error of this relationship,

198   p-values for each of these statistics, and then R^2 and adjusted R^2. Although R^2 and p are not measures of

199   statistical confidence per se, they are often used by ecologists in this way (Nakagawa & Cuthill, 2007;

200   Yoccoz, 1991), and thus can be used as a means to approximate 'conclusions' that a researcher might

201   make of the data. We use this fitting function within a moving window function that takes a provided data

202   frame and iterates through it at all possible subsets and intervals, feeding each interval to the fitting

203   function described above, and compiling the fit statistics for each into a single object.

204

205   **3 Calculation**

206   The moving window function is defined as follows. Let D represent the complete dataset, with $D_{t,r}$

207   representing a single observations of time t and response r. Let $Y = (y_1, y_2, …, y_n)$ represent the set of

208   unique values of t for which observations are recorded, where n is the total number of unique values of t.

209   D is partitioned into sequential subsets of size $S = (3, 4, …, n)$ to create windows $w_{Y,S}$ such that each

210   window

211                    $$w_{i,j} \subset D = \{ D_{t,r} \mid Y_i \leq t \leq Y_{i+Sj}, \forall\ Y_{i+Sj} \leq y_n\}, \text{ and } w_{y1,n} = D$$

212

213   For each $w_{i,j}$, we apply the fitting function described above, and compile the resultant fit statistics for

214   downstream analyses into a data frame. Then, we calculate several meta-statistics and produce

215   visualizations of trends from the resultant data frame.

216

217   First, we defined the slope of the longest time series (i.e. the slope of the linear regression of the whole

218   dataset, D) as the proxy for the 'true' trajectory of the data (as it represents the best information

219   available), along with the computed slope's standard deviation and standard error of the mean as

220   measures of the 'true' variability of the set. Meta-statistics are computed based on comparison to these

221   'true' statistics.

222

223   For all meta-statistics based on frequentist assumptions, we used a set of frequently used 'significance'

224   levels as defaults (i.e. an α=0.05 for line fit statistics) but also encoded the functions so that a user could

225   change these default values easily through supplying a function with different arguments. For each

226   relevant function, we allowed users to toggle via a function argument between these meta-statistics

227   based on the full set of windows tested, or only on the set of windows with statistically significant results,

228   as defined above.

7

229

230   We defined "stability time" as the number of time steps needed before a given proportion of slopes

231   (default = 95%) observed in a window of that length are within a certain number of standard deviations

232   (default = 1) of the true slope. These values were selected to mitigate the impact of outlying data and to

233   reflect industry standards. We computed absolute range (minimum and maximum values) of slope across

234   all windows, as well as relative range (minimum and maximum difference from the 'true' slope, computed

235   as the slope($w_{i,j}$) minus slope(D)). We also created functions that computed the proportion of windows

236   examining a dataset would produce particular results. The proportion of statistically significant slopes

237   produced by a given D measure the probability that a randomly selected window of time would produce a

238   'statistically significant' result. We defined the 'proportion wrong' as the proportion of windows producing

239   statistics that would lead to conclusions differing from those observed for the 'true' trend (i.e. if the true

240   trend was a positive slope, all windows suggesting a negative or non-significant zero-magnitude slope

241   were considered spurious, and so on). We provide functions to compute the proportion wrong for all

242   windows in combination, for each window length, and in the set of windows with lengths less than stability

243   time. In combination, these functions provide a standardized approach to asking the questions of how

244   long a system must be observed to make consistent conclusions about its trajectory, and the likelihood of

245   coming to misleading conclusions about a system if it is observed for less than that time period.

246

247   We created several visualization functions to enable a user to, for a given dataset D, quickly interpret

248   trends based on these meta-statistics, and compare trends in outputs across multiple datasets. A pyramid

249   plot (**Fig. 2A**) uses the data frame of summary statistics from the fits of all windows. It plots the computed

250   slope for each window on the x axis and the length of the window on the y-axis, resulting in a triangular or

251   funnel shaped cloud of points. By default, point size is scaled by the R^2 of the response-by-time

252   relationship within a given window and statistically significant points are demarcated by a circle, and non-

253   significant points given by an 'X'. All points are given with lines indicating their respective standard error.

254   A vertical dashed line indicates the slope of the longest time series, and two dotted vertical lines are

255   plotted at one standard deviation from this value, allowing a user to visually identify the stability time, that

256   is, the length of time required for the majority of windows to produce slopes within a certain interval of the

257   true slope.

258

259   The "wrongness" plot (**Fig. 2B**) examines the same data from a summarized perspective- it plots the

260   average R^2 value and proportion wrong on the y axis by number of years in a window on the x-axis,

261   allowing a user to visualize the relationship between misleading results and the 'confidence' in them for a

262   given D. Finally, the "broken stick" plot (**Fig. 3**) allows a user to visualize the raw time series from D

263   simultaneously with some of the results of the broken_window algorithm. The z-scaled response metric

264   (y-axis) is plotted by observation time (x-axis). The true slope of the entire dataset D is plotted as a solid

265   black line. Then, best fit lines for each window of a user-specified length (default=3-time steps) are

266   plotted, allowing a user to visualize the variation in trend at different points in the time series. Statistically

267   significant slopes are given by dashed red lines, non-significant slopes are indicated by dotted lines.

268   Finally, we created a function which layers and animates broken stick plots to visualize how window

269   slopes change given increasing window length.

270   The R script was developed in RStudio Version 1.2.5033 "Orange Blossom" running R 3.6.2 "Dark and

271   Stormy Night." The script, its development history and all code for case studies and figure generation, are

272   available on GitHub at https://github.com/cbahlai/broken_window.

273

274   **4 Results**

275   We demonstrated the utility of the broken_window algorithm using the firefly study which inspired

276   its development (Hermann et al., 2016). These data on firefly (beetles in the family Lampyridae, with

277   those captured primarily thought to belong to *Photinus pyralis*) captures on insect sticky traps were

278   collected 2004-2015 across 10 plant communities in southwestern Michigan. Complete sampling design

279   and treatments descriptions are provided in Hermann et al (2016). For the purpose of this demonstration,

280   we used the data collected at the perennial early secessional community plots, where fireflies were

281   relatively abundant and complete data were available. Data were subjected to cleaning and quality control

282   using scripts developed by Hermann et al (2016), and then compiled into a metric of total captures per

283   trap, by year (N=12) and replicate (N=6), for a total of 72 observations.

284   The broken_window algorithm produced 55 unique windows (1 sequence of 12 years of data, 2

285   sequences of 11 years of data, … , 10 sequences of 3 years of data). The full 12 year, 72 observation

286   dataset of the normalized response over time was found to have a non-significant slope (-0.01±0.03,

287   p=0.70) and low R^2 value (0.002) suggesting there is unlikely to be a linear trend with time in these data

288   (or, more specifically, we fail to reject the null hypothesis that there is no linear relationship between our

289   response and time) (**Fig. 2A**). Values computed for the slopes across the various windows ranged ±1.2

290   units around the true slope. The algorithm found a stability time of 7 years, that is, once seven years of

291   data were collected, slopes on >95% of windows tested from anytime in the study were within one

292   standard deviation of the slope of the longest series. Overall, nearly half (27/55) of the windows tested

293   found a statistically significant slope, and thus there was nearly a 50% chance a shorter sample leading

294   to a misleading conclusion. Although misleading slopes combined with significant p-values occurred for

295   window lengths longer than 7 years (**Fig. 2B**), they were much more common with window lengths shorter

296   than the stability time (68% of windows), yet these shorter windows were also more likely to be

297   accompanied by a $R^2 > 0.1$ (**Fig. 2B**). Although 3 of these 21 windows ≥7 years in length contained

298   statistically significant trends, after stability time, relative slope ranged from -0.14 to 0.17 z-scaled units

299   around the true slope (**Fig. 3**).

300

301   **5 Discussion**

9

302   Patterns observed in local scale, short-term ecology tend to be dominated by stochastic forces,

303   making generalizations, extrapolations and predictions difficult at larger scales, yet are essential to

304   capture fine-scale understanding of system dynamics (Chave, 2013; Willis & Birks, 2006). The

305   broken_window algorithm formalizes a framework for determining how long a system must be observed

306   before conclusions about its general trends can be reached, and the prevalence of misleading results that

307   occur prior to that time period. With our firefly case study, we found that trends observed prior to our

308   'stability time' of seven years had essentially even odds of being misleading: of three possible outcomes

309   for each window (slope more negative than overall trend, slope more positive than overall trend, slope the

310   same as overall trend), 2/3 of outcomes fell into the two former, and erroneous categories. In this case,

311   no net linear trend was observed in the firefly population data (**Fig. 1, 2A**). Interestingly, we observed that

312   in our case study, statistics commonly used as indicators of "strength" of relationship suggested more

313   uncertainty, and less 'confidence' in results from windows of longer length: p-values, on average, went

314   up, and R^2 values decreased on average as longer windows of the time series were examined (**Fig. 2C**).

315   This finding shines an important light on the reliability of these statistical tools as indicators of model

316   performance: although they provide measures of how well the data from a given window fit the selected

317   model at that time, they also inflate our confidence in what is often an inappropriate model fit to a

318   spurious or short-term trend (Nakagawa & Cuthill, 2007). Given the high likelihood that these

319   observations will vary by context, future work must consider how process characteristics, data availability,

320   and cultural precedent (i.e.: the history of use of a given approach in a scientific field) affect the selection

321   and interpretation of these models. Furthermore, it should explicitly examine data with different structures

322   to examine the relationship between time series shape and likelihood of erroneous conclusions at

323   differing study lengths.

324   In this paper, we demonstrate the utility of the broken_window algorithm in the context of a

325   simple, single population case study. However, this analytical approach has broad application which has

326   been applied by several colleagues in additional systems. In a recent study, Cusser et al (2020) applied

327   the algorithm to a thirty-year experiment comparing the sustainability and productivity attributes of an

328   agricultural cropping system under several management regimes. In this system, due to high variability

329   between treatments, 15 year observation periods were needed to detect consistent between-treatment

330   differences in yield and soil water availability, and at least 1/5 of all windows examined resulted in

331   spurious, statistically misleading trends (i.e. suggest the opposite relationship between management

332   treatments). In an expansion of this work, Cusser and colleagues (2021) used the broken window

333   algorithm to mine more than 100 additional long-term population datasets and found that ~50% of studies

334   had temporal dependencies between treatments that could not be reliably detected with fewer than 10

335   years of data. Furthermore, they linked the stability of the abiotic environment to the ability to detect

336   trends: simply, experiments taking place in more variable environments were more prone to spurious

337   trends and required more data and time to establish experimental differences. In another study, R.

338   Christie et al (2021) compiled 289 surveys of deer tick activity produced by public health departments and

10

339   researchers primarily in the northeast and Midwest United States and subjected each set of observations

340   to the broken_window algorithm. They found none of the survey data reached stability time in less than 5

341   years, indicating that shorter term studies may be insufficient to infer long term population dynamics.

342   Bruel and White (2021) used a similar approach to investigate the optimal sampling of sediment cores for

343   constructing phytoplankton communities. However, they examined the sampling effort required to detect

344   abrupt shifts (i.e., changepoints) in community structure, as opposed to simple linear trends over time.

345   This work highlights the need for future studies investigating the sampling required to detect patterns

346   beyond those from simple linear regression. Other related work has focused on estimating the length of

347   time series required to achieve high statistical power (White, 2019), and studying data-poor fisheries

348   (White & Bahlai, 2020): taken together, these tools will enable previous work to be mined to understand

349   the characteristics of trends common to those systems, and enable future studies to be designed to

350   maximize information value.

351   The broken_window algorithm uses the longest available study duration as a proxy for 'truth' as

352   its core assumption. However, long-term studies themselves are not immune to uncovering misleading

353   trends. Methodology, site selection, and periods of disturbance following the initiation of a long-term study

354   may inherently bias the apparent trajectory of a system (Fournier et al., 2019). This highlights the

355   importance not just of study duration, but of the selection of study starting and ending points: capturing an

356   outlying data point or a high or low in a system's natural variability near the beginning or end of the study

357   period will be highly influential on the statistical outcome, and thus the conclusions reached (Chatterjee &

358   Hadi, 1986; Fournier et al., 2019). Understanding and characterizing these highly influential observations

359   in the analysis process is essential to our interpretations of these ecological trajectories. Thus, it is

360   important to consider these biasing factors when using long-term data in algorithms like the one

361   presented herein: any statistical method is likely to be influenced by outlying or unlikely observations.

362   The broken_window algorithm uses a linear model as its underlying structure, which is the

363   simplest case of a relationship a response variable might take with time. However, many ecological

364   processes are not linear with time and may be better described with non-linear approaches (Bahlai &

365   Zipkin, 2020; Knape, 2016; Wauchope et al., 2019). In the initial deployment of this algorithm, we created

366   a tool for the simplest case that would be applicable under a whide variety of circumstances, but future

367   iterations should consider multiple underlying model structures, as well as contingencies for unevenly

368   spaced observations or missing data.

369   **6 Conclusions**

370   The ever-increasing availability of long-term data, fostered by the growth of technology that

371   enables automated collection and sharing of data products, and the infrastructure availability and

372   'maturity' of projects like the US (and international) Long Term Ecological Research networks (Brunt et

373   al., 2002) and more recently, the National Ecological Observatory Network (SanClements et al., 2020;

374   Schimel et al., 2007) present several key opportunities for new understanding of temporal processes in

375   ecology. Not only can these data be used to observe long-term processes in their respective systems,

376   these data can be used to contextualize the vast amount of data produced by shorter-term studies in our

377   field. Ecology, until relatively recently, was a field defined by data scarcity: studies took place at local

378   scales, over time periods manageable to small groups of researchers, and these shorter-term studies

379   remain the most common output in ecological research (Peters, 2010). Their work represents a huge

380   human undertaking, however, and it is critical that we are able to interpret the insights these observations

381   provide appropriately.

382

383   The broken_window algorithm provides a framework for understanding how ecological data

384   produced by different domains behaves at different temporal scales. Thus, this tool can be used to

385   synthesize data describing ecological processes, specifically examining how system properties (such as

386   landscape, site, seasonality, lifespan in the case of organisms, management regimes, cycles in

387   population trends) affect the likelihood of a spurious trend being observed. In future work, we will examine

388   data of differing structures to identify the characteristics of observation periods that are more likely to

389   produce misleading results, and conversely, the characteristics of time periods that are consistent with

390   longer system trends. This framework will support ongoing research efforts to separate trends in

391   ecological systems from natural variability, human biases and research-specific influences and underlying

392   processes, and provide critical insight into the scaling to temporal processes between short- and long-

393   term experimental designs.

394

Acknowledgements

408   **References**

409   Bahlai, C. A., & Zipkin, E. F. (2020). The Dynamic Shift Detector: An algorithm to identify

410   changes in parameter values governing populations. *PLOS Computational Biology*,

411   *16*(1), e1007542. https://doi.org/10.1371/journal.pcbi.1007542

412   Birkhead, T. (2014). Stormy outlook for long-term ecology studies. *Nature News*, *514*(7523),

413   405.

414   Brown, P. M. J., Frost, R., Doberski, J., Sparks, T. I. M., Harrington, R., & Roy, H. E. (2011).

415   Decline in native ladybirds in response to the arrival of Harmonia axyridis: Early

416   evidence from England. *Ecological Entomology*, *36*(2), 231–240.

417   https://doi.org/10.1111/j.1365-2311.2011.01264.x

418   Bruel, R., & White, E. R. (2021). Sampling requirements and approaches to detect ecosystem

419   shifts. *Ecological Indicators*, *121*, 107096. https://doi.org/10.1016/j.ecolind.2020.107096

420   Brunt, J. W., McCartney, P., Baker, K., & Stafford, S. G. (2002). The future of ecoinformatics in

421   long term ecological research. *Proceedings of the 6th World Multiconference on*

422   *Systemics, Cybernetics and Informatics: SCI*, 14–18.

423   Carey, C. C., & Cottingham, K. L. (2016). Cross-scale Perspectives: Integrating Long-term and

424   High-frequency Data into Our Understanding of Communities and Ecosystems. *The*

425   *Bulletin of the Ecological Society of America*, *97*(1), 129–132.

426   https://doi.org/10.1002/bes2.1205

427   Chatterjee, S., & Hadi, A. S. (1986). Influential Observations, High Leverage Points, and

428   Outliers in Linear Regression. *Statist. Sci.*, *1*(3), 379–393.

429   https://doi.org/10.1214/ss/1177013622

430   Chave, J. (2013). The problem of pattern and scale in ecology: What have we learned in

431   20 years? *Ecology Letters*, *16*(s1), 4–16. https://doi.org/10.1111/ele.12048

432   Chow, A. T., Chong, J.-H., Cook, M., & White, D. (2014). Vanishing Fireflies: A Citizen-Science

433   Project Promoting Scientific Inquiry and Environmental Stewardship. *Science Education*

434   *and Civic Engagement*, *6*(1), 23–31.

435   Christie, R., Whitney, K. S., Perrone, J., & Bahlai, C. A. (2021). Longer study length,

436   standardized sampling techniques, and broader geographic scope leads to higher

437   likelihood of detecting stable abundance patterns in long term deer tick (Ixodes

438   scapularis) studies. *BioRxiv*, 2021.03.06.434217.

439   https://doi.org/10.1101/2021.03.06.434217

440   Cusser, S., Bahlai, C., Swinton, S. M., Robertson, G. P., & Haddad, N. M. (2020). Long-term

441   research avoids spurious and misleading trends in sustainability attributes of no-till.

442   *Global Change Biology*, *26*(6), 3715–3725. https://doi.org/10.1111/gcb.15080

443   Cusser, S., Helms IV, J., Bahlai, C. A., & Haddad, N. M. (2021). How long do population level

444   field experiments need to be? Utilising data from the 40-year-old LTER network. *Ecology*

445   *Letters*, *n/a*(n/a). https://doi.org/10.1111/ele.13710

446   Daskalova, G. N., Phillimore, A. B., & Myers-Smith, I. H. (2021). Accounting for year effects and

447   sampling error in temporal analyses of invertebrate population and biodiversity change:

448   A comment on Seibold et al. 2019. *Insect Conservation and Diversity*, *14*(1), 149–154.

449   https://doi.org/10.1111/icad.12468

450   Didham, R. K., Basset, Y., Collins, C. M., Leather, S. R., Littlewood, N. A., Menz, M. H. M.,

451   Müller, J., Packer, L., Saunders, M. E., Schönrogge, K., Stewart, A. J. A., Yanoviak, S.

452   P., & Hassall, C. (2020). Interpreting insect declines: Seven challenges and a way

453   forward. *Insect Conservation and Diversity*, *13*(2), 103–114.

454   https://doi.org/10.1111/icad.12408

455   Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Comput. Surv.*, *45*(1), 1–34.

456   Evans, M. R., Norris, K. J., & Benton, T. G. (2012). Predictive ecology: Systems approaches.

457   *Philosophical Transactions of the Royal Society of London. Series B, Biological*

458   *Sciences*, *367*(1586), 163–169. PubMed. https://doi.org/10.1098/rstb.2011.0191

459   Fournier, A. M. V., White, E. R., & Heard, S. B. (2019). Site-selection bias and apparent

460   population declines in long-term studies. *Conservation Biology*, *33*(6), 1370–1379.

461   https://doi.org/10.1111/cobi.13371

462   Franklin, J. F., Bledsoe, C. S., & Callahan, J. T. (1990). Contributions of the Long-Term

463   Ecological Research Program. *Bioscience*, *40*(7), 509–523.

464   Grames, E., Montgomery, G., Haddaway, N., Dicks, L., Elphick, C., Matson, T., Nakagawa, S.,

465   Saunders, M., Tingley, M., White, T., Woodcock, P., & Wagner, D. (2019). *Trends in*

466   *global insect abundance and biodiversity: A community-driven systematic map protocol*.

467   https://doi.org/10.17605/OSF.IO/Q63UY

468   Hastings, A. (2004). Transients: The key to long-term ecological understanding? *Trends in*

469   *Ecology & Evolution*, *19*(1), 39–45. https://doi.org/10.1016/j.tree.2003.09.007

470   Hermann, S. L., Xue, S., Rowe, L., Davidson-Lowe, E., Myers, A., Eshchanov, B., & Bahlai, C.

471   A. (2016). Thermally moderated firefly activity is delayed by precipitation extremes.

472   *Royal Society Open Science*, *3*(12), 160712.

473   Higgins, K., Hastings, A., Sarvela, J. N., & Botsford, L. W. (1997). Stochastic Dynamics and

474   Deterministic Skeletons: Population Behavior of Dungeness Crab. *Science*, *276*(5317),

475   1431. https://doi.org/10.1126/science.276.5317.1431

476   Hughes, B. B., Beas-Luna, R., Barner, A. K., Brewitt, K., Brumbaugh, D. R., Cerny-Chipman, E.

477   B., Close, S. L., Coblentz, K. E., de Nesnera, K. L., Drobnitch, S. T., Figurski, J. D.,

478   Focht, B., Friedman, M., Freiwald, J., Heady, K. K., Heady, W. N., Hettinger, A.,

479   Johnson, A., Karr, K. A., … Carr, M. H. (2017). Long-Term Studies Contribute

480   Disproportionately to Ecology and Policy. *BioScience*, *67*(3), 271–281.

481   https://doi.org/10.1093/biosci/biw185

482   Jarvis, M. F., & Williams, M. (2016). Irreproducibility in Preclinical Biomedical Research:

483   Perceptions, Uncertainties, and Knowledge Gaps. *Trends in Pharmacological Sciences*,

484   *37*(4), 290–302. https://doi.org/10.1016/j.tips.2015.12.001

485   Knape, J. (2016). Decomposing trends in Swedish bird populations using generalized additive

486   mixed models. *Journal of Applied Ecology*, *53*(6), 1852–1861.

487   https://doi.org/10.1111/1365-2664.12720

488   Knapp, A. K., Smith, M. D., Hobbie, S. E., Collins, S. L., Fahey, T. J., Hansen, G. J. A., Landis,

489   D. A., La Pierre, K. J., Melillo, J. M., Seastedt, T. R., Shaver, G. R., & Webster, J. R.

490   (2012). Past, present, and future roles of long-term experiments in the LTER Network.

491   *Bioscience*, *62*(4), 377–389. ProQuest Research Library.

492   https://doi.org/10.1029/2008gb003336

493   Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur

494   Award Lecture. *Ecology*, *73*(6), 1943–1967. https://doi.org/10.2307/1941447

495   Lovett, G. M., Burns, D. A., Driscoll, C. T., Jenkins, J. C., Mitchell, M. J., Rustad, L., Shanley, J.

496   B., Likens, G. E., & Haeuber, R. (2007). Who needs environmental monitoring? *Frontiers*

497   *in Ecology and the Environment*, *5*(5), 253–260.

498   Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. McP., Elston, D. A., Scott, E. M., Smith,

499   R. I., Somerfield, P. J., & Watt, A. D. (2010). Long-term datasets in biodiversity research

500   and monitoring: Assessing change in ecological communities through time. *Special*

501   *Issue: Long-Term Ecological Research*, *25*(10), 574–582.

502   https://doi.org/10.1016/j.tree.2010.06.016

503   Nakagawa, S., & Cuthill, I. (2007). Effect size, confidence interval and statistical significance: A

504   practical guide for biologists. *Biological Reviews*, *82*, 591–605.

505   Orwig, D. A., Cobb, R. C., D'Amato, A. W., Kizlinski, M. L., & Foster, D. R. (2008). Multi-year

506   ecosystem response to hemlock woolly adelgid infestation in southern New England

507   forests. *Canadian Journal of Forest Research*, *38*(4), 834–843.

16

508   Peters, D. P. C. (2010). Accessible ecology: Synthesis of the long, deep, and broad. *Trends in*

509   *Ecology & Evolution*, *25*(10), 592–601. https://doi.org/10.1016/j.tree.2010.07.005

510   Rueda-Cediel, P., Anderson, K. E., Regan, T. J., Franklin, J., & Regan, H. M. (2015). Combined

511   Influences of Model Choice, Data Quality, and Data Quantity When Estimating

512   Population Trends. *PLOS ONE*, *10*(7), e0132255.

513   https://doi.org/10.1371/journal.pone.0132255

514   Sánchez-Bayo, F., & Wyckhuys, K. A. (2019). Worldwide decline of the entomofauna: A review

515   of its drivers. *Biological Conservation*, *232*, 8–27.

516   SanClements, M., Lee, R. H., Ayres, E. D., Goodman, K., Jones, M., Durden, D., Thibault, K.,

517   Zulueta, R., Roberti, J., Lunch, C., & Gallo, A. (2020). Collaborating with NEON.

518   *BioScience*, *70*(2), 107–107. https://doi.org/10.1093/biosci/biaa005

519   Schimel, D., Hargrove, W., Hoffman, F., & MacMahon, J. (2007). NEON: a hierarchically

520   designed national ecological network. *Frontiers in Ecology and the Environment*, *5*(2),

521   59–59. https://doi.org/10.1890/1540-9295(2007)5[59:NAHDNE]2.0.CO;2

522   Schneider, D. C. (2001). The Rise of the Concept of Scale in Ecology: The concept of scale is

523   evolving from verbal expression to quantitative expression. *BioScience*, *51*(7), 545–553.

524   https://doi.org/10.1641/0006-3568(2001)051[0545:TROTCO]2.0.CO;2

525   Suding, K. N., & Gross, K. L. (2006). The dynamic nature of ecological systems: Multiple states

526   and restoration trajectories. *Foundations of Restoration Ecology*, 190–209.

527   Sutherland, W. J., Freckleton, R. P., Godfray, H. C. J., Beissinger, S. R., Benton, T., Cameron,

528   D. D., Carmel, Y., Coomes, D. A., Coulson, T., Emmerson, M. C., Hails, R. S., Hays, G.

529   C., Hodgson, D. J., Hutchings, M. J., Johnson, D., Jones, J. P. G., Keeling, M. J., Kokko,

530   H., Kunin, W. E., … Wiegand, T. (2013). Identification of 100 fundamental ecological

531   questions. *Journal of Ecology*, *101*(1), 58–67. https://doi.org/10.1111/1365-2745.12025

532   Turchin, P. (2003). *Complex population dynamics: A theoretical/empirical synthesis* (Vol. 35).

533   Princeton University Press.

534   Voelkl, B., & Würbel, H. (2016). Reproducibility Crisis: Are We Ignoring Reaction Norms?

535   *Trends in Pharmacological Sciences*, *37*(7), 509–510.

536   https://doi.org/10.1016/j.tips.2016.05.003

537   Wagner, D. L. (2020). Insect declines in the Anthropocene. *Annual Review of Entomology*, *65*,

538   457–480.

539   Wagner, M., Heinrich, W., & Jetschke, G. (2006). Seed bank assembly in an unmanaged

540   ruderal grassland recovering from long-term exposure to industrial emissions. *Acta*

541   *Oecologica*, *30*(3), 342–352. https://doi.org/10.1016/j.actao.2006.06.002

542   Wauchope, H. S., Amano, T., Sutherland, W. J., & Johnston, A. (2019). When can we trust

543   population trends? A method for quantifying the effects of sampling interval and duration.

544   *Methods in Ecology and Evolution*, *10*(12), 2067–2078. https://doi.org/10.1111/2041-

545   210X.13302

546   White, E. R. (2019). Minimum Time Required to Detect Population Trends: The Need for Long-

547   Term Monitoring Programs. *BioScience*, biy144–biy144.

548   https://doi.org/10.1093/biosci/biy144

549   White, E. R., & Bahlai, C. A. (2020). Experimenting with the past to improve environmental

550   monitoring programs. *Frontiers in Ecology and Evolution*.

551   https://doi.org/10.3389/fevo.2020.572979

552   Willis, K. J., & Birks, H. J. B. (2006). What Is Natural? The Need for a Long-Term Perspective in

553   Biodiversity Conservation. *Science*, *314*(5803), 1261.

554   https://doi.org/10.1126/science.1122667

555   Wood, C. M., Loman, Z. G., McKinney, S. T., & Loftin, C. S. (2020). Testing prediction accuracy

556   in short-term ecological studies. *Basic and Applied Ecology*, *43*, 77–85.

557   https://doi.org/10.1016/j.baae.2020.01.003

558   Yoccoz, N. G. (1991). Use, Overuse, and Misuse of Significance Tests in Evolutionary Biology

559   and Ecology. *Bulletin of the Ecological Society of America*, *72*(2), 106–111.

560

561

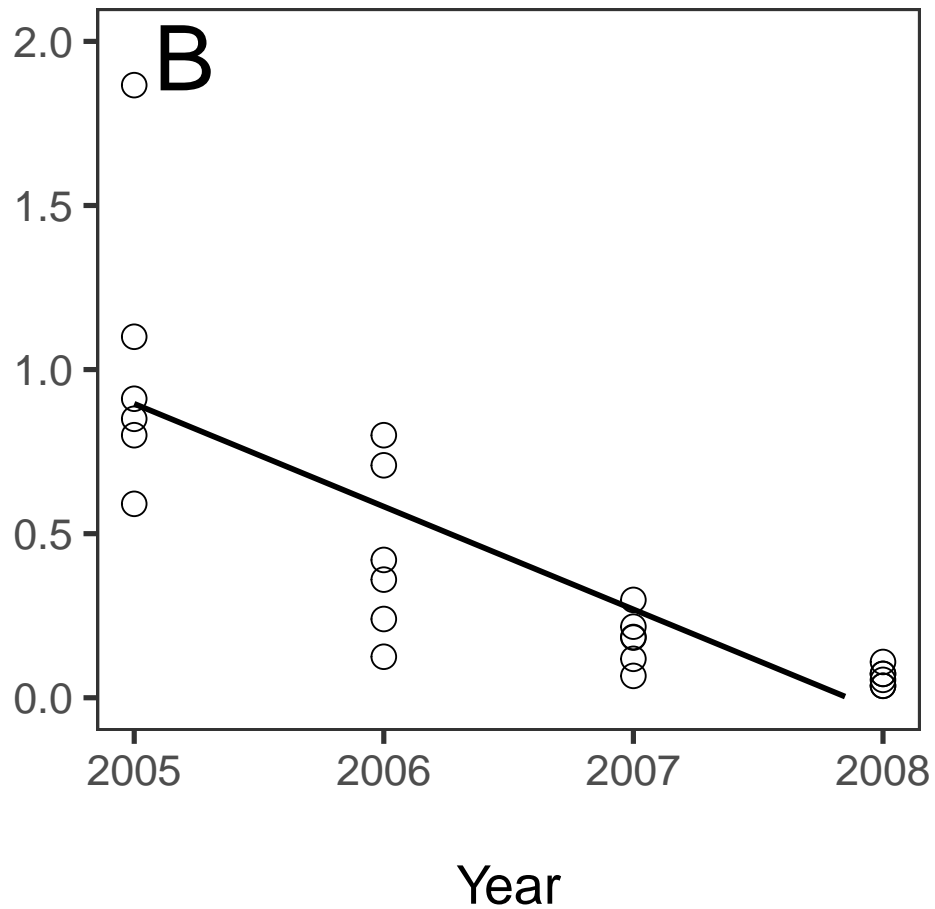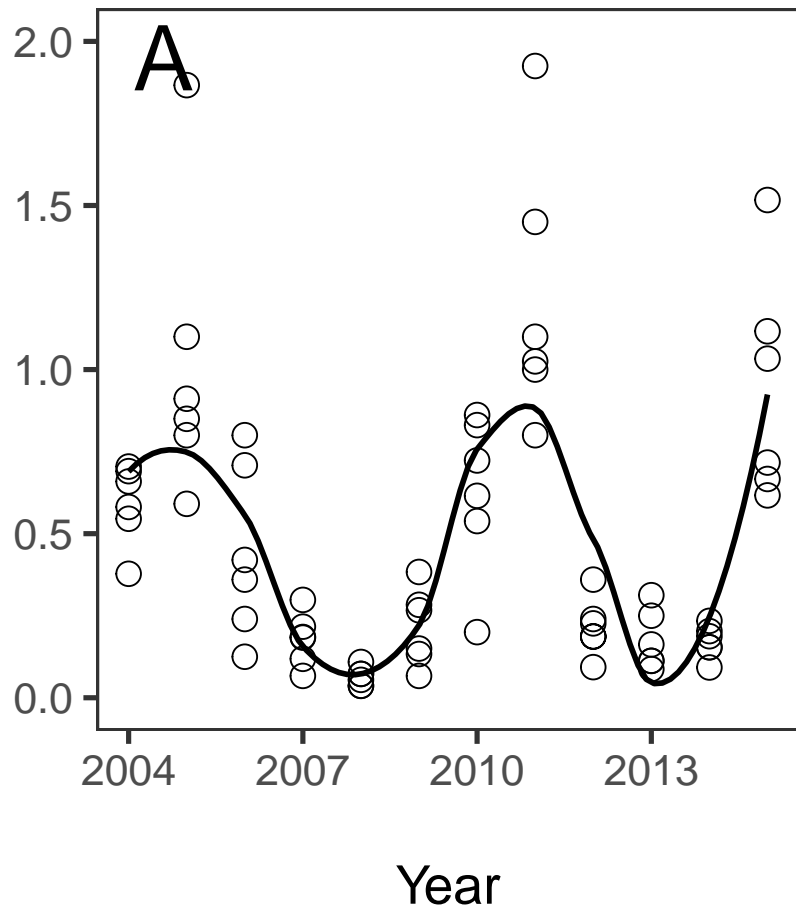562   **Figures**

563

564   **Figure 1: Same data, different observation periods, different conclusions.** Firefly populations

565   (reported as mean number of adults captured per trap) monitored in ten plant community treatments at

566   Kellogg Biological Station in southwestern Michigan cycle over an approximately 6 year period (panel A).

567   Yet, if sampling had only occurred over a 4 year period, we would conclude the population underwent a

568   steep (and statistically significant) decline in the four years from 2005-2008 (slope -0.31±0.05,

569   p=0.000003, R^2=0.633; panel B). Data and figures adapted from Hermann et al (2016).

570

571   **Figure 2: Core outputs of the broken_window algorithm:** Using the firefly data from the early

572   sucessional plant community presented in Hermann et al (2016), we are able to compile 55 possible

573   windows of three years or greater. **A)** The pyramid plot gives a distribution of possible conclusions. On

574   this plot, each point represents a window and its corresponding summary statistics for a linear

575   relationship between the response variable (in this case, z-scaled population density of fireflies) and time.

576   Point coordinates are defined by the slope and length of a window, and point size is scaled by the R^2

577   computed for that regression. The lines accpmanying each point represent standard error of the slope for

578   each point. Statistically significant relationships (in this case α=0.05) are plotted as black circles, and non-

579   significant slopes are plotted as red Xs. The vertical central dashed black line represents the slope of the

580   complete time series (here with 12 years of data) and the vertical dotted grey lines are placed at one

581   standard deviation in both the positive and negative direction from the 'true' slope. **B)** The 'wrongness

582   plot' visualizes the relationship between the likelihood of a spurious conclusion and statistical proxies for

583   'confidence' in a relationship. The proportion of windows where spurious slopes were observed by the

584   length of window are displayed as black circular points with blue solid smoothing line , and the average

585   R^2 value across windows of that length are given as orange triangular points with a dashed red

586   smoothing line. The grey dotted vertical line is placed at the 'stability time' of 7 years, after which the

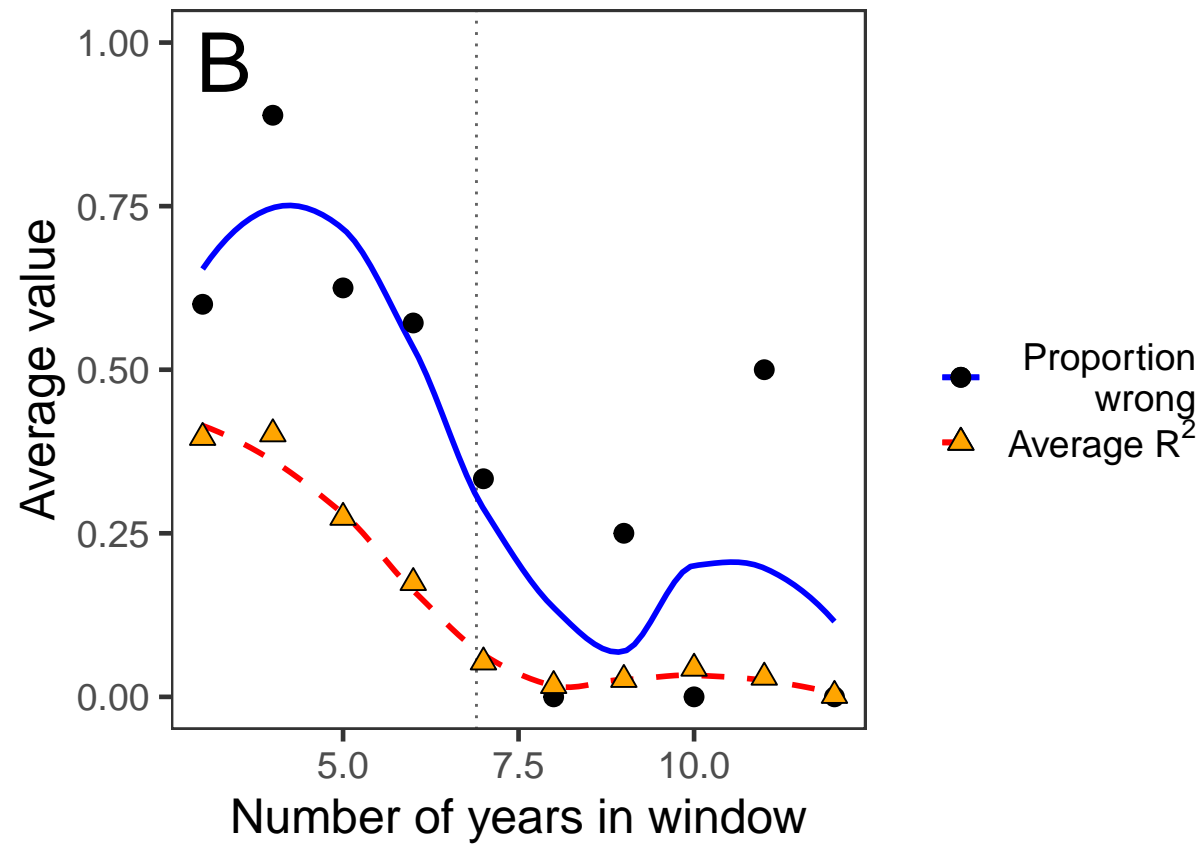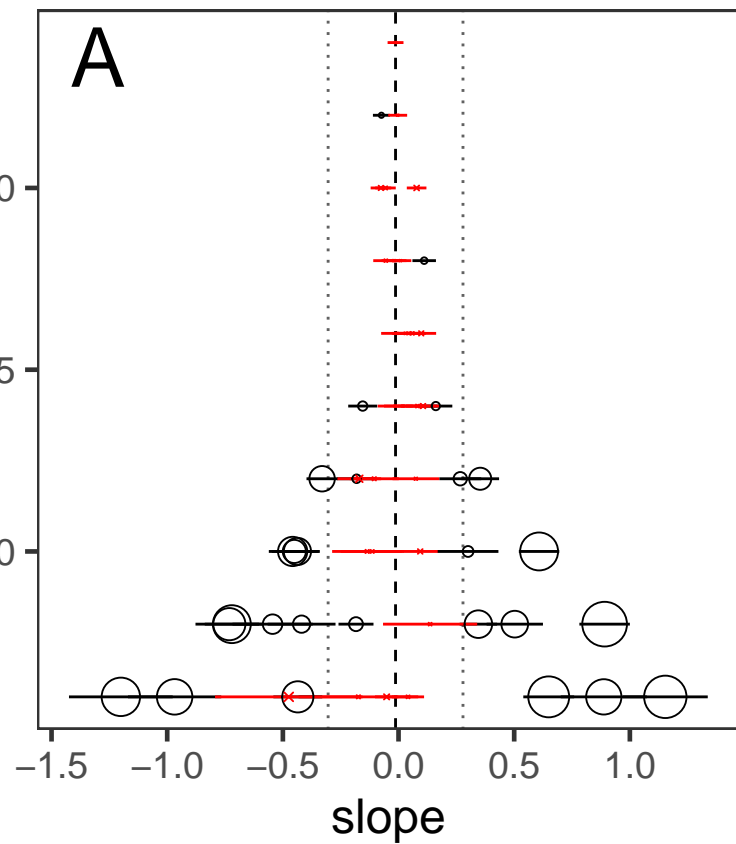587   slopes in 95% of the windows occur within one standard deviation of the 'true' slope.

588

589

590   **Figure 3: The broken stick plot allows a user to visualize the magnitude of difference between the**

591   **slopes produced at different window lengths**. Using the firefly data from the early successional plant

592   community from Hermann et al (2016), all of the nine panels presents the Z-scaled response of firefly

593   density over time, and a solid black line indicates the linear regression of the full data series (the 'true'

594   slope). The 95% confidence interval of this line is plotted in light blue. Within each panel, the linear

595   regressions for each window of a given length are plotted: regressions with a statistically significant slope

596   (at α=0.05) are given with red dashed lines, and non-significant regressions are plotted as grey dotted
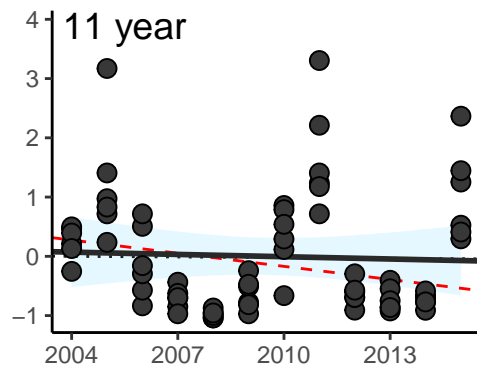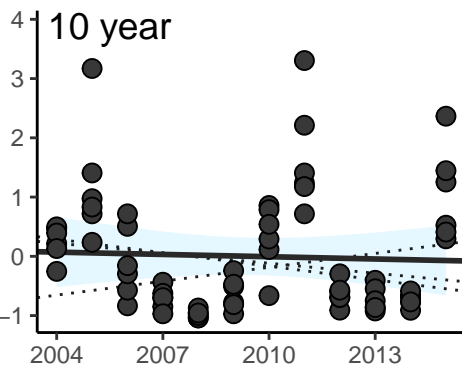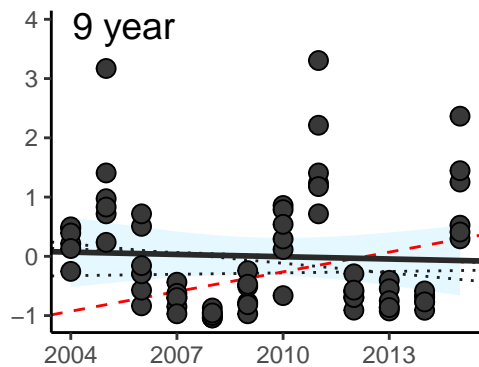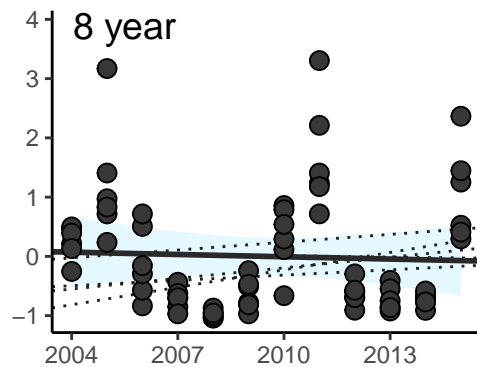
597   lines.

598