JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2012) **9**, 2527–2539
doi:10.1098/rsif.2012.0125
Published online 16 May 2012

Quantifying limits to detection of early warning for critical transitions

Carl Boettiger[1,*] and Alan Hastings[2]

[1]*Center for Population Biology, 1 Shields Avenue, and* [2]*Department of Environmental Science and Policy, University of California, Davis, CA 95616, USA*

Catastrophic regime shifts in complex natural systems may be averted through advanced detection. Recent work has provided a proof-of-principle that many systems approaching a catastrophic transition may be identified through the lens of early warning indicators such as rising variance or increased return times. Despite widespread appreciation of the difficulties and uncertainty involved in such forecasts, proposed methods hardly ever characterize their expected error rates. Without the benefits of replicates, controls or hindsight, applications of these approaches must quantify how reliable different indicators are in avoiding false alarms, and how sensitive they are to missing subtle warning signs. We propose a model-based approach to quantify this trade-off between reliability and sensitivity and allow comparisons between different indicators. We show these error rates can be quite severe for common indicators even under favourable assumptions, and also illustrate how a model-based indicator can improve this performance. We demonstrate how the performance of an early warning indicator varies in different datasets, and suggest that uncertainty quantification become a more central part of early warning predictions.

Keywords: early warning signals; tipping point; alternative stable states; likelihood methods

1. INTRODUCTION

There is an increasing recognition of the importance of regime shifts or critical transitions at a variety of scales in ecological systems [1–6]. Many important ecosystems may currently be threatened with collapse, including corals [7], fisheries [8], lakes [6] and semi-arid ecosystems [9]. Given the potential impact of these shifts on the sustainable delivery of ecosystem services [10] and the need for management to either avoid an undesirable shift or else to adapt to novel conditions, it is important to develop the ability to predict impending regime shifts based on early warning signs.

A number of particular systems have demonstrated the kinds of relationships that would produce regime shifts, including dynamics of coral reefs [11], and simple models of metapopulations with differing local population sizes [12]. In cases like these, one sensible approach to understanding whether a regime shift would be likely to fit the model is to use either a time series or else independent estimates of parameters. More generally, with a good model of the system, detail-oriented approaches could be useful [13]. In this treatment, we focus on the situation where these more detailed models are not available.

Indeed, for many ecological systems, specific models are not available and general approaches are needed [4,13] that do not depend on estimating the parameters of a known model of a specific system. This has led to a variety of approaches based on summary statistics [6,14–19] that look for generic signs of impending regime shifts. Here, we extend earlier work by providing estimates of the ability of different potential indicators to accurately signal impending regime shifts, and develop new approaches that both are more efficient and also lay bare some of the important assumptions underlying attempts to find general warning signs of regime shifts. We distinguish this question from the extensive literature involving a change-point analysis for the post hoc identification of if and when a regime shift has occurred [20–22], which is of little use if the goal is the advanced detection of the shift.

We begin by discussing the limitations of current approaches that rely on summary statistics and provide a description of assumptions through the introduction of a model-based approach to detect early warning signals. We then illustrate how stochastic differential equation (SDE) models can be used to reflect the uncertainty inherent in the detection of early warning signals. We caution against paradigms that are not useful for capturing uncertainty in a model-selection-based approach, such as information criteria. Finally, we use receiver-operating characteristics (ROC) [23,24] as a way to illustrate the sensitivity that different datasets and different indicators have in detecting early warning signals and use this to explore a number of examples. This approach provides a visualization of the types of errors that arise and how one can trade off between them, and is important for framing the problem as one focused on prediction.

*Author for correspondence (cboettig@ucdavis.edu).

Received 16 February 2012
Accepted 24 April 2012

2. THE SUMMARY STATISTICS APPROACH

Foundational work on early warning signals has operated under the often-implicit assumption that the system dynamics contain a saddle–node bifurcation by looking for patterns that are associated with this kind of transition. A saddle–node bifurcation occurs when a parameter changes and a stable equilibrium (node) and an unstable equilibrium (saddle) coalesce and disappear. The system then moves to a more distant equilibrium. Guckenheimer & Holmes [25] or any other textbook on dynamical systems will provide precise definitions and further explanation.

Typical patterns used as warning signals include an increasing trend in a summary statistic, such as variance [14], autocorrelation [15,16], skew [17] and spectral ratio [18]. While attractive for their simplicity, such approaches must confront numerous challenges. In this paper, we argue for a model-based approach to warning signals, and describe how this can be done in a way that best addresses these difficulties. We begin by enumerating several of the difficulties encountered in approaches lacking an explicit model.

2.1. Hidden assumptions

The underlying assumption that the system contains a saddle–node bifurcation can be easily overlooked in common summary-statistics-based approaches. For instance, variance may increase for reasons that do not signal an approaching transition [26,27]. Alternatively, variance may not increase as a bifurcation is approached [28,29]. Some classes of sudden transitions may exhibit no warning signals [30]. Like saddle–node bifurcations, transcritical bifurcations involve an eigenvalue passing through zero, and exhibit the patterns of critical slowing down and increased variance [5]. However, transcritical bifurcations involve a change in stability of a fixed point, rather than the sudden disappearance of a fixed point that has made critical transitions so worrisome. While no approach will be applicable to all classes of sudden transitions, it is certainly still useful to have an approach that detects transitions driven by saddle–node bifurcations, which have been found in many contexts [3].

Even when we can exclude or ignore other dynamics and restrict ourselves to systems that can produce a saddle-node bifurcation, approaches based on critical slowing down or rising variance [4,6,15] must further assume that a changing parameter has brought the system closer to the bifurcation. This assumption excludes at least three alternative explanations for the transition in system behaviour. The first possibility is that a large perturbation of the system state has moved the system into the alternative basin of attraction [3]. This is an exogenous forcing that does not arise from the system dynamics; so it is not the kind of event we can expect to forecast. (An example might be a sudden marked increase in fishing effort that pushes a harvested population past a threshold.) The second scenario is a purely noise-induced transition, a chance fluctuation that happens to carry the system across the boundary [31]. Livina *et al.* [28] indicate that such noise-induced transitions cannot be predicted through early warning signals—at least they are not expected to exhibit the same early warning patterns of increased variance and increased autocorrelation anticipated in the case of a saddle–node bifurcation. The third scenario is that the system does pass through a saddle–node bifurcation, but rather than gradually and monotonically approaching the critical point, the bifurcation parameter moves in a rapid or highly nonlinear way, making the detection of any gradual trend impossible.

2.2. Arbitrary windows

In addition to the assumption of a saddle–node bifurcation, the calculation of statistics that would be used to detect an impending transition is subject to several arbitrary choices. A basic difficulty arises from the need to assume a time-series is *ergodic*: that averaging over time is equivalent to averaging over replicate realizations, while trying to test if it is not. Theoretically, the increasing trend in variance, autocorrelation or other statistics is something that would be measured across an ensemble—across replicates. As true replicates are seldom available in systems for which developing warning signals would be most desirable, typical methods average across a single replicate using a moving window in time. The selection of the size of this window and whether and by how much to overlap consecutive windows varies across the literature. Lenton *et al.* [32] demonstrate that these differences can influence the results, and that the different choices each carry advantages and disadvantages.

In addition to introducing the challenge of selecting a window size, this ergodic assumption raises further difficulties. While appropriate for a system that is stationary, or changing slowly enough in the window that it may appear stationary, the assumption is at odds with the original hypothesis that the system is approaching a saddle–node bifurcation.

Further, certain statistics such as the critical slowing down measured by autocorrelation require data that is evenly sampled in time. Interpolating from existing data to create evenly spaced points is particularly problematic, as this introduces an artificial autocorrelation into the data.

2.3. No quantitative measures

Summary statistics typically invoke qualitative patterns such as an increase in statistic x, rather than a quantitative measure of the early warning pattern. This makes it difficult to compare between signals or to attribute a statistical significance to the detection. Some authors have suggested that Kendall's correlation coefficient, τ, could be used to quantify an increase [16,33] in autocorrelation or variance. Other measures of increase, such as Pearson's correlation coefficient, have also been proposed [5], while most of the literature simply forgoes quantifying the increase or estimating significance. While adequate in experimental systems that can compare patterns between controls and replicates [5,6], any real-world application of these approaches must be useful on a single time-series of observations. In these cases, a quantitative definition of a statistically significant detection is essential. Without this, we have no assurance that a purported detection is not, in fact, a false positive. By focusing primarily on examples known to be approaching a transition when testing warning signals, the probability of false positives has largely been overlooked.

2.4. Problematic null models

Specifying an appropriate null model is also difficult. Non-parametric null hypotheses seem to require the fewest assumptions but in fact can be the most problematic. For instance, the standard non-parametric hypothesis test with Kendall's τ rank correlation coefficient assumes only that the two variables are independent, but this is an assumption that is violated by the very experimental design: temporal correlations will exist in any finely enough sampled time series, and moving windows introduce temporal correlations in the statistics. Under such a test, any adequately large dataset will find a significant result, regardless of whether a warning signal exists. A similar problem arises when the points in the time series are reordered to create a null hypothesis—this destroys the natural autocorrelation in the time series. More promising parametric null models have been proposed, such as autoregressive models in Dakos *et al.* [16], bringing us closer to a model-based approach with explicit assumptions. Others have looked for alternative summary statistics where reasonable null models are more readily available, such as Seekkell *et al.*'s [19] proposal to test for conditional heteroscedasticity.

2.5. Summary-statistic approaches have less statistical power

Methods for the detection of early warning signals are continually challenged by inadequate data [4,6,15–17,34–36]. Despite the widespread recognition of the need for large datasets, there have been very few quantitative studies of power to determine how much data are required [37], how often a particular method would produce a false alarm or fail to detect a signal, and which tests will be the most powerful or sensitive. The Neyman–Pearson lemma demonstrates that the most powerful test between hypotheses compares the likelihood that the data were produced under each [38]. Such likelihood calculations require a model-based approach.

3. A MODEL-BASED APPROACH

Model-based approaches are beginning to play a larger role in early warning signal detection, though we have not as yet seen the direct fitting and simulation of models to compare hypotheses. Although choosing appropriate models without system-specific knowledge is challenging, much can be accomplished by framing the implicit assumptions into equations. Lade & Gross [13] introduce the idea of generalized models for early warning signals, and Kuehn [39] presents normal forms for bifurcation processes that can give rise to critical transitions. Carpenter & Brock [40] and Dakos *et al.* [29] start by assuming that the dynamics obey a generic SDE, but use this only to derive or define the summary statistics of interest.

In this section, we outline how the detection of early warning signals may be thought of as a problem of model choice. We next show that generic models can be constructed under the assumptions discussed earlier and estimated from the data in a maximum-likelihood framework. We highlight the disadvantages of comparing these estimates by information criteria, and instead introduce a simulation or bootstrapping approach rooted in Cox [41] and McLachlan [42] that characterizes the rate of missed detections and false alarms expected in the estimate.

3.1. Early warning signals as model choice

It may be useful to think of the detection of early warning signals as a problem of model choice rather than one of pattern recognition. The model choice approach attempts to frame each of the possible scenarios as structurally different equations, each with unknown parameters that must be estimated from the data. In any model choice problem, it is important to identify the goal of the exercise—such as the ability to generalize, to imitate reality or to predict [43]. In this case, generality is more important than realism or predictive capability: we will write down a general model that is capable of approximating a wide class of models in which regime shifts are characterized by a saddle–node bifurcation, and a second generic model that is capable of representing the behaviour of such systems when they are not approaching a bifurcation. These may be thought of as the hypothesis and null hypothesis, though they are in fact compound hypotheses, as we must first estimate the model parameters from the data. In this approach, it is not assumed that 'reality' is included in the models being tested, but that one of the models is a better approximation of the true dynamics than the other. System whose dynamics violate the assumptions common to both models, such as in the examples of Hastings & Wysham [30] where systems exhibit sudden transitions without warning, fall outside the set of cases where this approach would be valid; though the inability of either model to match the system dynamics could be an indication of such a violation.

3.2. Models

In the neighbourhood of a bifurcation, a system can be transformed into its *normal form* by a change of variables to facilitate analysis [25]. The normal form [25,39] for the saddle–node bifurcation is

$$\frac{\mathrm{d}x}{\mathrm{d}t} = r_t - x^2. \tag{3.1}$$

where x is the state variable and r_t our bifurcation parameter. We have added a subscript t to the bifurcation parameter as a reminder that it is the value which may be slowly varying in time and consequently moving the system closer to a critical transition or regime shift [4]. Transforming this canonical form to allow for an arbitrary mean in the state variable θ, the system near the bifurcation looks like $\mathrm{d}x/\mathrm{d}t = r_t - (\theta - x)^2$, with fixed point $\hat{x} = \sqrt{r_t} + \theta =: \phi(r_t)$. We expand around the fixed point and express as an SDE [44]:

$$\mathrm{d}X = \sqrt{r_t}(\phi(r_t) - X_t)\mathrm{d}t + \sigma\sqrt{\phi(r_t)}\mathrm{d}B_t, \qquad (3.2)$$

where B_t is the standard Brownian motion. This expression captures the behaviour of the system near the stable point as it approaches the bifurcation. Allowing the stochastic term to scale with the square root of ϕ follows from the assumption that of an internal-noise process, such as demographic stochasticity, that arises in deriving the SDE from a Markov process, see Kampen [45] or Black & McKane [46]. The square root could be removed for an external noise process, such as environmental noise. In practice, it will be difficult to discriminate between the square root and linear scaling in these applications, because the average value of the state changes little before the bifurcation.

As we discussed earlier, in this paradigm we must include an assumption on how the bifurcation parameter, r_t, is changing. We assume a gradual, monotonic change that we approximate to first order:

$$r_t = r_0 - mt. \qquad (3.3)$$

Detecting accelerating or otherwise nonlinear approaches to the bifurcation will generally require more power. When the underlying system is not changing, r_t is constant ($m = 0$) and equation (3.2) will reduce to a simple Ornstein–Uhlenbeck (OU) process,

$$\mathrm{d}X_t = r(\theta - X_t)\mathrm{d}t + \sigma\mathrm{d}B_t. \qquad (3.4)$$

This is the continuous time analogue of the first-order autoregressive model considered as a null model elsewhere [16,47].

3.3. Likelihood calculations

The probability $P(X|M)$ of the data X given the model M is the product of the probability of observing each point in the time series given the previous point and the length of the interval,

$$\log P(X|M) = \sum_i \log P(x_i|x_{i-1}, t_i). \qquad (3.5)$$

For (3.2) or (3.4), it is sufficient [44] to solve the moment equations for mean and variance, respectively,

$$\frac{\mathrm{d}}{\mathrm{d}t}E(x|M) = f(x) \qquad (3.6)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}V(x|M) = -\partial_x f(x)\,V(x|M) + g(x)^2. \qquad (3.7)$$

For the OU process, we can solve this in closed form over an interval of time t_i between subsequent observations

$$E(x_i|M = \mathrm{OU}) = X_{i-1}\mathrm{e}^{-rt_i} + \theta(1 - \mathrm{e}^{-rt_i}) \qquad (3.8)$$

and

$$V(x_i|M = \mathrm{OU}) = \frac{\sigma^2}{2r}(1 - \mathrm{e}^{-2rt_i}). \qquad (3.9)$$

For the time-dependent model, we have analytic forms only for the dynamical equations of these moments from equation (3.7), which we must integrate numerically over each time interval. The moments of equation (3.2) are given by

$$\frac{\mathrm{d}}{\mathrm{d}t}E(x_i|M = \mathrm{LSN}) = 2\sqrt{r(t)}(\sqrt{r(t)} + \theta - x_i) \qquad (3.10)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}V(x_i|M = \mathrm{LSN}) = -2\sqrt{r(t)}\,V(x_i) + \sigma^2(\sqrt{r(t)} + \theta). \qquad (3.11)$$

These are numerically integrated using `lsoda` routine available in `R` for the likelihood calculation.

3.4. Comparing models

Likelihood methods form the basis of much of modern statistics in both Frequentist and Bayesian paradigms. The ability to evaluate likelihoods directly by computation has made it possible to treat cases that do not conform to traditional assumptions more directly. The basis of likelihood comparisons has its roots in the Neyman–Pearson Lemma, which essentially asserts that comparing likelihoods is the most powerful test of a choice between two hypotheses [38], and motivates tests from the simple likelihood ratio test up through modern model adequacy methods.

The hypotheses considered here are more challenging than the original lemma provides for, as they are composite in nature: they specify two model forms (stable and changing stability) but with model parameters that must be first estimated from the data. Comparing models whose parameters have been estimated by maximum likelihood is first treated by Cox [41,48], and has since been developed in this simulation estimation of the null distribution [42], by parametric bootstrap estimate [49]. Cox's δ statistic (often called the deviance between models) is simply the difference between the log likelihoods of these maximum-likelihood estimates, defined as follows.

Let L_0 be the likelihood function for model 0, let $\theta_0 = \arg\max \theta_0 \in \Omega_0$, $(L_0(\theta_0|X))$ be the maximum-likelihood estimate for θ_0 given X and let $L_0 = L_0\theta_0|X$; and define L_1, θ_1, L_1 similarly for model 1. The statistic we will use is δ, defined to be twice the difference in log likelihood of observing the data under the two MLE models,

$$\delta = -2(\log L_0 - \log L_1). \qquad (3.12)$$

This approach has been applied to the problem of model adequacy [50] and model choice [51] in other contexts. We have extended the approach by generating the test distribution as well as a null distribution of the statistic δ.

3.5. Simulation-based comparisons

We perform the identical analysis procedure described earlier on each of these three datasets. First, we estimate parameters for the null and test model to each dataset by maximum likelihood. Comparing the

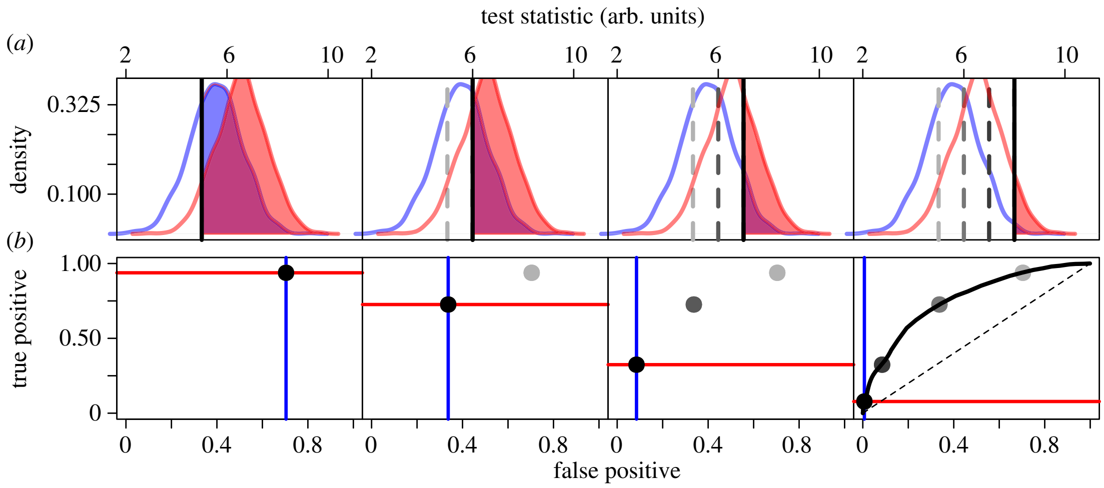

Figure 1. (*a*) The distributions of a hypothetical warning indicator are shown under the case of a stable system (blue) and a system approaching a critical transition (red). (*b*) Points along the ROC curve are calculated for each possible threshold indicated in (*a*). The false positive rate is the integral of the distribution of the test statistic under the stable system right of the threshold (blue shaded area, corresponding to blue vertical line). The true positive rate is the integral of the system approaching a transition left of the threshold (red shaded area, corresponds to the red line). Successive columns show the threshold increasing, tracing out the ROC curve. (Online version in colour.)

likelihood of these fits directly gives us only a minimal indication of which model fits better. To identify if these differences are significant, and by what probability they could arise as a false alarm or a missed event, we simulate 500 replicate time series from each estimated model.

The model parameters of both models are re-estimated on both families of replicates (the null and test, i.e. 2 × 2 × 500 fits). The differences in the likelihood values between the model estimates produced from the first set of simulations determines the null distribution for the deviance statistic δ. As the constant OU process model is nested within the time-heterogeneous model, these values are always positive, but tend to be not as large as those produced when the models are fit to the second family of data.

The extent to which these distributions overlap indicates our inability to distinguish between these scenarios. The tendency of the observed deviance to fall more clearly in the domain of one distribution or the other indicates the probability our observed data corresponds best with that model—either approaching a critical transition or remaining stable. While it trivial to assign a statistical significance to this observation based on how far into the tail of the null distribution it falls, for the reasons we discussed we prefer the more symmetric comparison of the probability that this value was observed in either distribution. We visualize the trade-off between false alarms and failed detection using the ROC curves introduced earlier.

3.6. Information criteria will not serve

One will commonly observe models representing alternative processes being compared through the use of various information criteria such as the Akaike information criterion. While tempting to apply in this situation, such approaches are not suited to this problem for several reasons. The first is that information criteria are not concerned with the model choice objective we have in mind, as they are typically applied to find an adequate model description without too many parameters that the system may be over-fit. More pointedly, information criteria have no inherent notion of uncertainty. Information criteria tests alone will not tell us our chances of a false alarm, of missing a real signal or how much data we need to be confident in our ability to detect transitions.

3.7. Beyond hypothesis testing

It is possible to frame the question of sensitivity, reliability and adequate data in the language of hypothesis testing. This introduces the need for selecting a statistical significance criterion. In the hypothesis testing framework, a false positive is a type I error, which is defined relative to this arbitrary statistical significance criterion, most commonly 0.05. By changing the criterion, one can increase or decrease the probability of the type I error at the cost of decreasing or increasing false negative or type II error, which must also be defined relative to this criterion.

The language of hypothesis testing is built around a bias that false positives are worse than false negatives, and consequently an emphasis on p-values rather than power. In the context of early warning signals, this is perilous—it suggests that we would rather fail to predict a catastrophe than to sound a false alarm. To avoid this linguistic bias and the introduction of an nuisance parameter on which to define statistical significance, we propose the use of ROC curves.

3.8. Receiver-operating characteristic curves

We illustrate the trade-off between false alarms and failed detection using ROC curves first developed in signal-processing literature [23,24]. The curves represent the corresponding false alarm rate at any detection sensitivity (true positive rate; figure 1). The closer these distributions are to one-another, the

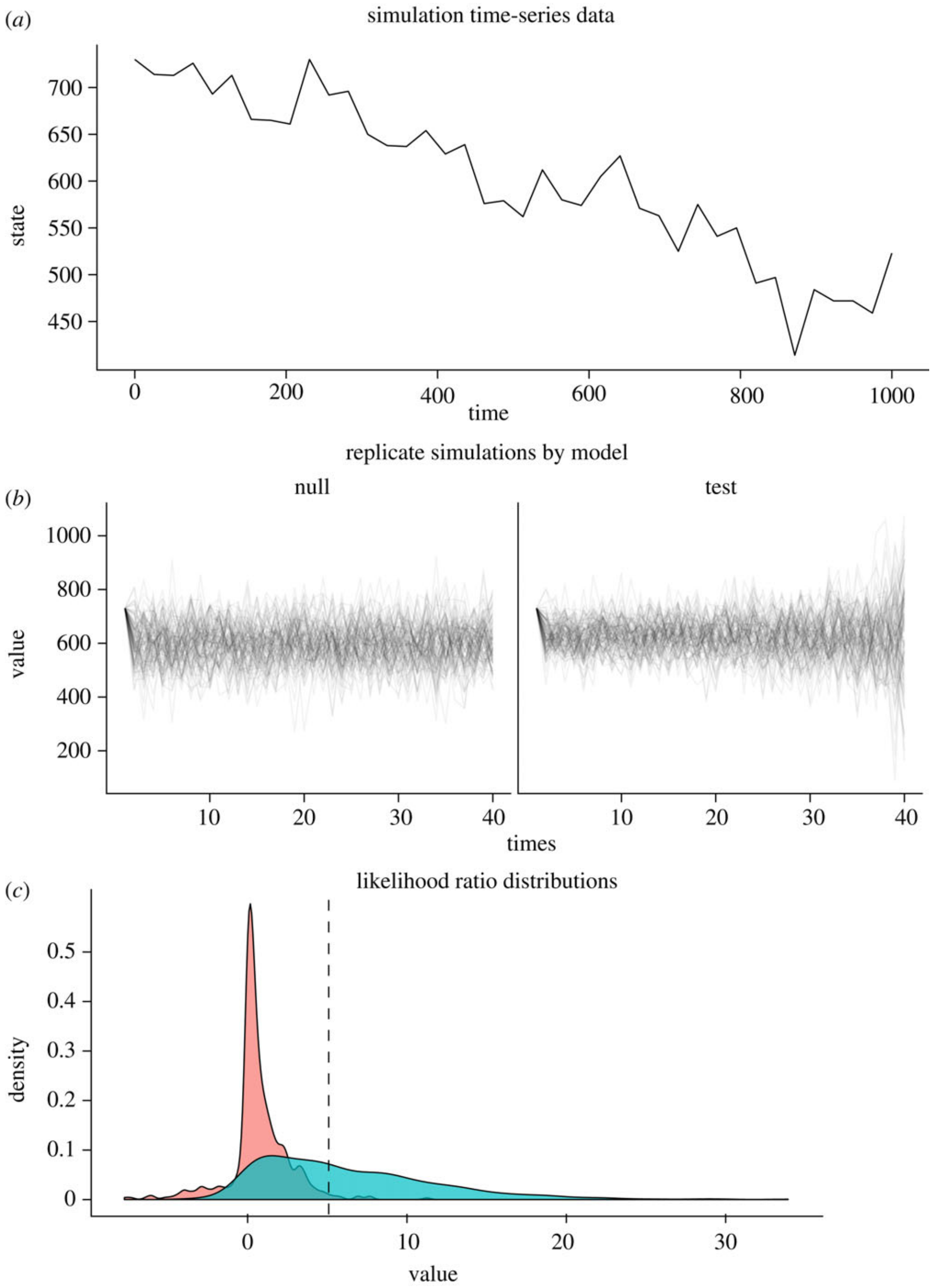

Figure 2. A model-based calculation of warning signals for the simulated data example. (*a*) The original time-series data on which model parameters for equations (3.2) and (3.4) are estimated. (*b*) Replicate simulations under the maximum-likelihood estimated (MLE) parameters of the null model, equation (3.4) and test model, equation (3.2). (*c*) The distribution of deviances (differences in log likelihood, equation (3.12)), when both null and test models are fit to each of the replicates from the null model, 'null', in red, and these differences when estimating for each of the replicates from the test model, in blue. The overlap of distributions indicate replicates that will be difficult to tell apart. The observed differences in the original data are indicated by the vertical line. (Online version in colour.)

more severe the trade-off. If the distributions overlap exactly, the ROC curve has a constant slope of unity. The ROC curve demonstrates this trade-off between accuracy and sensitivity. Different early-warning indicators will vary in their sensitivity to detect differences between stable systems and those approaching a critical transition, making the ROC curves a natural way to compare their performance. Because the shape of the curve will also depend on the duration and frequency of the time-series observations, we can use these curves to illustrate by how much a given increase in sampling effort can decrease the rate of false alarms or failed detections.

4. EXAMPLE RESULTS

We illustrate this approach on simulated data as well as several natural time-series that have been previously analysed for early warning signals. All data and code

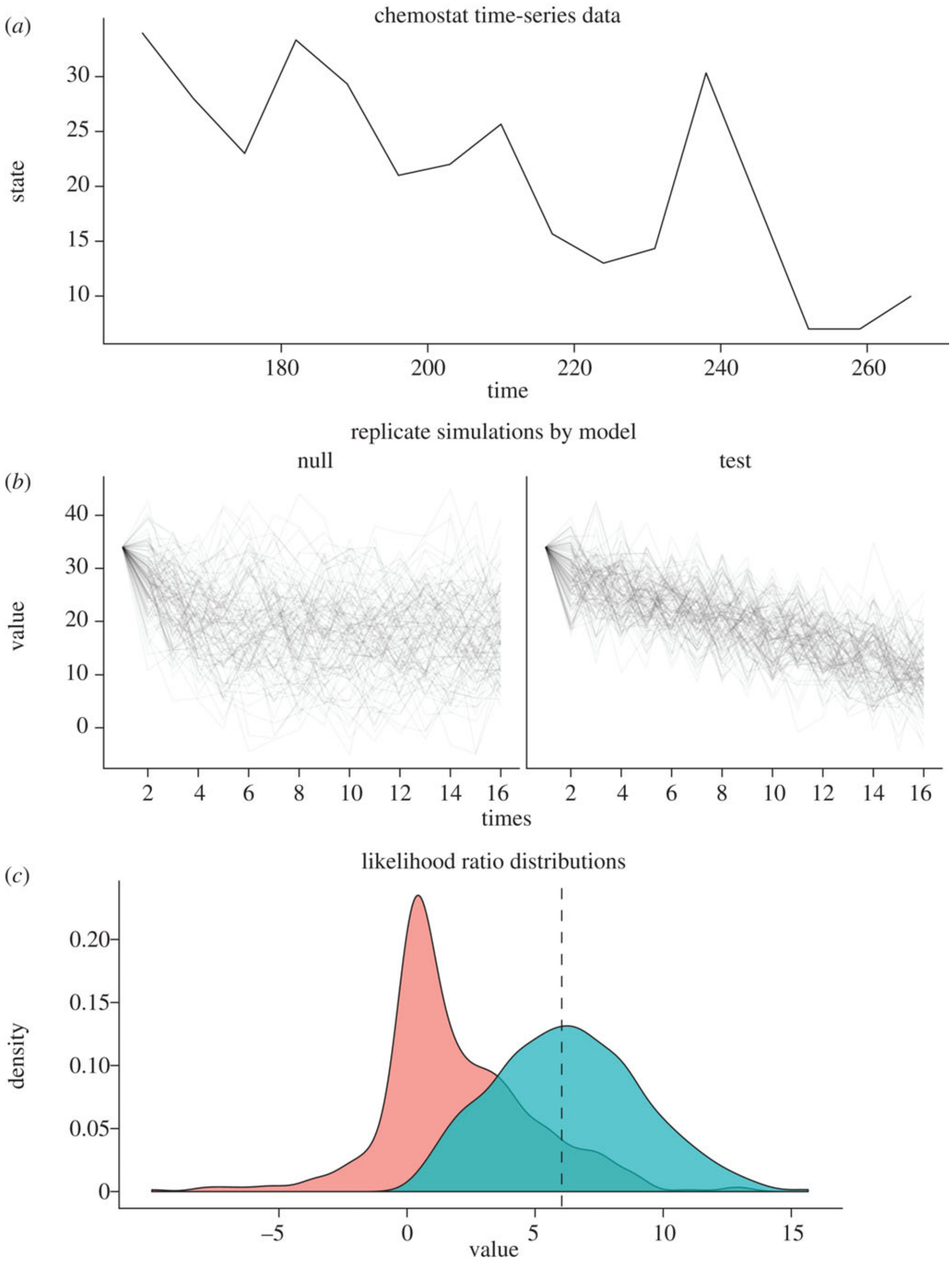

Figure 3. A model-based calculation of warning signals for the Daphnia data analysed in Drake & Griffen [5] (chemostat H6). Panels as in figure 2. (Online version in colour.)

for simulations and analysis are found in the accompanying R package, `earlywarning`.

4.1. Data

The simulation implements an individual, continuous-time stochastic birth–death process, with rates given by the master equation [44]

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = b_{n-1}P(n-1,t) + d_{n+1}P(n+1,t) - (b_n + d_n)P(n,t), \tag{4.1}$$

$$b_n = \frac{eKn^2}{n^2+h^2} \tag{4.2}$$

and

$$d_n = en + a_t, \tag{4.3}$$

where $P(n, t)$ is the probability of having n individuals at time t, b_n is the probability of a birth event occurring in a population of n individuals an d_n the probability of a death. e, K, h and a_t are parameters. This corresponds to the well-studied ecosystem model of over-exploitation [52,53], with stochasticity introduced directly through the demographic process. We select this model since it is has discrete numbers of individuals, nonlinear processes and the noise is driven by Poisson process of births and deaths instead of a Gaussian, and thus provides an illustration that our approach is robust to the violations of those assumptions in model (3.2).

This model is forced through a bifurcation by gradually increasing the a parameter, which increases can be thought of as an increasing toxicity of the environment (from $a_0 = 100$ increasing at constant rate of 0.09 units/

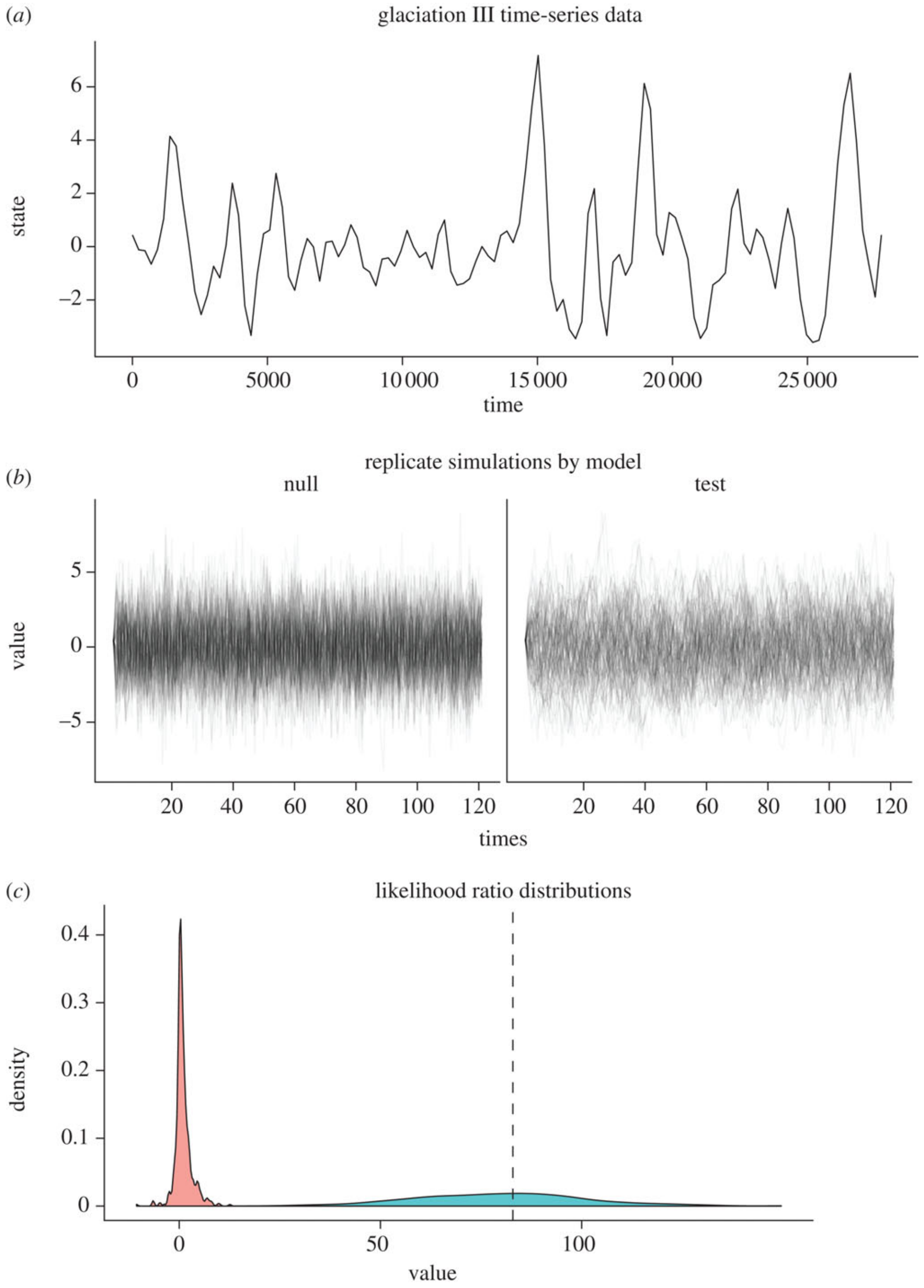

Figure 4. A model-based calculation of warning signals for the glaciation data analysed in Dakos *et al.* [16] (glaciation III). Panels as in figure 2. (Online version in colour.)

unit time). Other parameters are $Xo = 730$, $e = 0.5$, $K = 1000$, $h = 200$. We run this model over a time interval from 0 to 500 and sample at 40 evenly spaced time points, which were used for a subsequent analysis. This sampling frequency was chosen to be representative of reasonable sampling in biological time-series, and provides enough points to detect a signal while not too many that errors can be avoided entirely. For the convenience of the inquisitive reader, we have also provided a simple function in the associated R package where the user can vary the sampling scheme and parameter values and rerun this analysis. This time series is shown in figure 2*a*.

The first empirical dataset comes from the population dynamics of *Daphnia* living in the chemostat ‘H6’ in the experiments of Drake & Griffen [5]. This individual replicate was chosen as an example that showed a pattern of increasing variance over the 16 data points where the system was being manipulated towards a crash. This time series is shown in figure 3*a*.

Our second empirical dataset comes from the glaciation record seen in deuterium levels in Antarctic ice cores [54], as analysed by Dakos *et al.* [16]. The data are preprocessed by linear interpolation and de-trending by Gaussian kernel smoothing to be as consistent as possible with the original analysis. We focus on the third glaciation event, consisting of 121 sample points. The match is not exact because [16] estimates the de-trending window size manually, but the estimated correlations in the first-order auto-regression coefficients are in close agreement with that analysis. De-trending is intended to make the data consistent with the assumptions of the warning signal detection [16], which did not apply to the other datasets [5]. This time series is shown in figure 4*a*.

4.2. Analysis

The deviances δ observed are 5.1, 6.0, 83.9 for the simulation, the chemostat data and the glaciation data, respectively. On the basis of AIC score, each is large enough to reject the null hypothesis of a stable model with its one extra parameter, but this does not give the full picture of the anticipated error rates. The size of these differences reflects not only the magnitude of the difference in fit between the models but also the arbitrary units of the raw likelihoods, which are smaller for larger datasets. Consequently, the glaciation score reflects as much the greater length of its time series as it does anything else.

Our simulation approach can provide a better sense of the relative trade-off in error rates associated with these estimates. As described already (§3.1), we simulate 500 replicates under each model, shown in figures 2*b*–4*b*, and determine the distributions in likelihood ratio under each, shown in the lower panels. The observed deviance from the original data is also indicated (vertical line).

The ROC curves for each of these datasets are plotted in figure 5. While differences in the rate at which the system approaches a transition will also improve the ratio of true positives to false positives, here we see the best-sampled dataset, glaciation, with 121 points, also has the clearest signal with no observed errors in the 500 replicates of each type. Comparing the chemostat and simulation curves illustrate how the trade-off between false positives and true positives can vary between data. The chemostat signal, which estimates a relatively rapid rate of change but has less data, captures a higher rate of true positives for a given rate of false positives than the simulation dataset with a weaker rate of change but more data, for false positive rates above 20 per cent. However, the simulated set with more data performs better if lower false positive rates are desired.

5. COMPARING THE PERFORMANCE OF SUMMARY STATISTICS AND MODEL-BASED APPROACHES

Owing to the variety of ways in which early warning signals based on summary statistics are implemented and evaluated, it is difficult to give a straight-forward comparison between them and the performance of this model-based approach. However, by adopting one of one of the quantitative measures of a warning signal pattern, such as Kendall's τ [16,33,55], we are able to make a side-by-side comparison of the different summary statistics and the model-based approach in the context of false alarms and failed detections shown by the ROC curve. Values of τ near unity indicate a strongly increasing trend in the warning indicator, which is supposed to be indicative of an approaching transition. Values near zero suggest a lack of a trend, as expected in stable systems.

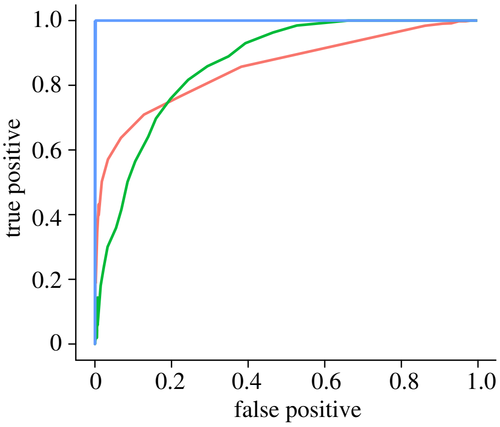

Figure 5. ROC curves for the simulation (red), chemostat (green) and glaciation (blue) data, computed from the distributions shown in figures 2*c*–4*c*. (Online version in colour.)

Figure 6 shows the time series for each dataset in columns and the early warning indicators of variance and autocorrelation computed over a sliding window for each. Kendall's correlation coefficient τ is calculated for each warning indicator and displayed on the graphs (inset). For comparison, the left-most column includes data simulated under a stable system, which nevertheless shows a chance increasing autocorrelation with a $\tau = 0.7$. We can adapt the approach we have described earlier to determine how often such a strong increase would appear by chance in a stable system as follows.

By estimating the stable and critical transition models from the data, and simulating 500 replicate datasets under each as in the earlier-mentioned analysis, we can then calculate the warning signals statistic over a sliding window of size equal to one-half the length of the time series, and compute the correlation coefficient τ measuring the degree to which the statistic shows an increasing trend. This results in a distribution of τ values coming from a model of a stable system, and a corresponding distribution of τ values coming from the model with an impending transition. These distributions are shown in figure 7. Contrary to the expectation that the replicates of the null model (stable system, equation (3.4)) would cluster around zero, while the test model, equation (3.2), would cluster around larger positive τ values, the observed τ values on the replicates extend evenly across the range. This results in a marked overlap and offers little ability to distinguish between the stable replicates and the replicates approaching a transition.

The use of box plots in figure 7 provide a convenient and familiar way to visualize the overlap between more than two distributions, though they lack the resolution of the overlapping density distributions in figures 2–4. The overlapping distributions are the natural representation from which to introduce the ROC curve, as in figure 1.

The ROC curves for these data (figure 8) show that the summary-statistic-based indicators frequently

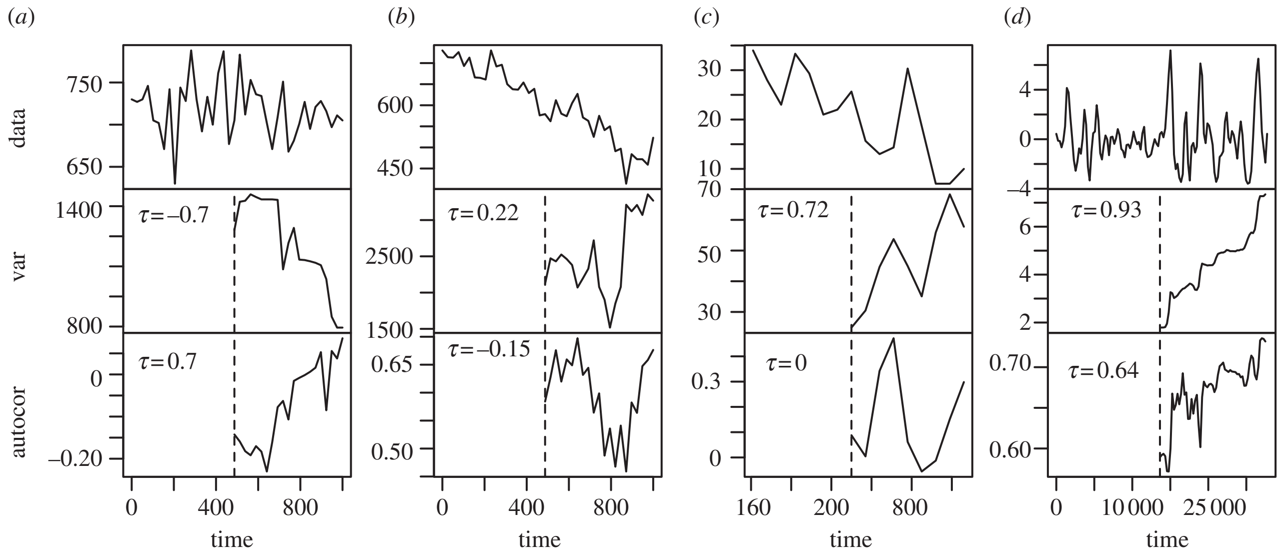

Figure 6. Early warning signals in simulated and empirical datasets. The first two columns are simulated data from (*a*) a stable system (stable), and (*b*) the same system approaching a saddle–node bifurcation (deteriorating). Empirical examples are from (*c*) *Daphnia magna* concentrations manipulated towards a critical transition (daphnia), and (*d*) deuterium concentrations previously cited as an early warning signal of a glaciation period (glaciation). Increases in summary statistics, computed over a moving window, have often been used to indicate if a system is moving towards a critical transition. The increase is measured by the correlation coefficient τ. Note that positive correlation does not guarantee that the system is moving towards a transition, as seen in the stable system, first column.

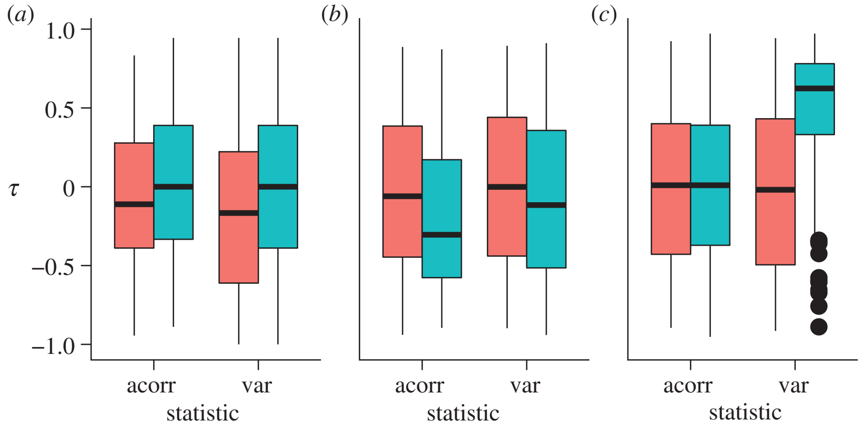

Figure 7. Box-plots of the distributions of Kendall's τ observed for the summary statistic methods variance and autocorrelation, applied to three different datasets (from figures 2–4). The distributions show extensive overlap, suggesting that it will be difficult to distinguish early warning signals by the correlation coefficient in these summary statistics. Red denotes null model; green, test. (*a*) chemostat, (*b*) glaciation, (*c*) simulation. (Online version in colour.)

lack the sensitivity to distinguish reliably between observed patterns from a stable or unstable system. The large correlations observed in the empirical examples (figure 6) are not uncommon in stable systems. It is notable that in both empirical examples, the summary statistics approach does little better than chance in distinguishing replicates that have been simulated from models (3.2) and (3.4), despite the fact that these models correspond to the assumptions of the summary statistics approaches. On the simulated data, the variance-based method approaches the true positive rate of our likelihood method at higher levels of false positives, but performs worse when the desired level of false positives is low. The ROC curve helps us to compare the performance of the different approaches at different tolerances. For instance, table 1 shows the fraction of true crashes caught at a 5 per cent false positive rate. We can instead set a desired True positive rate and read off the resulting number of false alarms, table 2.

6. DISCUSSION

The challenge of determining early warning signs for impending possible regime shifts requires real attention to the underlying statistical issues and other assumptions. Doing this, does, however, open up new possibilities for asking what the goal of detection should be, and for clearly identifying underlying assumptions. We consider alternative approaches based either on summary statistics or on a likelihood-based model choice. By assuming

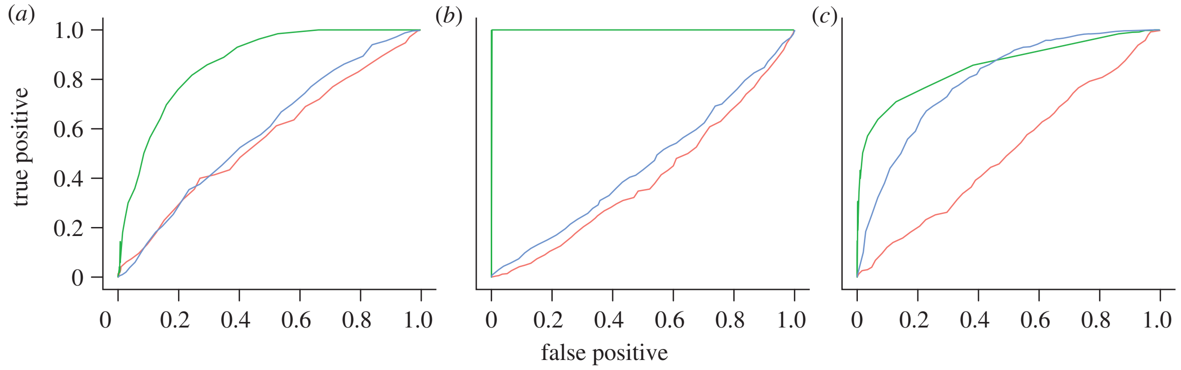

Figure 8. ROC curves compare the performance of the summary statistics variance (blue) and autocorrelation (red) against the likelihood-based (green) approach from figure 5 on each of three example datasets (figures 2–4). (*a*) chemostat, (*b*) glaciation, (*c*) simulation. (Online version in colour.)

Table 1. Fraction of *crashes detected* when the desired false alarm rate is fixed to 5%.

	variance (%)	likelihood (%)
simulation	25	61
chemostat	5.0	34
glaciation	5.4	100

Table 2. Fraction of *false alarms* when the desired detection rate is fixed to 90%.

	variance (%)	likelihood (%)
simulation	49	55
chemostat	81	35
glaciation	93	0

that the underlying model corresponds to a saddle–node bifurcation, our analysis presents a ‘best-case scenario’ for both summary statistic and likelihood-based approaches. Other literature has already begun to address the additional challenges posed when the underlying dynamics do not correspond to these models [30]. Our results illustrate that even in this best-case scenario, reliable identification of warning signals from summary statistics can be difficult.

We have used three examples to illustrate the performance of this approach in data from simulation, a chemostat experiment and paleo-atmospheric record; examples differing in sampling intensity and strength of signal of an approaching collapse. While the well-sampled geological data shows an unmistakable signal in this model-based approach, the uncertainty in the smaller simulated and experimental data forces a trade-off between errors.

As a way to clearly illustrate the choices involved in looking for warning signals while avoiding false alarms, we introduce an approach based on receiver operator curves. These curves illustrate the extent to which an potential warning signal mitigates the trade-off between missed events and false alarms. The extent of the difficulty in finding reliable indicators of impending regime shifts based on summary statistics becomes clear from the ROC curves of these statistics, where a 5 per cent false positive rate often corresponds to only a 5 per cent true positive rate, performing no better than the flip of a coin. By estimating the ROC curve for a given set of data, we can better avoid applying warning signals in cases of inadequate power. By taking advantage of the assumptions being made to write down a specific likelihood function, we can develop approaches that get the most information from the data available.

In any application of early warning signals, it is essential to address the question of model adequacy. Our approach formalizes the assumptions about the underlying process to match the assumptions of the other warning signals. As the bifurcation results from the principle eigenvalue passing through zero, the warning signal is expected in linear-order dynamics; estimation of the nonlinear model is less powerful and less accurate. The performance of this approach in the simulated data—which is nonlinear in its dynamics and driven with non-Gaussian noise introduced by the Poisson demographic events—demonstrates the accuracy under violation of these assumptions.

The conclusion is not simply that likelihood approaches are more reliable, but rather more broadly that warning signals should consider the inherent trade-off between sensitivity and accuracy, and must quantify how this trade-off depends on both the indicators used and the data available. The approach developed here estimates the risk of both failed detection and false alarms; concepts that are critical to prediction-based management. Using the methods, we have outlined when designing early warning strategies for natural systems can ensure that data collection has adequate power to offer a reasonable chance of detection.

We thank S. Schreiber, M. Holyoak, M. Baskett, A. Perkins, N. Ross, M. Holden and two anonymous reviewers for comments on the manuscript. This research was supported by funding from NSF grant EF 0742674 and a Computational Sciences Graduate Fellowship from the Department of Energy grant DE-FG02-97ER25308. Data and code are available at https://github.com/cboettig/earlywarning.

REFERENCES

1 Holling, C. S. 1973 Resilience and stability of ecological systems. *Annu. Rev. Ecol. Syst.* **4**, 1–23. (doi:10.1146/annurev.es.04.110173.000245)

2 Wissel, C. 1984 A universal law of the characteristic return time near thresholds. *Oecologia* **65**, 101–107. (doi:10.1007/BF00384470)

3 Scheffer, M., Carpenter, S. R., Foley, J. A., Folke, C. & Walker, B. 2001 Catastrophic shifts in ecosystems. *Nature* **413**, 591–596. (doi:10.1038/35098000)

4 Scheffer, M. *et al.* 2009 Early-warning signals for critical transitions. *Nature* **461**, 53–59. (doi:10.1038/nature08227)

5 Drake, J. M. & Griffen, B. D. 2010 Early warning signals of extinction in deteriorating environments. *Nature* **467**, 456–459. (doi:10.1038/nature09389)

6 Carpenter, J. 2011 May the best analyst win. *Science* **331**, 698–699. (doi:10.1126/science.331.6018.698)

7 Bellwood, D. R., Hughes, T. P., Folke, C. & Nyström, M. 2004 Confronting the coral reef crisis. *Nature* **429**, 827–833. (doi:10.1038/nature02691)

8 Berkes, F. *et al.* 2006 Globalization, roving bandits, and marine resources. *Science* **311**, 1557–1558. (doi:10.1126/science.1122804)

9 Kéfi, S., Rietkerk, M., Alados, C. L., Pueyo, Y., Papanastasis, V. P., Elaich, A. & de Ruiter, P. C. 2007 Spatial vegetation patterns and imminent desertification in Mediterranean arid ecosystems. *Nature* **449**, 213–217. (doi:10.1038/nature06111)

10 Folke, C., Carpenter, S. R., Walker, B., Scheffer, M., Elmqvist, T., Gunderson, L. & Holling, C. 2004 Regime shifts, resilience, and biodiversity in ecosystem management. *Annu. Rev. Ecol. Evol. Syst.* **35**, 557–581. (doi:10.1146/annurev.ecolsys.35.021103.105711)

11 Mumby, P. J., Hastings, A. & Edwards, H. J. 2007 Thresholds and the resilience of Caribbean coral reefs. *Nature* **450**, 98–101. (doi:10.1038/nature06252)

12 Hastings, A. 1991 Structured models of metapopulation dynamics. *Biol. J. Linnean Soc.* **42**, 57–71. (doi:10.1111/j.1095-8312.1991.tb00551.x)

13 Lade, S. J. & Gross, T. 2012 Early warning signals for critical transitions: a generalized modeling approach. *PLoS. Comput. Biol.* **8**, e1002360. (doi:10.1371/journal.pcbi.1002360)

14 Carpenter, S. R. & Brock, W. A. 2006 Rising variance: a leading indicator of ecological transition. *Ecol. Lett.* **9**, 311–318. (doi:10.1111/j.1461-0248.2005.00877.x)

15 Held, H. 2004 Detection of climate system bifurcations by degenerate fingerprinting. *Geophys. Res. Lett.* **31**, 1–4. (doi:10.1029/2004GL020972)

16 Dakos, V., Scheffer, M., van Nes, E. H., Brovkin, V., Petoukhov, V. & Held, H. 2008 Slowing down as an early warning signal for abrupt climate change. *Proc. Natl Acad. Sci. USA* **105**, 14 308–14 312. (doi:10.1073/pnas.0802430105)

17 Guttal, V. & Jayaprakash, C. 2008 Spatial variance and spatial skewness: leading indicators of regime shifts in spatial ecological systems. *Theoret. Ecol.* **2**, 3–12. (doi:10.1007/s12080-008-0033-1)

18 Biggs, R., Carpenter, S. R. & Brock, W. A. 2009 Turning back from the brink: detecting an impending regime shift in time to avert it. *Proc. Natl Acad. Sci. USA* **106**, 826–831. (doi:10.1073/pnas.0811729106)

19 Seekell, D. A., Carpenter, S. R. & Pace, M. L. 2011 Conditional heteroscedasticity as a leading indicator of ecological regime shifts. *Am. Nat.* **178**, 442–451. (doi:10.1086/661898)

20 Easterling, D. R. & Peterson, T. C. 1995 A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol.* **15**, 369–377. (doi:10.1002/joc.3370150403)

21 Rodionov, S. N. 2004 A sequential algorithm for testing climate regime shifts. *Geophys. Res. Lett.* **31**, 2–5. (doi:10.1029/2004GL019448)

22 Lenton, T. M., Myerscough, R. J., Marsh, R., Livina, V. N., Price, A. R. & Cox, S. J. Genie Team 2009 Using GENIE to study a tipping point in the climate system. *Phil. Trans. R. Soc. A* **367**, 871–884. (doi:10.1098/rsta.2008.0171)

23 Green, D. M. & Swets, J. A. 1989 *Signal detection theory and psychophysics.* Los Altos, CA: Peninsula Publication.

24 Keller, R. P., Lodge, D. M., Lewis, M. A. & Shogren, J. F. 2009 *Bioeconomics of invasive species: integrating ecology, economics, policy, and management.* Oxford, UK: Oxford University Press.

25 Guckenheimer, J. & Holmes, P. 1983 *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, vol. 42. Applied Mathematical Sciences. New York, NY: Springer.

26 Schreiber, S. 2003 Allee effects, extinctions, and chaotic transients in simple population models. *Theoret. Popul. Biol.* **64**, 201–209. (doi:10.1016/S0040-5809(03)00072-8)

27 Schreiber, S. J. & Rudolf, V. H. W. 2008 Crossing habitat boundaries: coupling dynamics of ecosystems through complex life cycles. *Ecol. Lett.* **11**, 576–87. (doi:10.1111/j.1461-0248.2008.01171.x)

28 Livina, V., Ditlevsen, P. & Lenton, T. 2012 An independent test of methods of detecting system states and bifurcations in time-series data. *Phys. A Stat. Mech. Appl.* **391**, 485–496. (doi:10.1016/j.physa.2011.08.025)

29 Dakos, V., van Nes, E. H., D'Odorico, P. & Scheffer, M. 2011 Robustness of variance and autocorrelation as indicators of critical slowing down. *Ecology* **93**, 264–271. (doi:10.1890/11-0889.1)

30 Hastings, A. & Wysham, D. B. 2010 Regime shifts in ecological systems can occur with no warning. *Ecol. Lett.* **13**, 464–472. (doi:10.1111/j.1461-0248.2010.01439.x)

31 Ditlevsen, P. D. & Johnsen, S. J. 2010 Tipping points: early warning and wishful thinking. *Geophys. Res. Lett.* **37**, 2–5. (doi:10.1029/2010GL044486)

32 Lenton, T. M., Livina, V. & Dakos, V. 2012 Early warning of climate tipping points from critical slowing down: comparing methods to improve robustness. *Phil. Trans. R. Soc. A* **370**, 1185–1204. (doi:10.1098/rsta.2011.0304)

33 Dakos, V., Kéfi, S., Rietkerk, M., Nes, E. H. V. & Scheffer, M. 2011 Slowing down in spatially patterned ecosystems at the brink of collapse. *Am. Nat.* **177**, E153–E166. (doi:10.1086/659945)

34 Inman, M. 2011 Sending out an SOS. *Nat. Climate Change* **1**, 180–183. (doi:10.1038/nclimate1146)

35 Scheffer, M. 2010 Complex systems: foreseeing tipping points. *Nature* **467**, 411–412. (doi:10.1038/467411a)

36 Bestelmeyer, B. T. 2011 Analysis of abrupt transitions in ecological systems. *Ecosphere* **2**, 129. (doi:10.1890/ES11-00216.1)

37 Contamin, R. & Ellison, A. M. 2009 Indicators of regime shifts in ecological systems: what do we need to know and when do we need to know it? *Ecol. Appl.* **19**, 799–816. (doi:10.1890/08-0109.1)

38 Neyman, J. & Pearson, E. 1933 On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A* **231**, 289–337. (doi:10.1098/rsta.1933.0009)

39 Kuehn, C. 2011 A mathematical framework for critical transitions: normal forms, variance and applications. (http://arxiv.org/abs/1101.2908)

40 Carpenter, S. & Brock, W. 2011 Early warnings of unknown nonlinear shifts: a nonparametric approach. *Ecology* **92**, 2196–2201. (doi:10.1890/11-0716.1)

41 Cox, D. R. 1961 Tests of separate families of hypotheses. In *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability, Berkeley, CA, 1961*, vol. 1. Berkeley, CA: University of California Press.

42 McLachlan, G. J. 1987 On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Stat.* **36**, 318. (doi:10.2307/2347790)

43 Levins, R. 1966 The strategy of model building in population biology. *Am. Sci.* **54**, 421–431.

44 Gardiner, C. 2009 *Stochastic methods: a handbook for the natural and social sciences*. Springer Series in Synergetics. New York, NY: Springer.

45 Kampen, N. V. 2007 *Stochastic processes in physics and chemistry*, 3rd edn. North Holland, The Netherlands: North-Holland Personal Library.

46 Black, A. J. & McKane, A. J. In press. Stochastic formulation of ecological models and their applications. *Trends Ecol. Evol.* (doi:10.1016/j.tree.2012.01.014)

47 Guttal, V. & Jayaprakash, C. 2008 Changing skewness: an early warning signal of regime shifts in ecosystems. *Ecol. Lett.* **11**, 450–460. (doi:10.1111/j.1461-0248.2008.01160.x)

48 Cox, D. R. 1962 Further results on tests of separate families of hypotheses. *J. R. Stat. Soc.* **24**, 406–424.

49 Efron, B. 1987 Better bootstrap confidence intervals. *J. Am. Stat. Assoc.* **82**, 171–185.

50 Goldman, N. 1993 Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**, 182–198. (doi:10.1007/BF00166252)

51 Huelsenbeck, J. P. & Bull, J. J. 1996 A likelihood ratio test to detect conflicting phylogenetic signal. *Syst. Biol.* **45**, 92–98. (doi:10.1093/sysbio/45.1.92)

52 Noy-Meir, I. 1975 Stability of grazing systems: an application of predator–prey graphs. *J. Ecol.* **63**, 459–481. (doi:10.2307/2258730)

53 May, R. M. 1977 Thresholds and breakpoints in ecosystems with a multiplicity of stable states. *Nature* **269**, 471–477. (doi:10.1038/269471a0)

54 Petit, J. R. *et al.* 1999 Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436. (doi:10.1038/20859)

55 Dakos, V., Nes, E. H., Donangelo, R., Fort, H. & Scheffer, M. 2009 Spatial correlation as leading indicator of catastrophic shifts. *Theoret. Ecol.* **3**, 163–174. (doi:10.1007/s12080-009-0060-6)