The Annals of Applied Statistics
2015, Vol. 9, No. 2, 754–776
DOI: 10.1214/15-AOAS828

GOODNESS OF FIT IN NONLINEAR DYNAMICS: MISSPECIFIED RATES OR MISSPECIFIED STATES?

BY GILES HOOKER[1] AND STEPHEN P. ELLNER[2]

Cornell University

This paper introduces diagnostic tests for the nature of lack of fit in ordinary differential equation models (ODEs) proposed for data. We present a hierarchy of three possible sources of lack of fit: unaccounted-for stochastic variation, misspecification of functional forms in rate equations, and omission of dynamic variables in the description of the system. We represent lack of fit by allowing a parameter vector to vary over time, and propose generic testing procedures that do not rely on specific alternative models. Instead, different sources for lack of fit are characterized in terms of nonparametric relationships among latent variables. The tests are carried out through a combination of residual bootstrap and permutation methods. We demonstrate the effectiveness of these tests on simulated data and on real data from laboratory ecological experiments and electro-cardiogram data.

1. Introduction. Recent statistical literature has seen substantial interest in the problem of fitting nonlinear continuous-time dynamical system models to data. Statistical problems include estimating parameters, determining parameter identifiability, experimental design, and testing goodness of fit. These topics have been approached from numerous perspectives and using various models, from deterministic models in the form of ordinary differential equations (ODEs) through stochastic models based on Wiener processes or finite population models such as branching processes. Techniques for fitting models include nonlinear least squares [Arora and Biegler (2004), Bates and Watts (1988), Bock (1983), Girolami and Calderhead (2011)], maximizing likelihoods for stochastic systems through particle filters [Ionides, Bretó and King (2006)] or via equivalent Bayesian methods [e.g., Golightly and Wilkinson (2011)], methods based on pre-smoothing [Bellman and Roth (1971), Ellner, Seifu and Smith (2002), Varah (1982), Wu, Xue and Kumar (2012)], mimicking forecast models [Pascual and Ellner (2000)] or indirect inference [Gouriéroux and Monfort (1997)], and fitting summary statistics [Ratmann et al. (2009), Reuman et al. (2006), Tien and Guckenheimer (2008), Wood (2010)]. Ramsay et al. (2007) combine the criteria from least squares and from pre-smoothing methods to achieve the advantages of each.

Received December 2013; revised December 2014.

[1]Supported in part by NSF Grants DEB-1353039 and DMS-10-53252. This work was partly carried out while visiting the Department of Mathematics and Statistics at the University of Melbourne.

[2]Supported in part by NSF Grants DEB-0813743 and DEB-125619.

Key words and phrases. Differential equation, diagnostics, goodness of fit, attractor reconstruction, bootstrap.

This paper presents an approach to model diagnostics for improving the fit of a dynamical systems model. Hooker (2009) proposed a goodness-of-fit test for ODE models using a likelihood ratio test. Here we assume that a proposed ODE model has been found to fit poorly, so the next goal is to distinguish among different potential sources of model misspecification. In particular, we suppose that the proposed model is an ODE

$$\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x}; t, \boldsymbol{\theta}) \tag{1}$$

in which $\mathbf{x} \in \mathbb{R}^d$ describes the state of the system and $\mathbf{f}(\mathbf{x}; t, \boldsymbol{\theta})$ describes how quickly the system changes at location $\mathbf{x}$ in the state-space, depending on a vector of model parameters $\boldsymbol{\theta}$ to be estimated. We assume that we have vector-valued data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ from this system observed at times $t_1, \ldots, t_n$, where $\mathbf{y}_i$ is related to $\mathbf{x}(t_i)$ by a known, possibly indirect, measurement process. If we find that the model cannot fit the data well, we then wish to improve the fit by changing the model in some way. Here, we develop testing methods to distinguish between three likely reasons for lack of fit, which would imply three different directions for improving the model:

1. Unmodeled disturbances unrelated to system dynamics, which if modeled as random suggests a probabilistic description of system dynamics.
2. Misspecification of the parametric form of $\mathbf{f}$.
3. Misspecification of the state vector $\mathbf{x}$, in particular, that the state vector $\mathbf{x}$ omits some variables that are needed to provide a full description of the system state.

The methods we propose can be used in combination with a variety of methods for parameter estimation in ordinary differential equations, as discussed below. The same ideas can be employed for model improvement in stochastic systems which propose a probabilistic model for the evolution of $\mathbf{x}$. However, applications to stochastic systems will require modifications to some of the details below and we will not examine these further.

Hooker (2009) notes that residuals from solutions to differential equation models give poor graphical indications of how lack of fit should be addressed. This is because the models describe the derivatives $d\mathbf{x}/dt$ rather than the (observed) state variables themselves. Instead, Hooker (2009) proposed estimating lack of fit in terms of *empirical forcing functions*. These are nonparametric functions $\mathbf{g}(t)$ which modify (1) to

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}\big(\mathbf{x}(t); t, \boldsymbol{\theta}\big) + \mathbf{g}(t) \tag{2}$$

in such a way that a good fit to the data is achieved. $\mathbf{g}(t)$ will thus represent both random disturbances to the system and deterministic lack of fit in $\mathbf{f}$.

The estimated $\mathbf{g}(t)$ can now be examined graphically by plotting its relationship to $\mathbf{x}(t)$, along with lagged values of both $\mathbf{x}$ and $\mathbf{g}$, although this can only be done comprehensively when $\mathbf{x}$ is relatively low dimensional. In ODE models, local (in time or state-space) disturbances to the system are usually modeled as affecting $d\mathbf{x}/dt$. These modify future values of $\mathbf{x}$, so the effects of the disturbances will persist over time in the observations. However, they can be accounted for locally in $\mathbf{g}$. Hooker (2009) provides approximate goodness-of-fit tests for the null hypothesis $\mathbf{g} \equiv 0$ based on a basis expansion, $\mathbf{g} = \Psi(t)D$ for a vector of basis functions $\Psi(t)$, and a coefficient matrix D.

In this paper, we take the same approach, but we model lack of fit in a more general way that includes the possibility of parameter values changing over time, producing the system

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}\big(\mathbf{x}(t); \boldsymbol{\theta}, \mathbf{g}(t)\big) \tag{3}$$

in which $\mathbf{g}(t)$ can modify $\mathbf{f}$ more generally than by additive forcing. In particular, we will examine allowing a parameter of interest to vary over time when doing so has a relevant, mechanistic interpretation. The calculations in Hooker (2009)—based on first-order Taylor expansions—can be readily extended to test $\mathbf{g}(t) \equiv 0$ in this more general model. This approach can be seen as encompassing the model (2) and we will use it throughout the paper.

Our new diagnostic tests provide more information about that nature of the lack of fit when $\mathbf{g}(t)$ is found to be significant. In particular, three nested possibilities for the properties of $\mathbf{g}(t)$ correspond to the alternatives listed above for how model (1) should be reformulated:

Case 1. Exogenous stochastic perturbations: if $\mathbf{g}(t)$ is independent of $\mathbf{x}(t)$, this suggests that $\mathbf{g}(t)$ be modeled as a stochastic process, but that the functional form of (1) is otherwise reasonable.

Case 2. Misspecification of $\mathbf{f}$: this is indicated by $\mathbf{g}(t)$ being at least partly determined by $\mathbf{x}(t)$. This would require $\mathbf{f}$ to be revised, as already discussed in Hooker (2009).

Case 3. Missing state variables: if $\mathbf{g}(t)$ depends not only on $\mathbf{x}(t)$ but also on past values $\mathbf{g}(t-\delta)$. These lags serve as surrogates for missing state variables such as additional species in an ecological model, additional chemical products in a reaction, or additional ion channels in a neuron. See Section 4 for further details.

We can motivate this sequence of tests by supposing the data in fact come from an ODE of the form

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \tilde{\mathbf{f}}(\mathbf{x}, y), \\ \frac{dy}{dt} &= k(\mathbf{x}, y), \end{aligned} \tag{4}$$

in which y represents a possible additional state variable and $\tilde{\mathbf{f}}$ represents the true law of motion that may differ from the assumed law of motion $\mathbf{f}$. Model (4) has both of the sources of error that we want to detect. Case 2 corresponds to $\tilde{\mathbf{f}}$ being a function of only $\mathbf{x}$, $\tilde{\mathbf{f}}(\mathbf{x}, y) = \tilde{\mathbf{f}}(\mathbf{x})$. We consider the additive form of lack of fit (2). Then we can write

$$\mathbf{g}(t) = \tilde{f}\big(\mathbf{x}(t)\big) - f\big(\mathbf{x}(t), \boldsymbol{\theta}\big),$$

so case 2 implies $\mathbf{g}(t)$ can be written as a function of $\mathbf{x}(t)$ only.

In case 3 we have

$$\mathbf{g}(t) = \tilde{f}\big(\mathbf{x}(t), y(t)\big) - f\big(\mathbf{x}(t), \boldsymbol{\theta}\big),$$

so the time derivative of $\mathbf{g}$ is given by

$$\begin{aligned}\frac{d\mathbf{g}(t)}{dt} &= \frac{d\mathbf{x}(t)}{dt}\left[\frac{d\tilde{\mathbf{f}}(\mathbf{x}(t), y(t))}{d\mathbf{x}} - \frac{d\mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta})}{d\mathbf{x}}\right] + \frac{dy(t)}{dt}\frac{d\tilde{\mathbf{f}}(\mathbf{x}(t), y(t))}{dy} \\ &= \tilde{f}\big(\mathbf{x}(t), y(t)\big)\left[\frac{d\tilde{\mathbf{f}}(\mathbf{x}(t), y(t))}{d\mathbf{x}} - \frac{d\mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta})}{d\mathbf{x}}\right] \\ &\quad + k\big(\mathbf{x}(t), y(t)\big)\frac{d\tilde{\mathbf{f}}(\mathbf{x}(t), y(t))}{dy}. \end{aligned} \tag{5}$$

If the map from $(\mathbf{x}, y)$ to $(\mathbf{x}, \mathbf{g})$ is invertible, then the expression above implies that $\frac{dg}{dt} = l(\mathbf{x}, \mathbf{g})$ for some function l. The complete dynamical system therefore has the form

$$\begin{aligned}\frac{d\mathbf{x}}{dt} &= \tilde{\mathbf{f}}(\mathbf{x}, \mathbf{g}, \boldsymbol{\theta}), \\ \frac{d\mathbf{g}}{dt} &= l(\mathbf{x}, \mathbf{g}). \end{aligned} \tag{6}$$

If case 2 holds, the second term in (5) is zero and the first term does not depend on y, meaning that $d\mathbf{g}/dt$ is only dependent on $\mathbf{x}$. This suggests testing for dependence of $d\mathbf{g}/dt$ on $\mathbf{g}$, after controlling for $\mathbf{x}$, as a way of distinguishing case 3 from case 2. However, we have found that this test is statistically less stable than testing whether the lagged quantity $\mathbf{g}(t - \delta)$ helps to predict $\mathbf{g}(t)$, after controlling for $\mathbf{x}(t)$. The rationale for this approach is explained more fully in Section 4.

This heuristic can be extended to the model (3) if $\mathbf{f}(\mathbf{x}(t); \boldsymbol{\theta}, \mathbf{g}(t))$ is an invertible function of $\mathbf{g}$ for every $\mathbf{x}$ and $\boldsymbol{\theta}$. However, we note that if this is not the case—for example, if $\mathbf{g}$ is too low dimensional—we will not be able to completely resolve lack of fit and this could make a case 2 misspecification appear as case 3. Apparent case 3 dependence can also result from stochastic fluctuations if the system evolves probabilistically.

We also note that (5) also indicates that there may be little power to detect case 3 dependence in some systems. In particular, if y is itself close to being a function

of $\mathbf{x}$—as we find to be the case in the chemostat experiments described below—it will be difficult or impossible to distinguish case 2 from case 3.

A system in which parameters are changing systematically (e.g., a steady upward trend) will also appear as a case 3 type misspecification, if there is sufficient power to distinguish case 3 from case 2. We believe that this is appropriate. Parameters that are changing systematically can be considered to have their own dynamics and are effectively additional state variables. Similar comments can be made about systems with stochastic dynamics.

In this paper, we develop tests to distinguish between each successive pair of possibilities. These tests need to account for sources of variation that include resampling methods for the $\mathbf{y}_t$ as well as examining the significance of an appropriate nonparametric regression. Our methods can be considered as nonlinear continuous-time extensions of methods to select the number of lags in linear time-series models and to test between models of parameter drift; unlike our case, such tests for linear models can be performed by likelihood ratio tests [see, e.g., Hamilton (1994)].

To provide a concrete example, we consider a model and data from experimental population ecology. In the actual experiments [Becks et al. (2010)] algae of the species *Chlamydomonas reinhardtii*, (C), are grown in a chemostat microcosm which is continuously supplied with nitrogen-limited medium. These algae are preyed upon by rotifers of the species *Brachionus calyciflorus*, (B), near-microscopic animals that feed on algae and reproduce asexually unless at high population density. As a candidate model for this system, we use a standard predator–prey model from the ecological literature, the Rosenzweig–MacArthur model:

$$\begin{aligned} \frac{dC}{dt} &= rC\left(1 - \frac{C}{K_C}\right) - \frac{pGCB}{K_B + pC}, \\ \frac{dB}{dt} &= \frac{\chi_B pGCB}{K_B + pC} - \delta B. \end{aligned} \tag{7}$$

Here dC/dt is the rate of change of the algal population. The first equation describes this change in terms of logistic growth (because algae are limited by resource constraints) with maximal growth rate r and carrying capacity K_C. This term represents algal birth rate minus deaths for causes unrelated to predation (in the actual experiments, washout from the chemostat is the main cause of algal mortality). The second term represents predation by rotifers. Predation occurs at maximum rate G but is reduced when algae are scarce, with K_B representing the algal density pC at which the predation rate is half of its maximum. The parameter p represents the fraction of algae available for predation, and is held at 1 for the moment. Later we will allow p to vary with time, in providing goodness-of-fit diagnostics. The equation for the rotifer growth rate dB/dt represents the conversion of consumed algae into rotifers with conversion rate χ_B, and rotifer mortality δB in proportion to their numbers. Numerically, it is advantageous to reexpress

this system in terms of log variables $\tilde{\mathbf{x}} = (\log C, \log B)$ with differential equation $d\tilde{\mathbf{x}}/dt = \mathbf{f}(\exp(\tilde{\mathbf{x}}); t, \boldsymbol{\theta})/\exp(\tilde{\mathbf{x}})$ and we have employed this below. Note that explicitly modeling washout from the chemostat will be confounded with parameters r, K_C, and δ and we have not included this in the model.

The experimental system was sampled once each day, and rofiters and algae in the sample were counted. Two samples were taken each day, from the top and bottom of the chemostat, to verify that the system was well mixed so that spatial variation in population densities does not need to be considered. The data we analyze are the average of the two daily samples. Plots of the time series and a fit to these data are given in the first panel of Figure 1; these data come from Becks et al. (2010), where the experimental methods are presented in detail.

A number of features are evident from these plots. Most evidently, solutions to the ODE have much more regular cycles than the observed time series. There is also a difference in phase relationships between the rotifers and algae. In the ODE solutions the rotifer peak is about $1/4$ cycle period delayed from the algal peak (because rotifer *population growth rate* peaks when algal density is at a maximum), but in the observed time series the delay is about $1/2$ the cycle period. A proposed explanation for this discrepancy [Yoshida et al. (2003)] is that the algae consist of two subpopulations: one of which does not get predated but pays a cost in reproducing less efficiently, so that the relative advantage of each subpopulation is determined by the rate of rotifer predation. Models incorporating subpopulation structure—hence expanding the state-vector to (C_1, C_2, B) for two algal populations—reproduce the out-of-phase dynamics [Yoshida et al. (2003)]. However, this does not rule out the possibility that the lack of fit is actually due to misspecifying the functional forms for the dynamics of the two-dimensional state vector (C, B).

In our examination below, we will allow p—the proportion of C that is edible—to vary over time. We examine whether this variation can be considered random (case 1), is partly determined by C and B (case 2), or also depends on its own past history, indicating a case 3 misspecification. Experimental evidence tells us that the right answer is case 3 [Yoshida et al. (2003)]: when the algal population is homogenous (all individuals are descended from a single cell), the dynamics are much more like the predictions of classical predator–prey models such as (7) and do not have a $1/2$-period delay.

To represent time-varying quantities $\mathbf{g}(t)$, we employ a basis expansion, $\mathbf{g}(t) = \Psi(t)D$ in which the coefficients D of the basis function $\Psi(t) = \psi_1(t), \ldots, \psi_K(t)$ are treated as additional parameters to be estimated. Because the addition D can make the system unidentifiable [e.g., Hooker (2009)], we employ a two-stage estimation procedure, first estimating fixed parameters $\boldsymbol{\theta}$ and then obtaining an estimate for D. Because estimating derivatives by differencing noisy data significantly increases the noise level and degrades performance, in all of the methods presented below, $\boldsymbol{\theta}$, $\mathbf{x}$, and D are estimated without the need to difference the data.

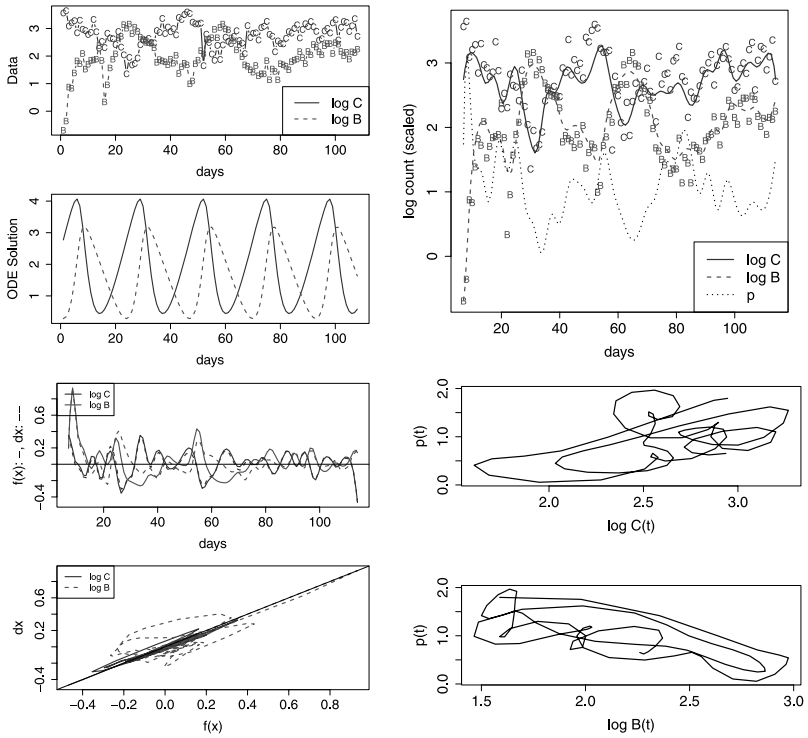

FIG. 1. *Diagnostics for the chemostat data. Top left*: *time series plot of log data* (*top*) *and solution to the Rosenzweig–MacArthur ODE on log scale with constant* $p(t)$ (*bottom*). *These plots allow a comparison between the qualitative behavior of the observed time series and of solutions to the ODE model, which produces phase relationships between* $B(t)$ *and* $C(t)$ *different from those in the data. Top right*: *estimated smooth trajectory* $\hat{\mathbf{x}}(t)$ *and time-varying* $p(t)$. *This allows a comparison of* $\hat{\mathbf{x}}(t)$ *with the data to ensure that our smoothing procedures reflect the data appropriately. Bottom left*: *comparison of* $d\mathbf{x}/dt$ (*dashed lines*) *and* $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}, p(t))$ (*solid lines*) *to ensure that these largely agree after estimating* $p(t)$. *Very large discrepancies relative to the size of* $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}, p(t))$ *would indicate that lack of fit has not been adequately addressed. The lower plot gives* $d\mathbf{x}/dt$ *plotted against* $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}, p(t))$ *for each of* $B(t)$ *and* $C(t)$ *to evaluate the relative size of these departures. Bottom right*: $p(t)$ *plotted against* $C(t)$ *and* $B(t)$. *The evident relationships in these graphs are a visual indicator that* $\mathbf{f}$ *has been incorrectly specified.*

While the ecological experiment described above provides a useful motivation, our diagnostics can be employed on a variety of systems. We explore by simulation the effectiveness of our methods in models for cardiac rhythms and chaotic dynamics as well as the Rosenzweig–MacArthur model above. These are investigated both in cases in which simulated data are generated from an ODE and also when a stochastic differential equation is used to generate noisy trajectories which are then observed with noise.

The rest of the paper is structured as follows. Section 2 details parameter estimation methods and visual diagnostics for lack of fit and Sections 3 and 4 provide

testing procedures for misspecification of $\mathbf{f}$ and $\mathbf{x}$, respectively. Section 5 evaluates these procedures in distinguishing van der Pol and Rössler systems from linear ODEs, while Sections 6 and 7 investigate these procedures with the nonlinear Rosenzweig–MacArthur and van der Pol sytems, respectively, along with applying them to real-world data. We conclude with some speculation about the power of these tests and further directions to be investigated.

2. Parameter estimation and visual diagnostics. In this section we describe a straightforward method of obtaining parameter estimates for use in the simulations below. Throughout this paper we assume that an ordinary differential equation of the form (1) has been proposed for a system under study in which $\mathbf{x}(t)$ is a d-dimensional vector and $\mathbf{f}(\mathbf{x}; t, \boldsymbol{\theta})$ takes values in $\mathbb{R}^d$. We further assume that we have observations $\mathbf{y}_i = \mathbf{x}(t_i) + \boldsymbol{\varepsilon}_i$ taken at times t_i in which each of the state variables is measured with error. This assumption allows us to use the gradient matching procedures described below, which we have chosen for the sake of clarity. However, the tests that we employ can be combined with alternative parameter estimation methods that do not require observations of all assumed state variables.

Gradient matching [Ellner, Seifu and Smith (2002)], also referred to as two-stage least squares in Wu, Xue and Kumar (2012), fits parameters of an ODE model via an initial smoothing step. It proceeds via the following two steps:

1. Fit a vector of smooth curves to the data $(t_i, \mathbf{y}_i)$ to obtain estimates $\hat{\mathbf{x}}(t)$ of the state variables and their time derivatives $d\hat{\mathbf{x}}/dt$. In our studies below we use smoothing splines as implemented in the `fda` package in `R` [Ramsay, Hooker and Graves (2009), see Section 5 for details], but alternatives such as local polynomial models [used in Ellner, Seifu and Smith (2002), Wu, Xue and Kumar (2012)] could also be employed.
2. Estimate parameters $\boldsymbol{\theta}$ by minimizing $\int [d\mathbf{x}/dt - \mathbf{f}(\hat{\mathbf{x}}(t); t, \boldsymbol{\theta})]^2 \, dt$.

The first step is implemented in many software packages and the second may be carried out efficiently with a Gauss–Newton iteration. Note that if $\mathbf{f}$ is linear in its parameters, the second step can be solved with a simple matrix inversion, a property exploited by Dattner and Klaassen (2013) and which also pertains in our examples. Importantly, we expect that this procedure will be relatively robust to model misspecification or disturbances that additively impact $d\mathbf{x}/dt$; this is in contradistinction to fitting solutions to (1) to observed data directly ("trajectory matching") where local disturbances of $d\mathbf{x}/dt$ can persist in deviations from the unperturbed solutions for a long time. This means that we expect to be able to better focus on sources of lack of fit. However, our tests described below can also be applied using trajectory matching as a parameter estimation method.

The gradient matching procedure can be readily extended to higher-order systems. In Section 7 we employ a second-order representation of the van der Pol equation in one state variable. Here step 2 is modified to fit the estimated second derivative of $\mathbf{x}(t)$ to a function of its values and its first derivative.

Gradient matching, while simple to implement and present, is limited in its applicability. Most importantly, it cannot be applied to systems in which some state variables are not directly measured. It also introduces bias when there are either relatively few observations or substantial observation noise. Generalized profiling, introduced in Ramsay et al. (2007), avoids both these complications by using the ODE model to improve the smooth in the first step. We have used generalized profiling with the chemostat example in Section 6 and provided a description of these methods in the supplementary material [Hooker and Ellner (2015)] along with a further set of simulations.

In our methods we first estimate $\hat{\boldsymbol{\theta}}$ in step 2 above with $\mathbf{g}(t) \equiv 0$. In order to estimate $\mathbf{g}$, we represent it by another basis expansion: $\mathbf{g}(t) = \Psi(t)D$. The coefficients D are now fit with $\hat{\boldsymbol{\theta}}$ held fixed by minimizing the gradient matching objective: $\int [d\mathbf{x}/dt - \mathbf{f}(\hat{\mathbf{x}}(t); t, \boldsymbol{\theta}, \Phi(t)D)]^2\, dt$. This two-stage estimation procedure is carried out to ensure the identifiability of parameters. Note that D is estimated within the gradient matching methodology so that the estimate $\mathbf{x}(t)$ will not correspond to an exact ODE solution.

We can now employ the estimate $\hat{\mathbf{g}}(t) = \Psi(t)\hat{D}$ to visually examine lack of fit. First, examining the discrepancy between $d\hat{\mathbf{x}}/dt$ and $\mathbf{f}(\hat{\mathbf{x}}; t, \boldsymbol{\theta}, \mathbf{g}(t))$ provides a visual diagnostic of whether time-varying parameters can account for lack of fit. The procedures we develop here are only appropriate when this is true, because they presume that some function $\mathbf{g}(t)$ exists that brings the model into line with the data. If so, we can first test whether $\hat{\mathbf{g}}(t)$ differs from being constant using the methods in Hooker (2009). Assuming it does (as we do here), we can then plot $\hat{\mathbf{g}}(t)$ versus $\hat{\mathbf{x}}(t)$ to look for consistent relationships that may indicate misspecification of the form of $\mathbf{f}$.

These visual diagnostics are demonstrated in Figure 1. The two panels at the top left show the data and a solution of the proposed Rosenzweig–MacArthur ODE model. The top right panel shows the smooth curves fitted to $C(t)$ and $B(t)$ in the first step of gradient matching, and the estimated $\mathbf{g}(t) = p(t)$. $p(t)$ appears to bear some relationship to both $B(t)$ and $C(t)$ (bottom right panels). The bottom left panels show that $d\mathbf{x}/dt$ and $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}, \mathbf{g})$ are fairly similar (i.e., their values lie near the 1:1 line in the bottom panel), but there remains some additional departure. This is because Rosenzweig–MacArthur is an "off the shelf" predator–prey model which is not mechanistically right for the chemostat system.

3. Tests for dependence between $\mathbf{g}(t)$ and $\mathbf{x}(t)$. For this paper we assume that $\mathbf{g}(t)$ has been shown to differ from zero, hence, the ODE mode (1) is misspecified. We next want to distinguish between the three alternative forms of misspecification listed in the Introduction. The first step is to distinguish between alternatives 1 and 2 by asking whether $\mathbf{g}(t)$ has a consistent relationship with $\mathbf{x}(t)$. If so, this indicates that the functional form of $\mathbf{f}$ has been misspecified [because replacing $\mathbf{g}(t)$ with a function of $\mathbf{x}(t)$ produces a different ODE model]. The visual diagnostics above can then indicate help to determine how $\mathbf{f}$ should be amended.

To determine whether $\mathbf{g}$ depends on $\mathbf{x}$, we assume a null hypothesis in which $\mathbf{g}(t)$ follows a smooth, stationary stochastic process with zero mean. We attempt to distinguish this from the alternative hypothesis of some dependence of $\mathbf{g}(t)$ on $\mathbf{x}(t)$. This alternative still allows for error due to genuine random disturbances, estimation errors, and other forms of misspecification. We conduct this test via a block-permutation test, using nonparametric estimates for the relationship between $\mathbf{g}(t)$ and $\mathbf{x}(t)$. We also account for the estimation of $\mathbf{g}(t)$ through a residual bootstrap.

Formally, our test can be stated as

$$\mathrm{H}_0 : E\big(\mathbf{g}(t)|\mathbf{x}(t)\big) \equiv 0 \quad \text{versus} \quad \mathrm{H}_A : E\big(\mathbf{g}(t)|\mathbf{x}(t)\big) \equiv \mathbf{h}\big(\mathbf{x}(t)\big)$$
$$\text{for some nonconstant function } \mathbf{h}(\cdot),$$

where $\mathbf{h}$ is assumed to be a sufficiently smooth function that nonparametric methods can be employed to estimate. This test could be conducted via a generalized likelihood ratio test [Fan and Yao (2003)], but we must account for the functional nature of $\mathbf{g}(t)$ and $\mathbf{x}(t)$ and their estimation.

To develop a testing procedure for H_0, we first propose a test statistic given by the form of an F-statistic. To calculate this, we estimate $\hat{\mathbf{h}}$ to fit the nonparametric regression model

$$\hat{\mathbf{g}}(t) = \mathbf{h}\big(\hat{\mathbf{x}}(t)\big) + \boldsymbol{\varepsilon}(t).$$

$\hat{\mathbf{h}}$ can be obtained by estimating values of $\hat{\mathbf{g}}$ at a dense set of time points $t_1, \ldots, t_K$, and then applying any smoothing method that minimizes squared error. In the simulations and examples below we set the t_j equal to the observation times in the data and estimated $\mathbf{h}$ by smoothing splines using 40 basis functions with the default settings in the `mgcv` package in `R` [Wood (2013)]. However, our methods are not specific to these choices.

We now propose the F-statistic

$$F = \frac{(1/K)\sum_{i=1}^{K} \|\hat{\mathbf{h}}(\hat{\mathbf{x}}(t_i)) - (1/K)\sum_{j=1}^{K} \hat{\mathbf{h}}(\hat{\mathbf{x}}(t_j))\|^2}{(1/K)\sum_{i=1}^{K} \|\hat{\mathbf{g}}(t_i) - \hat{\mathbf{h}}(\hat{\mathbf{x}}(t_i))\|^2} \tag{8}$$

as a measure of the strength of association between $\hat{\mathbf{g}}(t_i)$ and $\hat{\mathbf{x}}(t_i)$. F is analogous to the standard F-statistic for one-way ANOVA, with $\hat{\mathbf{x}}$ values regarded as "treatment" levels. Alternative measures such as mutual information could also be employed. We have chosen the F-statistic for its familiarity in statistical practice and because it can be readily extended to tests for missing state variables in Section 4.

We now need to compare F to its distribution if H_0 were true. We develop this distribution via a two-stage resampling method. For a fixed $\hat{\mathbf{g}}$ and $\hat{\mathbf{x}}$, a null distribution for F can be obtained by a permutation test: permute the values of $\hat{\mathbf{g}}(t_i)$ relative to $\hat{\mathbf{x}}(t_i)$ so that any relationship between $\mathbf{g}$ and $\mathbf{x}$ is destroyed, re-estimate

$\mathbf{h}(\mathbf{x})$, and re-calculate the F-statistic. Because of the continuity of $\hat{\mathbf{g}}(t)$, the values of $\hat{\mathbf{g}}(t_i)$ exhibit serial dependence over short time intervals, and we therefore permute these values in blocks. In addition, we must also account for the variability in the estimates of $\hat{\mathbf{g}}$ and $\hat{\mathbf{x}}$. This is done via a residual bootstrap, and the block-permutation test is conducted within each bootstrap. This procedure is sketched below, with specific details following:

1. Estimate $\hat{\mathbf{x}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\mathbf{g}}$ from the data.
2. Estimate $\hat{\mathbf{h}}$ to predict $\hat{\mathbf{g}}$ from $\hat{\mathbf{x}}$, by smoothing the values $(\mathbf{x}(t_j), \mathbf{g}(t_j))_{t=1}^{K}$. Use the fitted smooth to calculate $\mathbf{h}(t_j)$ values and the F-statistic in (8).
3. Evaluate a null distribution for F by a residual bootstrap. Loop over 1 to B_1:

   (a) Create new data by resampling the residuals $\boldsymbol{\varepsilon}_i = \mathbf{y}_i - \mathbf{x}(t_i)$ to create new data $\mathbf{y}_i^b = \mathbf{x}(t_i) + \boldsymbol{\varepsilon}^b$ where the superscript b indicates a resampled quantity.

   (b) Estimate $\hat{\mathbf{x}}^b$, $\hat{\boldsymbol{\theta}}^b$, and $\hat{\mathbf{g}}^b$ using the bootstrap data.

   (c) Estimate $\hat{\mathbf{h}}^b$ to predict $\hat{\mathbf{g}}^b$ from $\hat{\mathbf{x}}^b$ and calculate the F-statistic F_{0b} from (8).

   (d) (Permutation test): loop over $k = 1, \ldots, B_2$:

   (i) Permute blocks of the vector $\hat{\mathbf{g}}^b(t_1), \ldots, \hat{\mathbf{g}}^b(t_K)$ to create new values $\hat{\mathbf{g}}_1^{kb}, \ldots, \hat{\mathbf{g}}_K^{kb}$.

   (ii) Estimate $\hat{\mathbf{h}}^{kb}$ to predict the permuted $\hat{\mathbf{g}}^{kb}$ from the $\hat{\mathbf{x}}^b$ and calculate the F-statistic F_{kb}.

   (e) Measure the significance of F_{0b} by evaluating its p-value relative to the permutation distribution:

$$p_b = \frac{1}{B_1} \sum_{k=1}^{B_2} I(F_{0b} > F_{kb}).$$

4. Assess the significance of the test by rejecting H_0 if the average bootstrap p-value is less than α: $\sum_b p_p / B_1 < \alpha$.

We now elaborate on some of these steps to provide detail. In reverse order:

Step 4 rejects based on an average of p-values. This approach is also taken for tests based on random projections [Srivastava (2014)]. Under the null, the p_b should have a uniform distribution. Their average is thus not uniform—it should be more concentrated around $1/2$. Since the p_b are not plausibly independent, we cannot derive a null distribution for their average, and rejecting based on the original significance threshhold is at least conservative.

Step 3(d)(i). We employ blocks larger than the support of the basis functions $\Psi(t)$, so that the permutation does not remove the dependence among close-in-time $\tilde{\mathbf{g}}^b$ values due to the basis function representation. We also remove one half block at the beginning and end of time points, to avoid edge effects in estimating $\mathbf{g}$.

Step 3(b) is easily computed when parameters are estimated by gradient matching, particularly when $\mathbf{f}(\mathbf{x}; t, \boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$. However, this step can be computationally demanding for profiling methods. For this case, in the supplementary material [Hooker and Ellner (2015)] we provide a one-step bootstrap based on a Taylor series expansion.

4. Tests for missing dynamical variables. In addition to misspecifying the parametric form of $\mathbf{f}$, in dynamical systems the proposed model can also misspecify $\mathbf{x}$ by omitting important components of a system. One example of this is the presence of two visually indistinguishable subpopulations of algae in the chemostat system described in the Introduction. Another occurs in neural dynamics in which the voltage across the neuron cell membrane is governed by multiple ion channels [e.g., Tien and Guckenheimer (2008), and see Wilson (1999) for an overview]. Not all of the known channels are always necessary to describe the dynamics of a single neuron, so models often focus on a subset of channels, and lack of fit may result when too few channels are included in a model. Similar situations can arise in modeling chemical reactions or pharmacokinetics, if a model omits some reactions or reaction products.

In this section we assume that a model of the form (1) has been proposed, but the data actually correspond to a model of the form

$$\frac{d\mathbf{x}}{dt} = \tilde{\mathbf{f}}(\mathbf{x}, y, \boldsymbol{\theta}),$$
$$\frac{dy}{dt} = k(\mathbf{x}, y).$$

To determine whether the proposed model is misspecified in this way, we seek to evaluate evidence that the estimated forcing function $\mathbf{g}(t)$ has additional internal dynamics that are not accounted for by a functional dependence of $\mathbf{g}$ on $\mathbf{x}$.

As we observed above [see equation (6)], the difference between this kind of misspecification and the case 2 misspecification considered in the last section is that how $g(t)$ changes over time depends on g itself, not just on the putative state vector $\mathbf{x}$. However, we do not directly test for dependence of dg/dt on g. Instead, motivated by the literature on *attractor reconstruction* [see Abarbanel (1996), Kantz and Schreiber (2005) for an overview] that has developed around the Takens embedding theorem [Takens (1981)], we instead test for dependence of $g(t)$ on $g(t - \delta)$. The methods in this literature predominantly test for dependence on time-lagged state variables rather than derivatives because the results are generally more stable [e.g., Kantz and Schreiber (2005)]. Our experience is in line with this—estimated derivatives were more noisy, and our use of a basis expansion creates an unavoidable relationship between g and dg/dt. As a result, using derivatives instead of time-lagged variables decreased the power of our tests. The theorem's underlying attractor reconstruction does not necessarily hold in stochastic systems or systems far away from their limiting behavior [although see Stark

et al. (1997) for extensions]. But for our purposes this is not important. Testing for dependence of $g(t)$ on $g(t-\delta)$ in addition to $\mathbf{x}$ is simply a stable method for seeking evidence that g is a dynamically evolving state variable whose present state depends on its past. In contrast, if $g(t)$ is just a function of $\mathbf{x}(t)$, past values of g provide no additional information about its present value. This qualitative distinction and the tests we now propose do not depend on the existence of the transform l or on its invertibility.

The use of a basis expansion induces a relationship between $\mathbf{g}(t)$ and $\mathbf{g}(s)$ when $|s-t|$ is small. We therefore choose δ to be larger than the support of the B-spline basis used to estimate $\mathbf{g}(t)$, specifically twice the block length employed in the block permutation test. With this in mind, we can state our test of missing components explicitly as

$$\mathrm{H}_0 : Eg_i(t) \equiv \mathbf{h}_0\big(\mathbf{x}(t)\big) \quad \text{versus} \quad \mathrm{H}_A : Eg_i(t) \equiv \mathbf{h}_1\big(\mathbf{x}(t), g_i(t-\delta)\big).$$

We will approach this test using the same ideas as in the previous section. To do so, we construct smooths $\hat{\mathbf{h}}_0$ and $\hat{\mathbf{h}}_1$ corresponding to the two hypotheses above and calculate an F-statistic for the difference in predictions between these. Specifically, we define

$$F = \frac{(1/K)\sum_{i=1}^{K} \|\hat{\mathbf{h}}_1(\hat{\mathbf{x}}(t_i), \hat{\mathbf{g}}(t_i-\delta)) - \hat{\mathbf{h}}_0(\hat{\mathbf{x}}(t_i))\|^2}{(1/K)\sum_{i=1}^{k} \|\hat{\mathbf{g}}(t_i) - \hat{\mathbf{h}}_1(\hat{\mathbf{x}}(t_i), \hat{\mathbf{g}}(t_i-\delta))\|^2}. \tag{9}$$

For this we again use the functions in the `mgcv` package, but any smoothing method could be employed. We also need to modify the permutation test, which we do by permuting the residuals from the null model $\boldsymbol{\eta}(t) = \mathbf{g}(t) - \hat{\mathbf{h}}_0(\mathbf{x}(t))$ in blocks to create a data set in which H_0 is true.

To carry this out, we proceed following the procedure given in Section 3, modifying only the following steps:

3(c) Estimate $\hat{\mathbf{h}}_0^b$ to predict $\hat{\mathbf{g}}^b$ from $\hat{\mathbf{x}}^b$ and $\hat{\mathbf{h}}_1^b$ to predict $\hat{\mathbf{g}}^b$ from both $\hat{\mathbf{x}}^b$ and $\hat{\mathbf{g}}^b(t-\delta)$ and calculate the F-statistic F_{0b} from (9).

3(d) (Permutation test): loop over $k = 1, \ldots, B_2$:

(a) Permute blocks of the residual vector $\boldsymbol{\eta}^b(t_i) = \mathbf{g}(t_i) - \hat{\mathbf{h}}_0(\mathbf{x}(t_i))$ and add these to predictions to create $\hat{\mathbf{g}}_j^{kb} = \hat{\mathbf{h}}_0^b(\mathbf{x}(t_j)) + \boldsymbol{\eta}^{kb}(t_j)$. $\hat{\mathbf{g}}_1^{kb}, \ldots, \hat{\mathbf{g}}_K^{kb}$.

(b) Estimate $\hat{\mathbf{h}}_0^{kb}$ to predict $\hat{\mathbf{g}}^{kb}$ from $\hat{\mathbf{x}}^b$ and $\hat{\mathbf{h}}_1^{kb}$ to predict $\hat{\mathbf{g}}^{kb}$ from both $\hat{\mathbf{x}}^b$ and $\hat{\mathbf{g}}^{kb}(t-\delta)$ and calculate the F-statistic F_{0b} from (9).

This test can thus be run alongside the test in Section 3.

5. Simulation example: Linear systems versus van der Pol and Rössler systems. We have a set of four nested hypotheses concerning the misspecification of the system, which we can write as:

H0. $\mathbf{g}(t) \equiv 0$,
H1. $E[\mathbf{g}(t)|\mathbf{x}(t), \mathbf{g}(t-\delta)] \equiv 0$,
H2. $E[\mathbf{g}(t)|\mathbf{x}(t), \mathbf{g}(t-\delta)] = h(\mathbf{x}(t))$,
H3. $E[\mathbf{g}(t)|\mathbf{x}(t), \mathbf{g}(t-\delta)] = l(\mathbf{x}(t), \mathbf{g}(t-\delta))$.

In the previous sections we have proposed tests to distinguish H2 from H1 and H3 from H2. Hooker (2009) presents methods to distinguish H1 from H0. We now examine the performance of these tests using simulations and real data.

In our first experiment the proposed model is the 2-dimensional linear system

$$\frac{dx_1}{dt} = a_{11}x_1 + a_{12}, x_2,$$
$$\frac{dx_2}{dt} = a_{21}x_1 + a_{22}x_2$$

with the a_{ij} as unknown parameters. We examine three data-generating models:

1. Circular motion, which corresponds to the linear model with $(a_{11}, a_{12}, a_{21}, a_{22}) = (0, -1, 1, 0)$. In this case H0 is true, because the model is correctly specified.
2. The van der Pol oscillator [van der Pol (1927)]:

$$\frac{dx_1}{dt} = ax_2,$$
$$\frac{dx_2}{dt} = b\left(x_2 - x_1 - \frac{x_2^3}{3}\right),$$

in which misspecification appears as an additive term in the equation for x_2. In this case H2 is true. We take $(a, b) = (0.25, 4)$.
3. The Rössler system [Rössler (1976)]:

$$\frac{dx_1}{dt} = -x_2 - z,$$
$$\frac{dx_2}{dt} = x_1 + ax_2,$$
$$\frac{dz}{dt} = b + z(x - c).$$

In this case the true state vector includes a third variable, so H3 is true. We take $(a, b, c) = (0.2, 0.2, 3)$, and we also consider values $(a, b, c) = (0.2, 0.2, 5.7)$, parameter values classically chosen to produce chaotic dynamics.

For each of these we will examine data generated from the differential equation and data from a stochastic differential equation with additive noise corresponding to

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta})\, dt + \sigma\, d\mathbf{W}, \tag{10}$$

where $\mathbf{W}$ is a multivariate Wiener process with independent components. For the systems above, we took $\sigma^2 = 0.01$ for the linear and van der Pol models and $\sigma^2 = 0.004$ for the Rössler system. These choices gave us a range of stochastic variabilities without making the nonlinear systems diverge to infinity. For the Rössler system with chaotic parameter values, the stochastic system exhibits noticeably shorter-period oscillations; we therefore sped up the ODE experiments by multiplying the right-hand side of this system by a factor of 2, which gave periods similar to the stochastic version.

For each of these systems, we generated a set of observations by adding Gaussian noise to the state of the system:

$$\mathbf{y}_i = \mathbf{x}(t_i) + \boldsymbol{\varepsilon}_i,$$

where the t_i are taken to be 440 equally spaced time points from $t = 0$ to $t = 55$ and the $\boldsymbol{\varepsilon}_i$ are independent Gaussians with variances 0.25, 0.001, and 0.01 for the linear, van der Pol, and Rössler systems, respectively. For the Rössler system, only x_1 and x_2 were observed. In each case we estimated an empirical forcing function $g(t)$ that was added to the second state variable x_2. We used cubic B-splines with a second-derivative penalty to generate $\hat{\mathbf{x}}$ based on knots every 0.25 time intervals with penalty parameter 0.01; some undersmoothing at this step is recommended to reduce bias [Ellner, Seifu and Smith (2002)]. $g(t)$ was represented by a cubic B-spline with knots at integer time intervals from 0 to 55. Each simulation was repeated 200 times.

Visual diagnostics for lack of fit are given in Figure 2, which shows three-dimensional representations of the empirical relationship between $g(t)$, $x_1(t)$, and $x_2(t)$. For the linear data we see no relationship, correctly supporting H0. For the van der Pol data, we see a clear functional depence of g on (x_1, x_2), correctly supporting H2. For the Rössler and Chaotic data, there is no single-valued functional relationship. Rather, the plots suggest trajectories of a three- (or more) dimensional dynamical system, which correctly supports H3.

The power of our proposed tests for each of these systems is given in Table 1. For the linear system, the formal tests correctly do not detect any lack of fit, and for the van der Pol system the tests correctly reject H1 against H2 with high power, but do not reject H2 against H3. For the Rössler and Chaotic, H1 and H2 should both be rejected, but this does not always occur with high power. In these systems, the unequivocal evidence for presence of an unmeasured third state variable is that trajectories in the (x_1, x_2) plane cross each other, which cannot happen in any ODE with (x_1, x_2) as the only state variables. In these simulations, such crossings only occur in a limited region of the two-dimensional state space, and this may account for the reduction in power.

Overall, our tests are somewhat conservative for these test cases. We would expect that the power of our tests would increase with longer time intervals and more frequent data, but would likely decrease as the dimension of the systems under study increases. However, our tests do have reasonable power to detect relevant types of misspecification in these models.

TABLE 1

Power of goodness-of-fit test for case 2 (*misspecification of* **f**) *and case* 3 (*missing components in* **x**) *for data generated by the linear, van der Pol and Rössler ODE and SDE models following parameter estimation by gradient matching. These were estimated from* 200 *simulations for each model as described in Section* 5

		Linear dynamics	**van der Pol**	**Rössler**	**Chaotic**
ODE model	Case 2 (H2 v H1) test	0.06	1	1	1
	Case 3 (H3 v H2) test	0.005	0	0.48	1
SDE model	Case 2 (H2 v H1) test	0.01	1	1	0.91
	Case 3 (H3 v H2) test	0.005	0	0.915	0.68

6. Example: Chemostat models. In this section we present the application of these tests to assess evidence for evolution in the chemostat models described in the Introduction and shown in Figure 1, with the Rosenzweig–MacArthur model (7) as the proposed model. Because of the relative sparsity of the experimental data, we estimated model parameters using the profiling methods described in the supplementary material [Hooker and Ellner (2015)], rather than gradient matching as described in Section 2. All other aspects of testing the model remain the same.

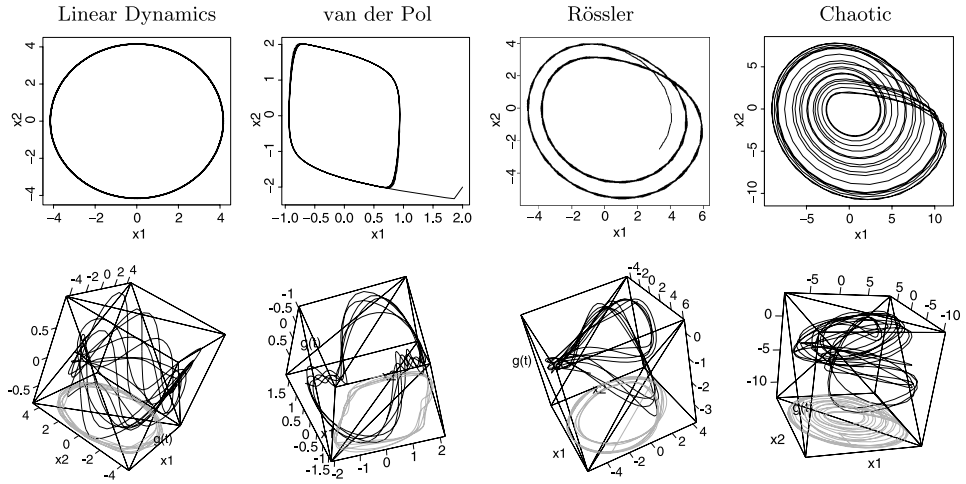

FIG. 2. *Diagnosing lack of fit for the linear model fitted to data generated by the linear, van der Pol, Rössler, and chaotic Rössler systems. Top row*: *phase plane plots of the state variables* x_1, x_2 *in the ODE* (*deterministic*) *model that were sampled to create the data series. Bottom*: *diagnostic plots of* $\hat{\mathbf{g}}(t)$ *plotted against* $x_1(t)$ *and* $x_2(t)$. *Black curves are the three-dimensional trajectories of the SDE, and the grey curves are their projections onto the* (x_1, x_2) *plane. A clear functional relationship is especially visible for the van der Pol example, suggesting correctly that* H2 *is true in this case.*

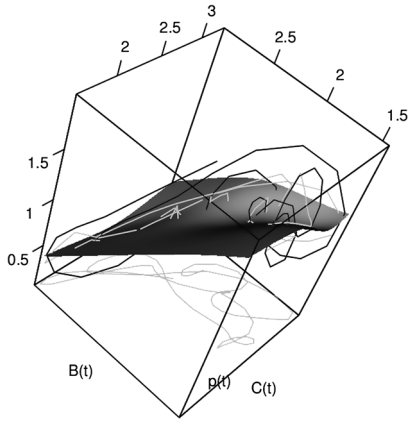

FIG. 3. *Visualization of the diagnostic tests for the Rosenzweig–MacArthur model applied to the chemostat data. Surface indicates predictions of* $p(t)$ *based only on* $(C(t), B(t))$; *Dark lines are* $p(t)$ *plotted against* $(C(t), B(t))$; *Light lines are predictions of* $p(t)$ *based also on* $p(t-\delta)$.

Figure 3 presents the estimated time-varying trait $p(t)$ plotted against the estimated $C(t)$ and $B(t)$ (represented by a cubic B-spline basis with knots every 0.5 days), along with a surface representing the smooth of this relationship, and predictions from a model that also includes $p(t-\delta)$ where $p(t)$ was parameterized by a cubic B-spline basis with knots every 3 days. There is apparent misspecification of $\mathbf{f}$ (H2 against H1), although the p-value for this (0.052) falls short of the traditional threshold for significance. There is insufficient evidence ($p = 0.45$) that the state variable is missing a component (H3 against H2), which could be produced by an additional algal subpopulation.

However, these results do not warrant the conclusion that evolution does not occur in this system, indeed, additional experiments proved that it does [Yoshida et al. (2003)]. The tests rely on the system producing behaviors in which this type of dependence can be readily uncovered. For this system, the power to detect such lack of fit is very low. To demonstrate this, we conducted a simulation study based on two plausible, more complex, stochastic models for the rotifer-algae system. Details of these models are in the supplementary material [Hooker and Ellner (2015)]. The salient distinction between the two models is that one of them includes two populations of algae, while the other does not. We again simulated 200 data sets from each and conducted the proposed tests. Figure 4 presents histograms of the p-values for each test along with example plots relating $p(t)$ to $B(t)$ and $C(t)$ in each model. Here we see that misspecification of $\mathbf{f}$ is detectable (p-value < 0.05 in 53% of the data sets) in the two-algal population model, but the test for missing state variables has very little power (0 out of 200 in both models). The diagnostic plots of Figure 4 are helpful in explaining why this is the case; the grey lines produce the design of covariates values for the case 2 regression of $p(t)$ on $(C(t), B(t))$. Here we see that while the model that incorporates multiple

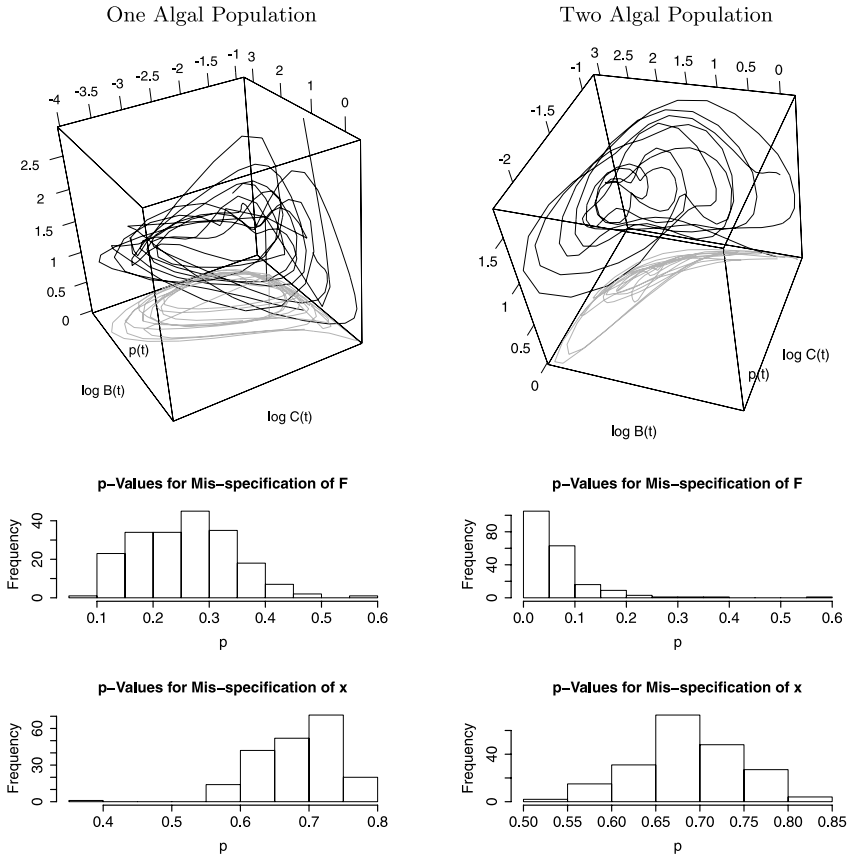

FIG. 4. *Top row*: *example diagnostic plots for a Rosenzweig–MacArthur model with only one algal population fitted to data from a chemostat system model with either one algal population* (*left*) *or two algal populations* (*right*). *Bottom*: *histograms of p-values tests for misspecification of the dynamics* **f** (*top*) *and misspecification of the state vector* (*bottom*), *based on* 200 *simulations*.

algal types produces cycles which are much more elongated, the cycles still do not cross (as they do in the Rössler system in Figure 2). This means that an appropriate nonlinear dependence of $p(t)$ on $(C(t), B(t))$ can capture all of the signal in this relationship, so adding $p(t-\delta)$ as a covariate will not improve predictive performance.

This example provides the important practical lesson that detection of missing state variables requires the system to behave in ways that cannot be replicated by *any dynamical model* that uses the current state space. In this case, there are mechanisms besides algal evolution that can generate the observed system behavior. Once the system is close to its stable periodic trajectory, the relative abundance of the two different algal types can be predicted from the rotifer abundance and total algal abundance (as seen in the functional relationships of p with B and C

in the bottom right panels of Figure 1). Inserting this dependence into the rotifer's feeding rate equation [where $p(t)$ has the largest effect] produces a two-variable model that can exhibit the kind of antiphase cycles seen in the experiment with two algal subpopulations. We hypothesize that this modification to the predator's feeding rate equation serves as a proxy for predator age structure, allowing the model to behave like models that can exhibit the kind of antiphase cycles seen in the experiment as a result of predator age structure. Independent experimental evidence tells us predator age structure is not the mechanism operating in these experiments [Hiltunen et al. (2014), Yoshida et al. (2003)], but from the time series alone it may not be possible to determine that the actual mechanism involves additional state variables.

We also undertook 200 simulations employing the ODE model (7), transformed to represent $\log C(t)$ and $\log B(t)$, to generate data along with additive Gaussian errors with variance 0.25. This provides a means of checking that the nonlinearity of these equations does not distort our tests. The levels of both tests were estimated from this simulation at 0, indicating that the test remains conservative in the presence of nonlinearities.

7. Example: Cardiogram data and the van der Pol system. In this section we present data from electro-cardiogram measurements obtained from the MIT-BIH Arrhythmia Database [subject 214, Goldberger et al. (2000), Moody and Mark (2001)], given in the first plot of Figure 5. For these data we employ an alternative formulation of the van der Pol model studied in Section 5 that is given as a second-order differential equation

$$\frac{d^2x}{dt^2} = a + b\frac{dx}{dt} + cx + dx^2 + ex\left(\frac{dx}{dt}\right)^2. \tag{11}$$

The van der Pol model places further restrictions on the parameters a, b, c, d, and e, but we leave these to be estimated independently. For this system we em-

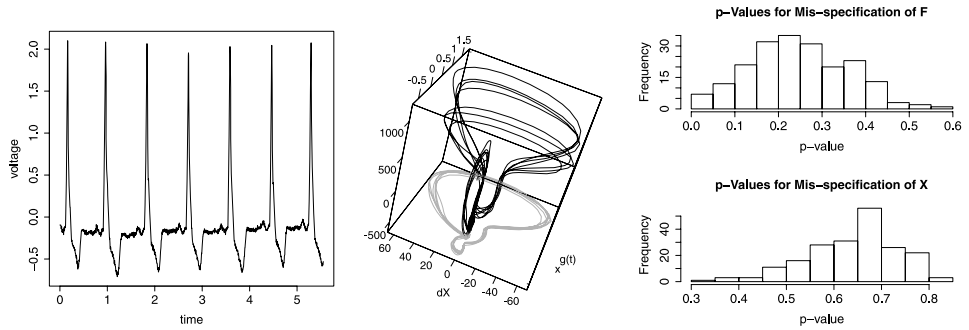

FIG. 5. *Left*: *electro-cardiogram data*. *Middle*: *diagnostic plots for the van der Pol model indicating both cases* 2 *and* 3 *misspecification*. *Right*: *histograms of p-values for data simulated from a van der Pol model without misspecification*.

ploy an extension of gradient matching to second-order ODE's by estimating two derivatives: $\hat{x}$, $d\hat{x}/dt$, and $d^2\hat{x}/dt^2$ using a cubic B-spline basis with 500 knots across the time interval. We then choose parameters to minimize

$$\int \left(\frac{d^2\hat{x}}{dt^2} - a - b\frac{d\hat{x}}{dt} - c\hat{x} - d\hat{x}^2 - e\hat{x}\left(\frac{d\hat{x}}{dt}\right)^2 \right)^2 dt.$$

This can be carried out by evaluating the estimated smooth and its derivatives at a fine grid of time points and then employing linear regression. Following this, the residuals are smoothed using an unpenalized cubic B-spline basis expansion with knots every 0.05 seconds—about 8 observations per knot—to obtain an estimated $g(t)$ as a lack of fit forcing function. The testing procedure proceeds as above with model misspecification obtained by relating $g(t)$ to $x(t)$ and dx/dt, and tests for missing state variables carried out by testing whether $g(t-\delta)$ provides additional predictive accuracy.

A visual display of the analysis for this system is given in Figure 5. The middle panel, in particular, plots the estimated $g(t)$ against $x(t)$ and dx/dt. Here we see a consistent relationship, but also an evident, nearly vertical "cycle" that is preserved across multiple heart beats. This cycle corresponds to the small, but consistent bump in the left-hand plot just before the main spike in voltage. It presents a visual indication of missing state variables, where knowledge of $g(t-\delta)$ can distinguish which part of the subcycle the system is in. To formally test this conclusion, we left off the first and last 100 time points in our testing procedures, and used blocks of size 50. Here both tests returned p-values of zero, indicating that both types of misspecification are present and confirming our visual impression.

To ensure that this effect was not an artifact of the estimation methodology, we conducted a simulation study employing solutions to (11) as the data, with additive observation noise, so that the fitted model is correctly specified. Histograms of p-values from both tests are given in the final plot of Figure 5. Although these are not uniformly distributed, the level of the test is at least conservative (0.035 for case 2, 0 for case 3).

8. Conclusions. This paper represents lack of fit in differential equation models as a series of nested hypotheses:

1. No lack of fit.
2. Unaccounted-for stochastic variation.
3. Misspecified right-hand side functions for the differential equation.
4. Missing or misspecified state variables that describe the system.

We presented tests to distinguish the third from the second and the fourth from the third of these. This nested structure is necessary for the last two possibilities, but nesting the second and third is not strictly required. However, we believe this nesting makes sense in analogy to regression model diagnostics which include a random error term. Lack of fit can alternatively be tested by proposing alternative parametric models and comparing model likelihoods; to our knowledge, this paper

is the first attempt to produce tests that distinguish between different kinds of lack of fit without explicitly modeling them.

Our tests rely on bootstrap and permutation methodologies in order to require as few assumptions as possible. This leads to their being conservative at the null hypothesis; it also makes conducting them computationally demanding. However, they are still capable of distinguishing meaningful differences between models, as our simulations indicate. While our methods are based on explicitly smooth models of dynamics, we have also demonstrated that these systems work well with nonsmooth diffusion processes.

The nonparametric nature of these tests can reduce their power. Moreover, some systems exhibit dynamics in which detecting a missing component is fundamentally difficult. As our ecological example indicates, genuinely three-dimensional systems can often be represented as two-dimensional systems, unless they have behavior that cannot be embedded in two dimensions, and this confounds the two tests that we propose. Methods to distinguish which systems will exhibit this type of confounding are an important direction for future research. More powerful tests can be based on specific alternative hypotheses. For example, the two-algal population model given in the supplementary material [Hooker and Ellner (2015)] provides better qualitative agreement with the data than does the elaborated one-algal model. However, neither model is exactly correct, and tests to distinguish between them while making few assumptions about the form of a stochastic model have yet to be developed.

There is also room to design experiments that would yield behavior in which missing state variables, such as the second algal population in the chemostat data, is more readily detected by the tests proposed here. Hooker, Lin and Rogers (2015) and Thorbergsson and Hooker (2013) present some experimental design methods for dynamical systems in which inputs are perturbed so that observations yield optimal information about parameters of interest. Mork work is needed to adapt these techniques to our tests. The power of our test for misspecified state variables also might be higher when several trajectories have been observed that have different initial values. The test fails when the trajectory of an n-dimensional system, projected onto $n-k$ dimensions, can be reproduced or approximated well by the solution of some $(n-k)$-dimensional dynamical system. This is especially likely if the observed trajectory is on or near a low-dimensional attractor for the dynamics and the dynamics are close to deterministic because of the Takens Embedding Theorem [Takens (1981)]. A second trajectory, with initial values far from the attractor, might require a higher-dimensional system or a different lower-dimensional system to reproduce it, and these would reveal that the system is actually higher dimensional.

SUPPLEMENTARY MATERIAL

Supplementary material for "Goodness of fit in nonlinear dynamics: Misspecified rates or misspecified states?" (DOI: 10.1214/15-AOAS828SUPP;

.pdf). This appendix provides supporting material which includes the following: details of the chemostat models used to generate data for Section 6 and background material on the generalized profiling methods of Ramsay et al. (2007), along with simulation experiments using this method instead of gradient matching.

REFERENCES

ABARBANEL, H. D. I. (1996). *Analysis of Observed Chaotic Data*. Springer, New York. MR1363486

ARORA, N. and BIEGLER, L. T. (2004). A trust region SQP algorithm for equality constrained parameter estimation with simple parameter bounds. *Comput. Optim. Appl.* **28** 51–86. MR2049675

BATES, D. M. and WATTS, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York. MR1060528

BECKS, L., ELLNER, S. P., JONES, L. E. and HAIRSTON, N. G. (2010). Reduction of adaptive genetic diversity radically alters eco-evolutionary community dynamics. *Ecol. Lett.* **13** 989–997.

BELLMAN, R. and ROTH, R. S. (1971). The use of splines with unknown end points in the identification of systems. *J. Math. Anal. Appl.* **34** 26–33. MR0277269

BOCK, H. G. (1983). Recent advances in parameter identification techniques for ODE. In *Numerical Treatment of Inverse Problems in Differential and Integral Equations* (*Heidelberg*, 1982) (P. Deuflhard and E. Harrier, eds.). *Progr. Sci. Comput.* **2** 95–121. Birkhäuser, Boston, MA. MR0714563

DATTNER, I. and KLAASSEN, C. A. J. (2013). Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. Preprint. Available at arXiv:1305.4126.

ELLNER, S. P., SEIFU, Y. and SMITH, R. H. (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology* **83** 2256–2270.

FAN, J. and YAO, Q. (2003). *Nonlinear Time Series*: *Nonparametric and Parametric Methods*. Springer, New York. MR1964455

GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. MR2814492

GOLDBERGER, A. L., AMARAL, L. A., GLASS, L., HAUSDORFF, J. M., IVANOV, P. C., MARK, R. G., MIETUS, J. E., PENG M. G.B., C.-K. and STANLEY, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation* **101** e215–e220.

GOLIGHTLY, A. and WILKINSON, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* **1** 1–14.

GOURIÉROUX, C. and MONFORT, A. (1997). *Simulation-Based Econometric Methods*. Oxford Univ. Press, Oxford.

HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton Univ. Press, Princeton, NJ. MR1278033

HILTUNEN, T., HAIRSTON, N. G. JR., HOOKER, G., JONES, L. E. and ELLNER, S. P. (2014). AUG. A newly discovered role of evolution in previously published consumer-resource dynamics. *Ecology Letters* **17** 915–923.

HOOKER, G. (2009). Forcing function diagnostics for nonlinear dynamics. *Biometrics* **65** 928–936. MR2649866

HOOKER, G. and ELLNER, S. P. (2015). Supplement to "Goodness of fit in nonlinear dynamics: Misspecified rates or misspecified states?" DOI:10.1214/15-AOAS828SUPP.

HOOKER, G., LIN, K. K. and ROGERS, B. (2015). Control theory and experimental design in diffusion processes. Under review. *Journal on Uncertainty Quantification*.

IONIDES, E. L., BRETÓ, C. and KING, A. A. (2006). Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **103** 18438–18443.

KANTZ, H. and SCHREIBER, T. (2005). *Nonlinear Time Series Analysis*, Cambridge Univ. Press, Cambridge. MR2040330

MOODY, G. B. and MARK, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Mag.* **20** 45–50.

PASCUAL, M. and ELLNER, S. P. (2000). Linking ecological patterns to environmental forcing via nonlinear time series models. *Ecology* **81** 2767–2780.

RAMSAY, J. O., HOOKER, G. and GRAVES, S. (2009). *Functional Data Analysis in R and Matlab*. Springer, New York.

RAMSAY, J. O., HOOKER, G., CAMPBELL, D. and CAO, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 741–796. MR2368570

RATMANN, O., ANDRIEU, C., WIUF, C. and RICHARDSON, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Nat. Acad. Sci. USA* **106** 10576–10581.

REUMAN, D. C., DESHARNAIS, R. A., COSTANTINO, R. F., AHMAD, O. S. and COHEN, J. E. (2006). Power spectra reveal the influence of stochasticity on nonlinear population dynamics. *Proceedings of the National Academies of Sciences* **103** 18660–18665.

RÖSSLER, O. E. (1976). An equation for continuous chaos. *Physics Letters* **57A(5)** 397–398.

SRIVASTAVA, R. K. (2014). An exact two-sample test in high dimensions using random projections. Preprint. Available at arXiv:1405.1792.

STARK, J., BROOMHEAD, D. S., DAVIES, M. E. and HUKE, J. (1997). Takens embedding theorems for forced and stochastic systems. *Nonlinear. Anal.* **30** 5303–5314. MR1726033

TAKENS, F. (1981). Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, *Warwick* 1980 (*Coventry*, 1979/1980). *Lecture Notes in Math.* **898** 366–381. Springer, Berlin. MR0654900

THORBERGSSON, L. and HOOKER, G. (2013). Experimental design for partially observed Markov decision processes. Preprint. Available at arXiv:1209.4019.

TIEN, J. H. and GUCKENHEIMER, J. (2008). Parameter estimation for bursting neural models. *J. Comput. Neurosci.* **24** 358–373. MR2399636

VAN DER POL, B. (1927). On relaxation-oscillations. *The London*, *Edinburgh and Dublin Philosophical Magazine and Journal of Science* **2** 978–992.

VARAH, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Statist. Comput.* **3** 28–46. MR0651865

WILSON, H. R. (1999). *Spikes*, *Decisions*, *and Actions*: *The Dynamical Foundations of Neuroscience*. Oxford Univ. Press, New York. MR1972484

WOOD, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102–U113.

WOOD, S. (2013). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. R package version 1.7-27.

WU, H., XUE, H. and KUMAR, A. (2012). Numerical discretization-based estimation methods for ordinary differential equation models via penalized spline smoothing with applications in biomedical research. *Biometrics* **68** 344–352. MR2959600

YOSHIDA, T., JONES, L. E., ELLNER, S. P., FUSSMANN, G. F. and HAIRSTON, N. G. (2003). Rapid evolution drives ecological dynamics in a predator–prey system. *Nature* **424** 303–306.

DEPARTMENT OF BIOLOGICAL STATISTICS
AND COMPUTATIONAL BIOLOGY
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-4201
USA
E-MAIL: gjh27@cornell.edu

DEPARTMENT OF ECOLOGY
AND EVOLUTIONARY BIOLOGY
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-4201
USA
E-MAIL: spe2@cornell.edu