Supporting Information

Deyle et al. 10.1073/pnas.1215506110

SI Text

Simplex Projection. Simplex projection (1) is a method for using state space reconstruction to forecast a time series variable $X(t)$. The basic idea is simple: Points nearby in state space evolve similarly in time. Simplex projection forecasts are a weighted average of the dynamics observed at points nearby in the reconstructed state space, $\boldsymbol{M}$. We make this notion explicit as follows.

State space reconstruction involves identifying time series point $X(t)$ with corresponding vectors, $\underline{x}(t)$, in a multidimensional space. In the case of univariate SSR, the coordinate axes of this multidimensional space are time lags of the variable X. Thus, the vector $\underline{x}(t)$ has components $[x_1(t), x_2(t), x_3(t), ..., x_E(t)] = [X(t), X(t-\tau), X(t-2\tau), ..., X(t-(E-1)\tau)]$. Here, E is the embedding dimension and τ is the step size of the time lag (both for the sardine and model analyses $\tau = 1$ y). In multivariate SSR with two time series variables $X(t)$ and $Y(t)$, some of the coordinate variables of the state space vectors $\underline{x}(t)$ will correspond to lags of Y rather than X. In this paper, we only experiment with multivariate embeddings that include a single lag of an environmental variable Y. Thus, the vector $\underline{x}(t)$ will have components $[x_1(t), x_2(t), x_3(t), ..., x_E(t)] = [X(t), X(t-\tau), X(t-2\tau), ..., X(t-(E-2)\tau), Y(t)]$.

Suppose we have time series of X and Y with n observations, i.e., $\{X(1), X(2), ..., X(n)\}$ and $\{Y(1), Y(2), ..., Y(n)\}$. These time series give $n - E + 1$ vectors (for the univariate case) in the reconstructed state space, $\boldsymbol{M}$. To predict future values of X starting at time t^*, we first determine the $E + 1$ nearest neighbors to the vector $\underline{x}(t^*)$ in $\boldsymbol{M}$. Note that $E + 1$ is the minimum number of points to surround $\underline{x}(t^*)$ in the E dimensional space. Let t_1 be the time index of the first nearest neighbor to $\underline{x}(t^*)$, t_2 the time index of the second nearest neighbor, and so on. The simplex projection of $X(t^*)$ p steps into the future is then

$$X(t^*+p)\Big|\mathbf{M}_X = \sum_{i=1}^{E} w_i X(t_i+p) \Big/ \sum_{j=1}^{E} w_j, \qquad \textbf{[S1]}$$

where the weighting applied to each neighbor, w_i, is given by

$$w_i = \exp\left(-\left\|\underline{x}(t^*) - \underline{x}(t_i)\right\| \Big/ \left\|\underline{x}(t^*) - \underline{x}(t_1)\right\|\right).$$

Here, $||\underline{x} - \underline{y}||$ is the Euclidian distance between two vectors in an E dimensional space. That is, $||\underline{x} - \underline{y}|| = \{[x_1 - y_1]^2 + [x_2 - y_2]^2 + \ldots + [x_E - y_E]^2\}^{1/2}$.

Thus, the algorithm for simplex projection can be summarized as follows:

i) Translate the time series data into vectors in the (multidimensional) reconstructed state space. In this paper we consider two types of SSR: univariate SSR, where the state space vector for time point t is given by $\underline{x}(t) = [X(t), X(t-\tau), X(t-2\tau), ..., X(t-(E-1)\tau)]$, and multivariate SSR with a single lag of an environmental variables, where $\underline{x}(t) = [X(t), X(t-\tau), X(t-2\tau), ..., X(t-(E-2)\tau), Y(t)]$. Note that the first few points in the time series will not have all of the necessary lags and so will not have a corresponding vector in the reconstructed state space.

ii) Pick a target time point, t^*, and identify the corresponding vector $\underline{x}(t^*)$. We seek to predict $X(t^*+p)$.

iii) Define the set of library vectors that will be used to predict the behavior of $\underline{x}(t^*)$. For short time series (like we examine in this work), it is best to use cross-validation. In this case, the library set will be all possible vectors formed from the time series except the target vector. Note that with longer time series, it is possible to split the time-series in half and use the first half to predict the second half (as in ref. 1).

iv) Compute the Euclidian distance between $\underline{x}(t^*)$ and each vector in the library, $d(\underline{x},\underline{y}) = ||\underline{x} - \underline{y}|| = \{[x_1 - y_1]^2 + [x_2 - y_2]^2 + ... + [x_E - y_E]^2\}^{1/2}$.

v) Identify the $E + 1$ library vectors that are closest (have the shortest Euclidian distance) to the target vector $\underline{x}(t^*)$ in the reconstructed state space. Define t_1 as the time index of the first nearest neighbor to $\underline{x}(t^*)$, t_2 as the time index of the second nearest neighbor, and so on. The time indices t_i are the points in history that the system was in a similar state to the target time, t^* (according to this state space reconstruction).

vi) Predict $X(t^*+p)$ using Eq. **S1** above.

vii) Repeat steps *ii*–*vi* for each vector in the state space.

Note that for multivariate embeddings (the components of the state space vectors $\underline{x}(t)$ correspond to lags of two or more time series variables), it is important to normalize each time series. Otherwise, if time series X and Y have very different magnitudes, the choice of neighbors and weighting function will be dominated by the larger time series. Alternatively, one could use more complicated forecasting schemes, where weighting is adjusted for each component variable, but this will increase the risk of overfitting.

Embedding Dimension. Takens's theorem and its multivariate generalization state that lag-coordinate reconstructions are valid approximations of the true system as long as sufficiently many coordinate dimensions are used (2–4). Thus, if d dimensions are sufficient for representing the system, $d + 1$ dimensions will work as well. In real systems with observation and process error, including too many dimensions adds uncertainty. In practice, we chose the embedding dimension E that gives maximum predictability, ensuring that E is sufficiently large to capture the dynamics of the system without including extraneous dimensions (1). Predictability can be measured using mean absolute error (MAE), root mean squared error (RMSE), or correlation (ρ) between predictions and observations. Usually these measures will agree.

For the California Cooperative Oceanic Fisheries Investigations (CalCOFI) survey data of *Sardinops sagax* ichthyoplankton abundance, the relationship between embedding dimension (E) and ρ is noisy. With short time series, correlation is more sensitive to outliers than MAE. Thus, we rely on MAE. Fig. S1 displays prediction skill (1 − MAE) as a function of embedding dimension (E). From this result, we chose an embedding dimension of $E = 3$.

S-Map. Once the optimal embedding dimension is determined using simplex projection, the S-map (sequential locally weighted global linear map) procedure (5) can be used to test for state-dependent dynamics. For the target time point t^*, a linear model $\mathbf{C}$ is used to predict the future value $X(t^*+p)$ (in this paper we only deal with p = 1-y forecasts) from the reconstructed state space vector $\underline{x}(t^*)$. That is,

$$\hat{X}(t^*+p) = C_0 + \sum_{j=0}^{E-1} \mathbf{C}_j x_j(t^*). \qquad \textbf{[S2]}$$

The linear model is fit to the other vectors in the state space. However, points that are close to the target point, $\underline{x}(t)$, are given

greater weighting. Specifically, the model **C** is the SVD solution to the equation

$$\mathbf{B} = \mathbf{A}\cdot\mathbf{C}, \qquad \text{[S3]}$$

where **B** is an n-dimensional vector of the weighted future values $X(\underline{t}_i)$ for each historical point, $\underline{t}_i$, given by

$$B_i = w\left(\left\|\underline{x}(\underline{t}_i) - \underline{x}(t^*)\right\|\right)X(\underline{t}_i + p), \qquad \text{[S4]}$$

and **A** is the $n \times E$ dimensional matrix give by

$$A_{ij} = w\left(\left\|\underline{x}(\underline{t}_i) - \underline{x}(t^*)\right\|\right)x_j(\underline{t}_i), \qquad \text{[S5]}$$

The weighting function w is defined by

$$w(d) = \exp\left(-\theta d/\overline{d}\right),$$

which is tuned by the nonlinear parameter $\theta \geq 0$ and normalized by the average distance between $\underline{x}(t^*)$ and the other historical points,

$$\overline{d} = \frac{1}{n}\sum_{j=1}^{n}\left\|\underline{x}(\underline{t}_j) - \underline{x}(t^*)\right\|.$$

As above, $||\underline{x} - \underline{y}||$ is the Euclidian distance between two vectors in the E-dimensional state space. Note that the model **C** is separately calculated (and thus potentially unique) for each time point, t.

The algorithm for S-map given a choice of the nonlinear parameter θ can be summarized as follows:

i) Translate the time series data into vectors in the multidimensional state space. In this paper we consider pure univariate SSR, $\underline{x}(t) = [X(t), X(t-\tau), X(t-2\tau), ..., X(t-(E-1)\tau)]$ and multivariate SSR with a single lag of an environmental variables, $\underline{x}(t) = [X(t), X(t-\tau), X(t-2\tau), ..., X(t-(E-2)\tau), Y(t)]$.
ii) Pick a target time point, t^*, and identify the corresponding state space vector $\underline{x}(t^*)$. We seek to predict $X(t^*+p)$.
iii) Define the set of library vectors that will be used to predict the behavior of $\underline{x}(t^*)$. For short time series (like we examine in this work), it is best to use cross-validation. In this case, the library set will be all possible vectors formed from the time series except the target vector.
iv) Compute the Euclidian distance between $\underline{x}(t^*)$ and each vector in the library, $d(\underline{x},\underline{y}) = ||\underline{x} - \underline{y}|| = \{[x_1 - y_1]^2 + [x_2 - y_2]^2 + \ldots + [x_E - y_E]^2\}^{1/2}$.
v) Use these distances to define the weighted vector **B** using Eq. **S4** and matrix **A** using Eq. **S5**.
vi) Solve Eq. **S3** using singular value decomposition (SVD) for the matrix **C**.
vii) Calculate $X(t^*+p)$ using Eq. **S2** above with **C** found above.
viii) Repeat steps *ii*–*vii* for each vector in the state space.

A time series is shown to have state-dependent dynamics if nonlinear models (that depend on the location in state space) produce better forecasts than linear models. Fig. S2 shows S-map results for the first-differenced CalCOFI ichthyoplankton survey abundance of Pacific sardine, using the embedding dimension determined from simplex projection (Fig. S1). Forecasts are substantially improved by accounting for nonlinear dynamics, indicating the Pacific sardine have nonlinear dynamics.

Three-Species Logistic Model. The general form for a three-species coupled logistic model for species X_1, X_2, and X_3 in continuous time is given by

$$\frac{dX_i}{dt} = X_i\left(r_i + \sum_{j=1}^{3}\alpha_{ij}X_j\right).$$

The parameter values were taken from figure 5 in ref. 6.

$$\alpha = \begin{pmatrix} -0.0020 & -0.4604 & -0.5051 \\ 0.2324 & -0.1920 & 2.3847 \\ 1.2949 & -0.0153 & -0.306 \end{pmatrix}; \; r = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} = \begin{pmatrix} 0.9675 \\ -2.4251 \\ -0.9736 \end{pmatrix}.$$

The data were generated using a fourth-order Runge–Kutta with integration step of $h = 0.01$ and initial conditions $X_1 = 5$, $X_2 = 0.001$, and $X_3 = 11$. Fig. 1 shows data for $t = (101, 102, \ldots, 300)$. The three red points correspond to time indices $t_1 = 108$, $t_2 = 143$, and $t_3 = 248$.

Expanded Model Example of Scenario Exploration. We repeat the analysis shown in Fig. 2*A* of the main text to show that the result is robust over a wide range of the growth rate parameter, r. The model structure is the same (Eq. 1 in the main text):

$$S(t+1) = S(t)\exp[(r + \varepsilon(t))(1 - S(t))]\exp(\psi T(t)),$$

where $\varepsilon(t)$ is a normally distributed random variable with mean(ε) = 0 and SD(ε) = 0.2. As in the analysis in the main text, we set $\psi = 0.3$. However, the growth parameter is varied between 1.8 and 2.8. The temperature $T(t)$ was modeled as red noise with mean 0 and SD 1 by applying a 10-y averaging window to white noise.

Using scenario exploration with multivariate SSR, we predict the effect on stock size S of a 10% increase in temperature, ΔT, relative to the SD σ_T of the temperature time series. We then compare the SSR predictions to the exact calculations with the model. As in Fig. 3*A* of the main text, we use the multivariate embedding $[S(t), S(t-1), S(t-2), S(t-3), T(t)]$ that contains lags of population abundance and temperature. We predict the effect that an increase in temperature ΔT at time t would have on the population abundance the following year, $t + 1$. That is, we make a nearest-neighbor forecast of the adult SSB for the state $[S(t), S(t-1), S(t-2), S(t-3), T(t) + \Delta T]$. Fig. S3 displays the results. For each value of r, we generated 10 time series of 50 y each with different initial conditions and realizations of T and $\varepsilon_{\text{proc}}$.

1. Sugihara G, May RM (1990) Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* 344(6268):734–741.
2. Takens F (1981) Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence, Warwick 1980*. Lecture Notes in Mathematics (Springer, Berlin), Vol 381, pp 366–381.
3. Sauer T, Yorke JA, Casdagli M (1991) Embedology. *J Stat Phys* 65(3-4):579–616.
4. Deyle ER, Sugihara G (2011) Generalized theorems for nonlinear state space reconstruction. *PLoS ONE* 6(3):e18295.
5. Sugihara G (1994) Nonlinear forecasting for the classification of natural time series. *Philos Trans R Soc Lond A* 348(1688):477–495.
6. Gardini L, Lupini R, Messia M (1989) Hopf bifurcation and transition to chaos in Lotka-Volterra Equation. *J Math Biol* 27(3):259–272.

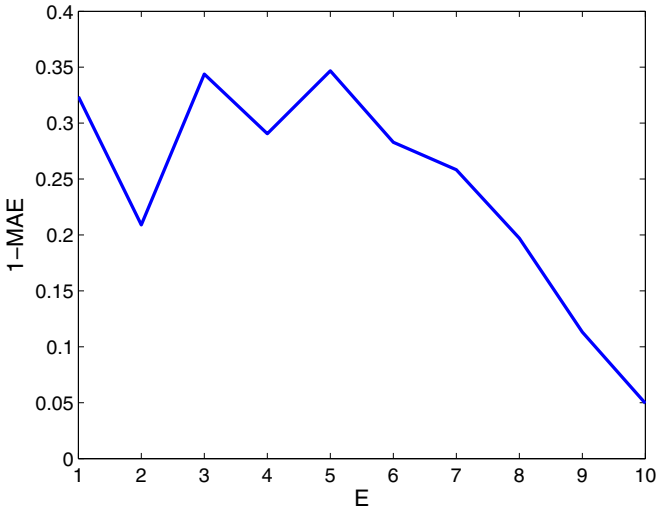

Fig. S1. Univariate prediction skill (1 – MAE) is shown as a function of embedding dimension (*E*) for normalized first differences of the CalCOFI *S. sagax* ichthyoplankton time series. MAE, mean absolute error.

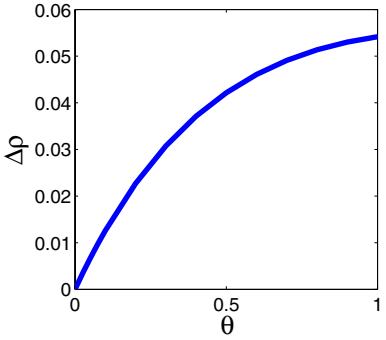

Fig. S2. S-map analysis of CalCOFI survey abundance of Pacific sardine ichthyoplankton. The improvement in forecast skill, $\Delta\rho$, of sardine ichthyoplankton abundance for nonlinear models compared with the completely linear model ($\theta = 0$) for increasing values of the nonlinear tuning parameter, θ. Here, forecast skill is measured by the Pearson's correlation coefficient between model predictions (using cross-validation) and the observed values. Increasing the nonlinear parameter substantially improves forecasts, suggesting that the dynamics of Pacific sardine are state-dependent (nonlinear).

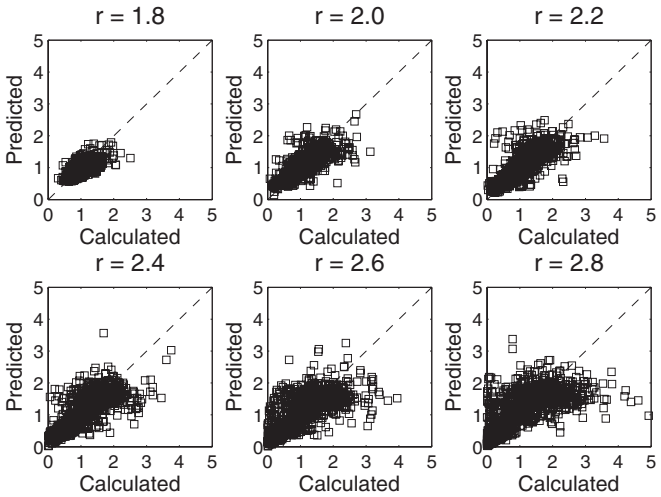

Fig. S3. Scenario exploration illustrated for short (50-y) time series generated with a simple Ricker model forced by temperature for a range of growth rates, *r*. For each time series point *t*, the effect of warming on $S(t+1)$ is predicted with multivariate SSR for warming of $\Delta T = 0.1\sigma_T$ (10% of the SD of *T*). Predictions are compared with the true value calculated with the model over 10 model realizations for each *r*.