**S2 Additional Simulations using AIC as a decision rule**

**Methods**

In this appendix, we present a detailed examination of the behavior of the regime shift detector’s (RSD) behavior using an alternate information criterion with less conservative behavior- AIC. In our initial calibrations of the RSD model, we found that, generally, AIC was more sensitive at picking up breaks in time series, but also tended to over-fit, likely because model complexity is not penalized as severely by this criterion than the AICc. However, we conducted the same complete simulation and case study analysis as described in the paper using AIC as well, and we present the results herein.

The AICc-based model in the RSD is a compromise between sensitivity and specificity. We found that, in general, results were better fit expectations when using a conservative information criterion to select results in simulations, and thus, we recommend users of the RSD model use AICc when examining new data. However, there may be cases where it is desirable to gain a more liberal estimate of changes in patterns: in this case, the more sensitive AIC can be used to rank break point combinations. We found final results of the RSD model were similar, regardless of the information criterion used, except that using AIC was more likely to indicate more candidate breaks, and weight both true and erroneous breaks with higher values (Appendix S2). Thus, this more sensitive approach may be most useful in the context of hypothesis generation, rather than as an explicit hypothesis test.

*Simulation study*

We examined the regime shift detector’s performance for all test scenarios. First, we evaluated the ability of the model to detect scenario initialization conditions within the set of equivalent break point combinations (Fig S2-1). We also examined the performance of the break-point weighting tool from the perspective of its average weightings of true and erroneous break points (Fig. S2-2).

In general, The AIC-based model sets were dramatically less likely to identify initiation conditions within the equivalently performing break point combination sets in comparison to the AICc based RSD. The performance associated the most complex parameterizations tested stayed about the same, but the performance associated with all parameterizations was <70% (Fig. S2-1). Throughout the analysis, we observed the same general patterns around the RSD’s performance in parameter space- a decreased performance with increasing experimental noise, extreme values of r, small changes in K, large changes in r, and longer time series length (Fig. S2-1).

When we conducted the weight analysis to determine the performance of the RSD using AIC on break weights, we found that, like with AICc, in the vast majority of parameterization cases, the average weight of a ‘true’ break (i.e. one that was intentionally simulated in the data) typically exceeded a value of 0.8, (Fig. S2- 2), but unlike the analysis with AICc, erroneous breaks had higher average weights, up to a value of ~0.4, so in this case, we suggest using this higher weight as the cutoff to classify a break as erroneous.

*Case study- ladybeetles*

As with the RSD analysis using AICc, the AIC based analysis found two break points, one occurring after 2000 and one occurring after 2005, in the top break point combination model. However, the AIC analysis did not find any additional break point combinations with equivalent performance. Break weight analysis suggested a weight of 0.95 for the 2000 break, and a weight of 0.64 for the break after 2005. In the context of the AIC analysis, we would thus conclude unambiguous support for the break occurring at 2000 and moderate evidence for the break at 2005.

*Case study- monarchs*

For the monarch butterflies, using AIC as the information criterion in the RSD resulted in a much more complex result, compared to AICc based analysis. In this analysis, a break point combination with two breaks (at 2003 and 2008) was selected as the top model (AIC=106.86), but there were four additional break point combinations with equivalent performance (1998, 2003, 2008, AIC=106.90; 1999, 2003, 2008, AIC=108.02; 2003, 2007, AIC=108.51; and 1998, 2003, 2007, AIC=108.54). Thus, the model suggested many more candidate break points (1998, 1999, 2003, 2007, and 2008) but the weighting analysis found variable strength of evidence for these breaks (0.41, 0.23, 0.81, 0.27 and 0.65, respectively), and thus we conclude there is weak evidence for a break at 1998, fairly strong evidence for a break at 2003, and moderate evidence for a break at 2008.

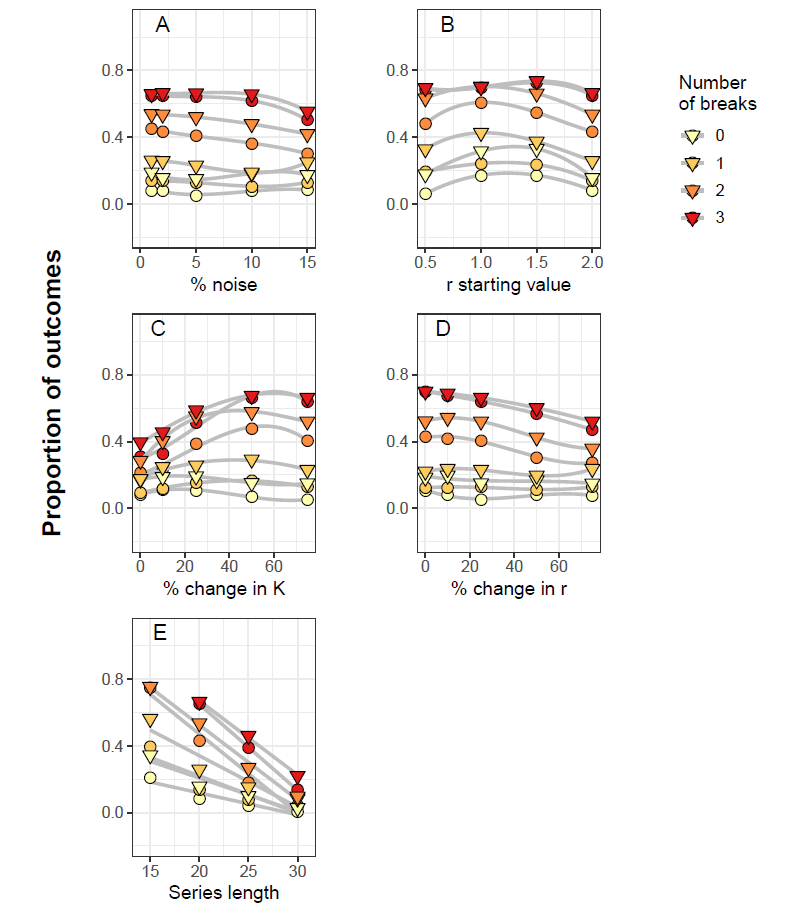


Figure S2-1: **Performance the regime shift detector model under varying conditions (Using AIC instead of AICc).** Proportion of results where initial conditions were detected by the top break point combination (circles) or within the equivalent model set (triangles) under A) varied noise (in the form of normally distributed error) B) varied starting values of the r constant, C) varied % changes in the K constant in the Ricker model D) varied % changes in r, the intrinsic rate of increase in the Ricker model and E) simulated time series length. Sets of 0, 1, 2 and 3 break points were randomly generated from within the set of possible values each scenario was iterated 250 times.

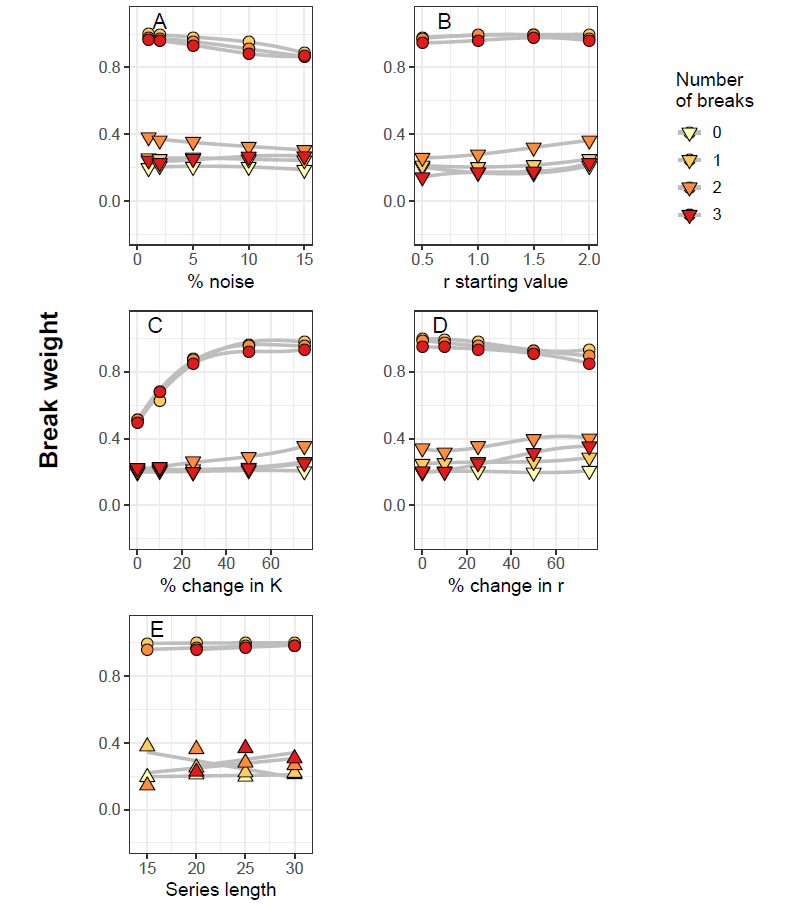


Figure S2-2: **Average break weight of break points found under varying parameterization conditions (Using AIC instead of AICc) .** Break weight is computed based on a modification of “Relative Variable Importance” formula, where each break point is multiplied by the Akaike weight of each model in which it appears. Average weights of break points identified by the regime shift detector model reflecting true parameterization conditions (circles) or erroneous breaks suggested by the model (triangles) under A) varied noise (in the form of normally distributed error) B) varied starting values of the r constant, C) varied % changes in the K constant in the Ricker model D) varied % changes in r, the intrinsic rate of increase in the Ricker model and E) simulated time series length. Sets of 0, 1, 2 and 3 break points were randomly generated from within the set of possible values each scenario was iterated 250 times.