**Supplementary Material**

Appendix 1. CCM implementation

We estimated the causal influence of spawners to recruits by using CCM with leave-one-out cross validation (Sugihara et al. 2012). This method is based on Takens' Theorem, which implies that if a variable *y* is influenced by a variable *x*, then the time series of *y* contains information about the history of *x*. To test whether a time series of spawners $\{S_t\}$ contained information about a time series of recruits $\{R_t\}$ we reconstructed an E-dimensional manifold using E successive time lags of $\{S_t\}$, each separated by a time step $\tau$: $< S_t, S_{t-\tau}, \ldots S_{t-(E-1)\tau} >$ and where E is referred as the embedding dimension, which is a property of a time series that helps to identify its complexity in terms of the number of variables. Then, we tested if nearby points in the manifold of $S_t$ can be used to identify nearby points in the manifold of $R_t$ and thereby forecast future trajectories. In our case, the value of $\tau$ was fixed at 1 for all time series, as some of them were not long enough to test larger values. E was selected for each time series as the one which resulted in the highest predictability, measured as the correlation between predicted and observed values (ρ). We tested the significance of the causal relationship detected by CCM by randomly shuffling the predicted time series 100 times and comparing the original ρ to the distribution of ρ values produced from the random shuffle. A relationship was considered significant when the original ρ value was higher than the 95$^{th}$ percentile of the distribution ($P < 0.05$).

## Appendix 2. GP EDM

Most approaches to nonlinear forecasting can be cast as tools for approximating the nonlinear mapping from past states to the future, including polynomials (e.g. Turchin and Ellner 1995), local linear regression (Sugihara 1994), support vector machines (e.g. Mukherjee et al. 1997), and neural networks (e.g. Bakker et al. 2000). Here, we use Gaussian process regression (Rasmussen and Williams 2006, Munch et al. 2017) with Automatic Relevance Determination (Neal 1996) to infer this mapping.

### *1. Gaussian process time-delay embedding*

Our goal is to estimate the unknown function $f$ that maps the history of observables into the future. To simplify notation, we'll use $\mathbf{x}_\mathrm{t} = \{S_t, \dots, S_{t-E}, y_{t-1}, \dots, y_{t-E}\}$ to represent the 'delay-coordinate vector' so that we are attempting to fit a model of the form $y_t = f(\boldsymbol{x}_t) + \varepsilon_t$ , $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ for $t \in \{E+1, \dots, T\}$.

We use a Gaussian process (GP) to model $f$. The GP is a continuous generalization of the multivariate normal distribution, completely defined in terms of its mean and covariance functions. GP models have been used widely in spatial statistics under the moniker Kriging (Cressie, 2015). In addition, they have been used in population modeling to estimate the form of density dependence (Munch et al., 2005), test for the presence of Allee effects (Sugeno and Munch, 2013), and as a tool to assess model misspecification (Thorson et al., 2014). Rasmussen and Williams (2006) is an excellent source for additional background on modeling with Gaussian processes.

In keeping with most GP applications, we set the prior mean to 0 to indicate that we do not have any *a priori* information on plausible shapes for the delay coordinate embedding. There are many choices for the covariance function or 'kernel' (see e.g. Rasmussen and Williams 2006). We used the squared exponential correlation function, $R(d) = exp[-d^2]$ where d is some measure of distance between inputs. In the present application the 'inputs' are the delay coordinate vectors, $\mathbf{x}_\mathrm{t}$ and we need to specify the covariance between $f$ evaluated at the delay coordinates for two different times, e.g. $f(\mathbf{x}_\mathrm{t})$ and $f(\mathbf{x}_\mathrm{s})$ for times t and s, respectively**.** We build the covariance function for this L-dimensional input from the product of 1-dimensional lag-specific squared exponentials. Specifically, the covariance between $f(\mathbf{x}_\mathrm{t})$ and $f(\mathbf{x}_\mathrm{s})$ is given by $\Sigma(\mathbf{x_t}, \mathbf{x_s}) = \tau^2 \prod_{\mathrm{i}=1}^{E_{max}} \mathrm{R}(\phi_i | \{\mathbf{x_t}\}_i - \{\mathbf{x_s}\}_i | / r_i)$ where the factor $r_i = \max_\mathrm{t}\{\mathbf{x_t}\}_i - \min_\mathrm{i}\{\mathbf{x_t}\}_i$ scales the distance to stay in [0,1] and the $\phi_i$'s control the 'wiggliness' of $f$ in the direction of the $i^{th}$ time-lag (i.e. larger values of $\phi$ allow more local extrema on the interval [0,1]). The product is taken over all lags from 1 to the maximum embedding dimension, $E_{max}$. In light of the relatively short time series available in the RAM database, we set the maximum embedding dimension $E_{max}$ to 5.

Note that when $\phi_i = 0$, $f$ is constant in the $i^{th}$ direction. Thus, to facilitate identification of a parsimonious model, we append a penalty on $\phi$ to the log likelihood of the form $penalty = -{\phi_i}^2/\pi$ (see Munch et al 2017 for details from a Bayesian perspective) This approach has been taken in the machine learning literature where it is referred to as automatic relevance determination (Neal, 1996). This penalty effectively identifies relevant lags for simulated data with delayed density dependence (Munch et al. 2017). The parameter $\sigma_\varepsilon^2$ controls the magnitude of the process noise and approximation error while $\tau^2$controls the variance in $f$ at a given point. We constrain both to the interval (0, 2Var(y)]. Estimates for the GP parameters were found by maximizing the marginal likelihood using the R-prop algorithm which provides stable and efficient convergence for GP regression (Blum and Reidmiller, 2013; Nocedal and Wright, 1999).

Appendix 3. Alternative models and fitting methods

*Parametric SR Models*

We fit three standard SR models and two linear autoregressive models to the database. The SR models are all given by $R=\alpha S g(S)$ where g(S) is $\exp(-\beta S)$, $(1+\beta S)^{-1}$ , and $(1-\beta\gamma S)^{1/\gamma}$ for the Ricker (1954), Beverton-Holt (Beverton and Holt 1957), and Schnute (1985) models respectively (see Quinn and Deriso (1999) for background and derivations). Note that the Schnute model is a 3-parameter model that contains Ricker and Beverton-Holt as special cases (for $\gamma = 0$ and -1 respectively). We constrain the parameters to the relevant ranges for each: $\alpha$ ($\alpha >0$), $\beta$ ($\beta \geq 0$), and $\gamma$ ($-1 \leq \gamma \leq 0$). After taking logs and appending noise, each model has the form $y = \ln(R/S) = \ln\alpha+\ln[g(S)]+\varepsilon$, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$

*Nonparametric SR*

To control for additional flexibility in EDM relative to the SR models, we fit the GP with the same specification as in S1, but with the restriction that it has the current stock size as its sole input, i.e. $y_t = f(S_t) + \varepsilon_t$ , $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$.

*Autoregressive models*

To control for additional input dimensions in EDM relative to traditional SR models, we fit linear autoregressive models of the form

$$y_t = \ln \alpha + \boldsymbol{x}_t{}^T \boldsymbol{\theta} + \varepsilon_t \text{ , } \varepsilon_t \sim N(0, \sigma_\varepsilon^2),$$

where, as before, $\mathbf{x}_t = \{S_t, \dots, S_{t-E}, y_{t-1}, \dots, y_{t-E}\}^T$ and the vector $\boldsymbol{\theta} = \{\theta_0, \dots \theta_{2E}\}^T$ collects the autoregressive parameters. As in EDM, this model $E_{max}$ was set to 5. Initially the AR model was fit without constraints on $\boldsymbol{\theta}$. However, this is a relatively large number of parameters (12) to be estimated given that some of the time series were as small as 20 observations. The GP EDM approach uses an ARD prior to trim out irrelevant inputs and regularize the model. To ensure that the AR and GP-EDM approaches were treated similarly, we also fit the AR model with a penalty on the coefficients of the form $penalty = -0.1\boldsymbol{\theta}^T\boldsymbol{\theta}$.

The AR model has the Ricker SR model as a special case (all $\theta_i$=0 except $\theta_0$). In addition, the case where $\theta_0$, $\theta_1$, and $\theta_{E+1}$ are non-zero is equivalent a Ricker model with autocorrelated errors. To see this, write the Ricker model as $y_t = \ln \alpha - \beta x_t + z_t$ where $z_t$is AR(1) noise which can be constructed as $z_t = \rho z_{t-1} + \varepsilon_t$ where $\rho$ is the autocorrelation coefficient and $\varepsilon_t$is white noise. Using a time shift to replace $z_t$, we find that $y_t = \ln \alpha - \beta x_t + \rho(y_{t-1} - \ln \alpha + \beta x_{t-1}) + \varepsilon_t$ which has the indicated form after a tiny bit of algebra.

## Appendix 4. Supplementary analyses of the RAM database

Figure S1 shows the results comparing GP-EDM prediction error to Ricker and Schnute models. Methods are as described in the main text and Appendix 1.

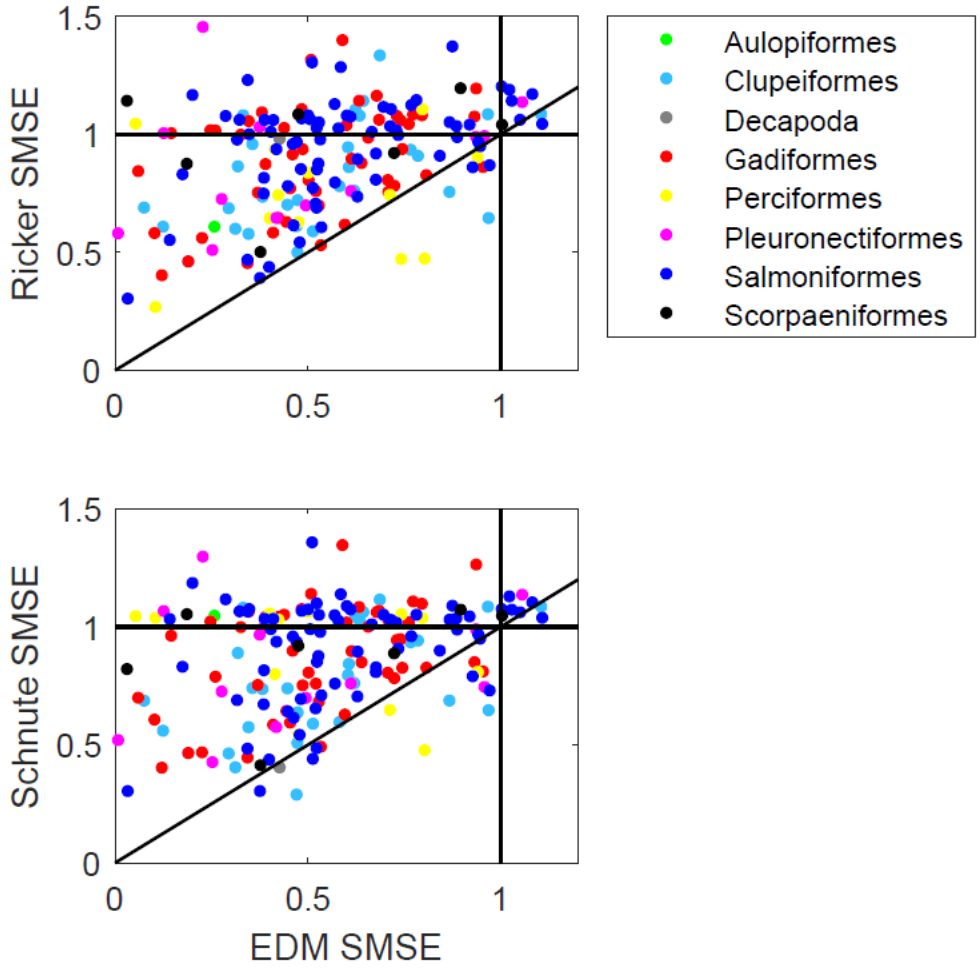


Figure S1. Comparison of EDM v. Ricker and Schnute models. In each panel the axes indicate the scaled mean-square error (SMSE, i.e. the variance in predictions estimated by leave-one-out cross validation divided by the total variance in y) obtained with the method indicated. Thus the horizontal and vertical lines at 1 indicate the SMSE expected using the only the sample mean as our prediction. The diagonal is the 1:1 line. The horizontal axis in each panel is the scaled prediction error obtained using EDM and the vertical axes are for top) the Ricker SR model and bottom) the Schnute SR model.

## Appendix 5. 'Regime' models and EDM

To illustrate the possibility of having both distinct 'regimes' and predictable dynamics, we use the map $\{ x_{n+1} = x_n \exp[r(x_n - a)(1 - x_n/K)] + R_n exp[-\mu x_n]$, where $x_n$ is population size, r is growth rate, K is the carrying capacity and a is the Allee effect threshold below which residents go extinct in the absence of immigration. The second term represents the supply of immigrants with $\log[R_n] \sim N(m, \sigma^2)$ and resident-dependent mortality, $\mu$.

We iterated this map for 100 steps with $r$=0.125, a=27, K=54, m=60, $\sigma^2$=0.1, and $\mu$=1. (figure S2). A piecewise constant 'regime' model was fit to the time series assuming that there were only two regimes. For comparison, we fit the first 30 data points with GP EDM as described in the main text and Appendix 1 with $E_{max}$ = 5. We used the fitted model to make step-ahead predictions for the remaining 70 points (Figure S2, panel B). The two-regime model explains 73.6% of the variance. EDM predictions explain 98.3%.

*A second example*

The preceding example is just one of many models that exhibit apparent regimes and predictable dynamics. The canonical model for bistability in discrete time is the Duffing map which is given by $\{ x_{n+1} = ax_n - by_n - x_n^3, y_{n+1} = x_n \}$. We repeat the above comparison of EDM and piece-wise constant models for the Duffing map. We iterated this map for 250 steps with a=2.77, b=0.2 and discarded the first half to avoid transients (figure S3). A piecewise constant 'regime' model was fit to the time series assuming that there were only two regimes. For comparison, we fit the first 30 data points with GP EDM as described in the main text and Appendix 1 with $E_{max}$ = 5. We used the fitted model to make step-ahead predictions for the remaining 95 points (Figure S3, panel B).

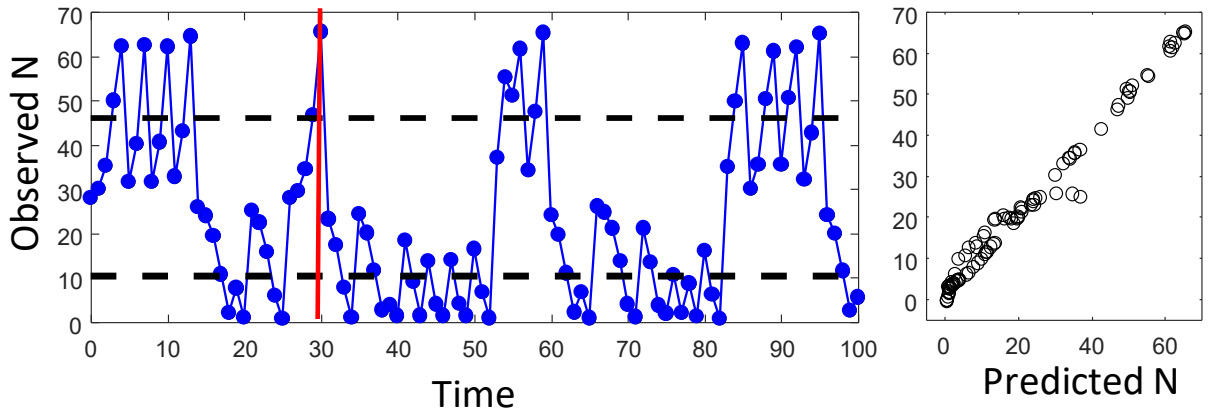


Figure S2. Apparent regimes in population dynamics Allee effects and immigration. A. The blue points are the time series of population sizes generated by the model. The dashed line is the pieceswise constant 'regime' model model which explains 73.4% of the variance. B. After training the GP EDM on the first 30 points from the series (points to the left of the red line in panel A) the predictions for the remaining 70 points explain 98.3% of the variance.

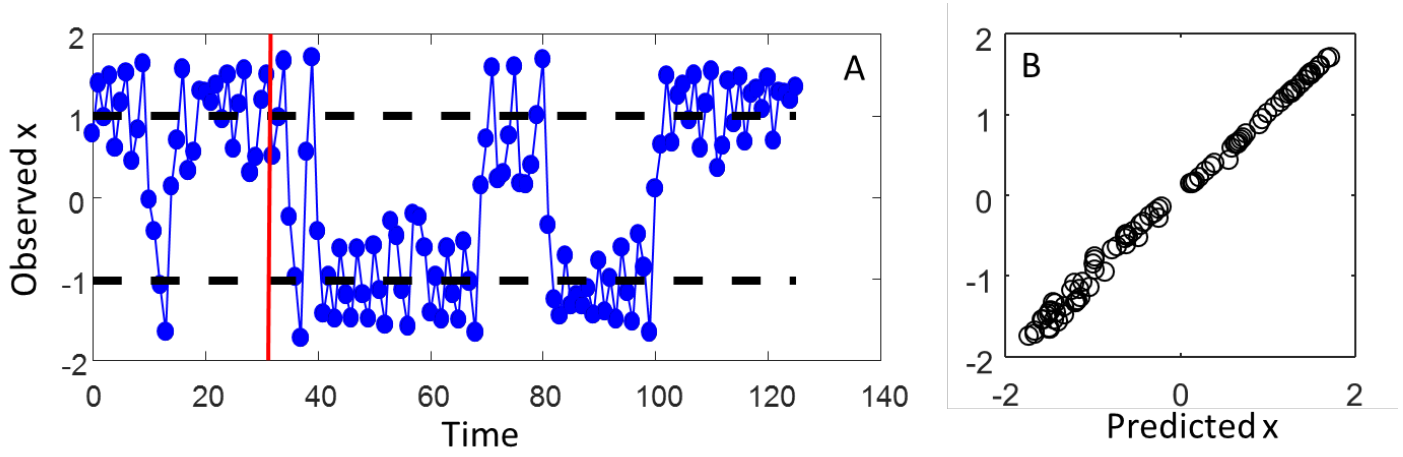


Figure S3. Apparent regimes in the Duffing map. A. The blue points are the time series of x from the Duffing map { $x_{n+1} = ax_n - by_n - x_n^3$, $y_{n+1} = x_n$with a=2.77, b=0.2}. The dashed line is the pieceswise constant 'regime' model model which explains ~80% of the variance. B. After training the GP EDM on the first 30 points from the series (points to the left of the red line in panel A) the predictions for the remaining 90 points explain 99.5% of the variance.

**Supplementary References**


Beverton, RJH and Holt, SJ (1957) On the Dynamics of Exploited Fish Populations. Chapman and Hall, London.

Blum, M. & Riedmiller, M.A. (2013). Optimization of Gaussian process hyperparameters using Rprop, *In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Belgium, 24-26.

Boettiger, C., Mangel, M. & Munch, S.B., (2015). Avoiding tipping points in fisheries management through Gaussian process dynamic programming. *Proceedings of the Royal Society of London B: Biological Sciences*, 282, 20141631

Cressie, N., (2015). *Statistics for Spatial Data*. John Wiley & Sons.

Ellner, S.P., & turchin, P. (1995). Chaos in a noisy world: new methods and evidence from time-series analysis. *The American Naturalist*, 145, 343-375

Mukherjee, S., Osuna, E. & Girosi, F., (1997). Nonlinear prediction of chaotic time series using support vector machines. In *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Workshop*, 511-520. IEEE.

Munch, S.B., Kottas, A. & Mangel, M., (2005). Bayesian nonparametric analysis of stock recruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, *62*, 1808-1821.

Munch, S.B., Poynor, V. & Arriaza, J.L., (2017). Circumventing structural uncertainty: A Bayesian perspective on nonlinear forecasting for ecology. *Ecological Complexity.* 32:134-143.

Neal, R.M., (1996). *Bayesian Learning for Neural Networks*, Springer-Verlag, New York

Nocedal, J. & Wright, S. (1999). *Numerical Optimization*. Springer-Verlag, New York.

Quinn, T.J. & Deriso, R.B., (1999). *Quantitative Fish Dynamics*. Oxford University Press.

Rasmussen, C.E. & Williams, C.K., (2006). *Gaussian Processes for Machine Learning.* Cambridge: MIT press.

Ricker, WE (1954) Stock and recruitment. J. Fish. Res. Bd. Can. 11:559-623

Schnute, J (1985) A general theory for the analysis of catch and effort data. Can. J. Fish. Aquat. Sci. 42:414-429

Sugeno, M. & Munch, S.B., (2013b). A semiparametric Bayesian method for detecting Allee effects. *Ecology*, 94, 1196-1204

Sugihara, G. (1994) Nonlinear forecasting for the classification of natural time series. *Philosophical Transactions of the Royal Society of London, Series A* 348, 477-495

Sugihara, G., May, R., Ye, H., Hsieh, C.H., Deyle, E., Fogarty, M. & Munch, S.B., (2012). Detecting causality in complex ecosystems. *Science*, 338, 496-500.

Thorson, J.T., Ono, K. & Munch, S.B., (2014a). A Bayesian approach to identifying and compensating for model misspecification in population models. *Ecology*, 95, 329-341.