**S2 Additional Simulations**

**Methods**

In the main manuscript, we present ‘outcome based’ results (Figure 1), i.e. given the model’s output, what is the probability of correctly identifying the input scenario? In this appendix, we present the converse: given known input conditions, what is the accuracy of the model output, and what errors are most likely to occur?

We created a function to simulate time series data following Ricker dynamics under set break point combination conditions. The function takes values for of start year, number of years to simulate, % noise, a starting population N, starting values for K and r, a break point combination, a % change in K and a % change in r to be simulated for each break point. Percent noise was included as a means to simulate sampling error that would be observed in a real sampling plan, and was simulated by creating a continuous interval from 100% minus noise to 100% plus noise, randomly sampling from within that interval, and multiplying the predicted observation N(t+1) by the resultant value. Change of K and r at each break point were randomly selected by the script to either increase or decrease by the given % change.

The simulated data were fed into a function that tested if the regime shift detector model was able to identify the breaks as set for the simulation by comparing the input conditions to those output by the bestmodel function (Appendix S1) in the regime shift detector script file. Results of comparing the input to the output were encoded as follows:

1. model was successful at detecting all break points and simulation conditions
2. model identified all simulated breaks, but also found one or more ‘extra’ breaks
3. script missed one of the simulated breaks, but all others found were correct
4. script identified the correct number of breaks, but one or more breaks were mismatched
5. no correct breaks were identified by the script, or breaks were identified in a no-break scenario

A base scenario was constructed, with start year =1, number of years = 25, a starting population of 3000, a sampling error of up to 5%, a starting value for K = 2000, a starting value for r =2, a % change at each break point of 40 and 20% for K and r respectively, and a set of 0, 1, 2, or 3 break points randomly selected from within the possible values defined by start year and number of years.

**Detailed results**

When varied sampling error was simulated as ‘noise’ (Fig. noise\_sim), the script’s ability to detect starting conditions dropped as percent noise increased, with the exception of no-break scenarios, which were generally correctly identified at a rate of approximately 60%, regardless of simulated sampling error (Fig. noise\_sim A). Outcomes involving the script finding extra breaks were most common in scenarios initiated with only one break (Fig. noise\_sim B), while outcomes where one break was missed by the script only occurred in scenarios initiated with three breaks, increasing with sampling error, and then plateauing at about 20% of outcomes above 30% sampling error (Fig. noise\_sim C). Outcomes identifying the correct number of breaks but misidentifying one break’s location peaked at around 20% sampling error in scenarios initiated with three break points, and at approximately 40% sampling error in scenarios initiated with two break points (Fig. noise\_sim D), and total failure to identify initial conditions generally increased with sampling error, with the exception of scenarios initiated with no break points (Fig. noise\_sim E).

Modifying the length of time series that a scenario affected the ability of the model to identify the starting conditions (Fig. Nyears), with model performance decreasing slightly with length of time series (Fig. Nyears A). In general, increasing time series length increased the probability that the script would correctly identify the break points from the initial conditions, but also ‘find’ additional break points (Fig. Nyears B) or find a break in a scenario that was not initiated with any breaks (Fig. Nyears E). Other erroneous results were rare (Figs. Nyears C, D).

The effect of modifying regime shift size (i.e. the magnitude of the parameter chages) on the model’s ability to detect conditions with which the scenarios were initiated was examined by modifying the % change in r and K at the given break point combination (Figs. changeK, changeR). The script was best able to identify initial conditions when the value for K was shifted by 40% (Fig. changeK A) with extra break points more frequently erroneously detected in scenarios initiated with larger changes of K at break points (Fig. changeK B). Complete failure to identify break points was most common in scenarios with small shifts in K (Fig. changeK E); missed breaks occurred rarely in 3 break scenarios regardless of the shift in K (Fig. changeK C) and misidentified breaks occurred occasionally in scenarios with 2 or 3 breaks and very large or very small shifts in K (Fig. changeK D.) The efficiency of the model responded differently to modifications of the size of shifts in r: instead of an intermediate optimum shift as observed for K, smaller shifts involving changes in r were most easily detected by the model (Fig. changeR A). The model was more likely to erroneously find additional breaks, miss breaks, or misidentify breaks as the percent change in r increased (Fig. changeR B, C,D). Complete failure to identify correct break combinations increased slightly with increases in r for scenarios initiated with 1, 2, or 3 break points, but error rates remained constant regardless of shift in r in the zero-break scenarios (Fig. changeR E)

**Insert figures directly into appendix below here**

Figure ChangeK: **Performance of Regime Shift Detector Script under conditions of varying K.** Proportion of results with a given outcome under varying % changes in the K constant in the Ricker model at four simulated break point scenarios. Sets of 0, 1, 2 and 3 break points were randomly generated from within the set of possible values, and data were simulated with random 5% sampling error and a 20% shift of r at the given break point. Each series consisted of 25 years of simulated data and each scenario was iterated 500 times. Lines joining points represent a third order polynomial GAM representing the best fit, with standard error. Possible outcomes were A) Successful identification of all break points; B) One extra break point identified; C) One break point missed; D) Correct number of breaks found, but one or more break points misidentified; and E) Complete failure to identify the correct break point combination by the regime shift detector script.

Figure ChangeR: **Performance of Regime Shift Detector Script under conditions of varying r.** Proportion of results with a given outcome under varying % changes in the r constant in the Ricker model at four simulated break point scenarios. Sets of 0, 1, 2 and 3 break points were randomly generated from within the set of possible values, and data were simulated with random 5% sampling error and a 40% shift of K at the given break point. Each series consisted of 25 years of simulated data and each scenario was iterated 500 times. Lines joining points represent a third order polynomial GAM representing the best fit, with standard error. Possible outcomes were A) Successful identification of all break points; B) One extra break point identified; C) One break point missed; D) Correct number of breaks found, but one or more break points misidentified; and E) Complete failure to identify the correct break point combination by the regime shift detector script.

Figure noise\_sim: **Performance of Regime Shift Detector Script under conditions of varying sampling error.** Proportion of results with a given outcome under varying % in sampling error (‘noise’), modeled as randomly generated values selected from a continuous interval within a given % noise, for each observation generated in a simulation. Sets of 0, 1, 2 and 3 break points were randomly generated from within the set of possible values, and data were simulated with a 20% shift of r and a 40% shift of K at the given break point. Each series consisted of 25 years of simulated data and each scenario was iterated 500 times. Lines joining points represent a third order polynomial GAM representing the best fit, with standard error. Possible outcomes were A) Successful identification of all break points; B) One extra break point identified; C) One break point missed; D) Correct number of breaks found, but one or more break points misidentified; and E) Complete failure to identify the correct break point combination by the regime shift detector script.

Figure Nyears: **Performance of Regime Shift Detector Script under varied time series length.** Proportion of results with a given outcome under varied simulation length in years. Sets of 0, 1, 2 and 3 break points were randomly generated from within the set of possible values, and data were simulated with a 20% shift of r and a 40% shift of K at the given break point, generated with a 5% random noise to simulate sampling error. Each series consisted of 25 to 33 years of simulated data and each scenario was iterated 500 times. Lines joining points represent a third order polynomial GAM representing the best fit, with standard error. Possible outcomes were A) Successful identification of all break points; B) One extra break point identified; C) One break point missed; D) Correct number of breaks found, but one or more break points misidentified; and E) Complete failure to identify the correct break point combination by the regime shift detector script.