INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Cenci S, Saavedra S. 2018 Uncertainty quantification of the effects of biotic interactions on community dynamics from nonlinear time-series data. *J. R. Soc. Interface* **15**: 20180695.
http://dx.doi.org/10.1098/rsif.2018.0695

Received: 17 September 2018
Accepted: 4 October 2018

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
bioinformatics, systems biology

Keywords:
biotic interactions, model averaging, multivariate time series, uncertainty quantification, statistical inference, Jacobian matrix

Author for correspondence:
Serguei Saavedra
e-mail: sersaa@mit.edu

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4267439.

THE ROYAL SOCIETY PUBLISHING

Uncertainty quantification of the effects of biotic interactions on community dynamics from nonlinear time-series data

Simone Cenci and Serguei Saavedra

Department of Civil and Environmental Engineering, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

SC, 0000-0001-6843-468X; SS, 0000-0003-1768-363X

Biotic interactions are expected to play a major role in shaping the dynamics of ecological systems. Yet, quantifying the effects of biotic interactions has been challenging due to a lack of appropriate methods to extract accurate measurements of interaction parameters from experimental data. One of the main limitations of existing methods is that the parameters inferred from noisy, sparsely sampled, nonlinear data are seldom uniquely identifiable. That is, many different parameters can be compatible with the same dataset and can generalize to independent data equally well. Hence, it is difficult to justify conclusive assertions about the effect of biotic interactions without information about their associated uncertainty. Here, we develop an ensemble method based on model averaging to quantify the uncertainty associated with the effect of biotic interactions on community dynamics from non-equilibrium ecological time-series data. Our method is able to detect the most informative time intervals for each biotic interaction within a multivariate time series and can be easily adapted to different regression schemes. Overall, this novel approach can be used to associate a time-dependent uncertainty with the effect of biotic interactions. Moreover, because we quantify uncertainty with minimal assumptions about the data-generating process, our approach can be applied to any data for which interactions among variables strongly affect the overall dynamics of the system.

1. Introduction

Biotic interactions are important regulators of the dynamics of natural systems [1–5]. For example, interactions between primary producers in the ocean can alter the chemistry of their environment [6]. Similarly, microbial interactions influence evolutionary responses to new environments [7], and facilitate emergent phenomena such as interaction-induced production of metabolites [8]. Yet, the relative importance of the effects of biotic and abiotic factors on community dynamics still remains a matter of debate [9–11].

One of the main limiting factors in our understanding of the effects of biotic interactions on community dynamics is that, contrary to abiotic parameters, reliable experimental measurements of biotic parameters are seldom available [12,13]. Hence, a large number of data-driven statistical methods have been proposed to infer the effect of biotic interactions on community dynamics from species abundance data [14–18]. Broadly speaking, these methods can be divided into three categories: (i) statistical or mechanistic parametric approaches, (ii) non-parametric approaches based on correlations and co-occurrence of species (or taxa) in several equilibrium samples [19–21], and (iii) data-driven non-parameteric approaches based on the theory of nonlinear state-space reconstruction [17].

The first two approaches have presented a number of limitations: models are either too simplistic to capture the true dynamics (small deviations from the assumed to the true model can have significant impacts on the inference outcome [22]) or too complex to be structurally identifiable (i.e. it is not possible to associate a unique value with the model parameters from empirical measurements) [23–26]. Similarly, correlation- or co-occurrence-based approaches have

long being criticized for not being able to capture causation [22], and for being prone to identify mirage correlations ubiquitous in nonlinear systems [17].

Instead, non-parametric state-space approaches based on attractor reconstruction have offered a promising alternative [17]: they do not require linear equilibrium or mechanistic assumptions, and their parameters (used as a proxy for the effects of biotic interactions on community dynamics [15,17,22]) are structurally identifiable. However, structural identifiability does not imply that parameters can be uniquely identified given that noise, missing data or sampling rates in empirical data can introduce confounding effects [27]. Specifically, distinct sets of parameters can be equally compatible with the observed data and can be used to construct models that generalize equally well on independent data, even if the model parameters themselves are structurally identifiable. Hence, there is unavoidable uncertainty associated with the parameters inferred from empirical data that, if caused by the data-collection process or inherent to the data-generating process, cannot be avoided regardless of the inference method used [27,28].

Importantly, uncertainty on inferred parameters (the existence of more than one single explanation of the empirical data) translates into an uncertainty of the scientific conclusions that are drawn based on them. This may have significant consequences when, for example, conclusions are used to inform policy [29], to develop novel drugs [30] or to validate theories [11]. Yet, current non-parametric state-space approaches do not provide a methodology to quantify the level of uncertainty associated with the inferred interactions, or a measure of how this uncertainty changes in time when dynamics are nonlinear and interactions are time dependent [17]. To fill this gap, we develop a methodology to associate an uncertainty level with the temporal effect of biotic interactions on community dynamics when inferred from multivariate nonlinear time-series data using non-parameteric approaches.

Following our proposed methodology, we assess temporal variations of uncertainty about the effect of biotic interactions on community dynamics by measuring how many different explanations (sets of parameters) are equally compatible with the same observational data at any given point in time. To validate our methodology, we study a chaotic synthetic time series for which we know the ground truth and use it to test the validity of the methodology. Then, as a case study, we investigate the uncertainty associated with the effect of biotic interactions in a marine microbial community from two independent datasets: the Bermuda Atlantic time series (BATS) and the Hawaii ocean time series (HOTS).

2. Methods

2.1. Background

Throughout this work, we assumed that data are non-equilibrium time series of species abundances generated by nonlinear population dynamics models of the form

$$\frac{d\boldsymbol{x}}{dt} = \mathcal{F}(\boldsymbol{x}, \boldsymbol{\beta}), \tag{2.1}$$

where $\mathcal{F}$ is an unknown and unspecified vector field (defining the dynamics of the system), $\boldsymbol{x} \in \mathbb{R}^d$ is the state vector (e.g. the abundances of d species) and $\boldsymbol{\beta} \in \mathbb{R}^q$ are the q parameters of the model (e.g. the birth and death rates of species). Importantly, the vector field (or model) $\mathcal{F}$ does not need to be purely deterministic, but we assumed the existence of a manifold attractor [31,32]. That is, (2.1) has a steady state (e.g. chaotic, fixed point, limit cycle) to which any initial trajectory converges (this is the only assumption on the data-generating process).

Here, the goal is to infer the Jacobian ($\mathcal{J}$) of $\mathcal{F}$ from a realization of (2.1), and to associate an uncertainty level with its coefficients. Then, following standard approaches in population dynamics [15,17,33], we used the Jacobian matrix of $\mathcal{F}$ as an approximation for the local effect of biotic interactions on the dynamics of (2.1). In fact, the Jacobian matrix (which is formally the matrix of partial derivatives of the vector field $\mathcal{F}$ with respect to the state variables $\boldsymbol{x}$) provides an estimate of the local change of growth rate of a species as a result of a change in the abundance of another species [17].

Because ecological time series are rarely at equilibrium [33], Jacobian coefficients are state-space dependent. Hence, both their values and their associated uncertainty are a function of the state of a system. Time-varying Jacobian coefficients can be inferred from non-equilibrium time-series data using either parametric or non-parametric approaches. Using parametric approaches, the Jacobian coefficients are computed analytically from an assumed model after inference of its parameters (performed using, for example, state-space Bayesian approaches or Kalman filters [34–37]). Using non-parametric approaches, the Jacobian coefficients are inferred directly from the data with minimal assumptions (such as stationarity and distribution of the noise) on the data-generating process [17,38–40]. Because true equations for population dynamics (and other complex interacting systems) are rarely known [41], here we inferred Jacobian coefficients using non-parametric approaches. Specifically, we used the S-map algorithm [32], which has been shown to outperform other existing methods for Jacobian inference [17].

2.2. Statistical inference of the effect of biotic interactions from multivariate time series

The S-map is a weighted local regression model [17]. The weights are state-space-dependent kernel functions. To limit potential problems of overfitting and singularities in the regression procedure of the S-map, we imposed an elastic-net regularization function—a convex mixture of L_1 and L_2 penalty terms [42,43]. Specifically, for each point X^* on the manifold attractor (i.e. for each $t^* \in \{1, \ldots, n\}$ with n number of observations) and for each $i \in \{1, \ldots, d\}$ (with d number of dimension of the system, i.e. number of species), we solved the following minimization problem:

$$\min_{\mathcal{J}_i \in \mathbb{R}^d} (Y_i - X\mathcal{J}_i)^{\mathrm{T}} K(X, X^*, \theta)(Y_i - X\mathcal{J}_i) + \lambda(\alpha \|\mathcal{J}_i\|_2^2 + (1-\alpha)\|\mathcal{J}_i\|_1). \tag{2.2}$$

In the above equation, $\mathcal{J}_i$ is row i of the Jacobian $\mathcal{J}$ and X is an $(n-1) \times d$ data matrix where the point removed from the data matrix is the target point $x(t^*)$. Note that for the analysis of empirical data, where the true dimension of the attractor is unknown, d is not necessarily equal to the number of observed species in the system but to the number of species plus an embedding. In addition, Y_i is the variable to be predicted (i.e. $Y_i = X_{t+1,i}\ \forall X_i \neq X_i^*$), $K(X, X^*, \theta)$ is a kernel function, and λ and α are two regularization parameters. The coefficients $\mathcal{J}_i$ that minimize (2.2) are the rows of the Jacobian matrix—vectors on the tangent space of the manifold attractor of the data-generating process (i.e. (2.1)). That is, (2.2) provides the parameters of the linearization of the dynamics along each point on the manifold attractor. This linearization depends upon the curvature of the attractor at each point through the kernel function. Motivated by the analysis of chaotic time series [32], a typical choice for the entries $K(\boldsymbol{x}, \boldsymbol{x}^*, \theta)$ of the kernel matrix (K) is an exponentially decaying function with a tuning parameter θ (chosen with

cross-validation) that sets the level of nonlinearity of the fit

$$K(\boldsymbol{x}, \boldsymbol{x}^*, \theta) = e^{-\theta(\|x-x^*\|/\bar{d})}, \tag{2.3}$$

where x^* is the target point and $\bar{d}$ is the average distance of each point x to x^*. Note that the parameter θ measures the nonlinearity of $\mathcal{F}$ [32]. The choice of the kernel function (e.g. tri-cubic or Epanechnikov kernel) is not unique but depends on the structure of the time series. Recent work has shown that the S-map provides a good approximation of the Jacobian coefficients in nonlinear dynamical systems [17,33].

It is important to note that the solution of (2.2) is unique for any given set of λ, α, and kernel parameters. Indeed, (2.2) is a strictly convex problem for any $\alpha > 0$ [42]. Hence, the parameters inferred using the regularized S-map are structurally identifiable and their optimum value can be found, for example, by means of cross-validation [44]. However, the existence of a unique minimum of the loss function (2.2) is a weak condition for identifiability. In fact, while cross-validation applied to a strictly convex loss function selects one single model, the training-error and validation-error landscapes of (2.2) can be degenerate around the minimum error. That is, generally, different sets of λ, α, and kernel parameters can provide solutions (inferred parameters) with validation and training errors close to the minimum. However, as we will see in the next section, it cannot be guaranteed that degeneracy in training and validation errors necessarily translates into a degeneracy of the inferred parameters. Hence, different sets of parameters may explain the observational data equally well, leaving the uncertainty of which set of parameters to choose as a true explanation of the data. In the next section, we develop an algorithm to associate an uncertainty level with the Jacobian coefficients inferred from the S-map for each point along the manifold attractor.

2.3. A model average algorithm to associate an uncertainty level with the effect of biotic interactions

To associate an uncertainty level with the Jacobian matrix of (2.1) inferred from the S-map at any given time t, we proposed an algorithm based on model averaging. Specifically, for each point in time, we fixed the ratio α of L_1 to L_2 norms and solved (2.2) using leave-one-out cross-validation. Then, we changed α and solved again (2.2). We repeated this for $\alpha \in [0, 1]$ with steps of $\delta\alpha = 0.01$. From this ensemble of solutions (i.e. ensemble of Jacobian coefficients), we selected the subset $\mathcal{M}$ of parameters that can be used to construct local linear models that exhibit minimum training and test errors (within a threshold that we fixed at 95% of the minimum training and test error, respectively). If the intersection is empty, we only considered models with optimum test error. This selection procedure allowed us to discard parameters (and therefore models) that either do not fit or overfit the data.

The fitting procedure above has a Bayesian interpretation. In fact, a particular choice of α in (2.2) corresponds to the assumption of a specific prior distribution on the parameters [42,43]. Then, the subset $\mathcal{M}$ is an ensemble of models equally compatible with the observed data, but with both prior and posterior distributions of the parameters that are not necessarily the same. In other words, we assumed uncertainty on the prior distribution of the parameters, which is translated into an uncertainty in the posterior. Finally, from the optimal ensemble of parameters $\mathcal{M}$, we calculated an expected value of the Jacobian coefficients with a weighted average of the parameters in the ensemble (i.e. $\mathcal{L}_m$):

$$\mathbb{E}[\mathcal{J}(t)|X] = \sum_{m\in\mathcal{M}} \mathbb{E}[\mathcal{J}(t)|\mathcal{L}_m, X]\mathbb{P}[\mathcal{L}_m|X] \quad \forall t \in [t_o, t_f]. \tag{2.4}$$

Note that the choice of the weights is arbitrary. For example, in a Bayesian framework, one typically includes in the weighted average all the models and uses the posterior probability of each model as weight [44]. Instead, in an information-theoretic framework, one can use the likelihood of a model given the data applying any information-theoretic measure, such as Akaike's information criterion [45]. Here, we restricted the weighted average to the subset M. We assumed that the weights are proportional to the probability that the model explains the data using a normalized fraction of explained variance in the training set (recall that we have already selected those models with optimal generalization skills). This procedure allowed us to take into account all the best solutions of (2.2) in the computation of the effect of biotic interactions, and to discard solutions that either do not explain the data or generalize poorly on independent data. Using the weighted average, we constructed a new Jacobian matrix at each time t as $\mathcal{J}_{\text{ens}}(t) = \mathbb{E}[\mathcal{J}(t)|X]$. Note that while the training and test errors in the ensemble are all approximately the same, the coefficients over which we averaged can be significantly different.

An important advantage of computing the Jacobian from an ensemble of models is that we can associate an uncertainty level with the elements of $\mathcal{J}$. That is, we can compute the standard errors of the Jacobian coefficients and use them to construct confidence intervals. Then, using these errors, one can associate an uncertainty level with any quantity expressed as a function of the Jacobian elements (e.g. eigenvalues, Lyapunov exponents). In the following, we used the coefficient of variation (i.e. standard deviation over the mean) in order to compare the uncertainty across Jacobian coefficients with different means. It is important to stress that the coefficient of variation cannot be used to construct confidence intervals, but it provides a more intuitive measure about how many different explanations (Jacobian coefficients) are compatible with the same dataset. Figure 1 shows a schematic of the proposed methodology. In the next sections, we describe the application of our methodology on synthetic and empirical time-series data.

2.4. Analysis of synthetic time-series data

Firstly, we tested our approach on synthetic data for which we know the ground truth of the Jacobian coefficients. As an illustrative example of nonlinear dynamics, we generated synthetic data using a chaotic four-species Lotka–Volterra population dynamics model

$$\dot{\boldsymbol{x}} = \boldsymbol{r}\boldsymbol{x}(1 - \mathcal{A}\boldsymbol{x}), \tag{2.5}$$

where $\boldsymbol{r}$ is a vector of intrinsic growth rates and $\mathcal{A}$ is the interaction matrix. Following previous work [46], we set the parameters used to generate chaotic trajectories from (2.5). To mimic sparsity in the time series, we sampled data every 200 data points after a numerical integration of (2.5). Then, we compared the parameters inferred using (2.2) and the model-averaging method with the analytical Jacobian of (2.5) computed at each point along the manifold attractor. To perform this analysis, we fixed the length of the training set to 100 data points and the length of the test set to 10 data points. The number of data points is chosen according to the standard length of ecological time series. To obtain reliable statistics, we repeated this process for 20 randomly sampled time intervals of the generated time series. Finally, we explored how the uncertainty level associated with different parameters changes in time from this purely deterministic nonlinear dynamical system.

2.5. Analysis of empirical time-series data

To illustrate the applicability of our methodology, we performed our uncertainty analysis on the effect of biotic interactions from two empirical datasets. The first dataset is a time series from a marine microbial community located at 19.225° N–39.455° N, 59.649° W–74.6° W. The data are publicly available at http://batsftp.bios.edu/BATS/bottle/bval_bottle.txt. The dataset includes the abundance of four species (*Prochlorococcus*,

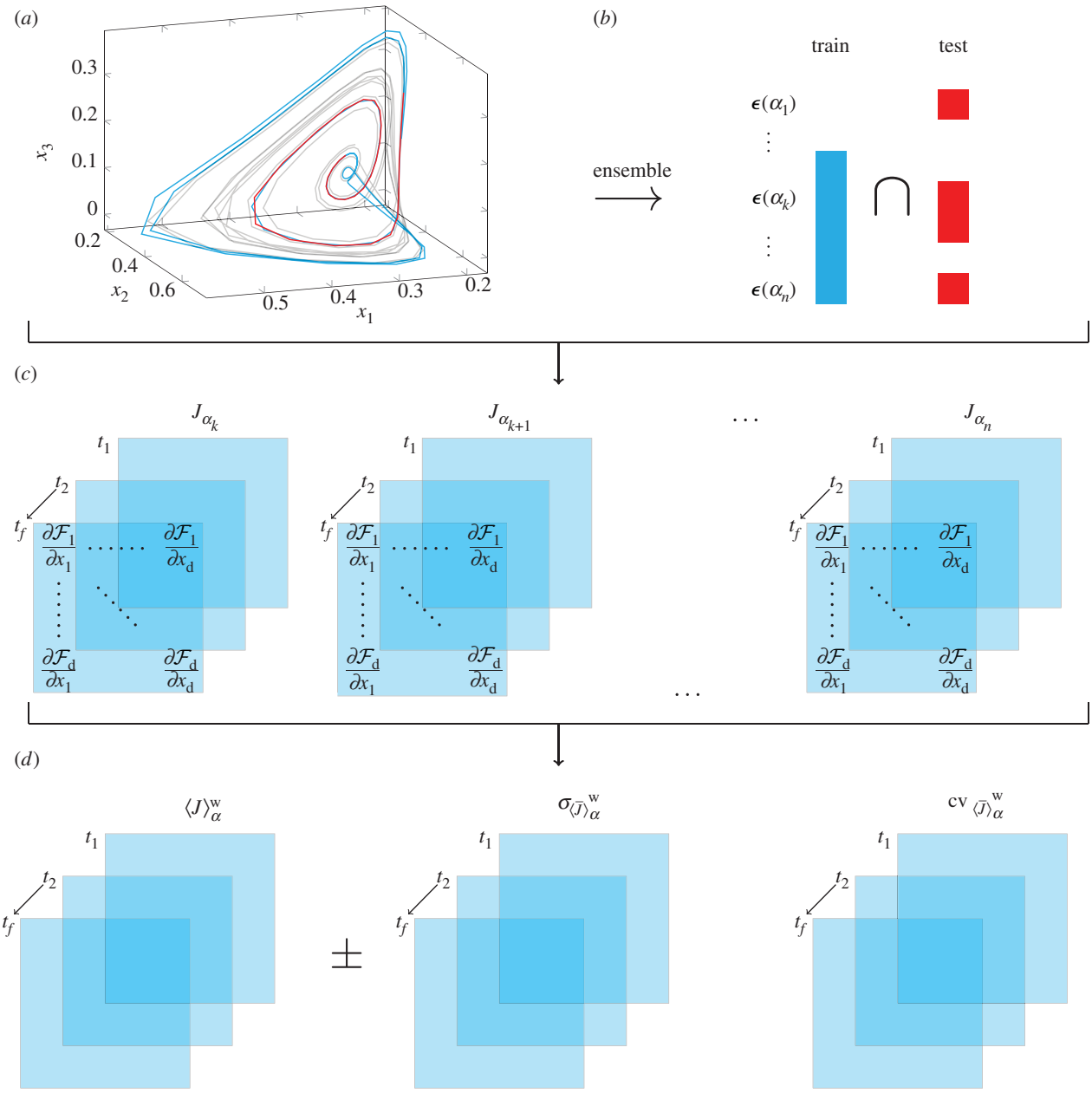

Figure 1. Schematic picture of the proposed ensemble methodology used to associate a level of uncertainty with the effects of biotic interactions. Based on a multivariate time series (*a*), we inferred the Jacobian matrices at each point $\boldsymbol{x}(t_i)$ of a training set (light blue) on the manifold attractor $\mathcal{M}$. The quality of the model is then tested on an independent test set (dark red). We used (2.2) to perform the inference of the Jacobian coefficients. Then, we repeated the inference by changing the ratio α of L_2 to L_1 norms and running cross-validation to select the optimal λ and bandwidth of the kernel. This generated an ensemble of optimal models, from which we selected those with the best performance on the training and test sets—i.e. minimum $\epsilon(\alpha_i)$ within a threshold (coloured squares in (*b*)). Finally, using these time series of Jacobians (one for each model in the intersection (*c*)), we computed their weighted average, standard error and coefficient of variation (*d*).

Synechococcus, picoeukaryotes and nanoeukaryotes) as well as environmental parameters, such as temperature, depth of sampling and salinity. Other ecological variables such as nutrients have been omitted from the analysis as they included many missing values. We analysed data collected approximately twice per day from 28 September 2011 to 18 October 2011 across the ocean surface down to approximately 200 m. The second dataset is a time series of a similar microbial community with *Prochlorococcus*, *Synechococcus*, picoeukaryotes and heterotrophic bacteria located at 22°45′ N, 158°00′ W—approximatively 100 km north of Oahu, Hawaii. The data are publicly available at http://hahana.soest.hawaii.edu/hot/hot-dogs/index.html. We analysed data collected approximately once per month from 2006 to 2016.

In both datasets, we focused on the upper layer of the ocean, i.e. from the surface to a depth of 50m, which is the region of the ocean dominated by high light-adapted clades of *Prochlorococcus* (i.e. eMIT9312 and eMED4) [47]. After sub-setting the data within this region, we aggregated the collected variables (i.e. abundance of species and environmental parameters) and excluded the remaining missing points. This subdivision of data yielded a time series of 30 data points for the Bermuda dataset, out of which 28 were used for training and two for testing. Similarly, this subdivision yielded a time series of 93 data points for the Hawaii dataset, out of which 91 were used for training and two for testing.

Using the pre-processed data, we performed a variable selection: we standardized variables to have zero mean and unitary variance in the training set. Following standard approaches, we standardized the test set using the mean and variance of the original training set [48]. Then, we performed a causality test to

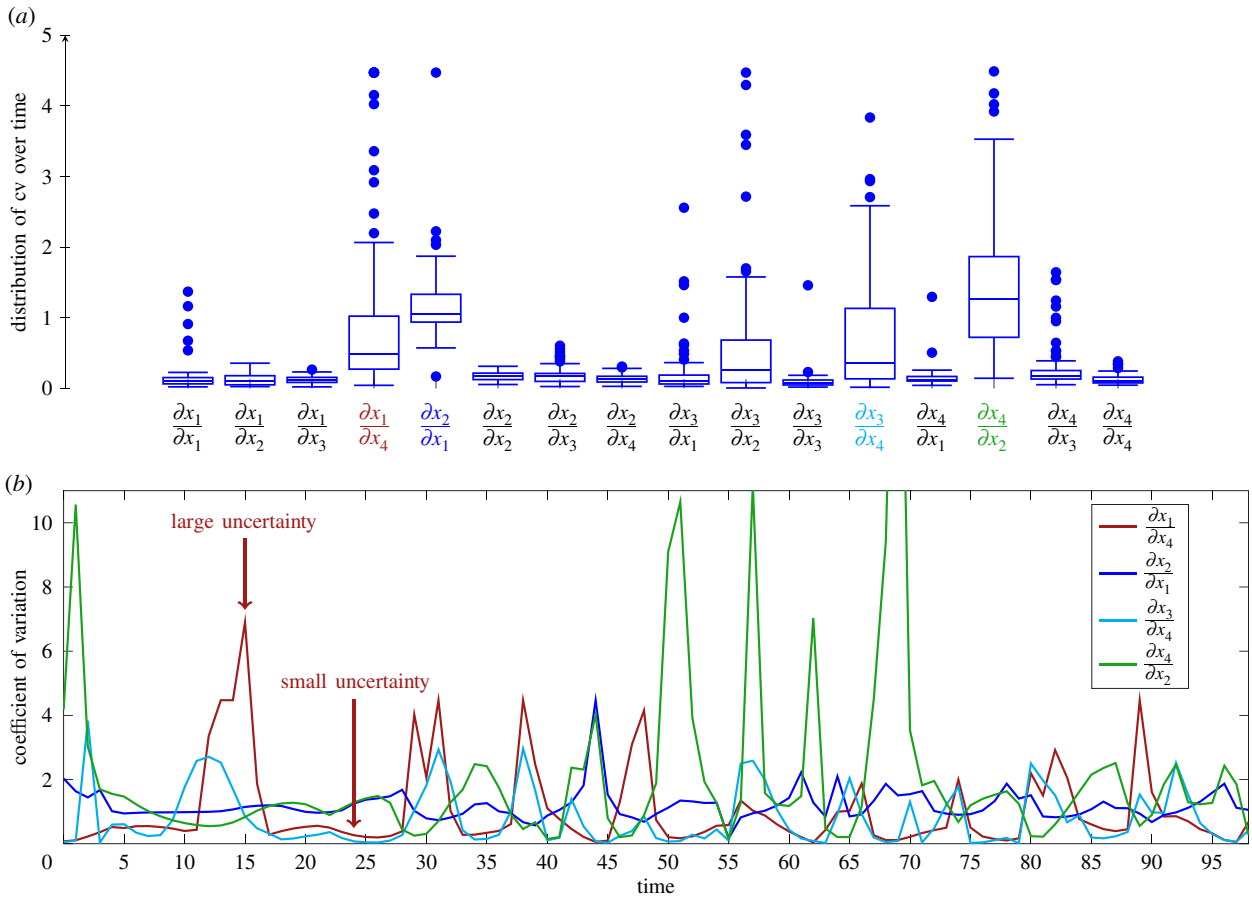

Figure 2. Uncertainty analysis on synthetic data. The figure illustrates that the uncertainty level associated with the Jacobian coefficients of nonlinear time series exhibits a strong time dependency even in a perfect deterministic setting. (*a*) The distribution of the coefficient of variation of the Jacobian coefficients. We used the coefficient of variation to compare the level of uncertainty across coefficients with a different mean. The distribution is computed for each element of the Jacobian over the time interval of the training set. The coloured labels correspond to the four elements with the largest coefficients of variation. For these four elements, panel (*b*) shows how their uncertainty level changes across time, i.e. it is weak or strong depending on the position on the manifold attractor.

reduce the risk of including spurious correlations and to select a model system within which variables were causally related. Specifically, we used a convergent cross-mapping (CCM) test [49]. Using this new subset of variables, we selected the kernel function (from a list of exponential, Epanechnikov, tri-cubic and Matern), the combination of predictors, and the time lags (i.e. embedding dimension) that maximized the out-of-sample forecasting skills of the regularized S-map (2.2). We measured the forecasting skills by the root mean squared error (RMSE), which we compared against the RMSE of the naive forecast (i.e. the forecast that assumes that the variables on the test set are equal to the last point in the training set [48]). Finally, we ran the same analysis described in the previous sections to analyse the uncertainty level associated with the effect of biotic interactions (Jacobian coefficients) for these two microbial communities.

3. Results

3.1. Analysis of synthetic data

We tested our proposed methodology on a (synthetic) chaotic time series. We found that, for this particular time series, our model-averaging method provides a better inference of the Jacobian coefficients than simple cross-validation. We measured the quality of the inference using the Pearson correlation coefficient [32] between inferred and analytical Jacobians (see electronic supplementary material, figures S1 and S2). Then, we looked at the distribution of the coefficient of variation of the Jacobian coefficients within the ensemble of optimum models (recall that, as discussed in the Methods section, we look at the coefficient of variation to fairly compare uncertainty across coefficients with different means). Surprisingly, figure 2*a* shows that, while the value of some elements of the Jacobian matrix was consistently estimated across each model in the ensemble (i.e. we observed a small coefficient of variation), the value of other coefficients changed significantly from model to model (i.e. large coefficient of variation). These findings reveal that not all the inferred effects of biotic interactions (Jacobian coefficients) can be equally trusted [24,50]. Recall that we are inferring coefficients from a purely deterministic, noise-free, time series. Hence, this effect is not due to the quality of the data, but to intrinsic properties of the dynamics.

Moreover, we found that the uncertainty level associated with the elements of the Jacobian matrix changes in time. For example, figure 2*b* shows how the uncertainty level associated with the least consistently inferred Jacobian coefficients changes along each of the points on the manifold attractor. In this figure, one can observe regions of small uncertainty (i.e. small coefficients of variation) followed by peaks of strong uncertainty (i.e. large coefficients of variation). These findings suggest that the inferred effects of biotic interactions can be trusted more in certain periods of time than in other periods. This also suggests that our method can be used to detect when the data are more informative about the inferred effects of biotic interactions.

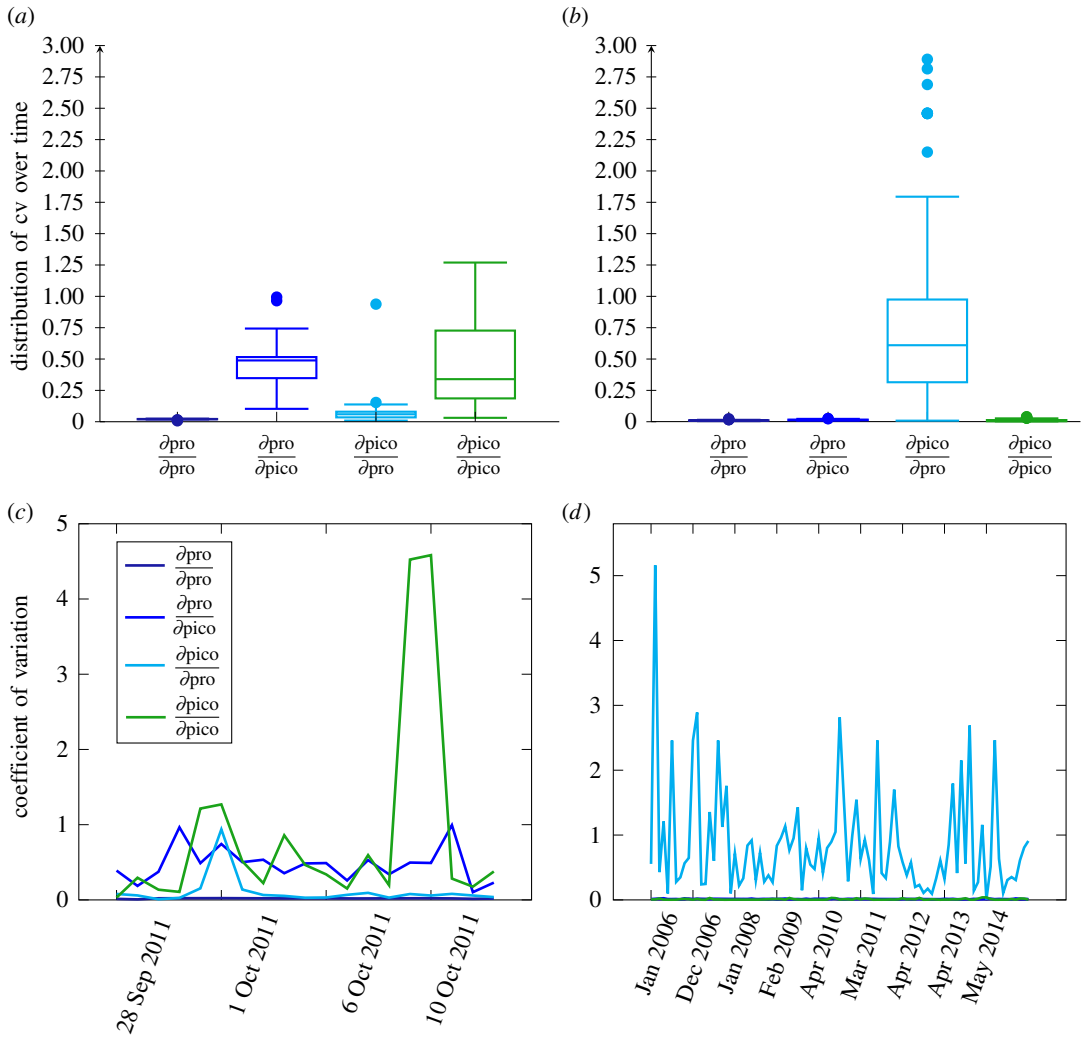

Figure 3. Uncertainty analysis on empirical data. (*a*,*b*) The coefficient of variation of the empirical Jacobian coefficients from each of the top models from the ensemble (figure 1). The distributions were computed for each element of the Jacobian over the time interval of the training set. (*c*,*d*) Temporal changes of the coefficient of variation along the manifold attractor.

3.2. Analysis of empirical data

3.2.1. Bermuda Atlantic time series.

Focusing on the BATS dataset, the CCM test showed a significantly low causal relation between *Prochlorococcus* and nanoeukaryotes, as well as between *picoeukaryotes* and nanoeukaryotes. The strongest causal relation was found within the sub-system made up of *Prochlorococcus*, *Synechococcus* and picoeukaryotes. Electronic supplementary material, figure S3, shows the CCM analysis. Using the time series of *Prochlorococcus*, *Synechococcus* and picoeukaryotes, the model system with the smallest out-of-sample forecast error ($\mathrm{RMSE}_{\mathrm{test,smap}} = 0.17$, $\mathrm{RMSE}_{\mathrm{test,naive}} = 0.31$) was found for exponentially decaying kernels and a combination of *Prochlorococcus* (one time lag or embedding dimension $E = 1$) and picoeukaryotes (two time lags). We excluded the abundance of *Synechococcus* from our analysis since adding it as a predictor variable (with a suitable change in the embedding dimensions) increased both the in- and out-of-sample errors of *Prochlorococcus* and picoeukaryotes. Using this two-species model system, we inferred the interactions using (2.2) and the procedure described in §2. This resulted in an ensemble of 21 models with the lowest test and training errors ($R^2_{\mathrm{training}} \sim 0.82$), which were used for the uncertainty analysis.

Figure 3*a* shows the distribution of the coefficient of variation of the Jacobian coefficients of this two-species model system. Similar to the results shown for the synthetic data, we found results consistent with theoretical expectations. Specifically, some Jacobian coefficients (effect of biotic interactions on community dynamics) are inferred with more confidence than others. Moreover, in line with the results of the analysis on synthetic data, the uncertainty level associated with each coefficient also changed across time. That is, there were regions of the manifold attractor of this model system in which all Jacobian coefficients had equally low uncertainty (resulting in a strong confidence about the selected average model), and other regions in which some coefficients had large uncertainty (providing a low confidence) (figure 3*c*). Electronic supplementary material, figure S5, shows the Jacobian coefficients with confidence intervals constructed using the standard error.

3.2.2. Hawaii ocean time series.

Shifting our focus to the HOTS dataset, we centred our uncertainty analysis on the interactions between *Prochlorococcus* and picoeukaryotes in order to compare their effects against the BATS dataset. In this time series, the causal relation of the two phytoplankton species (computed with the CCM test) was significantly lower than in the BATS dataset. Nevertheless, this model system provided the best out-of-sample

forecast ($RMSE_{test,smap} = 0.17$, $RMSE_{test,naive} = 0.92$), which was found for an embedding dimension (time lag) of $E = 2$ for each variable using a tri-cube kernel function [44]. The maximum explained variance in the training set was $R^2_{training} \sim 0.7$. This resulted in an ensemble of eight optimal models, which were used for the uncertainty analysis.

Figure 3*b* shows the distribution of the coefficient of variation of the Jacobian coefficients of this two-species model system. Different from the BATS dataset, most of the Jacobian coefficients have a significantly low level of uncertainty. Specifically, only the effect of *Prochlorococcus* on picoeukaryotes exhibited a large coefficient of variation, but because this coefficient is very close to zero the actual standard error is very low. Additionally, figure 3*d* shows that, contrary to the BATS dataset, the temporal pattern of uncertainty on the coefficients was relatively homogeneous across time. Electronic supplementary material, figure S6, shows the Jacobian coefficients with confidence intervals constructed using the standard error.

Finally, we observed that the temporal pattern of uncertainty did not change dramatically when we included additional variables in the regression analysis. For example, in electronic supplementary material, figure S4, we show that the distribution of the coefficient of variation of only one of the Jacobian elements changed significantly by adding the abundance of *Synechococcus*. Recall, however, that *Synechococcus* was excluded from the original analysis because its inclusion reduced the performance on both the training and test sets.

4. Conclusion

Understanding the effects of biotic interactions on community dynamics has been challenging because of the difficulty in accurately estimating interaction parameters and their associated uncertainty from empirical observations. To tackle this problem, standard approaches have used conditional least-squares estimates of MAR(1) parameters as a proxy for biotic interactions, and have used either parametric bootstrapping or profile likelihood methods to estimate their uncertainty [15,51]. However, because these approaches rely on equilibrium assumptions and equilibrium dynamics are rarely observed in natural ecosystems [32,33,41,49,52,53], their applicability on empirical data has been limited. Similarly, other approaches have used non-parameteric estimates of Jacobian coefficients as a proxy for the effects of biotic interactions on community dynamics [17,38]. However, while these approaches can deal with non-equilibrium dynamics, they lack a consistent framework for quantifying the uncertainty associated with their results.

To address the limitations above, here we have developed a novel data-driven approach based on model averaging to quantify the uncertainty level associated with the local effect of biotic interactions (Jacobian coefficients) on community dynamics across time from a multivariate nonlinear time series. We have quantified the uncertainty of these interactions based on the number of equally valid explanations compatible with a particular dataset. Importantly, the confidence intervals constructed using our approach, which is local in time, are time dependent (see electronic supplementary material, figures S5 and S6). This is an important advantage because, even in the presence of noise, strong nonlinearities or poor quality of the data, our approach provides a clear intuitive methodology to identify regions of the data that are strongly identifiable and from which conclusions can be asserted with stronger confidence.

We have found three main results derived from our proposed methodology. Firstly, by averaging out different posterior distributions, our methodology can provide better inference of the effects of biotic interactions than previous methodologies [17] (see electronic supplementary material, figures S1 and S2). Note that the inference quality of our methodology, or any other statistical inference algorithm, cannot be tested from empirical data as the ground truth is unknown. Cross-validation provides a means to choose a particular set of parameters, but low validation errors do not necessarily guarantee a good accuracy on the inferred parameters as inferring and forecasting are two separate tasks [48,54].

Secondly, using two marine microbial communities as case studies, we have found that the uncertainty associated with the effect of biotic interactions changes significantly across time (figure 3*a*,*b*). Moreover, we have shown that different interactions can have significantly different levels of uncertainty. This result implies that, while some of the single interactions can be trusted, the whole Jacobian matrix can have a large associated uncertainty. In fact, we have shown that this can also be true for noise-free synthetic datasets (figure 2). This is an important point to bear in mind because if the aim of a study is, for example, to investigate the stability of a community from the inferred Jacobian matrix [15,33], then even a small uncertainty associated with an element of the Jacobian can translate into a large uncertainty on its eigenvalues (see electronic supplementary material, figures S7–S9).

Thirdly, we have found that in both synthetic and empirical time series the pattern of uncertainty can be considerably different across time. These differences happen within the same model system sampled at different locations and at different time scales. Hence, our method can also be used to choose from a number of datasets the one from which parameter inference is more reliable.

Overall, we have proposed a methodology to associate a level of uncertainty with Jacobian coefficients of nonlinear systems inferred from empirical data. As a case study, we have analysed data from population biology for which Jacobian coefficients can be used as a proxy for the local effect of biotic interactions on the dynamics of the community. However, because our approach only relies on the assumption that data are generated by processes that are not purely stochastic, it can be of practical use for other disciplines where time-series data are expected to align with this assumption, such as in finance or economics.

Data accessibility. The code supporting the results is available on GitHub at https://github.com/MITEcology/Royal.Society.Interface-Cenci-Saavedra. The Bermuda Atlantic time series (BATS) is publicly available at: http://batsftp.bios.edu/BATS/bottle/bval_bottle.txt. The Hawaii ocean time series (HOT) is publicly available at: http://hahana.soest.hawaii.edu/hot/hot-dogs/index.html.

Authors' contributions. S.C. and S.S. designed the study and wrote the manuscript, S.C. performed the study, S.S. supervised the study.

Competing interests. The authors declare no competing financial interests.

Funding. Funding was provided by the MIT Research Committee Funds and the Mitsui Chair (S.S.).

Acknowledgements. We thank Chuliang Song for insightful discussions.

Reference

1. Tyc O, van den Berg M, van Veen JA, Raaijmakers JM, de Boer W, Garbeva P. 2014 Impact of interspecific interactions on antimicrobial activity among soil bacteria. *Front. Microbiol.* **5**, 567. (doi:10.3389/fmicb.2014.00567)
2. Crowther TW, Thomas SM, Maynard DS, Baldrian P, Covey K, Frey SD, van Diepen LTA, Bradford MA. 2015 Biotic interactions mediate soil microbial feedbacks to climate change. *Proc. Natl Acad. Sci. USA* **112**, 7033–7038. (doi:10.1073/pnas.1502956112)
3. Abrudan MI, Smakman F, Grimbergen AJ, Westhoff S, Miller EL, van Wezel GP, Rozen DE. 2015 Socially mediated induction and suppression of antibiosis during bacterial coexistence. *Proc. Natl Acad. Sci. USA* **112**, 11 054–11 059. (doi:10.1073/pnas.1504076112)
4. Ho A, Angel R, Veraart AJ, Daebeler A, Jia Z, Kim SY, Kerckhof F-M, Boon N, Bodelier PLE. 2016 Biotic interactions in microbial communities as modulators of biogeochemical processes: methanotrophy as a model system. *Front. Microbiol.* **7**, 1285. (doi:10.3389/fmicb.2016.01285)
5. Enke TN, Leventhal GE, Metzger M, Saavedra J, Cordero OX. 2018 Microscale ecology regulates particulate organic matter turnover in model marine microbial communities. *Nat. Commun.* **9**, 2743. (doi:10.1038/s41467-018-05159-8)
6. Amin SA *et al.* 2015 Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* **522**, 98–101. (doi:10.1038/nature14488)
7. Lawrence D, Fiegna F, Behrends V, Bundy JG, Phillimore AB, Bell T, Barraclough TG. 2012 Species interactions alter evolutionary responses to a novel environment. *PLoS Biol.* **10**, 1–11. (doi:10.1371/journal.pbio.1001330)
8. Watrous J *et al.* 2012 Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci. USA* **109**, E1743–E1752. (doi:10.1073/pnas.1203689109)
9. Fargione J, Brown CS, Tilman D. 2003 Community assembly and invasion: an experimental test of neutral versus niche processes. *Proc. Natl Acad. Sci. USA* **100**, 8916–8920. (doi:10.1073/pnas.1033107100)
10. Houlahan JE *et al.* 2007 Compensatory dynamics are rare in natural ecological communities. *Proc. Natl Acad. Sci. USA* **104**, 3273–3277. (doi:10.1073/pnas.0603798104)
11. Mutshinda CM, O'Hara RB, Woiwod IP. 2009 What drives community dynamics? *Proc. R. Soc. B* **276**, 2923–2929. (doi:10.1098/rspb.2009.0523)
12. Wootton JT, Emmerson M. 2005 Measurement of interaction strength in nature. *Annu. Rev. Ecol. Evol. Syst.* **36**, 419–444. (doi:10.1146/annurev.ecolsys.36.091704.175535)
13. Fuhrman JA, Cram JA, Needham DM. 2015 Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.* **13**, 133–146. (doi:10.1038/nrmicro3417)
14. Lamon EC, Carpenter SR, Stow CA. 1998 Forecasting PCB concentrations in Lake Michigan salmonids: a dynamic linear model approach. *Ecol. Appl.* **8**, 659–668. (doi:10.1890/1051-0761(1998)008[0659:FPCILM]2.0.CO;2)
15. Ives AR, Dennis B, Cottingham KL, Carpenter SR. 2003 Estimating community stability and ecological interactions from time-series data. *Ecol. Monogr.* **73**, 301–330. (doi:10.1890/0012-9615(2003)073[0301:ECSAEI]2.0.CO;2)
16. Jiang X, Hu X, Xu W, Park EK. 2015 Predicting microbial interactions using vector autoregressive model with graph regularization. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **12**, 254–261. (doi:10.1109/TCBB.2014.2338298)
17. Deyle ER, May RM, Munch SB, Sugihara G. 2016 Tracking and forecasting ecosystem interactions in real time. *Proc. R. Soc. B* **283**, 20152258. (doi:10.1098/rspb.2015.2258)
18. Marbach D *et al.* 2012 Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804. (doi:10.1038/nmeth.2016)
19. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C. 2012 Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, 1–17. (doi:10.1371/journal.pcbi.1002606)
20. Friedman J, Alm EJ. 2012 Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, 1–11. (doi:10.1371/journal.pcbi.1002687)
21. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. 2015 Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, 1–25. (doi:10.1371/journal.pcbi.1004226)
22. Xiao Y, Angulo MT, Friedman J, Waldor MK, Weiss ST, Liu Y-Y. 2017 Mapping the ecological networks of microbial communities. *Nat. Commun.* **8**, 2042. (doi:10.1038/s41467-017-02090-2)
23. Bellman R, Aström K. 1970 On structural identifiability. *Math. Biosci.* **7**, 329–339. (doi:10.1016/0025-5564(70)90132-X)
24. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. 2009 Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929. (doi:10.1093/bioinformatics/btp358)
25. Villaverde AF, Barreiro A, Papachristodoulou A. 2016 Structural identifiability of dynamic systems biology models. *PLoS Comput. Biol.* **12**, 1–22. (doi:10.1371/journal.pcbi.1005153)
26. Saccomani MP, Thomaseth K. 2016 *Structural vs practical identifiability of nonlinear differential equation models in systems biology*, pp. 31–41. Cham, Switzerland: Springer International Publishing.
27. Angulo MT, Moreno JA, Lippner G, Barabási A-L, Liu Y-Y. 2017 Fundamental limitations of network reconstruction from temporal data. *J. R. Soc. Interface* **14**, 20160966(doi:10.1098/rsif.2016.0966)
28. Hines KE, Middendorf TR, Aldrich RW. 2014 Determination of parameter identifiability in nonlinear biophysical models: a Bayesian approach. *J. Gen. Physiol.* **143**, 401–416. (doi:10.1085/jgp.201311116)
29. Milner-Gulland EJ. 2012 Interactions between human behaviour and ecological systems. *Phil. Trans. R. Soc. B* **367**, 270–278. (doi:10.1098/rstb.2011.0175)
30. Bollenbach T. 2015 Antimicrobial interactions: mechanisms and implications for drug discovery and resistance evolution. *Curr. Opin Microbiol.* **27**, 1–9. (doi:10.1016/j.mib.2015.05.008)
31. Casdagli M, Eubank S, Farmer J, Gibson J. 1991 State space reconstruction in the presence of noise. *Physica D* **51**, 52–98. (doi:10.1016/0167-2789(91)90222-U)
32. Sugihara G. 1994 Nonlinear forecasting for the classification of natural time series. *Phil. Trans. R. Soc. Lond. B* **348**, 477–495. (doi:10.1098/rsta.1994.0106)
33. Ushio M, Hsieh C-h, Masuda R, Deyle ER, Ye H, Chang C-W, Sugihara G, Kondoh M. 2018 Fluctuating interaction network and time-varying stability of a natural fish community. *Nature* **554**, 360–363. (doi:10.1038/nature25504)
34. So P, Ott E, Dayawansa WP. 1994 Observing chaos: deducing and tracking the state of a chaotic system from limited observation. *Phys. Rev. E* **49**, 2650–2660. (doi:10.1103/PhysRevE.49.2650)
35. Meyer R, Christensen N. 2000 Bayesian reconstruction of chaotic dynamical systems. *Phys. Rev. E* **62**, 3535–3542. (doi:10.1103/PhysRevE.62.3535)
36. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, Sander C, Xavier JB. 2013 Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol.* **9**, 1–11. (doi:10.1371/journal.pcbi.1003388)
37. Bucci V *et al.* 2016 Mdsine: microbial dynamical systems inference engine for microbiome time-series analyses. *Genome Biol.* **17**, 121. (doi:10.1186/s13059-016-0980-6)
38. Holmes EE, Ward EJ, Wills K. 2012 Marss: multivariate autoregressive state-space models for analyzing time-series data. *R J.* **4**, 11–19.
39. Ting J, D'Souza A, Vijayakumar S, Schaal S. 2008 A Bayesian approach to empirical local linearization for robotics. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA'08), Pasadena, CA, 19–23 May 2008*, pp. 2860–2865. New York, NY: IEEE.
40. Ghosh A, Mukhopadhyay S, Roy S, Bhattacharya S. 2014 Bayesian inference in nonparametric dynamic state-space models. *Stat. Methodol.* **21**, 35–48. (doi:10.1016/j.stamet.2014.02.004)

41. Perretti CT, Munch SB, Sugihara G. 2013 Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. *Proc. Natl Acad. Sci. USA* **110**, 5253–5257. (doi:10.1073/pnas.1216076110)
42. Zou H, Hastie T. 2005 Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**, 301–320. (doi:10.1111/j.1467-9868.2005.00503.x)
43. Li Q, Liny N. 2010 The Bayesian elastic net. *Bayesian Anal.* **5**, 151–170. (doi:10.1214/10-BA506)
44. Hastie T, Tibshirani R, Friedman J. 2001 *The elements of statistical learning*. Springer Series in Statistics. New York, NY: Springer.
45. Burnham KP, Anderson DR. 2002 *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. New York, NY: Springer.
46. Vano JA, Wildenberg JC, Anderson MB, Noel JK, Sprott JC. 2006 Chaos in low-dimensional Lotka–Volterra models of competition. *Nonlinearity* **19**, 2391–2404. (doi:10.1088/0951-7715/19/10/006)
47. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. 2006 Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737–1740. (doi:10.1126/science.1118052)
48. Shmueli G, Bruce PC, Patel NR. 2017 *Data mining for business analytics: concepts, techniques, and applications in R*. Hoboken, NJ: Wiley.
49. Sugihara G, May R, Ye H, Hsieh C-h, Deyle E, Fogarty M, Munch S. 2012 Detecting causality in complex ecosystems. *Science* **338**, 496–500. (doi:10.1126/science.1227079)
50. St John PC, Doyle FJ. 2013 Estimating confidence intervals in predicted responses for oscillatory biological models. *BMC Syst. Biol.* **7**, 71. (doi:10.1186/1752-0509-7-71)
51. Knight K. 2000 *Mathematical statistics*. Boca Raton, FL: Chapman and Hall/CRC.
52. Benincà E, Huisman J, Heerkloss R, Jöhnk KD, Branco P, Van Nes EH, Scheffer M, Ellner SP. 2008 Chaos in a long-term experiment with a plankton community. *Nature* **451**, 822–825. (doi:10.1038/nature06512)
53. Bjørnstad ON, Grenfell BT. 2001 Noisy clockwork: time series analysis of population fluctuations in animals. *Science* **293**, 638–643. (doi:10.1126/science.1062226)
54. Shmueli G. 2010 To explain or to predict? *Stat. Sci.* **25**, 289–310. (doi:10.1214/10-STS330)