



## NOTE

# Quantitative argument for long-term ecological monitoring

Alfredo Giron-Nava<sup>1</sup>, Chase C. James<sup>1</sup>, Andrew F. Johnson<sup>1</sup>, David Dannecker<sup>2</sup>,  
Bethany Kolody<sup>1</sup>, Adrienne Lee<sup>2</sup>, Maitreyi Nagarkar<sup>1</sup>, Gerald M. Pao<sup>3</sup>, Hao Ye<sup>1</sup>,  
David G. Johns<sup>4</sup>, George Sugihara<sup>1,\*</sup>

<sup>1</sup>Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

<sup>2</sup>Division of Biological Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

<sup>3</sup>Salk Institute for Biological Studies, 10010 N Torrey Pines Road, La Jolla, CA 92037, USA

<sup>4</sup>Sir Alister Hardy Foundation for Ocean Science, The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK

**ABSTRACT:** Although it seems obvious that with more data, the predictive capacity of ecological models should improve, a way to demonstrate this fundamental result has not been so obvious. In particular, when the standard models themselves are inadequate (von Bertalanffy, extended Ricker etc.) no additional data will improve performance. By using time series from the Sir Alister Hardy Foundation for Ocean Science Continuous Plankton Recorder, we demonstrate that long-term observations reveal both the prevalence of nonlinear processes in species abundances and an improvement in out-of-sample predictability as the number of observations increase. The empirical results presented here quantitatively demonstrate the importance of long-term temporal data collection programs for improving ecosystem models and forecasts, and to better support environmental management actions.

**KEY WORDS:** Long-term monitoring · Predictability · Nonlinearity · Time series · Population dynamics · Ecological data

## INTRODUCTION

Although there is a strong consensus for the benefits of long-term ecological monitoring and the creation of long-term datasets, there are significant fiscal and political challenges facing the continuation of these efforts. Given typical funding horizons and the inherent costs associated with long-term research programs (Callahan 1984, Magurran et al. 2010), it is no surprise that the majority of ecological research occurs within short time frames (Magnuson 1990).

Ecosystem dynamics, however, may not be decipherable from observations on short timescales. Indeed, there is a growing view that nonlinear and non-stationary dynamics are common and may require

longer time series for study (Sugihara & May 1990, Hsieh et al. 2005, Sugihara et al. 2012, Glaser et al. 2014). While most ecological models can be fit to short time series, they may subsequently fail when used to make predictions into the future or to explain changes that occur over the long term (Pilkey & Pilkey-Jarvis 2007, Evans et al. 2012). The challenge of models to predict well has led some to suggest that scientific efforts, and particularly those in fisheries science, should be shifted away from forecasting as an attainable goal (Schindler & Hilborn 2015). This view contradicts our natural expectation that ecosystems behave in nonrandom ways and that with more data (i.e. longer time series), models should generally improve. Although we commonly acknowledge out-

\*Corresponding author: gsugihara@ucsd.edu

of-sample prediction to be the ultimate test for understanding, many current parametric ecosystem models fail to pass this test (Magurran et al. 2010). A possible reason is that these models are not well suited for describing nonlinear and nonstationary dynamics. Insofar as nonlinearity is to be expected in ecosystems, it is almost guaranteed that improperly formulated models will fail to achieve successful predictions, no matter how much additional data are provided.

If one were to use a minimal nonparametric nonlinear framework to model ecosystems, to what extent would short-term datasets limit predictive capabilities? In other words, what is the value of long-term ecosystem monitoring for gaining a predictive understanding of ecosystem processes? We investigate this question using a unique dataset and nonparametric analytical approach. The data are from the continuous plankton recorder (CPR) program at the Sir Alister Hardy Foundation for Ocean Science (SAHFOS). The analytical approach is minimal in that it allows the data to inform ecological dynamics with no assumptions about underlying equations—so-called empirical dynamic models (EDMs) (Sugihara et al. 2012, DeAngelis & Yurek 2015). Our objectives are straightforward and 2-fold: (1) to explore the relationship between time series length and the ability to detect nonlinearity, and (2) to examine the ability to predict future population abundances a month ahead as data availability/length increases.

## MATERIALS AND METHODS

### Dataset and data completeness

We use time series from the SAHFOS program, the longest and most spatially extensive planktonic ecosystem dataset currently available for the Atlantic Ocean. Our data, recorded by the CPR (Hays et al. 2005), comprise monthly average abundances of phytoplankton and zooplankton taxa in the southern North Sea (55° to 58° N, 3° to 11° W) from 1958 to 2013, with sampling methods described in depth in Richardson et al. (2006). Data are available from the SAHFOS repository at [http://doi.sahfos.ac.uk/doi-library/data-for-zooplankton-and-phytoplankton-from-the-1\)-southern-north-sea-and-2\)-the-irish-sea.aspx](http://doi.sahfos.ac.uk/doi-library/data-for-zooplankton-and-phytoplankton-from-the-1)-southern-north-sea-and-2)-the-irish-sea.aspx). Because of the practical limitations of identifying taxa and making counts, the time series within the SAHFOS repository count organisms at a variety of taxonomic levels (but typically either species or genus). We used the data as they are provided; thus, the time series refer to taxa rather than species.

We use 2 distinct measures to quantify the amount of data in each time series: time series length and data availability. Time series length refers to the number of data points in a time series, whereas data availability is defined as the number of nonzero values within a time series. We use the term data availability because we recognize that zeros could represent an absence of an organism or a lack of detection, and we are unable to distinguish between the two. Our analyses that examine the effect of time series length use subsampled segments from the time series with the fewest nonzero values or greatest data availability 'Effect of time-series length'.

### Effect of data availability

We investigate the degree of nonlinearity and predictability using 2 methods of time series analysis: simplex projection (Sugihara & May 1990) and S-maps (Sugihara 1994). Simplex projection is used to assess the one-step-ahead predictive skill for each time series and is measured by the Pearson correlation coefficient ( $\rho$ ) between predictions and observations. Following Glaser et al. (2014), to minimize overfitting, we use leave-one-out cross validation, where the point being forecast is excluded from the data used to construct the forecast. Different values for the embedding dimension ( $E$ ) were tested (ranging from  $E = 1$  to  $E = 10$ ), and the optimal embedding dimension was selected as that which produced the highest  $\rho$ . We note that the optimal  $E$  is a property of the data (how noisy and how long the time series are) as well as the underlying system (how complex it is in terms of numbers of variables) and indicates the number of lags that best resolves one-step-ahead forecasts. While  $E$  can be informative about the complexity of the system or the number of drivers (Liu et al. 2012, Glaser et al. 2014), we caution against over-interpretation.

S-maps are used to test for nonlinear state dependence (Sugihara 1994). This involves demonstrating curvature in the attractor and is quantified by the improvement in  $\rho$  ( $\Delta\rho$ ) obtained with a nonlinear versus a linear model. Statistical significance for  $\Delta\rho$  was determined using a randomization test. For each time series, we generated 100 surrogate time series and calculated a  $\Delta\rho$  for each surrogate using the previously determined  $E$ . This creates a null distribution that can be compared to the original (nonshuffled)  $\Delta\rho$ . Nonlinearity is indicated when the  $\Delta\rho$  statistic for the original time series is greater than the 0.95 quantile of the null distribution.

### Effect of time-series length

To examine the effect of time series length as a property independent of taxonomic identity, we focused on the 23 time series with the greatest data availability ( $\geq 461$  nonzero data points) and which were identified as significantly nonlinear. For each time series, we investigated the effect of varying the data length on predictability using randomly selected contiguous segments of length: 25, 50, 100, 200, 400. The subsampling was repeated 25 times at each length for each species (so that 575 [ $23 \times 25$ ] subsamples were used to compute forecast performance at each length). Predictions were made using a form of leave-one-out cross validation where the predicted point is excluded to minimize overfitting. We then tested for nonlinear state dependence using S-maps as described in the previous subsection.

### Significance of nonlinearity as a function of data completeness

To test whether data availability or time series length are significant predictors of nonlinearity, we treated data availability and time series length by binning them into discrete categories (Fig. 1) and recorded the detection of nonlinearity as a binary response. Then, we fit a logistic regression using the R glm function (binomial family).

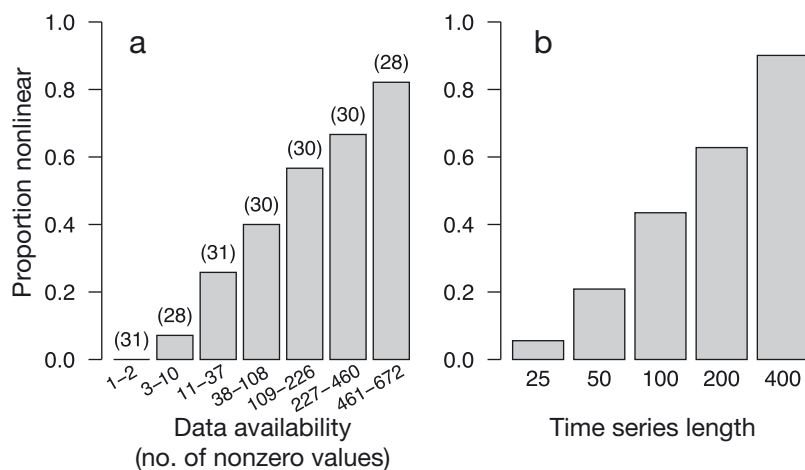


Fig. 1. Effect of data availability and time series length on nonlinearity detection. (a) Proportion of taxa identified as significantly nonlinear as a function of data availability (number of nonzero values). Numbers above each bar represent the number of taxa in that bin. (b) Proportion of taxa identified as significantly nonlinear as a function of time series length. Each bar corresponds to 575 subsamples of a specified length created by generating 25 subsamples from each of the 23 time series with the greatest data availability

## RESULTS AND DISCUSSION

The S-map analysis shows that nonlinear dynamics are more readily identified as data availability increases (Fig. 1a;  $p < 0.01$ ; logistic regression,  $df = 207$ ). For the 90 taxa with the lowest data availability ( $\leq 37$  nonzero data points out of 672), only 11 % (10 taxa) showed significant nonlinear dynamics. In contrast, of the 90 taxa with 38 to 460 nonzero values, 57 % (51 taxa) showed significant nonlinear dynamics. Finally, among the 28 taxa with the highest data availability ( $\geq 461$  nonzero values), 82 % (23 taxa) showed significant nonlinear dynamics. In other words, the time series that are most complete also show stronger evidence for nonlinear dynamics. To test whether this effect could be driven by the specific taxa that happen to appear most often in the data, we also analyzed subsampled time series (artificially shortened from the 23 nonlinear taxa with the most data availability), finding a similar pattern (Fig. 1b;  $p < 0.01$ ; logistic regression,  $df = 23$ ). It has been recognized that aggregating taxon levels can obscure nonlinear dynamics detection (Liu et al. 2014); however, we have shown that nonlinearity detection increases with time series length regardless of the taxonomic resolution. The subsampling procedure here obviates any taxon-specific effect related to aggregation that could reduce nonlinearity in more aggregated (linearly summed) groupings (Sugihara et al. 1999).

A similar advantage with increased data holds for prediction; in general, greater data availability also corresponds to higher forecast skill (Fig. 2a). Using the same subsamples as in the previous paragraph, Fig. 2b shows that predictability also increases with time series length. Thus, we expect that as more data are collected, populations will be more readily identified as nonlinear and that the longer time series will enable better predictions. However, we note that even at the longest time series lengths ( $\geq 600$  nonzero values in Fig. 2a or 400 time points in Fig. 2b), the level of forecast skill can vary substantially. These differences in predictability could partially reflect differences in the natural response times (e.g. generation times), leading to differences in the density of points on the attractor (Table 1), with shorter generation times producing denser attractors;

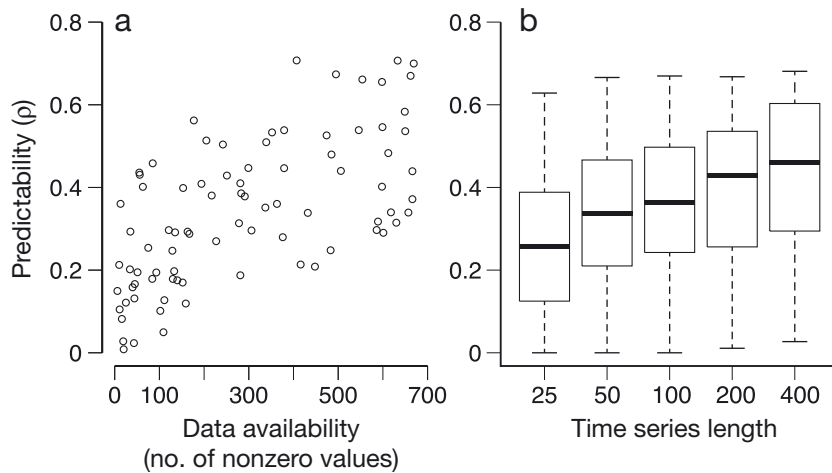


Fig. 2. Effect of data availability and time series length on predictability ( $\rho$ ) for (a) each time series (circles) and (b) different lengths of subsampled time series. Each boxplot corresponds to 575 subsamples as in Fig. 1b. Bold lines: median; boxes: interquartile range; whiskers: minimum and maximum values

however, they could also reflect exogenous effects, such as environmental drivers, that are not captured in the abundance time series and may therefore need to be included explicitly in forecast models (see Dixon et al. 1999, Deyle et al. 2013, and Harford et al. 2017 for examples).

Finally, we note that in addition to leave-one-out cross validation, for the 23 taxa with the highest data availability, nearly identical out-of-sample prediction results are obtained with 2-fold cross validation. Here each half of the data series is used to predict the other half, and the predictions for the entire series are combined to calculate performance (Fig. 3). This demonstrates the robustness of the cross validation results and shows that the dynamics are stationary or essentially the same in both halves (Sugihara & May 1990).

Table 1. Taxa whose time series have more than 600 nonzero values and the Pearson correlation coefficient ( $\rho$ ) from nonlinear prediction (from Fig. 2a)

Taxon	$\rho$
<i>Oithona</i> sp.	0.71
Total copepods	0.70
<i>Para-pseudocalanus</i> sp.	0.67
<i>Acartia</i> sp.	0.58
Chaetognaths	0.54
Chlorophyll index	0.53
<i>Calanus</i> I–IV	0.48
<i>Calanus finmarchicus</i>	0.44
Fish larvae	0.34
<i>Calanus helgolandicus</i>	0.34
<i>Temora longicornis</i>	0.31
Total hyperiids	0.29

Documenting that ecosystems are capable of exhibiting predictability and that their dynamics are nonlinear is of considerable interest to agencies involved in data collection and to resource managers who must use those data. The reasonable expectation is that predictive models will improve with more data. However, if a model is a poor description of a system, no amount of additional data will improve predictability (MacNally 2000, DeAngelis & Yurek 2015). For example, fishery stock prediction has been a difficult challenge despite continual model refinements and new data inputs (Ward et al. 2014, Schindler & Hilborn 2015, Ye et al. 2015). The limitations of current ecosystem models to predict out of

sample may arise in part from observational error in the data. However, we believe that a more significant problem is the mismatch between the traditional assumptions of equilibrium and linearity (DeAngelis & Waterhouse 1987) and the increasing observations that many ecological systems are not in equilibrium but are demonstrably nonlinear (Fromentin & Powers

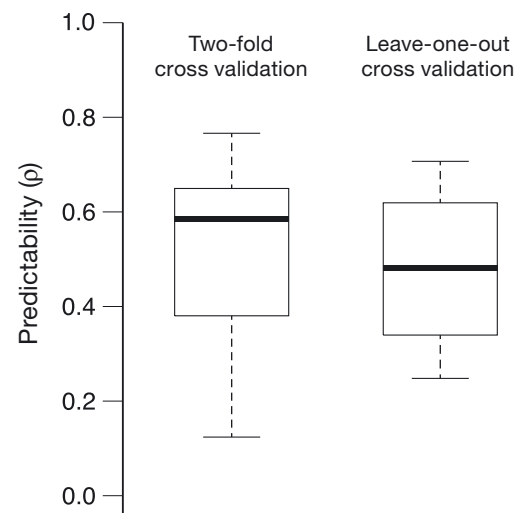


Fig. 3. Comparison of the predictive skill ( $\rho$ ) for the 23 species with the highest data availability under 2 different cross validation schemes: (1) 2-fold cross validation, where each half of the data series is used to predict the other half, and the predictions for the entire series are combined to calculate performance; and (2) leave-one-out cross validation. The 2 schemes produce results that do not differ significantly from each other ( $p < 0.05$ , paired  $t$ -test), showing that the dynamics are stationary in both halves of the data. Boxplots show the distribution of predictability values. Bold lines: median; boxes: interquartile range; whiskers: minimum and maximum values

2005, Hsieh et al. 2005, Glaser et al. 2014). Although short time series can be challenging to the identification of nonlinear dynamics and construction of empirical models, there have been several recent advances on this front (Hsieh et al. 2008, Glaser et al. 2011, Clark et al. 2015, Ye & Sugihara 2016).

Data-driven approaches where causal variables and functional relationships are determined empirically may offer a viable alternative to inductive equation-based approaches. For example, Sugihara et al. (2012) introduce an EDM method for using time series to identify the causal drivers of ecosystem dynamics, and several others (Dixon et al. 1999, Deyle et al. 2013, Harford et al. 2017) provide examples of incorporating these environmental effects into EDMs to forecast future ecosystem states—including apparently random events such as red tides (McGowan et al. 2017). These approaches do not rely *a priori* on hypothesized equations but instead infer relationships deductively as they appear in the data. With continued monitoring and longer time series, the ability of such techniques to describe nonlinear behavior will improve our understanding of ecological mechanisms, where unraveling the interdependence between environmental factors and endogenous population dynamics is certain to be critical for managing ecosystems in the context of climate change (Pershing et al. 2015). Thus, as predictive data-driven approaches continue to gain traction, investments in long-term data collection will surely yield long-term payoffs.

## CONCLUSIONS

The CPR at SAHFOS is a premiere example of how continuous long-term projects help to create new perspectives in ecology. By providing robust, long-term datasets, this program captures variability in species abundance that occurs on temporal scales ranging from interannual cycles to multidecadal shifts (Barton et al. 2003). Furthermore, SAHFOS, in collaboration with partners from around the globe, is now involved in conducting plankton counts in nearly every major ocean basin on earth. In addition to providing a long-term description of population dynamics, these spatially broad environmental monitoring programs can provide regional insights into global problems. Beyond the common sense value of collecting these data, our analyses provide a quantitative justification for continued support of these programs. These data are critical for predictability and understanding, which are particularly important

given increasing threats to global ecosystems, such as human exploitation pressures and climate change (Halpern et al. 2008).

**Data archive.** The data supporting the results of this study are available at [http://192.171.193.159/doi-library/data-for-zooplankton-and-phytoplankton-from-the-1\)-southern-north-sea-and-2\)-the-irish-sea.aspx](http://192.171.193.159/doi-library/data-for-zooplankton-and-phytoplankton-from-the-1)-southern-north-sea-and-2)-the-irish-sea.aspx).

**Acknowledgements.** Thanks to Martin Edwards, Darren Stevens, and SAHFOS for providing the authors with CPR data. This analysis was supported by US Department of Defense–Strategic Environmental Research and Development Program 15 RC-2509; Lenfest Ocean Program 0002 8335; the Sugihara Family Trust; the Deutsche Bank–Jameson Complexity Studies Fund; and the McQuown Fund and McQuown Chair in Natural Sciences, University of California San Diego. A.G.N. was funded by CONACYT (CVU 579904) and Fulbright Garcia-Robles (LASPAU ID 20140963) doctoral program fellowships. A.F.J. was supported by NSF grant DEB-1632648 (2016). B.K. and M.N. were funded by NSF grant DGE-1144086. The authors declare no conflicts of interest.

## LITERATURE CITED

- ✦ Barton AD, Greene CH, Monger BC, Pershing AJ (2003) The continuous plankton recorder survey and the North Atlantic Oscillation: interannual- to multidecadal-scale patterns of phytoplankton variability in the North Atlantic Ocean. *Prog Oceanogr* 58:337–358
- ✦ Callahan JT (1984) Long-term ecological research. *Bio-science* 34:363–367
- ✦ Clark AT, Ye H, Isbell F, Deyle ER, Cowles J, Tilman GD, Sugihara G (2015) Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology* 96:1174–1181
- ✦ DeAngelis DL, Waterhouse JC (1987) Equilibrium and non-equilibrium concepts in ecological models. *Ecol Monogr* 57:1–21
- ✦ DeAngelis DL, Yurek S (2015) Equation-free modeling unravels the behavior of complex ecological systems. *Proc Natl Acad Sci USA* 112:3856–3857
- ✦ Deyle ER, Fogarty M, Hsieh CH, Kaufman L and others (2013) Predicting climate effects on Pacific sardine. *Proc Natl Acad Sci USA* 110:6430–6435
- ✦ Dixon PA, Milicich MJ, Sugihara G (1999) Episodic fluctuations in larval supply. *Science* 283:1528–1530
- ✦ Evans MR, Norris KJ, Benton TG (2012) Predictive ecology: systems approaches. *Philos Trans R Soc Lond B* 367: 163–169
- ✦ Fromentin JM, Powers JE (2005) Atlantic bluefin tuna: population dynamics, ecology, fisheries and management. *Fish Fish* 6:281–306
- ✦ Glaser SM, Ye H, Maunder M, MacCall A, Fogarty M, Sugihara G (2011) Detecting and forecasting complex nonlinear dynamics in spatially structured catch-per-unit-effort time series for North Pacific albacore (*Thunnus alalunga*). *Can J Fish Aquat Sci* 68:400–412
- ✦ Glaser SM, Fogarty MJ, Liu H, Altman I and others (2014) Complex dynamics may limit prediction in marine fisheries. *Fish Fish* 15:616–633



- ✦ Halpern BS, Walbridge S, Selkoe KA, Kappel CV and others (2008) A global map of human impact on marine ecosystems. *Science* 319:948–952
- ✦ Harford WJ, Karnauskas M, Walter JF, Liu H (2017) Non-parametric modeling reveals environmental effects on bluefin tuna recruitment in Atlantic, Pacific, and Southern Oceans. *Fish Oceanogr* (in press), doi:10.1111/fog.12205
- ✦ Hays GC, Richardson AJ, Robinson C (2005) Climate change and marine plankton. *Trends Ecol Evol* 20:337–344
- ✦ Hsieh CH, Glaser SM, Lucas AJ, Sugihara G (2005) Distinguishing random environmental fluctuations from ecological catastrophes for the north Pacific Ocean. *Nature* 435:336–340
- ✦ Hsieh CH, Anderson C, Sugihara G (2008) Extending nonlinear analysis to short ecological time series. *Am Nat* 171:71–80
- Liu H, Fogarty MJ, Glaser SM, Altman I and others (2012) Nonlinear dynamic features and co-predictability of the Georges Bank fish community. *Mar Ecol Prog Ser* 464: 195–207
- ✦ Liu H, Fogarty MJ, Hare JA, Hsieh CH and others (2014) Modeling dynamic interactions and coherence between marine zooplankton and fishes linked to environmental variability. *J Mar Syst* 131:120–129
- ✦ MacNally R (2000) Regression and model-building in conservation biology, biogeography and ecology: the distinction between—and reconciliation of—‘predictive’ and ‘explanatory’ models. *Biodivers Conserv* 9: 655–671
- Magnuson JJ (1990) Uncovering the processes hidden because they occur slowly or because effects lag years behind causes. *Ecol Res* 40:495–501
- ✦ Magurran AE, Baillie SR, Buckland ST, Dick JM and others (2010) Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends Ecol Evol* 25:574–582
- ✦ McGowan JA, Deyle ER, Ye H, Carter ML and others (2017) Predicting coastal algal blooms in southern California. *Ecology* 98:1419–1433
- ✦ Pershing AJ, Alexander MA, Hernandez CM, Kerr LA and others (2015) Slow adaptation in the face of rapid warming leads to collapse of the Gulf of Maine cod fishery. *Science* 350:809–812
- Pilkey OH, Pilkey-Jarvis L (2007) Useless arithmetic: why environmental scientists can't predict the future. Columbia University Press, New York, NY
- ✦ Richardson AJ, Walne AW, John AWG, Jonas TD and others (2006) Using continuous plankton recorder data. *Prog Oceanogr* 68:27–74
- ✦ Schindler DE, Hilborn R (2015) Prediction, precaution, and policy under global change. *Science* 347:953–954
- ✦ Sugihara G (1994) Nonlinear forecasting for the classification of natural time series. *Philos Trans R Soc Lond A* 348: 477–495
- ✦ Sugihara G, May R (1990) Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* 344:734–741
- ✦ Sugihara G, Casdagli M, Habjan E, Hess D, Dixon P, Holland G (1999) Residual delay maps unveil global patterns of atmospheric nonlinearity and produce improved local forecasts. *Proc Natl Acad Sci USA* 96:14210–14215
- ✦ Sugihara G, May R, Ye H, Hsieh CH, Deyle E, Fogarty M, Munch S (2012) Detecting causality in complex ecosystems. *Science* 338:496–500
- ✦ Ward RJ, Holmes EE, Thorson JT, Collen B (2014) Complexity is costly: a meta-analysis of parametric and non-parametric methods for short-term population forecasting. *Oikos* 123:652–661
- ✦ Ye H, Sugihara G (2016) Information leverage in interconnected ecosystems: overcoming the curse of dimensionality. *Science* 353:922–925
- ✦ Ye H, Beamish RJ, Glaser SM, Grant SCH and others (2015) Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proc Natl Acad Sci USA* 112:E1569–E1576

*Editorial responsibility: Myron Peck,  
Hamburg, Germany*

*Submitted: December 5, 2016; Accepted: April 9, 2017  
Proofs received from author(s): May 12, 2017*